ON GRADIENT DESCENT WITH STOCHASTIC ROUNDING ON FIXED-POINT GRIDS FOR QUADRATIC OBJECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gradient descent on quadratic objectives whose iterates are constrained to a fixed-point grid with spacing Δ via stochastic rounding, by analyzing the hitting time to reach a grid minimizer, i.e., a minimizer over the grid. We first consider objectives whose minimizer lies on the grid and show that unlike gradient descent over the reals, the hitting time is not determined by the spectrum of the Hessian. Specifically, for a Hessian with a fixed spectrum and a random eigenbasis, the hitting time is $e^{\Omega(d)}$ in the dimension d with high probability, and the Hamming distance to the grid minimizer and the optimization error remain of order d and $\Delta^2 d$, respectively, e.g., for Gaussian least squares with $n = \Theta(d)$ samples. In contrast, a diagonal Hessian with the same spectrum gives an $O(\log d)$ hitting time, which we generalize to a sufficient condition satisfied by diagonally dominant Hessians and by Gaussian least squares with $n = \Omega(d^2/\log\log d)$. This logarithmic upper bound matches an $\Omega(\log d)$ lower bound that holds for all Hessians with a unique grid minimizer when the initial Hamming distance is $\Omega(d)$. However, a large Hamming distance does not always imply a large optimization error: for a class of rank-deficient Hessians, e.g., Gaussian least squares with $n = o(d)$, the optimization error drops to $O(\Delta^2 n)$ within $O(\log(d/n))$ iterations while the Hamming distance remains of order d for a number of iterations polynomial in d/n. Finally, if the minimizer of the objective lies off the grid, the exponential lower bound can persist, whereas the location of the minimizer alone can change the hitting time from logarithmic to exponential.

1 INTRODUCTION

Optimization theory is central to modern machine learning, where models are trained by first-order methods such as gradient descent (GD). The convergence of GD is well understood on the reals. For example, for strongly convex and smooth objectives, GD reduces the distance to the minimizer by a factor ε within $O(\kappa \log(1/\varepsilon))$ iterations, where κ is the condition number, and analogous rates for the optimization error are known for convex and nonconvex objectives (Nesterov, 2013; Bubeck, 2015). In these results, the dimension d enters only through the condition number, the smoothness constant, and the initial distance, e.g., the convergence speed of GD on a quadratic objective is determined by the spectrum of the Hessian.

However, computers can only represent a finite set of numbers, and every update must be rounded to a representable value. This issue is most pronounced in low-precision training, which reduces the memory and computation costs of large models and is now supported by hardware (Gupta et al., 2015; Micikevicius et al., 2018; Wang et al., 2018; Sun et al., 2019; Tseng et al., 2025; NVIDIA et al., 2026). When the precision is low, an update can be smaller than half of the spacing between representable numbers. Round-to-nearest then removes the update entirely, and the iterates can stop moving before reaching the minimizer (Gupta et al., 2015; Zamirai et al., 2020; Xia et al., 2024). Stochastic rounding (SR) avoids this issue by rounding up or down with probabilities proportional to the distances to the two neighboring representable numbers, i.e., the rounded update is unbiased (Gupta et al., 2015; Connolly et al., 2021; Croci et al., 2022). Hence, SR is widely used when updates are applied in low precision (Gupta et al., 2015; Wang et al., 2018; Tseng et al., 2025; Ozkara et al., 2025; Zhao et al., 2025), and existing convergence analyses focus on SR.

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.

A few works have analyzed GD with SR under finite precision. Li et al. (2017) and Li and De Sa (2019) consider stochastic gradient descent (SGD) on a fixed-point grid and bound the expected optimization error of an averaged or randomly selected iterate by the rate of SGD on the reals plus additional terms due to rounding. This term depends on the grid spacing Δ, e.g., it is of order $\Delta\sqrt{d}$ in Li et al. (2017) and of order Δ^2 times the square of an ℓ_1 condition number in Li and De Sa (2019). A term of order $\Delta\sqrt{d}$ also appears in the recent analysis of mini-batch SGD with SR by Liu et al. (2025), and a term linear in d appears for Adam with SR by Ozkara et al. (2025). Xia et al. (2024) and Xia et al. (2025) consider GD under floating-point and fixed-point arithmetic, respectively, and show that the rate on the reals is preserved as long as each update is large compared to the spacing. In these works, rounding is typically treated as a perturbation of the exact update, and its effect is bounded from above. However, such upper bounds cannot determine whether the additional term is necessary, how long they take to reach a grid minimizer, or whether the spectrum still governs the convergence speed. For example, the bounds of Li et al. (2017) and Li and De Sa (2019) allow an error of order $\Delta^2 d$ on a quadratic objective, but they do not tell whether GD with SR attains this error, i.e., differs from a grid minimizer in $\Omega(d)$ coordinates, or reaches a grid minimizer after a few iterations. To our knowledge, existing analyses do not provide lower bounds that distinguish these behaviors for SR-GD on a fixed-point grid.

1.1 Contributions

In this work, we consider the simplest setting in which these questions can be studied exactly: GD on a quadratic objective $f(\mathbf{w}) = \frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{A}(\mathbf{w}-\mathbf{w}^\star)$ with a symmetric positive semidefinite $\mathbf{A}$, whose update is rounded to a fixed-point grid with spacing Δ by SR at every iteration. We call this iteration SR-GD. The iterates of SR-GD form a Markov chain on the grid, and we analyze the hitting time, i.e., the number of iterations until the chain reaches a grid minimizer, a minimizer of f over the grid. We first present our results for the case where f has an on-grid solution, i.e., f attains its minimum on the grid (Section 3).

We first show that unlike GD, the hitting time of SR-GD is not determined by the spectrum of $\mathbf{A}$.

- If $\mathbf{A}$ has a fixed spectrum in $[a, b]$ with eigenvalue variance $\Omega(1)$ and a random eigenbasis, the hitting time is $e^{\Omega(d)}$ with high probability (Theorem 1). This is tight, as the hitting time is $e^{O(d)}$ whenever the condition number of $\mathbf{A}$ is bounded (Theorem 2).
- Specifically, with high probability, the Hamming distance between the iterate and $\mathbf{w}^\star$ increases at every iteration until it reaches $\Omega(d)$, which takes $O(\log\log d)$ iterations, even from an initial Hamming distance of one. Then the distance stays there for $e^{\Omega(d)}$ iterations.
- Hence, even from an initial Hamming distance of $O(1)$, the optimization error increases from $O(\Delta^2)$ to $\Omega(\Delta^2 d)$ and stays there over this horizon, i.e., the $\Delta^2 d$ error in the existing upper bounds of Li et al. (2017) and Li and De Sa (2019) can be attained. In contrast, GD contracts the distance to $\mathbf{w}^\star$ by a constant factor at every iteration for any eigenbasis.
- For Gaussian least squares with n samples, where $\Omega(d) \le n \le o(d^2)$, the hitting time is $e^{\Omega(\min\{d, d^2/n\})}$, which is exponential in d for $n = \Theta(d)$ (Theorem 4). For example, $n = \lceil d^{3/2} \rceil$ gives an $e^{\Omega(\sqrt{d})}$ hitting time, even though the condition number is $1 + O(d^{-1/4})$.

We next show that the hitting time can be $O(\log d)$ under the same spectrum in $[a, b]$ via contraction.

- If some potential function that vanishes only at $\mathbf{w}^\star$ and is $O(d)$ on the grid contracts in expectation by a dimension-independent factor at each iteration, the hitting time is $O(\log d)$ (Proposition 5). This is tight: the hitting time is $\Omega(\log d)$ for any $\mathbf{A}$ with a unique on-grid solution if the initial Hamming distance is $\Omega(d)$ (Theorem 6).
- We provide sufficient conditions on $\mathbf{A}$ for such a contraction (Section 3.2), mainly $\big\||\mathbf{J}|\big\|_2 < 1$ for $\mathbf{J} = \mathbf{I} - \eta\mathbf{A}$, which unlike the GD condition $\|\mathbf{J}\|_2 < 1$ depends on the eigenbasis. For example, a diagonal $\mathbf{A}$ with the spectrum in $[a, b]$ satisfies it, i.e., rotating the eigenbasis alone changes the hitting time from $O(\log d)$ to $e^{\Omega(d)}$ or vice versa.
- For Gaussian least squares, the hitting time is $O(\log d)$ for $n = \Omega(d^2/\log\log d)$ and polynomial in d for $n = \Omega(d^2/\log d)$ (Theorem 7). Together with the previous results, this identifies qualitatively distinct regimes as n grows: an $e^{\Omega(d)}$ hitting time at $n = \Theta(d)$, a $d^{\Theta(1)}$ hitting time at $n = \Theta(d^2/\log d)$, and a $\Theta(\log d)$ hitting time for $n = \Omega(d^2/\log\log d)$.

We then show that the optimization error can be small even at a large Hamming distance if the rank of $\mathbf{A}$ is $o(d)$, unlike the slow case above, where the Hamming distance lower bounds the optimization error since the restricted eigenvalue of $\mathbf{A}$ is bounded below.

- If $\mathbf{A}$ has rank $r = o(d)$ with nonzero eigenvalues of order d/r and a random eigenbasis, the optimization error of an SR-GD iterate drops to $O(\Delta^2 r)$ within $O(\log(d/r))$ iterations from an initial Hamming distance $\Omega(d)$, while the Hamming distance remains at least half of its initial value for $(d/r)^{1-\epsilon}$ iterations with probability $1 - o(1)$ (Theorem 8). Hence, the optimization error becomes small exponentially earlier than the hitting time.
- For Gaussian least squares with $n = o(d)$ samples, this gives an optimization error $O(\Delta^2 n)$ within $O(\log(d/n))$ iterations, while the squared Euclidean distance to $\mathbf{w}^\star$ stays $\Omega(\Delta^2 d)$ for $(d/n)^{1-\epsilon}$ iterations (Corollary 9).

We finally consider the case where f has an off-grid solution (Section 4), where a grid minimizer need not be absorbing, since the gradient need not vanish there. We show that the exponential lower bound above extends to off-grid solutions when each grid minimizer has at most a small constant fraction of coordinates on the grid boundary (Theorem 10), whereas the $O(\log d)$ upper bound does not. Specifically, the location of $\mathbf{w}^\star$ can change the hitting time from $O(\log d)$ to $e^{\Omega(d)}$ with the same $\mathbf{A}$, unique grid minimizer, and initializer (Theorem 12).

2 Preliminaries

2.1 Notation

For a vector $\mathbf{z}$ and a matrix $\mathbf{B}$, z_i and B_{ij} denote the i-th coordinate and the (i,j)-th entry, respectively. Let $\mathbb{N}$ denote the positive integers, and write $[m] \triangleq \{1, \dots, m\}$ for $m \in \mathbb{N}$. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality, and for $\mathcal{Z} \subset \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$, we write $\mathcal{Z} - \mathbf{v} \triangleq \{\mathbf{z} - \mathbf{v} : \mathbf{z} \in \mathcal{Z}\}$. For $\mathbf{z} \in \mathbb{R}^d$, define $\|\mathbf{z}\|_0 \triangleq |\{i \in [d] : z_i \neq 0\}|$. We write $\mathbf{0}$ for a zero vector or matrix, $\mathbf{1}$ for an all-one vector, and $\mathbf{I}_d$ for the $d \times d$ identity matrix, with dimensions understood from context. We use $\langle \cdot, \cdot \rangle$ for the standard inner product and $\|\cdot\|_p$ for vector ℓ_p norms and their induced matrix norms, where $1 \le p \le \infty$. Absolute values of vectors and matrices are entrywise. We write $\mathbf{u} \le \mathbf{v}$ for vectors if $u_i \le v_i$ for every i, and $\mathbf{B} \preceq \mathbf{C}$ for symmetric matrices if $\mathbf{C} - \mathbf{B}$ is positive semidefinite. We also use $<$ and $\prec$ similarly. All logarithms are natural unless a base is specified.

For a symmetric matrix $\mathbf{B}$, $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ denote its smallest and largest eigenvalues. For $\mathbf{B} \succ \mathbf{0}$, define $\kappa(\mathbf{B}) \triangleq \lambda_{\max}(\mathbf{B})/\lambda_{\min}(\mathbf{B})$. We also use the smallest eigenvalue restricted to a set $\mathcal{Z} \subset \mathbb{R}^d$ containing a nonzero vector, $\lambda_{\min}(\mathbf{B}; \mathcal{Z}) \triangleq \inf_{\mathbf{z} \in \mathcal{Z} \setminus \{\mathbf{0}\}} \mathbf{z}^\top \mathbf{B}\mathbf{z}/\|\mathbf{z}\|_2^2$, so that $\lambda_{\min}(\mathbf{B}; \mathbb{R}^d) = \lambda_{\min}(\mathbf{B})$. We use tr, $\ker$, and range for the trace, nullspace, and column space, respectively. The notation $\mathrm{diag}(a_1, \dots, a_d)$ denotes the diagonal matrix with these entries, whereas $\mathrm{diag}(\mathbf{B})$ sets the off-diagonal entries of a square matrix $\mathbf{B}$ to zero. Let $\mathcal{O}(d) \triangleq \{\mathbf{Q} \in \mathbb{R}^{d \times d} : \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_d\}$ be the orthogonal group. Then, the uniform distribution over $\mathcal{O}(d)$ ($\mathrm{Unif}(\mathcal{O}(d))$) is the Haar measure on $\mathcal{O}(d)$, i.e., the unique probability measure on $\mathcal{O}(d)$ that is invariant under multiplication by any fixed orthogonal matrix.

2.2 Fixed-point grid and stochastic rounding

For $M \in \mathbb{N}$ and $\Delta > 0$, define the grid and its interior grid points by

$$\mathcal{G} \triangleq \{j\Delta : j \in \mathbb{Z},\ |j| \le M\}, \quad \mathcal{G}^\circ \triangleq \mathcal{G} \setminus \{\pm M\Delta\}, \quad \mathcal{W} \triangleq \mathcal{G}^d, \quad \mathcal{W}^\circ \triangleq (\mathcal{G}^\circ)^d.$$

We fix M and Δ throughout the paper. Let $\Pi(u) \triangleq \min\{M\Delta, \max\{-M\Delta, u\}\}$ be the clipping map. For $u \in \mathbb{R}$, let $\lfloor u \rfloor_{\mathcal{G}}$ and $\lceil u \rceil_{\mathcal{G}}$ be the largest grid point at most $\Pi(u)$ and the smallest grid point at least $\Pi(u)$, respectively. Stochastic rounding (SR) is the random map

$$\mathrm{SR}(u) \triangleq \begin{cases} \lfloor u \rfloor_{\mathcal{G}} & \text{with probability } (\lceil u \rceil_{\mathcal{G}} - \Pi(u))/\Delta, \\ \lceil u \rceil_{\mathcal{G}} & \text{otherwise,} \end{cases}$$

so that $\mathrm{SR}(u) = \Pi(u)$ if $\Pi(u) \in \mathcal{G}$ and $\mathbb{E}[\mathrm{SR}(u)] = \Pi(u)$. For vectors, clipping and rounding act coordinatewise. Rounding uses independent randomness across coordinates and iterations.

2.3 GRADIENT DESCENT WITH STOCHASTIC ROUNDING ON QUADRATIC OBJECTIVES

Quadratic objectives and GD. We study convex quadratic objectives of the following form:

$$f(\mathbf{w}) \triangleq \tfrac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{A}(\mathbf{w}-\mathbf{w}^\star), \qquad \mathbf{A} \succeq \mathbf{0}, \qquad \lambda_{\max}(\mathbf{A}) > 0, \tag{1}$$

where $\mathbf{w}^\star \in \mathbb{R}^d$ is a (possibly non-unique) minimizer of f, and $\mathbf{A}$ is the Hessian of f. To minimize f over the continuous space $\mathbb{R}^d$, gradient descent (GD) moves the iterate along the negative gradient with a step size $\eta > 0$. We fix $S \in (0,1)$ and use $\eta \triangleq S/\lambda_{\max}(\mathbf{A})$, for which GD converges for every initializer. Starting from $\mathbf{u}^{(0)} \in \mathbb{R}^d$, GD iterates

$$\mathbf{u}^{(t+1)} \triangleq \mathbf{u}^{(t)} - \eta\mathbf{A}(\mathbf{u}^{(t)}-\mathbf{w}^\star) = \mathbf{w}^\star + \mathbf{J}(\mathbf{u}^{(t)}-\mathbf{w}^\star), \qquad \mathbf{J} \triangleq \mathbf{I}_d - \eta\mathbf{A}.$$

Here, GD satisfies $\|\mathbf{u}^{(t)}-\mathbf{w}^\star\|_2 \le (1-S/\kappa(\mathbf{A}))^t\|\mathbf{u}^{(0)}-\mathbf{w}^\star\|_2$ and $f(\mathbf{u}^{(t)}) \le (1-S/\kappa(\mathbf{A}))^{2t} f(\mathbf{u}^{(0)})$ for $\mathbf{A} \succ \mathbf{0}$, i.e., its convergence speed is determined by the spectrum of $\mathbf{A}$.

GD with stochastic rounding (SR-GD). SR-GD is GD whose iterates are constrained to the grid $\mathcal{W}$. Since the GD update is not on the grid in general, SR-GD rounds it back to $\mathcal{W}$ by stochastic rounding. This models low-precision training in which the weights are stored in a fixed-point format (Gupta et al., 2015; Li and De Sa, 2019). Starting from $\mathbf{w}^{(0)} \in \mathcal{W}$, the iterates satisfy

$$\mathbf{w}^{(t+1)} \triangleq \mathrm{SR}\big(\mathbf{w}^{(t)} - \eta\mathbf{A}(\mathbf{w}^{(t)}-\mathbf{w}^\star)\big) = \mathrm{SR}\big(\mathbf{w}^\star + \mathbf{J}(\mathbf{w}^{(t)}-\mathbf{w}^\star)\big). \tag{2}$$

Grid minimizers and optimization error. Since the iterates lie in $\mathcal{W}$, we measure the optimization error relative to the minimum over the grid. Define

$$f^\star_{\mathcal{W}} \triangleq \min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w}), \qquad \mathcal{M} \triangleq \arg\min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w}),$$

and call each element of $\mathcal{M}$ a grid minimizer. The optimization error at $\mathbf{w} \in \mathcal{W}$ is $f(\mathbf{w}) - f^\star_{\mathcal{W}}$. We say that f has an *on-grid solution* if $f^\star_{\mathcal{W}} = 0$, i.e., f attains its minimum on the grid, and an *off-grid solution* otherwise. For an off-grid solution, measuring the error relative to $f(\mathbf{w}^\star) = 0$ would give a trivial lower bound instead, e.g., if every coordinate of every minimizer is at distance $\Delta/2$ from $\mathcal{G}$, then $f(\mathbf{w}) \ge \lambda_{\min}(\mathbf{A})\Delta^2 d/8$ for every $\mathbf{w} \in \mathcal{W}$.

Hitting time and Hamming distance. Analogous to the convergence of GD, we analyze the hitting time of SR-GD, i.e., the number of iterations until SR-GD reaches a grid minimizer, which is the first iteration with zero optimization error. To this end, we use the Hamming distance to $\mathcal{M}$, since each coordinate of an iterate either matches a grid minimizer or not. Define

$$h(\mathbf{w}) \triangleq \min_{\mathbf{v}\in\mathcal{M}} \|\mathbf{w}-\mathbf{v}\|_0, \qquad h^{(t)} \triangleq h(\mathbf{w}^{(t)}), \qquad \tau_{\mathcal{M}} \triangleq \inf\{t \ge 0 : h^{(t)} = 0\},$$

with $\inf \varnothing \triangleq \infty$. The Hamming distance also lower bounds the optimization error, since each incorrect coordinate deviates by at least Δ. For an on-grid solution with $\mathbf{w}^\star \in \mathcal{W}$, every $\mathbf{w} \in \mathcal{W}$ satisfies

$$f(\mathbf{w}) \ge \frac{1}{2}\lambda_{\min}(\mathbf{A}; \mathcal{W}-\mathbf{w}^\star)\|\mathbf{w}-\mathbf{w}^\star\|_2^2 \ge \frac{\Delta^2}{2}\lambda_{\min}(\mathbf{A}; \mathcal{W}-\mathbf{w}^\star)h(\mathbf{w}). \tag{3}$$

We note that for every $\mathbf{A}$, SR-GD reaches $\mathcal{M}$ in finite time almost surely and in expectation; see Proposition 20 in Appendix.

3 SR-GD WITH ON-GRID SOLUTIONS

In this section, we consider f with an on-grid solution, where each statement specifies a minimizer $\mathbf{w}^\star$ of f on the grid. Throughout this section, $\mathbf{A}$ denotes the matrix in the quadratic objective f, which is the Hessian of f (see Eq. (1)). We first show that SR-GD is slow for most Hessians with a given spectrum (Section 3.1) and that it is fast under contraction (Section 3.2), by analyzing how the Hamming distance changes at each iteration. We then show that the optimization error can be small at a large Hamming distance for rank-deficient Hessians (Section 3.3), i.e., the Hamming distance does not always determine the optimization error. Each result is applied to Gaussian least squares: for $\mathbf{X} \in \mathbb{R}^{n\times d}$ with independent $\mathcal{N}(0,1)$ entries and $\mathbf{w}^\star$ fixed independently of $\mathbf{X}$,

$$f(\mathbf{w}) \triangleq \frac{1}{2n}\|\mathbf{X}(\mathbf{w}-\mathbf{w}^\star)\|_2^2 = \frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \frac{\mathbf{X}^\top\mathbf{X}}{n}(\mathbf{w}-\mathbf{w}^\star). \tag{4}$$

For a fixed $\mathbf{A}$, probabilities are over rounding, and for a random $\mathbf{A}$, the source of randomness is specified in each statement. All constants, including those hidden in O, Ω, Θ, depend only on M, S and the constants in the corresponding statement.

3.1 EXPONENTIAL HITTING TIMES AND PERSISTENT OPTIMIZATION ERROR

We first show that SR-GD is slow for most Hessians with a given spectrum: if the eigenbasis is random, the hitting time is exponential in d with high probability, even though all eigenvalues lie between two fixed positive constants, for which GD converges within $O(\log d)$ iterations. The following theorem establishes this behavior for a Haar-random eigenbasis, where the proof is deferred to Appendix B.1.3.

Theorem 1. *Let $0 < a < b$ and $v > 0$ be constants, and fix $\mathbf{w}^\star \in \mathcal{W}^\circ$. Let $\mathbf{\Lambda} \triangleq \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ be deterministic, with $\lambda_i \in [a, b]$ for every $i \in [d]$ and $d^{-1}\sum_{i=1}^d (\lambda_i - \bar{\lambda})^2 \geq v^2$, where $\bar{\lambda} \triangleq d^{-1}\operatorname{tr}(\mathbf{\Lambda})$. Let $\mathbf{A} \triangleq \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} \sim \operatorname{Unif}(\mathcal{O}(d))$. There exist $\rho \in (0,1)$, $c > 0$ and $d_0 \in \mathbb{N}$, depending only on M, S, a, b, v, such that the following holds for every $d \geq d_0$. For every $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$, with probability at least $1 - e^{-c\sqrt{d}}$ over $\mathbf{U}$ and rounding,*

$$h^{(t+1)} \geq \min\left\{\rho d,\ 2\sqrt{\rho d\, h^{(t)}}\right\} \quad \text{for every } 0 \leq t < e^{\rho d}.$$

In particular, $\tau_{\mathcal{M}} > e^{\rho d}$ on this event.

Since $\mathbf{A} \succ \mathbf{0}$ in Theorem 1, $\mathbf{w}^\star$ is the unique grid minimizer and $h(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|_0$. By Theorem 1, the Hamming distance increases from h to at least $2\sqrt{\rho d\, h}$ at each iteration until it reaches ρd, which takes at most $\lceil \log_2(1 + \frac{1}{2}\log_2(\rho d))\rceil = O(\log\log d)$ iterations even when $h^{(0)} = 1$. Then the Hamming distance stays at least ρd through time $e^{\rho d}$. Since $\mathbf{A} \succeq a\mathbf{I}_d$, Eq. (3) gives $f(\mathbf{w}^{(t)}) \geq \frac{a}{2}\Delta^2 h^{(t)}$, and hence, the optimization error increases from $O(\Delta^2)$ when $h^{(0)} = O(1)$ to $\Omega(\Delta^2 d)$ and stays $\Omega(\Delta^2 d)$ over the same horizon. This shows that the $\Delta^2 d$ error in the upper bounds of Li et al. (2017) and Li and De Sa (2019) can be attained and persist for exponentially many iterations. In contrast, GD satisfies $\|\mathbf{u}^{(t)} - \mathbf{w}^\star\|_2 \leq (1 - Sa/b)^t \|\mathbf{u}^{(0)} - \mathbf{w}^\star\|_2$ for every eigenbasis and reduces the optimization error to ε within $O(\log(d/\varepsilon))$ iterations. We present the explicit constants ρ, c, d_0 for Theorem 1 in Appendix B.1.3.

Tightness of Theorem 1. The following theorem shows that the exponential lower bound in Theorem 1 is tight up to constants in the exponent. Its proof is given in Appendix B.2.2.

Theorem 2. *Let $K \geq 1$ be a constant, and fix $\mathbf{w}^\star \in \mathcal{W}$. There exists $\beta > 0$, depending only on M, S, K, such that for every $d \geq 1$, every $\mathbf{A} \succ \mathbf{0}$ with $\kappa(\mathbf{A}) \leq K$, and every initializer $\mathbf{w}^{(0)} \in \mathcal{W}$,*

$$\Pr(\tau_{\mathcal{M}} > e^{\beta d} \mid \mathbf{A}, \mathbf{w}^{(0)}) \leq e^{-d}.$$

Theorem 2 holds for every $\mathbf{A} \succ \mathbf{0}$ with $\kappa(\mathbf{A})$ bounded by a constant, i.e., for every eigenbasis, whereas the lower bound of Theorem 1 holds for most eigenbases. Under the assumptions of Theorem 1, the hitting time is therefore $e^{\Theta(d)}$ with probability at least $1 - e^{-\Omega(\sqrt{d})}$ over $\mathbf{U}$ and rounding.

Proof sketch of Theorem 1. To prove Theorem 1, we examine how one SR-GD update changes the number of coordinates that differ from $\mathbf{w}^\star$, which equals the Hamming distance h when $\mathbf{w}^\star$ is the unique grid minimizer. Since $\mathbf{w}^\star \in \mathcal{W}^\circ$, the i-th coordinate after one update from $\mathbf{w}$ differs from $w_i^\star$ exactly when rounding moves it away from $w_i^\star$, which has probability $\min\{1, |(\mathbf{J}(\mathbf{w} - \mathbf{w}^\star))_i|/\Delta\}$ (see Eq. (2)). Hence, for fixed $\mathbf{A}$ and $\mathbf{w} \in \mathcal{W}$,

$$\bar{h}(\mathbf{w}) \triangleq \mathbb{E}[\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|_0 \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}] = \sum_{i=1}^d \min\left\{1, \frac{|(\mathbf{J}(\mathbf{w} - \mathbf{w}^\star))_i|}{\Delta}\right\}.$$

For GD, the error $\mathbf{u} - \mathbf{w}^\star$ contracts since $\|\mathbf{J}\|_2 < 1$. However, under SR, the off-diagonal part $\mathbf{K} \triangleq \mathbf{J} - \operatorname{diag}(\mathbf{J})$ of $\mathbf{J}$ creates new errors: for $\mathbf{z} \triangleq \mathbf{w} - \mathbf{w}^\star$, a coordinate with $z_i = 0$ becomes incorrect after one update with probability $\min\{1, |(\mathbf{K}\mathbf{z})_i|/\Delta\}$, since $(\mathbf{J}\mathbf{z})_i = (\mathbf{K}\mathbf{z})_i$ for such i. The following proposition lower bounds $\bar{h}(\mathbf{w})$ when $\mathbf{K}$ spreads every sparse $\mathbf{z}$ over many coordinates.

Proposition 3. *Let $\rho \triangleq (1 - S^2)/(144M^2)$, let $\mathbf{w}^\star \in \mathcal{W}^\circ$, and suppose that for some $m > 0$ and $r \in [d]$,*

$$\|\mathbf{K}\mathbf{z}\|_1 \geq m\|\mathbf{z}\|_2 \quad \text{for every } \mathbf{z} \in \mathcal{W} - \mathbf{w}^\star \text{ with } 1 \leq \|\mathbf{z}\|_0 \leq r. \tag{5}$$

If $k \triangleq \rho \min\{m^2, r\} \geq 1$, then for every $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$,

$$\bar{h}(\mathbf{w}) \geq 4\sqrt{k \min\{\|\mathbf{w} - \mathbf{w}^\star\|_0, k\}}. \tag{6}$$

Although Eq. (5) concerns only $\mathbf{z}$ with $\|\mathbf{z}\|_0 \le r$, Eq. (6) holds at every $\mathbf{w} \neq \mathbf{w}^\star$. For $\|\mathbf{z}\|_0 \le r$, Eq. (5) lower bounds the expected number of coordinates with $z_i = 0$ that become incorrect, and for $\|\mathbf{z}\|_0 > r$, a direct bound $\bar{h}(\mathbf{w}) \ge \Omega(\|\mathbf{z}\|_0)$ suffices. We present the full proof of Proposition 3 in Appendix B.3. We also note that ρ can be improved to $\Omega(1/\log M)$; see Appendix F.

Under the spectral assumptions of Theorem 1, Eq. (5) holds with $m = \Theta(\sqrt{d})$ and $r = \Theta(d)$ with probability at least $1 - e^{-\Omega(d)}$ over $\mathbf{U}$, since a Haar-random rotation spreads each row of $\mathbf{K}$ over all coordinates with entries of order $1/\sqrt{d}$, so that $\|\mathbf{K}\mathbf{z}\|_1 = \Theta(\sqrt{d}\|\mathbf{z}\|_2)$ for sparse $\mathbf{z}$. Here, the eigenvalue variance is necessary, since $\mathbf{K} = \mathbf{0}$ if $\mathbf{\Lambda}$ is a multiple of $\mathbf{I}_d$. Hence, Proposition 3 gives Eq. (6) with $k = \Theta(d)$, and since $\mathbf{w}^\star$ is the unique grid minimizer, $h^{(t+1)}$ is a sum of independent Bernoulli variables with mean $\bar{h}(\mathbf{w}^{(t)}) \ge 4\sqrt{k \min\{h^{(t)}, k\}}$ conditional on $\mathbf{A}$ and $\mathbf{w}^{(t)}$. Chernoff bounds and a union bound over e^k iterations then give Theorem 1. We present the full proof of Theorem 1 in Appendix B.1.3.

Gaussian least squares with $\Omega(d) \le n \le o(d^2)$. Proposition 3 also applies to Gaussian least squares (Eq. (4)). Since the off-diagonal entries of $\mathbf{A}$ are of order $1/\sqrt{n}$, Eq. (5) holds with $m = \Theta(\min\{d/\sqrt{n}, \sqrt{d}\})$ and $r = \Theta(d)$ in Proposition 3 with high probability over $\mathbf{X}$, and hence, Proposition 3 gives $k = \Theta(\min\{d, d^2/n\})$, which grows with d whenever $n = o(d^2)$.

Theorem 4. *Let $R > 0$ be a constant, and fix $\mathbf{w}^\star \in \mathcal{W}^\circ$. There exist $\rho \in (0,1)$, $a, c > 0$ and $d_0 \in \mathbb{N}$, depending only on M, S, R, such that the following holds for every $d \ge d_0$ and $d/R \le n \le \rho d^2$. Set $k \triangleq \rho \min\{d, d^2/n\}$. For every $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$, with probability at least $1 - e^{-c\sqrt{k}}$ over $\mathbf{X}$ and rounding, $\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \ge a$ and*

$$h^{(t+1)} \ge \min\left\{k,\, 2\sqrt{k h^{(t)}}\right\} \quad \textit{for every } 0 \le t < e^k.$$

In particular, $\tau_{\mathcal{M}} > e^k$ on this event.

In Theorem 4, $\mathbf{w}^\star$ is the unique grid minimizer with probability one over $\mathbf{X}$, even when $n < d$ (Lemma 25 in Appendix). Theorem 4 gives the same conclusions as Theorem 1 with ρd replaced by k, where $k = \Theta(d)$ for $n = \Theta(d)$. Here, the lower bound e^k grows with d if and only if $n = o(d^2)$: for $n = \Theta(d^2)$, Theorem 4 gives only a constant lower bound, and we will show in Section 3.2 that the hitting time is then $O(\log d)$. In particular, the optimization error stays $\Omega(\Delta^2 k)$ by Eq. (3), even when $n < d$, where $\mathbf{A}$ is rank deficient but its restricted eigenvalue is bounded below.

Theorem 4 also shows that the hitting time can be exponential in a power of d even as $\kappa(\mathbf{A})$ approaches one, which is the regime where GD is fastest. For $n = \lceil d^{3/2} \rceil$, the hitting time is at least $e^{\Omega(\sqrt{d})}$ with probability at least $1 - e^{-\Omega(d^{1/4})}$, while $\kappa(\mathbf{A}) \le \left(\frac{1+2d^{-1/4}}{1-2d^{-1/4}}\right)^2 = 1 + O(d^{-1/4})$ with probability at least $1 - 2e^{-d/2}$ (Lemma 22 in Appendix). We give the explicit constants ρ, a, c, d_0 for Theorem 4 and present its full proof in Appendix B.4.2.

3.2 Logarithmic hitting times via contraction

The exponential hitting times in Section 3.1 come from the repulsion in Proposition 3: the expected Hamming distance grows whenever it is small. SR-GD can instead contract, as GD does under $\|\mathbf{J}\|_2 < 1$. To capture this, we call $\psi : \mathcal{W} \to [0, \infty)$ a potential function if $\psi(\mathbf{w}^\star) = 0$ and $\psi(\mathbf{w}) \ge 1$ at every other grid point. If ψ contracts in expectation under one SR-GD update, then $\mathbb{E}[\psi(\mathbf{w}^{(t)})] \le q^t \psi(\mathbf{w}^{(0)})$ by iterating, and Markov's inequality bounds $\Pr(\tau_{\mathcal{M}} > t)$ since $\psi \ge 1$ off $\mathbf{w}^\star$. This is the multiplicative drift argument (Hajek, 1982; Doerr et al., 2010), and it gives the following. The proof of Proposition 5 is given in Appendix C.1.

Proposition 5. *Let $\mathbf{w}^\star \in \mathcal{W}$ be the unique grid minimizer, let ψ be a potential function, and let $q \in (0,1)$. Suppose that for every $\mathbf{w} \in \mathcal{W}$, $\mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] \le q\psi(\mathbf{w})$. Then, for every initializer $\mathbf{w}^{(0)} \in \mathcal{W}$ and integer $t \ge 0$,*

$$\Pr(\tau_{\mathcal{M}} > t \mid \mathbf{w}^{(0)}) \le \psi(\mathbf{w}^{(0)}) q^t.$$

Proposition 5 implies that if $\psi(\mathbf{w}) \le \xi d$ for every $\mathbf{w} \in \mathcal{W}$ with a constant $\xi > 0$, then for every constant $c > 0$, we have an $O(\log d)$ hitting time with probability at least $1 - d^{-c}$, where the constant factor in $O(\log d)$ depends only on c, ξ, q. This matches GD, which is within one grid spacing of $\mathbf{w}^\star$ in $O(\log d)$ iterations when $\kappa(\mathbf{A}) = O(1)$.

Diagonal Hessians. We first apply Proposition 5 to the diagonal Hessian $\mathbf{A} = \mathbf{\Lambda}$ in Theorem 1, which shows that with the eigenvalues held fixed, a different eigenbasis yields $O(\log d)$ hitting times. Consider the potential $\psi(\mathbf{w}) \triangleq \|\mathbf{w} - \mathbf{w}^\star\|_1/\Delta \le 2Md$. Before clipping and rounding, the i-th coordinate of the SR-GD update lies between w_i and $w_i^\star$. Then, stochastic rounding gives

$$\mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] = \Delta^{-1} \sum_{i=1}^{d} (1 - \eta\lambda_i)|w_i - w_i^\star| \le (1 - Sa/b)\psi(\mathbf{w}).$$

Together with Theorem 1, a Haar-random rotation of the eigenbasis alone increases the hitting time from $O(\log d)$ to $e^{\Omega(d)}$ with high probability for every fixed initializer $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$. To our knowledge, such a separation in hitting times is not captured by existing convergence bounds (Li et al., 2017; Li and De Sa, 2019; Xia et al., 2024; 2025).

General Hessians. We next give sufficient conditions on $\mathbf{A}$ for Proposition 5, whose proofs are given in Appendix C.3. The main condition is $\big\||\mathbf{J}|\big\|_2 \le \gamma$ for a constant $\gamma \in (0, 1)$, under which a weighted ℓ_1 potential $\langle \mathbf{y}, |\mathbf{w} - \mathbf{w}^\star| \rangle/\Delta$ with $\mathbf{y} \ge \mathbf{1}$ contracts and is $O(d)$ (Proposition 35 in Appendix); here, $|\mathbf{J}|$ is a matrix with $|\mathbf{J}|_{ij} = |\mathbf{J}_{ij}|$. Unlike the GD condition $\|\mathbf{J}\|_2 < 1$ for linear convergence, this condition depends on the eigenbasis: the diagonal Hessian above has $\big\||\mathbf{J}|\big\|_2 \le 1 - Sa/b$, whereas its Haar-random rotation has $\big\||\mathbf{J}|\big\|_2 = \Theta(\sqrt{d})$ with high probability (Appendix C.3.1). The following Hessians yield $O(\log d)$ hitting times.

- Diagonally dominant Hessians: $A_{ii} - \sum_{j \ne i} |A_{ij}| \ge \delta\lambda_{\max}(\mathbf{A})$ for every $i \in [d]$ and a constant $\delta > 0$. Then, $\big\||\mathbf{J}|\big\|_2 \le 1 - S\delta$.
- Hessians with nonpositive off-diagonal entries: $\mathbf{A} \succ \mathbf{0}$, $A_{ij} \le 0$ for $i \ne j$, and $\kappa(\mathbf{A}) = O(1)$. Then, $|\mathbf{J}| = \mathbf{J}$ and $\big\||\mathbf{J}|\big\|_2 = 1 - S/\kappa(\mathbf{A})$. This includes dense Hessians such as $(a + b)\mathbf{I}_d - (b/d)\mathbf{1}\mathbf{1}^\top$ for constants $a, b > 0$.
- Near-diagonal Hessians: $\|\mathbf{J}\|_2 \le \gamma$ and $\sqrt{d}\,\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 = O(1)$. Here, $\big\||\mathbf{J}|\big\|_2 < 1$ may fail, and we use a potential built from the exponential moment of $\tau_{\mathcal{M}}$ instead (Appendix C.3.2).
- Block-diagonal Hessians: $\|\mathbf{J}\|_2 \le \gamma$ and blocks of constant size. We use a similar potential (Appendix C.3.3).

Lower bound on hitting time. The logarithmic dependence on d cannot be improved when $h^{(0)} = \Omega(d)$: each iteration corrects at most a constant fraction of the incorrect coordinates in expectation, for every Hessian. The condition on $h^{(0)}$ is necessary, e.g., for the diagonal Hessian above, $h^{(0)} = 1$ gives $\psi(\mathbf{w}^{(0)}) \le 2M$ and hence an $O(1)$ hitting time with constant probability by Proposition 5.

Theorem 6. *Let $\omega \in (0, 1]$ be a constant, and fix $\mathbf{w}^\star \in \mathcal{W}$. There exist $\alpha, c > 0$ and $d_0 \in \mathbb{N}$, depending only on M, S, ω, such that, for every $d \ge d_0$, every $\mathbf{A}$ for which $\mathbf{w}^\star$ is the unique grid minimizer, and every $\mathbf{w}^{(0)} \in \mathcal{W}$ with $h^{(0)} \ge \omega d$,*

$$\Pr(\tau_{\mathcal{M}} > \alpha \log d \mid \mathbf{w}^{(0)}) \ge 1 - e^{-c\sqrt{d}}.$$

We give the explicit constants and the proof of Theorem 6 in Appendix C.4.

Gaussian least squares with $n = \Omega(d^2/\log d)$. For $n \ge Rd^2$ with a constant $R > 0$, $\mathbf{A}$ satisfies the near-diagonal condition above with probability at least $1 - 2e^{-d/2}$ over $\mathbf{X}$, since $\|\mathbf{A} - \mathbf{I}_d\|_2 = O(\sqrt{d/n})$ (Appendix C.5.1). For smaller n, the same potential still contracts, but its maximum grows with d^2/n: the proof bounds the hitting time by $e^{O(d^2/n)}(d^2/n + \log(1/\varepsilon)) + O(\log d)$ with probability at least $1 - \varepsilon$ over rounding (Appendix C.5.2). Substituting n gives the following theorem, where the proof is given in Appendix C.5.3.

Theorem 7. *Let $R > 0$ be a constant, and fix $\mathbf{w}^\star \in \mathcal{W}$. There exist $R_0 > 0$, depending only on M, S, and $\alpha, p > 0$ and $d_0 \in \mathbb{N}$, depending only on M, S, R, such that the following holds for every $d \ge d_0$ and $n \in \mathbb{N}$. With probability at least $1 - 2e^{-d/2}$ over $\mathbf{X}$, for every $\mathbf{w}^{(0)} \in \mathcal{W}$ and $\varepsilon \in (0, 1)$,*

(i) $\Pr(\tau_{\mathcal{M}} > \alpha \log(d/\varepsilon) \mid \mathbf{X}, \mathbf{w}^{(0)}) \le \varepsilon$ *if* $n \ge Rd^2$,
(ii) $\Pr(\tau_{\mathcal{M}} > \alpha \log d \mid \mathbf{X}, \mathbf{w}^{(0)}) \le 1/\log d$ *if* $n \ge R_0 d^2/\log\log d$, *and*
(iii) $\Pr(\tau_{\mathcal{M}} > \alpha d^p \log(d/\varepsilon) \mid \mathbf{X}, \mathbf{w}^{(0)}) \le \varepsilon$ *if* $n \ge Rd^2/\log d$.

Together with Theorem 6, Cases (i) and (ii) give $\tau_{\mathcal{M}} = \Theta(\log d)$ when $h^{(0)} = \Omega(d)$, with probability at least $1 - O(d^{-c})$ for every constant $c > 0$ and $1 - O(1/\log d)$, respectively. In Case (i), the constant factor in $O(\log d)$ depends on c. Case (iii) gives a polynomial upper bound down to $n = \Omega(d^2/\log d)$, where Theorem 4 gives a matching polynomial lower bound: for $n = \Theta(d^2/\log d)$, $k = \Theta(\log d)$ and hence $\tau_{\mathcal{M}} = d^{\Theta(1)}$ with probability at least $1 - e^{-\Omega(\sqrt{\log d})}$ over $\mathbf{X}$ and rounding, for every fixed $\mathbf{w}^\star \in \mathcal{W}^\circ$ and $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$ (Appendix C.5.4). In summary, as n grows from $\Theta(d)$ to $\Theta(d^2)$, the hitting time decreases from $e^{\Omega(d)}$ through $d^{\Theta(1)}$ at $n = \Theta(d^2/\log d)$ to $\Theta(\log d)$ for $h^{(0)} = \Omega(d)$. We give the explicit constants for Theorem 7 in Appendix C.5.3.

3.3 Small optimization error at large Hamming distance

The results in Section 3.1 translate a large Hamming distance into a large optimization error via Eq. (3), which requires the restricted eigenvalue $\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star)$ to be bounded below. However, this translation can fail for rank-deficient $\mathbf{A}$, where SR-GD can have a small optimization error while the Hamming distance stays proportional to d. The proof of the following theorem is given in Appendix D.3.

Theorem 8. *Let $0 < a \le b$ and $\omega \in (0, 1]$ be constants, assume $M \ge 2$, and fix $\mathbf{w}^\star \in \mathcal{W}$. There exist $\rho \in (0, 1)$ and $\alpha, \ell, c > 0$, depending only on M, S, a, b, ω, such that the following holds for every $d \ge 1$ and integer $1 \le r \le \rho d$. Let $\mathbf{A} \triangleq \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} \sim \mathrm{Unif}(\mathcal{O}(d))$ and $\mathbf{\Lambda} \triangleq \mathrm{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$ with $\lambda_i \in [ad/r, bd/r]$ for every $i \in [r]$. For every $\mathbf{w}^{(0)} \in \mathcal{W}$ with $h^{(0)} \ge \omega d$ and at least $d/2$ coordinates in $\mathcal{G}^\circ$, and every integer $T \ge t_\star \triangleq \lceil \alpha \log(2d/r) \rceil$, with probability at least $1 - c(\sqrt{r/d} + rT/d)$ over $\mathbf{U}$ and rounding,*

$$\min_{0 \le t \le t_\star} f(\mathbf{w}^{(t)}) \le \ell \Delta^2 r \quad \text{and} \quad \min_{0 \le t \le T} h^{(t)} \ge \frac{h^{(0)}}{2}.$$

In Theorem 8, $\mathbf{w}^\star$ is the unique grid minimizer with probability one over $\mathbf{U}$, since $\ker(\mathbf{A}) \cap (\mathcal{W} - \mathbf{w}^\star) = \{\mathbf{0}\}$ almost surely (Lemma 25 in Appendix). For $r = o(d)$, some iterate within $O(\log(d/r))$ iterations from $h^{(0)} = \Omega(d)$ has optimization error $O(\Delta^2 r) = o(\Delta^2 d)$, even though $d^{-1}\,\mathrm{tr}(\mathbf{A}) \in [a, b]$ as in Theorem 1, i.e., the rank of $\mathbf{A}$ rather than its average eigenvalue separates the two cases. Meanwhile, taking $T = \lfloor (d/r)^{1-\epsilon} \rfloor$ for a constant $\epsilon \in (0, 1)$ gives $\tau_{\mathcal{M}} > (d/r)^{1-\epsilon}$ with probability $1 - o(1)$, i.e., the optimization error becomes $O(\Delta^2 r)$ exponentially earlier than the first hit of $\mathbf{w}^\star$. Hence, for rank-deficient $\mathbf{A}$, neither the Hamming distance nor the hitting time determines the optimization error. A similar phenomenon occurs for GD over $\mathbb{R}^d$, whose objective converges geometrically to zero while $\mathbf{u}^{(t)} - \mathbf{w}^\star$ keeps its component in $\ker(\mathbf{A})$.

Gaussian least squares with $n = o(d)$. For $n \le d/16$, $\mathbf{A}$ has rank n almost surely, and its nonzero eigenvalues lie in $[d/(4n), 9d/(4n)]$ with probability at least $1 - 2e^{-d/32}$. Conditioned on them, $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{U} \sim \mathrm{Unif}(\mathcal{O}(d))$. Hence, Theorem 8 applies with $r = n$ (see Appendix D.4).

Corollary 9. *Let $\omega \in (0, 1]$ be a constant, assume $M \ge 2$, and fix $\mathbf{w}^\star \in \mathcal{W}$. There exists $\rho \in (0, 1/16]$, depending only on M, S, ω, such that for every $d \ge 1$ and integer $1 \le n \le \rho d$, the conclusions of Theorem 8 hold for Gaussian least squares with $r = n$, under the same assumptions on $\mathbf{w}^{(0)}$ and T.*

For $n = o(d)$, Corollary 9 gives an SR-GD iterate with optimization error $O(\Delta^2 n)$ within $O(\log(d/n))$ iterations. However, for every $T = o(d/n)$ (e.g., $T = \lfloor (d/n)^{1-\epsilon} \rfloor$ for a constant $\epsilon \in (0, 1)$), $h^{(t)} = \Omega(d)$ for all $t \le T$ with probability $1 - o(1)$. Hence, $\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|_2^2 \ge \Delta^2 h^{(t)} = \Omega(\Delta^2 d)$ for all $t \le T$, and T can be much longer than the $O(\log(d/n))$ iterations above. That is, SR-GD attains a small optimization error while remaining far from $\mathbf{w}^\star$. This does not contradict Theorem 4 which also covers $n < d$, where the error stays $\Omega(\Delta^2 d)$ for $n \ge d/R$ with constants depending on R. Indeed, for $n = d/R$, Corollary 9 shows that these constants must decay as R grows. We give the explicit constants for Corollary 9 in Appendix D.4.

4 SR-GD with off-grid solutions

In this section, we consider f with an off-grid solution, i.e., $f^\star_{\mathcal{W}} > 0$, where each statement specifies a minimizer $\mathbf{w}^\star \notin \mathcal{W}$ of f. Here, two important properties of the on-grid case are lost: (*i*) the grid minimizers need not be unique, and (*ii*) a grid minimizer $\mathbf{v} \in \mathcal{W}^\circ$ is absorbing only if $\mathbf{A}(\mathbf{v} - \mathbf{w}^\star) = \mathbf{0}$, since otherwise the GD step from $\mathbf{v}$ does not vanish and rounding can leave $\mathbf{v}$.

Exponential lower bound. The following theorem extends Theorem 1 to off-grid solutions, with the Hamming distance measured to $\mathcal{M}$. Its proof is given in Appendix E.1.3.

Theorem 10. *Let $0 < a < b$, $v > 0$ and $\theta \in (0, 1/8)$ be constants, and fix $\mathbf{w}^\star \in \mathbb{R}^d \setminus \mathcal{W}$. Let $\mathbf{\Lambda}$ and $\mathbf{A} \triangleq \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be as in Theorem 1. There exist $\rho \in (0, 1)$, $c > 0$, and $d_0 \in \mathbb{N}$, depending only on M, S, a, b, v, θ, such that for every $d \geq d_0$, with probability at least $1 - e^{-cd}$ over $\mathbf{U}$, the following holds. If*

$$\max_{\mathbf{v} \in \mathcal{M}} |\{i \in [d] : |v_i| = M\Delta\}| \leq \theta d, \tag{7}$$

then for every $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \mathcal{M}$, with probability at least $1 - e^{-c\sqrt{d}}$ over rounding,

$$h^{(t+1)} \geq \min\{\rho d, 2\sqrt{\rho d\, h^{(t)}}\} \quad \text{for every } 0 \leq t < e^{\rho d}.$$

Theorem 10 gives the same conclusions as Theorem 1 with the Hamming distance measured to $\mathcal{M}$: under Eq. (7), the Hamming distance reaches ρd within $O(\log \log d)$ iterations and stays there through time $e^{\rho d}$. The upper bound in Theorem 2 also extends up to a logarithmic factor in the exponent: $\tau_{\mathcal{M}} \leq e^{O(d \log d)}$ with probability at least $1 - e^{-d}$ over rounding (Proposition 45 in Appendix). However, unlike the on-grid case, a large Hamming distance to $\mathcal{M}$ does not by itself imply a large optimization error. For example, let $\mathbf{A} = \mathbf{I}_d$ and $\mathbf{w}^\star = (\Delta/2 - \epsilon)\mathbf{1}$ for a small $\epsilon > 0$. Then, $\mathbf{0}$ is the unique grid minimizer and $\Delta\mathbf{1}$ has Hamming distance d from it, but $f(\Delta\mathbf{1}) - f^\star_{\mathcal{W}} = d\Delta\epsilon$ is arbitrarily small, since $\mathbf{w}^\star$ is almost at the center of a grid cell and both corners are nearly equidistant from it. On the other hand, the implication holds if $\mathbf{w}^\star$ is close to the grid: if $\mathbf{A} \succ \mathbf{0}$ and $f^\star_{\mathcal{W}} \leq \lambda_{\min}(\mathbf{A})\Delta^2 k/32$, which holds, e.g., when some $\mathbf{v} \in \mathcal{W}$ satisfies $\|\mathbf{v} - \mathbf{w}^\star\|_2^2 \leq \Delta^2 k/(16\kappa(\mathbf{A}))$, then every $\mathbf{w} \in \mathcal{W}$ with $h(\mathbf{w}) \geq k$ satisfies $f(\mathbf{w}) - f^\star_{\mathcal{W}} \geq \lambda_{\min}(\mathbf{A})\Delta^2 k/4$ (Lemma 44 in Appendix).

Sufficient conditions for Eq. (7). In Theorem 10, the role of the condition $\mathbf{w}^\star \in \mathcal{W}^\circ$ in the on-grid case (Theorem 1) is replaced by Eq. (7), which allows a fraction of coordinates of grid minimizers to have boundary values. The following proposition gives two sufficient conditions in terms of $\mathbf{w}^\star$ alone. Its proof is given in Appendix E.2.2.

Proposition 11. *Let $K \geq 1$ and $\theta \in (0, 1)$ be constants, and suppose $\kappa(\mathbf{A}) \leq K$. Each of the following conditions implies Eq. (7):*

(i) $M > \frac{1}{2}\sqrt{K/\theta}$ *and* $\|\mathbf{w}^\star\|_\infty \leq (M - \frac{1}{2}\sqrt{K/\theta})\Delta$*;*
(ii) $\|\mathbf{w}^\star - \mathbf{u}\|_\infty \leq \Delta/(1 + \sqrt{K/\theta})$ *for some* $\mathbf{u} \in \mathcal{W}^\circ$.

Both conditions keep $\mathbf{w}^\star$ away from the boundary of $[-M\Delta, M\Delta]^d$, directly or through a nearby interior grid point, and every grid minimizer then lies close to $\mathbf{w}^\star$ since $\kappa(\mathbf{A}) \leq K$.

Hitting time and location of $\mathbf{w}^\star$. In contrast, the $O(\log d)$ hitting time under contraction depends on the location of $\mathbf{w}^\star$. The following theorem considers $\mathbf{A}$ with $\|\mathbf{A} - \mathbf{I}_d\|_\infty \leq C < 1/3$, for which $\big\||\mathbf{J}|\big\|_2 \leq 1 - S(1 - C)/(1 + C) < 1$, i.e., the contraction condition in Section 3.2 holds and the hitting time is $O(\log d)$ for every on-grid solution. Its proof is given in Appendix E.4.2.

Theorem 12. *Let $C \in (0, 1/3)$ be a constant. There exist $q, \rho \in (0, 1)$ and $a, c > 0$ depending only on S, C such that the following holds for every $d \geq 1$, every $\mathbf{A}$ with $\|\mathbf{A} - \mathbf{I}_d\|_\infty \leq C$, and $\mathbf{w}^{(0)} = \Delta\mathbf{1}$. In both cases below, $\mathbf{0}$ is the unique grid minimizer.*

(i) If $\mathbf{w}^\star = \varepsilon\Delta\mathbf{1}$ with $0 \leq \varepsilon \leq 1/(8d)$, then $\Pr(\tau_{\mathcal{M}} > t) \leq dq^t$ for every integer $t \geq 0$.
(ii) If $\mathbf{w}^\star = \Delta\mathbf{1}/4$, then, with probability at least $1 - e^{-cd}$ over rounding,

$$h^{(t)} \geq \rho d \quad \text{and} \quad f(\mathbf{w}^{(t)}) - f^\star_{\mathcal{W}} \geq a\Delta^2 d \quad \text{for every } 0 \leq t \leq e^{cd}.$$

Case (i) includes the on-grid solution $\varepsilon = 0$, and the $O(\log d)$ hitting time survives as long as the GD step at $\mathbf{0}$, $\eta\mathbf{A}\mathbf{w}^\star$, has ℓ_1 norm $O(\Delta)$, since the weighted ℓ_1 potential of Section 3.2 then still contracts. In Case (ii), this step moves every coordinate by about $S\Delta/4$, so each coordinate of an iterate near $\mathbf{0}$ rounds away from $\mathbf{0}$ with a constant probability, and the Hamming distance stays $\Theta(d)$. Unlike Section 3.1, this repulsion comes from the rounding noise at the grid minimizer, not from the off-diagonal part of $\mathbf{J}$, and the optimization error stays $\Omega(\Delta^2 d)$ although $f^\star_{\mathcal{W}} = \Theta(\Delta^2 d)$. The lower bound in Case (ii) is tight: the hitting time is $e^{\Theta(d)}$ with probability at least $1 - e^{-\Omega(d)}$ (Proposition 47 in Appendix).

AI USE STATEMENT

In this work, we utilized large language models (LLMs) to assist in three primary capacities. First, LLMs were employed to polish the academic writing, improve clarity, and correct grammatical errors. Second, AI-assisted search tools were used to navigate the literature and identify relevant related work. Finally, we used LLMs as an auxiliary tool to verify the correctness of our mathematical proofs. We emphasize that the core theoretical frameworks and proofs were originally formulated by the authors. All AI-generated suggestions and verifications were rigorously reviewed, and the authors take full responsibility for the originality, accuracy, and correctness of all contents and mathematical claims presented in this work.

REPRODUCIBILITY STATEMENT

The problem setup and stochastic rounding rule are specified in Section 2, with additional assumptions stated alongside the corresponding results in Section 3 and Section 4. Appendices A–F provide complete proofs and supporting lemmas, including explicit choices of constants and derivations of the probability bounds.

REFERENCES

Ikhlef Bechar. A bernstein-type inequality for stochastic processes of quadratic forms of gaussian variables. *arXiv preprint arXiv:0909.3595*, 2009.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

Michael P Connolly, Nicholas J Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021.

Matteo Croci, Massimiliano Fasi, Nicholas J Higham, Theo Mary, and Mantas Mikaitis. Stochastic rounding: implementation, error analysis and applications. *Royal Society Open Science*, 9(3): 211631, 2022.

Benjamin Doerr, Daniel Johannsen, and Carola Winzen. Multiplicative drift analysis. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 1449–1456, 2010.

Eugene Gover. Congruence and metrical invariants of zonotopes. *arXiv preprint arXiv:1401.4749*, 2014.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning (ICML)*, 2015.

Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Zheng Li and Christopher M De Sa. Dimension-free bounds for low-precision training. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Taowen Liu, Marta Andronic, Deniz Gunduz, and George Anthony Constantinides. Training with fewer bits: Unlocking edge LLMs training with stochastic rounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Elizabeth S Meckes. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

NVIDIA, Felix Abecassis, Anjulie Agrusa, Dong Ahn, Jonah Alben, Stefania Alborghetti, Michael Andersch, Sivakumar Arayandi, Alexis Bjorlin, Aaron Blakeman, Evan Briones, Ian Buck, Bryan Catanzaro, Muya Chang, Jinhang Choi, Mike Chrzanowski, Eric Chung, Victor Cui, Steve Dai, Bita Darvish Rouhani, Carlo del Mundo, Deena Donia, Burc Eryilmaz, Henry Estela, Abhinav Goel, Oleg Goncharov, Yugi Guvvala, Robert Hesse, Russell Hewett, Herbert Hum, Ujval Kapasi, Brucek Khailany, Mikail Khona, Nick Knight, Alex Kondratenko, Ronny Krashinsky, Ben Lanir, Simon Layton, Michael Lightstone, Daniel Lo, Paulius Micikevicius, Asit Mishra, Tim Moon, Deepak Narayanan, Chao Ni, Abhijit Paithankar, Satish Pasumarthi, Ankit Patel, Mostofa Patwary, Ashwin Poojary, Gargi Prasad, Sweta Priyadarshi, Yigong Qin, Xiaowei Ren, Oleg Rybakov, Charbel Sakr, Sanjeev Satheesh, Stas Sergienko, Pasha Shamis, Kirthi Shankar, Nishant Sharma, Mohammad Shoeybi, Michael Siu, Misha Smelyanskiy, Darko Stosic, Dusan Stosic, Bor-Yiing Su, Frank Sun, Nima Tajbakhsh, Shelby Thomas, Przemek Tredak, Evgeny Tsykunov, Gandhi Vaithilingam, Aditya Vavre, Rangharajan Venkatesan, Roger Waleffe, Qiyu Wan, Hexin Wang, Mengdi Wang, Lizzie Wei, Hao Wu, Evan Wu, Keith Wyss, Ning Xu, Jinze Xue, Charlene Yang, Yujia Zhai, Ruoxi Zhang, Jingyang Zhu, and Zhongbo Zhu. Pretraining large language models with NVFP4. *arXiv preprint arXiv:2509.25149*, 2026.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Xiao Sun, Jungwook Choi, Chia-Yu Chen, Naigang Wang, Swagath Venkataramani, Vijayalakshmi Viji Srinivasan, Xiaodong Cui, Wei Zhang, and Kailash Gopalakrishnan. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Albert Tseng, Tao Yu, and Youngsuk Park. Training llms with mxfp4. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, and Kailash Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Lu Xia, Stefano Massei, Michiel E Hochstenbach, and Barry Koren. On stochastic roundoff errors in gradient descent with low-precision computation. *Journal of Optimization Theory and Applications*, 200(2):634–668, 2024.

Lu Xia, Stefano Massei, and Michiel E Hochstenbach. On the convergence of the gradient descent method with stochastic fixed-point rounding errors under the polyak–łojasiewicz inequality. *Computational Optimization and Applications*, 90(3):753–799, 2025.

Pedram Zamirai, Jian Zhang, Christopher R Aberger, and Christopher De Sa. Revisiting bfloat16 training. *arXiv preprint arXiv:2010.06192*, 2020.

Kaiyan Zhao, Tsuguchika Tabaru, Kenichi Kobayashi, Takumi Honda, Masafumi Yamazaki, and Yoshimasa Tsuruoka. Direct quantized training of language models with stochastic rounding. In *Asian Conference on Machine Learning (ACML)*, 2025. arXiv:2412.04787.

A SHARED ROUNDING, PROBABILITY, AND DISTRIBUTIONAL ESTIMATES

A.1 STOCHASTIC ROUNDING AND QUADRATIC IDENTITIES

For quadratic SR-GD, the Hessian is sampled once and held fixed. Define

$$\mathcal{F}_t \triangleq \sigma(\mathbf{A}, \mathbf{w}^{(0)}, \dots, \mathbf{w}^{(t)}). \tag{A.1}$$

The random draws used to implement rounding at time $t+1$ are independent of $\mathcal{F}_t$ and of one another. Conditional on $\mathcal{F}_t$, the coordinates of $\mathbf{w}^{(t+1)}$ are therefore independent.

Lemma 13. *Fix $u \in \mathbb{R}$, put $v = \Pi(u)$, and let $Z = \mathrm{SR}(u)$. For every $g \in \mathcal{G}$,*

$$\mathbb{E}[Z] = v, \qquad \mathbb{E}[|Z-g|] = |v-g| \le |u-g|, \tag{A.2}$$

$$\Pr(Z \neq g) = \min\{1, |v-g|/\Delta\} \le \min\{1, |u-g|/\Delta\}, \tag{A.3}$$

$$\mathrm{Var}(Z) \le \Delta^2/4, \qquad \mathrm{Var}(Z) \le \Delta|v-g| \le \Delta|u-g|, \tag{A.4}$$

$$\mathbb{E}[(Z-g)^2] = (v-g)^2 + \mathrm{Var}(Z). \tag{A.5}$$

Here $\mathrm{Var}(Z) = \mathbb{E}[(Z-\mathbb{E}[Z])^2]$. The inequality in Eq. (A.3) is an equality if $g \in \mathcal{G}^\circ$. More explicitly,

$$\Pr(\mathrm{SR}(u) \neq g) = \begin{cases} \min\{1, |u-g|/\Delta\}, & g \in \mathcal{G}^\circ, \\ \min\{1, \max\{M\Delta - u, 0\}/\Delta\}, & g = M\Delta, \\ \min\{1, \max\{u + M\Delta, 0\}/\Delta\}, & g = -M\Delta. \end{cases}$$

For deterministic $\mathbf{u} \in \mathbb{R}^d$, $\mathbf{w}^\star \in \mathcal{W}$, and $\mathbf{0} \le \mathbf{y}$,

$$\mathbb{E}\left[\sum_{i=1}^d y_i |\mathrm{SR}(\mathbf{u})_i - w_i^\star|\right] = \sum_{i=1}^d y_i |\Pi(u_i) - w_i^\star| \le \sum_{i=1}^d y_i |u_i - w_i^\star|. \tag{A.6}$$

Proof. For $v = a + \theta\Delta$ with $a, a+\Delta \in \mathcal{G}$ and $\theta \in [0,1]$,

$$\Pr(Z = a) = 1-\theta, \qquad \Pr(Z = a+\Delta) = \theta, \qquad \mathbb{E}[Z] = v, \qquad \mathrm{Var}(Z) = \theta(1-\theta)\Delta^2.$$

Since g is not in the interior of $(a, a+\Delta)$, $x \mapsto |x-g|$ is affine on $[a, a+\Delta]$ and $|v-g| \ge \Delta \min\{\theta, 1-\theta\}$. Consequently,

$$\mathbb{E}[|Z-g|] = |v-g| \le |u-g|, \qquad \mathrm{Var}(Z) \le \min\{\Delta^2/4, \Delta|v-g|\}.$$

If $|v-g| < \Delta$, the two-point law gives $\Pr(Z \neq g) = |v-g|/\Delta$; otherwise $\Pr(Z \neq g) = 1$. For $g \in \mathcal{G}^\circ$, clipping changes neither this probability nor $\min\{1, |u-g|/\Delta\}$. The two boundary cases follow by projection. The variance identity proves Eq. (A.5), and summing the absolute-distance identity gives Eq. (A.6). □

If $\mathcal{M} = \{\mathbf{w}^\star\}$, then $h(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|_0$. Conditional independence of the coordinates of $\mathbf{w}^{(t+1)}$ implies that $h^{(t+1)}$ is a sum of independent Bernoulli variables conditional on $\mathcal{F}_t$, with

$$\mathbb{E}[h^{(t+1)} \mid \mathcal{F}_t] = \bar{h}(\mathbf{w}^{(t)}). \tag{A.7}$$

For $\mathbf{w}^\star \in \mathcal{W}^\circ$, summing Eq. (A.3) over the coordinates gives

$$\bar{h}(\mathbf{w}) = \sum_{i=1}^d \min\left\{1, \frac{|(\mathbf{J}(\mathbf{w} - \mathbf{w}^\star))_i|}{\Delta}\right\}.$$

A.2 CONDITIONAL EXPECTATIONS OF HAMMING DISTANCE

Throughout this subsection, assume $\mathbf{w}^\star \in \mathcal{W}$. The spectral theorem gives

$$(1-S)\mathbf{I}_d \preceq \mathbf{J} \preceq \mathbf{I}_d, \qquad 1 - S \le J_{ii} \le 1, \qquad \|\mathbf{J}\|_2 = 1 - \eta\lambda_{\min}(\mathbf{A}) \le 1. \tag{A.8}$$

Since both $\mathbf{A}$ and $\mathrm{diag}(\mathbf{A})$ lie between $\mathbf{0}$ and $\lambda_{\max}(\mathbf{A})\mathbf{I}_d$ in the positive-semidefinite order, $\|\mathbf{K}\|_2 \le \eta\lambda_{\max}(\mathbf{A}) = S$. Consequently,

$$\|\mathbf{J}\mathbf{z} - \mathbf{z}\|_2 \le S\|\mathbf{z}\|_2, \qquad \|\mathbf{K}\mathbf{z}\|_2 \le S\|\mathbf{z}\|_2 \quad (\mathbf{z} \in \mathbb{R}^d). \tag{A.9}$$

For $\mathbf{z} \in \mathcal{W} - \mathbf{w}^\star$, each nonzero coordinate satisfies $\Delta \le |z_i| \le 2M\Delta$, so

$$\Delta^2\|\mathbf{z}\|_0 \le \|\mathbf{z}\|_2^2 \le 4M^2\Delta^2\|\mathbf{z}\|_0, \qquad \|\mathbf{z}\|_1 \le 2M\Delta\|\mathbf{z}\|_0. \tag{A.10}$$

Lemma 14. *For*

$$\beta \triangleq \frac{1-S^2}{4M^2}, \tag{A.11}$$

we have

$$\bar{h}(\mathbf{w}) \geq \beta \|\mathbf{w}-\mathbf{w}^\star\|_0 \qquad (\mathbf{w} \in \mathcal{W}). \tag{A.12}$$

Proof. Set $\mathbf{z} = \mathbf{w} - \mathbf{w}^\star$ and $\mathbf{u} = \Pi(\mathbf{w}^\star + \mathbf{J}\mathbf{z}) - \mathbf{w}^\star$. By Eq. (A.3), $p_i = \min\{1, |u_i|/\Delta\}$ satisfies $\bar{h}(\mathbf{w}) = \sum_i p_i$. Since $\Pi(\mathbf{w}) = \mathbf{w}$, nonexpansiveness of projection and Eq. (A.9) give $\|\mathbf{z}-\mathbf{u}\|_2 \leq \|\mathbf{z}-\mathbf{J}\mathbf{z}\|_2 \leq S\|\mathbf{z}\|_2$. Therefore,

$$(1-S^2)\|\mathbf{z}\|_2^2 \leq \|\mathbf{z}\|_2^2 - \|\mathbf{z}-\mathbf{u}\|_2^2 = \sum_i (2z_i u_i - u_i^2).$$

If $p_i < 1$, then $|z_i| \leq 2M\Delta$ gives $2z_iu_i - u_i^2 \leq 4M\Delta^2 p_i \leq 4M^2\Delta^2 p_i$. If $p_i = 1$, the same expression is at most $z_i^2 \leq 4M^2\Delta^2$. Summing and using $\|\mathbf{z}\|_2^2 \geq \Delta^2\|\mathbf{z}\|_0$ proves the claim. □

Lemma 15. *Suppose in addition that* $\mathbf{w}^\star \in \mathcal{W}^\circ$, $m > 0$, $r \geq 1$, *and*

$$\|\mathbf{K}\mathbf{z}\|_1 \geq m\|\mathbf{z}\|_2 \quad \textit{for all } \mathbf{z} \in \mathcal{W} - \mathbf{w}^\star \textit{ with } 1 \leq \|\mathbf{z}\|_0 \leq r. \tag{A.13}$$

Let $\beta = (1-S^2)/(4M^2)$.

(i) *For every* $\mathbf{w} \in \mathcal{W}$ *with* $1 \leq \|\mathbf{w}-\mathbf{w}^\star\|_0 \leq \min\{r, m^2/9\}$,

$$\bar{h}(\mathbf{w}) \geq \frac{m}{3}\sqrt{\|\mathbf{w}-\mathbf{w}^\star\|_0}. \tag{A.14}$$

(ii) *Every real* $1 \leq k \leq (\beta/36)\min\{r, m^2\}$ *satisfies*

$$\bar{h}(\mathbf{w}) \geq 4\sqrt{k \min\{\|\mathbf{w}-\mathbf{w}^\star\|_0, k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}). \tag{A.15}$$

Proof. We first convert the assumed ℓ_1 lower bound into a lower bound on $\bar{h}(\mathbf{w})$ when $\mathbf{w}-\mathbf{w}^\star$ has small support. For larger support, Lemma 14 supplies the required bound. The conversion uses the following inequality for the truncated coordinate magnitudes. For $\mathbf{x} \neq \mathbf{0}$, put $q_i = \min\{1, |x_i|/\Delta\}$. Cauchy–Schwarz gives

$$\|\mathbf{x}\|_1^2 \leq \left(\sum_i q_i\right)\left(\sum_{q_i>0} x_i^2/q_i\right), \qquad x_i^2/q_i \leq \Delta|x_i| + x_i^2 \quad (q_i > 0).$$

Hence

$$\sum_i \min\{1, |x_i|/\Delta\} \geq \frac{\|\mathbf{x}\|_1^2}{\Delta\|\mathbf{x}\|_1 + \|\mathbf{x}\|_2^2}. \tag{A.16}$$

Set $\mathbf{z} = \mathbf{w}-\mathbf{w}^\star$, $h = \|\mathbf{z}\|_0$, and $u = \|\mathbf{z}\|_2/\Delta \geq \sqrt{h}$. For $1 \leq h \leq \min\{r, m^2/9\}$, apply Eq. (A.16) to $\mathbf{x} = \mathbf{K}\mathbf{z}$. The fraction increases in $\|\mathbf{x}\|_1$ and decreases in $\|\mathbf{x}\|_2^2$. Also, $(\mathbf{J}\mathbf{z})_i = (\mathbf{K}\mathbf{z})_i$ if $z_i = 0$. Omitting the h nonzero coordinates of $\mathbf{z}$ therefore gives

$$\begin{aligned}\bar{h}(\mathbf{w}) &\geq \frac{m^2 u}{m + S^2 u} - h \geq \frac{m^2\sqrt{h}}{m + S^2\sqrt{h}} - h \\ &\geq \left(\frac{1}{1+S^2/3} - \frac{1}{3}\right) m\sqrt{h} \geq \frac{m}{3}\sqrt{h}.\end{aligned}$$

The second inequality uses monotonicity in u; the last two use $\sqrt{h} \leq m/3$ and $S < 1$. This proves (i).

For (ii), $k \leq \beta\min\{r, m^2\}/36$ and $\beta \leq 1/4$ imply

$$4\sqrt{k} \leq m/3, \qquad \beta r \geq 4k, \qquad \beta m^2/9 \geq 4k.$$

If $1 \leq h \leq \min\{r, m^2/9\}$, part (i) gives $\bar{h}(\mathbf{w}) \geq 4\sqrt{kh}$. Otherwise Lemma 14 gives $\bar{h}(\mathbf{w}) \geq \beta h \geq 4k$. □

A.3 Bounds for stochastic processes

A.3.1 A conditional-mean bound for an on-grid minimizer

Lemma 16. *Fix* **A** *with* $\mathcal{M} = \{\mathbf{w}^\star\}$ *and an initializer* $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$. *Suppose that, for some real* $k \geq 1$,

$$\bar{h}(\mathbf{w}) \geq 4\sqrt{k \min\{h(\mathbf{w}), k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}). \tag{A.17}$$

With probability at least $1 - \frac{99}{100} e^{-\sqrt{k}/16}$ *over rounding,*

$$h^{(t+1)} \geq \min\{k, 2\sqrt{k h^{(t)}}\} \quad (0 \leq t < e^k), \qquad \tau_{\mathcal{M}} > e^k. \tag{A.18}$$

All time indices are integers.

Proof. We require $h^{(t+1)} \geq 2\sqrt{k h^{(t)}}$ when $1 \leq h^{(t)} < k$, and $h^{(t+1)} \geq k$ when $h^{(t)} \geq k$. These two inequalities imply the asserted recurrence. To bound the probability of their first failure, define

$$\mathcal{B}_t \triangleq \{1 \leq h^{(t)} < k,\ h^{(t+1)} < 2\sqrt{k h^{(t)}}\} \cup \{h^{(t)} \geq k,\ h^{(t+1)} < k\}, \qquad \mathcal{E}_t \triangleq \bigcap_{s=0}^{t-1} \mathcal{B}_s^c.$$

Here $\mathcal{E}_0 = \Omega$ and $\mathcal{E}_t \in \mathcal{F}_t$. On $\mathcal{E}_t$, the distance remains positive through time t, and $\mathcal{B}_t$ is the event that the next transition violates the applicable inequality. Set $T = \lceil e^k \rceil$, so that the integer indices $0 \leq t < e^k$ are exactly $0, \ldots, T-1$.

Conditional on $\mathcal{F}_t$, the distance $h^{(t+1)}$ is a sum of independent Bernoulli variables with mean $\mu_t \triangleq \bar{h}(\mathbf{w}^{(t)})$. For any such sum Z with conditional mean μ, exponential Markov's inequality gives

$$\Pr(Z \leq a \mid \mathcal{F}_t) \leq \exp\{\lambda a + \mu(e^{-\lambda} - 1)\} \qquad (\lambda > 0), \tag{A.19}$$

where a is $\mathcal{F}_t$-measurable. Taking $(a, \lambda) = (\mu/2, \log 2)$ gives

$$\Pr(Z \leq \mu/2 \mid \mathcal{F}_t) \leq e^{-(1-\log 2)\mu/2} \leq e^{-\mu/8}. \tag{A.20}$$

Taking $(a, \lambda) = (k, \log 4)$ gives

$$\Pr(Z < k \mid \mathcal{F}_t) \leq e^{k \log 4 - 3\mu/4} \leq e^{-(3-2\log 2)k} \qquad (\mu \geq 4k). \tag{A.21}$$

Put $\alpha \triangleq 3 - 2\log 2$. The conditional-mean assumption and Eqs. (A.20) and (A.21) give

$$\Pr(h^{(t+1)} < 2\sqrt{k h^{(t)}} \mid \mathcal{F}_t) \leq e^{-\sqrt{k h^{(t)}}/2} \qquad (1 \leq h^{(t)} < k), \tag{A.22}$$

$$\Pr(h^{(t+1)} < k \mid \mathcal{F}_t) \leq e^{-\alpha k} \qquad (h^{(t)} \geq k). \tag{A.23}$$

The first inequality controls the transitions before the distance reaches k; the second controls subsequent transitions. On $\mathcal{E}_t$, each transition that remains below k at least doubles the distance. There are therefore fewer than $1 + \log_2 k$ transitions starting below k before its first occurrence. For $k \geq 16$, this number is at most $\sqrt{k} + 1$.

For $2 \leq k \leq 4$, either $h^{(0)} \geq k$, or $\mathcal{E}_1$ implies $h^{(1)} \geq k$, since $h^{(0)} \geq 1$ and $2\sqrt{k} \geq k$. For $4 < k < 16$, on $\mathcal{E}_2$ the distance reaches k within at most two transitions, since $2\sqrt{k(2\sqrt{k})} \geq k$. If $k = 1$, then $h^{(0)} \geq k$. The first conditional failure bound is at most $e^{-\sqrt{k}/2}$, because the distance is at least one. Before a first failure, this bound is needed at most once for $2 \leq k \leq 4$, twice for $4 < k < 16$, and $\sqrt{k} + 1$ times for $k \geq 16$. The second conditional failure bound is at most $e^{-\alpha k}$ and is needed at most T times. For $k = 1$, only the second bound is needed. Using these counts and $T \leq e^k + 1$ in

$$\Pr(\mathcal{E}_T^c) = \sum_{t=0}^{T-1} \mathbb{E}[\mathbb{1}\{\mathcal{E}_t\} \Pr(\mathcal{B}_t \mid \mathcal{F}_t)]$$

gives the following bounds:

$$e^{\sqrt{k}/16} \Pr(\mathcal{E}_T^c) \leq \begin{cases} (1 + e^{-1}) e^{-(\alpha-1)+1/16}, & k = 1, \\ e^{-7\sqrt{2}/16} + (1 + e^{-2}) e^{-2(\alpha-1)+\sqrt{2}/16}, & 2 \leq k \leq 4, \\ 2e^{-7/8} + (1 + e^{-4}) e^{-4(\alpha-1)+1/8}, & 4 < k < 16, \\ 5e^{-7/4} + (1 + e^{-16}) e^{-16(\alpha-1)+1/4}, & k \geq 16. \end{cases} \tag{A.24}$$

Indeed, $(1+e^{-k})e^{-(\alpha-1)k+\sqrt{k}/16}$ decreases for $k \geq 1$, and $(x+1)e^{-7x/16}$ decreases for $x \geq 4$. Each displayed bound is less than $99/100$.

It remains to consider $1 < k < 2$. If $h^{(0)} \geq k$, only the second conditional bound is needed, giving

$$e^{\sqrt{k}/16}\Pr(\mathcal{E}_T^c) \leq (1+e^{-k})e^{-(\alpha-1)k+\sqrt{k}/16} \leq (1+e^{-1})e^{-(\alpha-1)+1/16} < \frac{99}{100}.$$

Otherwise $h^{(0)} = 1$ and $\mathcal{B}_0 = \{h^{(1)} \leq 2\}$. Taking $(a, \lambda) = (2, \log 2)$ and $(1, \log 4)$ in Eq. (A.19), respectively, gives

$$\Pr(\mathcal{B}_0) \leq 4e^{-\mu_0/2} \leq 4e^{-2}, \qquad \Pr(h^{(t+1)} \leq 1 \mid \mathcal{F}_t) \leq 4e^{-3\mu_t/4} \leq 4e^{-3k} \quad (h^{(t)} \geq k).$$

On $\mathcal{B}_0^c$, the distance is at least three. Summing over the first subsequent failure and using $T-1 \leq e^k$ gives

$$\Pr(\mathcal{E}_T^c \mid \mathcal{B}_0^c) \leq 4(T-1)e^{-3k} \leq 4e^{-2k} \leq 4e^{-2}.$$

Consequently,

$$\Pr(\mathcal{E}_T^c) \leq 1-(1-4e^{-2})^2 < \frac{99}{100}e^{-\sqrt{2}/16} \leq \frac{99}{100}e^{-\sqrt{k}/16}.$$

Together with Eq. (A.24), this proves

$$\Pr(\mathcal{E}_T^c) \leq \frac{99}{100}e^{-\sqrt{k}/16} \qquad (k \geq 1). \tag{A.25}$$

On $\mathcal{E}_T$, the asserted recurrence holds, and $h^{(t)} > 0$ for $0 \leq t \leq T$. Therefore $\tau_{\mathcal{M}} > T \geq e^k$. □

The recurrence also bounds the first time at which $h^{(t)} \geq k$. As long as $h^{(t)} < k$, taking logarithms and iterating yields

$$\log_2(k/h^{(t)}) \leq 2^{-t}\big(\log_2(k/h^{(0)}) + 2\big) - 2.$$

Since $h^{(0)} \geq 1$, every trajectory satisfying Eq. (A.18) therefore satisfies

$$h^{(t)} \geq k \qquad \big(\log_2(1+\tfrac{1}{2}\log_2 k) \leq t \leq e^k,\ t \in \mathbb{Z}\big). \tag{A.26}$$

If $h^{(0)} \geq k$, the same conclusion holds from time zero.

A.3.2 Stopped contraction

Lemma 17. *Let $(Z_t)_{t\geq 0}$ be a nonnegative integrable process adapted to $(\mathcal{F}_t)_{t\geq 0}$, with deterministic Z_0, and let $\tau \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ be a stopping time. Fix $a > 0$ and $q \in (0,1)$. Suppose that, for every integer $t \geq 0$, almost surely on $\{\tau > t\}$,*

$$Z_t \geq a, \qquad \mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] \leq qZ_t.$$

Then, for every integer $t \geq 0$,

$$\mathbb{E}[Z_t \mathbb{1}\{\tau > t\}] \leq q^t Z_0, \qquad \Pr(\tau > t) \leq \min\{1, q^t Z_0/a\}. \tag{A.27}$$

Moreover,

$$\mathbb{E}[\tau] \leq \frac{\log \max\{1, Z_0/a\}}{-\log q} + \frac{1}{1-q} < \infty.$$

Proof. We set the process to zero after τ, so that only the assumed contraction before τ is used. Since $\{\tau > t+1\} \subseteq \{\tau > t\}$, nonnegativity and the conditional-mean inequality give

$$\mathbb{E}[Z_{t+1}\mathbb{1}\{\tau > t+1\} \mid \mathcal{F}_t] \leq \mathbb{1}\{\tau > t\}\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] \leq qZ_t\mathbb{1}\{\tau > t\}.$$

Iteration proves the expectation bound in Eq. (A.27). On $\{\tau > t\}$, $Z_t \geq a$, so Markov's inequality proves the probability bound. In particular, $\tau < \infty$ almost surely.

To sum this bound, put $u \triangleq \log \max\{1, Z_0/a\}/(-\log q)$. Then $Z_0/a \leq q^{-u}$, and

$$\mathbb{E}[\tau] = \sum_{t\geq 0}\Pr(\tau > t) \leq \lceil u \rceil + \frac{q^{\lceil u \rceil - u}}{1-q} \leq u + \frac{1}{1-q}.$$

The last inequality follows from convexity: $q^x \leq 1-x(1-q)$ for $0 \leq x \leq 1$, applied at $x = \lceil u \rceil - u$. □

A.3.3 HITTING-TIME BOUNDS FROM TRANSITION PROBABILITIES

Lemma 18. *Let $(X_t)_{t\ge 0}$ be a time-homogeneous Markov chain on a finite nonempty set $\mathcal{X}$, and let $\varnothing \neq \mathcal{M} \subseteq \mathcal{X}$. Define $\tau_{\mathcal{M}} \triangleq \inf\{t \ge 0 : X_t \in \mathcal{M}\}$. Suppose that $B \in \mathbb{N}$ and $p \in (0,1]$ satisfy*

$$\Pr(\tau_{\mathcal{M}} \le B \mid X_0 = x) \ge p \qquad (x \in \mathcal{X} \setminus \mathcal{M}). \tag{A.28}$$

Then, for every $x \in \mathcal{X}$ and integer $j \ge 0$,

$$\Pr(\tau_{\mathcal{M}} > jB \mid X_0 = x) \le (1-p)^j, \qquad \mathbb{E}[\tau_{\mathcal{M}} \mid X_0 = x] \le B/p. \tag{A.29}$$

For every $\zeta \in (0,1)$,

$$\Pr\big(\tau_{\mathcal{M}} \le B\,\lceil p^{-1}\log(1/\zeta)\rceil \,\big|\, X_0 = x\big) \ge 1-\zeta. \tag{A.30}$$

The set $\mathcal{M}$ need not be absorbing.

Proof. For $\mathcal{F}_t = \sigma(X_0,\dots,X_t)$, the Markov property gives

$$\Pr(\tau_{\mathcal{M}} > (j+1)B \mid \mathcal{F}_{jB}) \le (1-p)\mathbb{1}\{\tau_{\mathcal{M}} > jB\}.$$

Iteration proves the probability bound, and

$$\mathbb{E}[\tau_{\mathcal{M}} \mid X_0 = x] = \sum_{t\ge 0} \Pr(\tau_{\mathcal{M}} > t \mid X_0 = x) \le B \sum_{j\ge 0} (1-p)^j = B/p,$$

where $(1-p)^0 = 1$, also for $p = 1$. Finally, $(1-p)^j \le e^{-pj}$ proves Eq. (A.30). □

A.4 FINITE EXPECTED HITTING TIMES

A.4.1 SELECTION OF A ROUNDING OUTCOME

Lemma 19. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, let $\eta > 0$, and let $\mathbf{w}, \mathbf{v} \in \mathcal{W}$. Set $\mathbf{u} = \mathbf{w} - \eta\nabla f(\mathbf{w})$. If $\Pi(u_i) \in \mathcal{G}$, set $r_i = \Pi(u_i)$. Otherwise, let $a_i < \Pi(u_i) < a_i + \Delta$ be consecutive grid points and set*

$$r_i \triangleq \begin{cases} a_i, & v_i \le a_i, \\ a_i + \Delta, & v_i \ge a_i + \Delta. \end{cases} \tag{A.31}$$

Then $\Pr(\mathrm{SR}(\mathbf{u}) = \mathbf{r}) > 0$, and $\mathbf{r} = (r_1,\dots,r_d)$ minimizes

$$Q_{\mathbf{w}}(\mathbf{x}) \triangleq \|\mathbf{x}-\mathbf{u}\|_2^2 + \Delta\|\mathbf{x}-\mathbf{v}\|_1 \qquad (\mathbf{x} \in \mathcal{W}). \tag{A.32}$$

In particular,

$$\nabla f(\mathbf{w})^\top(\mathbf{r}-\mathbf{w}) + \frac{\|\mathbf{r}-\mathbf{w}\|_2^2}{2\eta} + \frac{\Delta}{2\eta}\big(\|\mathbf{r}-\mathbf{v}\|_1 - \|\mathbf{w}-\mathbf{v}\|_1\big) \le 0. \tag{A.33}$$

Proof. For $q_i(x) = (x-u_i)^2 + \Delta|x-v_i|$ and adjacent $a, a+\Delta \in \mathcal{G}$,

$$q_i(a+\Delta) - q_i(a) = \begin{cases} 2\Delta(a-u_i), & a+\Delta \le v_i, \\ 2\Delta(a+\Delta-u_i), & a \ge v_i. \end{cases} \tag{A.34}$$

If $a_i < u_i < a_i + \Delta$, these differences change sign at a_i if $v_i \le a_i$, and at $a_i+\Delta$ if $v_i \ge a_i+\Delta$. If $u_i \in \mathcal{G}$, they are nonpositive to its left and nonnegative to its right. If $u_i > M\Delta$, all differences are negative; if $u_i < -M\Delta$, all are positive. Thus $q_i(r_i) \le q_i(x)$ for every $x \in \mathcal{G}$. The selected endpoint has positive rounding probability, and independent rounding gives $\Pr(\mathrm{SR}(\mathbf{u}) = \mathbf{r}) > 0$. Summing over coordinates proves the minimization statement; expanding $Q_{\mathbf{w}}(\mathbf{r}) \le Q_{\mathbf{w}}(\mathbf{w})$ gives Eq. (A.33). □

A.4.2 FINITE EXPECTED HITTING TIME FOR A FIXED CONVEX SMOOTH OBJECTIVE

Proposition 20. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a fixed convex differentiable function. Suppose that $L > 0$ satisfies*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{u})\|_2 \le L\|\mathbf{x} - \mathbf{u}\|_2 \qquad (\mathbf{x}, \mathbf{u} \in \mathbb{R}^d).$$

Fix $0 < \eta < 1/L$ and consider the SR-GD update

$$\mathbf{w}^{(t+1)} = \mathrm{SR}\big(\mathbf{w}^{(t)} - \eta\nabla f(\mathbf{w}^{(t)})\big), \qquad \mathbf{w}^{(0)} \in \mathcal{W}. \tag{A.35}$$

With $\mathcal{M}$ and $\tau_{\mathcal{M}}$ defined as in Section 2.3, for every initializer,

$$\Pr(\tau_{\mathcal{M}} < \infty \mid \mathbf{w}^{(0)}) = 1, \qquad \mathbb{E}[\tau_{\mathcal{M}} \mid \mathbf{w}^{(0)}] < \infty. \tag{A.36}$$

Probabilities and expectations concern rounding for the fixed objective; $\mathcal{M}$ may contain multiple or boundary grid points. For a quadratic with random Hessian and the normalized step size, Eq. (A.36) holds conditionally on the Hessian; its expectation over Hessians need not be finite.

Proof. It suffices to find, from every nonoptimal grid point, a finite sequence of SR-GD transitions that reaches $\mathcal{M}$ and has positive probability. Since the grid is finite, the lengths of these sequences have a common upper bound and their probabilities have a positive lower bound. Lemma 18 then gives finite expected hitting time. To construct the sequences, we use Lemma 19 and smoothness to obtain strict decrease of the function defined below at each selected transition before reaching $\mathcal{M}$. If $\mathcal{M} = \mathcal{W}$, then $\tau_{\mathcal{M}} = 0$. Otherwise fix $\mathbf{v} \in \mathcal{M}$ and, for each $\mathbf{w} \in \mathcal{W}$, let $\mathbf{u} = \mathbf{w} - \eta\nabla f(\mathbf{w})$ and let $\mathbf{r}(\mathbf{w})$ be the rounding outcome in Lemma 19. Write $\mathbf{r} = \mathbf{r}(\mathbf{w})$. By L-smoothness,

$$f(\mathbf{r}) \le f(\mathbf{w}) + \nabla f(\mathbf{w})^\top(\mathbf{r} - \mathbf{w}) + \frac{L}{2}\|\mathbf{r} - \mathbf{w}\|_2^2. \tag{A.37}$$

Define

$$\psi(\mathbf{w}) \triangleq f(\mathbf{w}) - f(\mathbf{v}) + \frac{\Delta}{2\eta}\|\mathbf{w} - \mathbf{v}\|_1 \ge 0.$$

The selected outcome satisfies

$$\nabla f(\mathbf{w})^\top(\mathbf{r} - \mathbf{w}) + \frac{\|\mathbf{r} - \mathbf{w}\|_2^2}{2\eta} + \frac{\Delta}{2\eta}(\|\mathbf{r} - \mathbf{v}\|_1 - \|\mathbf{w} - \mathbf{v}\|_1) \le 0$$

by Lemma 19. Combining this with Eq. (A.37) yields

$$\psi(\mathbf{r}) \le \psi(\mathbf{w}) - \frac{1 - \eta L}{2\eta}\|\mathbf{r} - \mathbf{w}\|_2^2. \tag{A.38}$$

If $\mathbf{w} \notin \mathcal{M}$, convexity gives

$$\nabla f(\mathbf{w})^\top(\mathbf{w} - \mathbf{v}) \ge f(\mathbf{w}) - f(\mathbf{v}) > 0.$$

Thus $\nabla_i f(\mathbf{w})(w_i - v_i) > 0$ for some i. If $w_i > v_i$, then $u_i < w_i$ and $v_i \le w_i - \Delta$, so the rule selects $r_i \le w_i - \Delta$, also after clipping. The opposite case gives $r_i \ge w_i + \Delta$. Therefore $\mathbf{r} \ne \mathbf{w}$, and Eq. (A.38) is strict because $\eta L < 1$.

As long as the current state is outside $\mathcal{M}$, the selected transition changes the state and strictly decreases ψ. Thus the sequence cannot revisit any state. If it failed to reach $\mathcal{M}$, we could continue selecting transitions and obtain infinitely many distinct states in the finite set $\mathcal{W}$, a contradiction. It therefore reaches $\mathcal{M}$ within $b \triangleq |\mathcal{W}| - 1 = (2M + 1)^d - 1$ transitions. Set

$$q_* \triangleq \min_{\mathbf{w} \in \mathcal{W} \setminus \mathcal{M}} \Pr\big(\mathbf{w}^{(1)} = \mathbf{r}(\mathbf{w}) \mid \mathbf{w}^{(0)} = \mathbf{w}\big) > 0, \qquad p \triangleq q_*^b > 0.$$

Finiteness gives $q_* > 0$, and following the selected transitions gives $\Pr(\tau_{\mathcal{M}} \le b \mid \mathbf{w}^{(0)} = \mathbf{w}) \ge q_*^b$ for every $\mathbf{w} \notin \mathcal{M}$. Apply Lemma 18 with $B = b$ and $p = q_*^b$ to obtain

$$\Pr(\tau_{\mathcal{M}} > jb \mid \mathbf{w}^{(0)}) \le (1 - p)^j \qquad (j \ge 0), \tag{A.39}$$

and

$$\mathbb{E}[\tau_{\mathcal{M}} \mid \mathbf{w}^{(0)}] \le \frac{b}{p} = \frac{b}{q_*^b} < \infty.$$

The tail bound implies $\Pr(\tau_{\mathcal{M}} < \infty \mid \mathbf{w}^{(0)}) = 1$. □

A.5 Gaussian distributions and finite-grid minimizers

A.5.1 Gaussian norm and spectral bounds

Lemma 21. *Let $p \geq 1$ be an integer and let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Then*

$$\Pr(\|\mathbf{g}\|_1 < p/8) \leq e^{-p/16}. \tag{A.40}$$

Proof. We have $\mathbb{E}[\|\mathbf{g}\|_1] = p\sqrt{2/\pi}$. The function $\mathbf{x} \mapsto \|\mathbf{x}\|_1$ is $\sqrt{p}$-Lipschitz, so Gaussian concentration (Vershynin, 2012) gives, for every $t > 0$,

$$\Pr\Big(\|\mathbf{g}\|_1 < p\sqrt{2/\pi} - \sqrt{2pt}\Big) \leq e^{-t}. \tag{A.41}$$

Taking $t = p/16$ and using $\sqrt{2/\pi} - 1/\sqrt{8} > 1/8$ proves the claim. □

Lemma 22. *Let $n, d \in \mathbb{N}$ and let $\mathbf{X} \in \mathbb{R}^{n\times d}$ have independent standard Gaussian entries, and let $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$. Write $s_{\max}(\mathbf{X})$ and $s_{\min}(\mathbf{X})$ for the largest and smallest singular values. For every $u \geq 0$,*

$$\Pr\big(s_{\max}(\mathbf{X}) > \sqrt{n} + \sqrt{d} + u\big) \leq e^{-u^2/2}, \tag{A.42}$$

$$\Pr\left(\lambda_{\max}(\mathbf{A}) > \left(1 + \sqrt{d/n} + u/\sqrt{n}\right)^2\right) \leq e^{-u^2/2}. \tag{A.43}$$

If $n \geq d$, then for every $u \geq 0$,

$$\Pr\big(s_{\min}(\mathbf{X}) < \sqrt{n} - \sqrt{d} - u\big) \leq e^{-u^2/2}. \tag{A.44}$$

Consequently,

$$\Pr\left(\lambda_{\min}(\mathbf{A}) < \left(\max\{1 - \sqrt{d/n} - u/\sqrt{n}, 0\}\right)^2\right) \leq e^{-u^2/2}. \tag{A.45}$$

If $n < d$, the matrix $\mathbf{A}$ has rank n almost surely. Its smallest positive eigenvalue, denoted by $\lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))$, satisfies, for every $u \geq 0$,

$$\Pr\left(\lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A})) < \left(\max\{\sqrt{d/n} - 1 - u/\sqrt{n}, 0\}\right)^2\right) \leq e^{-u^2/2}. \tag{A.46}$$

If $n \leq d/16$, all n positive eigenvalues belong to $[d/(4n), 9d/(4n)]$ with probability at least $1 - 2e^{-d/32}$.

Proof. The Gaussian expectation bounds and concentration in Vershynin (2012) for the 1-Lipschitz functions $s_{\max}$ and $-s_{\min}$ give Eqs. (A.42) and (A.44). Squaring and dividing by n gives the eigenvalue bounds; nonpositive lower thresholds give empty events. If $n < d$, the first n-column minor is nonsingular almost surely. Apply the lower bound to $\mathbf{X}^\top$, using $\lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A})) = s_{\min}(\mathbf{X}^\top)^2/n$. For $n \leq d/16$, applying the singular-value bounds to $\mathbf{X}^\top$ with deviation $\sqrt{d}/4$ gives, with probability at least $1 - 2e^{-d/32}$,

$$\sqrt{d}/2 \leq s_{\min}(\mathbf{X}^\top) \leq s_{\max}(\mathbf{X}) \leq 3\sqrt{d}/2.$$

Squaring and dividing by n proves the final assertion. □

Lemma 23. *Let $n, d \in \mathbb{N}$ and let $\mathbf{X} \in \mathbb{R}^{n\times d}$ have independent standard Gaussian entries and let $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$. Fix $\mathbf{w}^\star \in \mathbb{R}^d$ before drawing $\mathbf{X}$. With probability at least $1 - e^{-d}$ over $\mathbf{X}$,*

$$\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \geq \max\left\{\lambda_{\min}(\mathbf{A}), e^{-1}\big(e(2M+1)\big)^{-2d/n}\right\}. \tag{A.47}$$

For every fixed $R > 0$, if $n \geq d/R$, this implies

$$\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \geq \mu_R, \qquad \mu_R \triangleq e^{-1}\big(e(2M+1)\big)^{-2R} > 0, \tag{A.48}$$

on the same event of probability at least $1 - e^{-d}$.

Proof. We first obtain a lower bound on a deterministic finite set $\mathcal{Z} \subset \mathbb{R}^d$, then take $\mathcal{Z} = \mathcal{W} - \mathbf{w}^\star$. Let $N \in \mathbb{N}$ satisfy $|\mathcal{Z}| \le N$ and fix $\zeta \in (0,1)$. For fixed $\mathbf{z} \neq \mathbf{0}$, put $V = \|\mathbf{X}\mathbf{z}\|_2^2/\|\mathbf{z}\|_2^2 \sim \chi_n^2$, where χ_n^2 denotes the chi-square distribution with n degrees of freedom. Since $\mathbb{E}[e^{-\theta V}] = (1+2\theta)^{-n/2}$, Markov's inequality with $\theta = (u^{-1}-1)/2$ gives, for $0 < u < 1$,

$$\Pr(V \le nu) \le e^{\theta n u}(1+2\theta)^{-n/2} = \exp\{-n(u - 1 - \log u)/2\}. \tag{A.49}$$

For $u_0 = e^{-1}(\zeta/N)^{2/n}$,

$$u_0 - 1 - \log u_0 = u_0 + \frac{2}{n}\log(N/\zeta) \ge \frac{2}{n}\log(N/\zeta).$$

For each $\mathbf{z} \neq \mathbf{0}$, $\Pr(\mathbf{z}^\top \mathbf{A}\mathbf{z} < u_0\|\mathbf{z}\|_2^2) \le \zeta/N$; for $\mathbf{z} = \mathbf{0}$, the inequality holds deterministically. A union bound therefore gives, with probability at least $1-\zeta$,

$$\mathbf{z}^\top \mathbf{A}\mathbf{z} \ge e^{-1}(\zeta/N)^{2/n}\|\mathbf{z}\|_2^2 \qquad \text{for every } \mathbf{z} \in \mathcal{Z}. \tag{A.50}$$

Take $\mathcal{Z} = \mathcal{W} - \mathbf{w}^\star$, $N = (2M+1)^d$, and $\zeta = e^{-d}$. Combining with the deterministic Rayleigh lower bound $\lambda_{\min}(\mathbf{A})$ proves Eq. (A.47); $d/n \le R$ then gives Eq. (A.48). □

A.5.2 Gaussian action on a fixed displacement

Lemma 24. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ have independent standard Gaussian entries, where $n, d \in \mathbb{N}$, and set $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$. Fix a deterministic nonzero $\mathbf{z} \in \mathbb{R}^d$ and set*

$$\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|_2, \qquad Z = \frac{\|\mathbf{X}\mathbf{z}\|_2^2}{n\|\mathbf{z}\|_2^2}.$$

Then $nZ \sim \chi_n^2$, and there is a random vector $\mathbf{g}$, independent of Z, such that

$$\mathbf{A}\mathbf{z} = Z\mathbf{z} + \|\mathbf{z}\|_2\sqrt{Z/n}\,\mathbf{g}, \qquad \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \mathbf{u}\mathbf{u}^\top).$$

Proof. Complete $\mathbf{u}$ to $[\mathbf{u}, \mathbf{V}] \in \mathcal{O}(d)$ and write $\mathbf{X}[\mathbf{u}, \mathbf{V}] = [\mathbf{a}, \mathbf{H}]$; these entries are independent standard Gaussians. Then

$$\mathbf{A}\mathbf{z} = \frac{\|\mathbf{z}\|_2}{n}\left(\mathbf{u}\|\mathbf{a}\|_2^2 + \mathbf{V}\mathbf{H}^\top\mathbf{a}\right).$$

Almost surely $\mathbf{a} \neq \mathbf{0}$, and $\mathbf{H}^\top\mathbf{a}/\|\mathbf{a}\|_2 \mid \mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d-1})$. Thus $Z = \|\mathbf{a}\|_2^2/n$ and $\mathbf{g} = \mathbf{V}\mathbf{H}^\top\mathbf{a}/\|\mathbf{a}\|_2$ give the representation. The conditional distribution of $\mathbf{g}$ does not depend on $\mathbf{a}$, so $\mathbf{g}$ is independent of $\mathbf{a}$ and hence of Z. Also, $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_d - \mathbf{u}\mathbf{u}^\top$, which gives the stated covariance.

□

A.5.3 Kernel avoidance and multiplicity of grid minimizers

Lemma 25. *Fix $\mathbf{w}^\star \in \mathbb{R}^d$ and the grid $\mathcal{W}$ before sampling either of the following Hessians:*

(i) $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$, where $n \in \mathbb{N}$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$ has independent standard Gaussian entries;
(ii) $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ is Haar and $\mathbf{\Lambda} \succeq \mathbf{0}$ is deterministic and nonzero.

Almost surely,

$$\ker(\mathbf{A}) \cap (\mathcal{W} - \mathcal{W}) = \{\mathbf{0}\}. \tag{A.51}$$

If $\mathbf{w}^\star \in \mathcal{W}$, then $\mathcal{M} = \{\mathbf{w}^\star\}$. If $\mathbf{V}$ consists of the first r columns of a single $\mathbf{U} \sim \mathrm{Unif}(\mathcal{O}(d))$, where $1 \le r \le d$, then the kernel identity holds almost surely simultaneously for every $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with $\mathbf{D} \succ \mathbf{0}$.

Proof. For each fixed nonzero $\mathbf{z}$, $\mathbf{X}\mathbf{z}$ is a nondegenerate Gaussian vector, so $\Pr(\mathbf{X}\mathbf{z} = \mathbf{0}) = 0$. For Haar $\mathbf{U}$, $\mathbf{U}^\top\mathbf{z}$ is uniform on the sphere of radius $\|\mathbf{z}\|_2$; its projection onto any fixed nonzero eigenspace vanishes with probability zero. A finite union over $(\mathcal{W} - \mathcal{W}) \setminus \{\mathbf{0}\}$ proves Eq. (A.51). The same event works for every $\mathbf{D} \succ \mathbf{0}$, since $\ker(\mathbf{V}\mathbf{D}\mathbf{V}^\top) = \ker(\mathbf{V}^\top)$. For on-grid $\mathbf{w}^\star$, $f(\mathbf{w}) = 0$ if and only if $\mathbf{w} - \mathbf{w}^\star \in \ker(\mathbf{A})$, proving uniqueness. □

Lemma 26. *Under the Haar model in Lemma 25, suppose that $\mathbf{\Lambda}$ is not a multiple of $\mathbf{I}_d$. Almost surely, $|\mathcal{M}| \le 2$, and distinct grid points with $f(\mathbf{w}) = f(\mathbf{v})$ satisfy $\mathbf{w} + \mathbf{v} = 2\mathbf{w}^\star$. Consequently, $2\mathbf{w}^\star \notin \Delta\mathbb{Z}^d$ implies almost-sure uniqueness.*

Proof. We show that two fixed grid points can have the same objective value with positive probability only if their midpoint is $\mathbf{w}^\star$. For distinct fixed grid points $\mathbf{w}, \mathbf{v}$, put $\mathbf{u} = \mathbf{w} - \mathbf{w}^\star$, $\mathbf{z} = \mathbf{v} - \mathbf{w}^\star$, and $\mathbf{B} = \mathbf{u}\mathbf{u}^\top - \mathbf{z}\mathbf{z}^\top$. The equality $f(\mathbf{w}) = f(\mathbf{v})$ is equivalent to $\operatorname{tr}(\mathbf{AB}) = 0$.

Suppose $\mathbf{B} \neq \mathbf{0}$. The function $\mathbf{U} \mapsto \operatorname{tr}(\mathbf{U\Lambda U}^\top \mathbf{B})$ is real analytic on $\mathcal{O}(d)$. Write $b_1, \dots, b_d$ and $\lambda_1, \dots, \lambda_d$ for the eigenvalues of $\mathbf{B}$ and $\mathbf{\Lambda}$, respectively. If the function vanished identically, they would satisfy

$$\sum_{i=1}^{d} b_i \lambda_{\pi(i)} = 0 \qquad \text{for every permutation } \pi \text{ of } [d].$$

Indeed, one can choose $\mathbf{U}$ to align the eigenvectors of $\mathbf{\Lambda}$ with those of $\mathbf{B}$ in any prescribed order. Choose p, q with $\lambda_p \neq \lambda_q$. For arbitrary distinct i, j, choose a permutation assigning λ_p, λ_q to these two positions, and then interchange the assignments. Subtracting the two identities gives

$$(b_i - b_j)(\lambda_p - \lambda_q) = 0.$$

Hence every b_i equals b_1. Since $\operatorname{tr}(\mathbf{\Lambda}) > 0$, the identity $0 = b_1 \operatorname{tr}(\mathbf{\Lambda})$ implies $b_1 = 0$ and then $\mathbf{B} = \mathbf{0}$, a contradiction.

Right multiplication by the reflection across the hyperplane orthogonal to an eigenvector of $\mathbf{\Lambda}$ exchanges the two connected components of $\mathcal{O}(d)$ without changing $\mathbf{U\Lambda U}^\top$. Thus the function is not identically zero on either component. On a connected real analytic manifold, the zero set of a real analytic function that is not identically zero has measure zero with respect to smooth volume. Haar measure on each component of $\mathcal{O}(d)$ is proportional to its Riemannian volume. Therefore $\Pr(f(\mathbf{w}) = f(\mathbf{v})) = 0$ if $\mathbf{B} \neq \mathbf{0}$.

If $\mathbf{B} = \mathbf{0}$, distinctness gives $\mathbf{z} = -\mathbf{u}$, hence $\mathbf{w} + \mathbf{v} = 2\mathbf{w}^\star$. A finite union over pairs of grid points therefore shows that, almost surely, $f(\mathbf{w}) = f(\mathbf{v})$ with $\mathbf{w} \neq \mathbf{v}$ implies $\mathbf{v} = 2\mathbf{w}^\star - \mathbf{w}$, so $|\mathcal{M}| \le 2$. If $2\mathbf{w}^\star \notin \Delta\mathbb{Z}^d$, this equality is impossible for grid points. □

A.6 Projection reduction for small-ball events

Lemma 27. *Let $1 \le p \le N$ be integers, let vol_p denote p-dimensional Lebesgue volume, and fix deterministic $\mathbf{P} \in \mathbb{R}^{p \times N}$, $\mathbf{a} \in \mathbb{R}^N$, and $\eta_0, D > 0$ with $\mathbf{PP}^\top = \mathbf{I}_p$ and $\mathbf{Pa} = \mathbf{0}$. Let $\mathcal{K} \subset \mathbb{R}^N$ be a deterministic nonempty compact convex set with $\mathcal{K} = -\mathcal{K}$. Suppose that $\mathbf{y} \in \mathbb{R}^N$ is a random vector and that $\mathbf{Py}$ has a density with respect to p-dimensional Lebesgue measure bounded by D almost everywhere. Then*

$$\{\exists \eta \ge \eta_0 : \mathbf{a} - \eta\mathbf{y} \in \mathcal{K}\} \subseteq \{\mathbf{Py} \in \eta_0^{-1}\mathbf{P}\mathcal{K}\},$$

and

$$\Pr(\exists \eta \ge \eta_0 : \mathbf{a} - \eta\mathbf{y} \in \mathcal{K}) \le D\eta_0^{-p} \operatorname{vol}_p(\mathbf{P}\mathcal{K}). \tag{A.52}$$

Proof. If $\mathbf{a} - \eta\mathbf{y} \in \mathcal{K}$ for some $\eta \ge \eta_0$, then $\mathbf{Pa} = \mathbf{0}$ gives $-\eta\mathbf{Py} \in \mathbf{P}\mathcal{K}$. Since $\mathcal{K} = -\mathcal{K}$, also $\eta\mathbf{Py} \in \mathbf{P}\mathcal{K}$. The set $\mathbf{P}\mathcal{K}$ is convex and contains $\mathbf{0}$, so

$$\eta_0\mathbf{Py} = \frac{\eta_0}{\eta}(\eta\mathbf{Py}) \in \mathbf{P}\mathcal{K}.$$

This proves the event inclusion simultaneously for every $\eta \ge \eta_0$. The containing set has volume $\eta_0^{-p} \operatorname{vol}_p(\mathbf{P}\mathcal{K})$. Integrating the density of $\mathbf{Py}$, which is bounded by D, proves Eq. (A.52). □

B EXPONENTIAL HITTING TIMES FOR ON-GRID TARGETS

B.1 HAAR EIGENBASES

Fix $0 \le a < b$ and $v > 0$. Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ be deterministic, with $\lambda_i \in [a, b]$ and $d^{-1}\sum_i (\lambda_i - \bar{\lambda})^2 \ge v^2$, where $\bar{\lambda} = d^{-1}\,\mathrm{tr}(\mathbf{\Lambda})$. Let $\mathbf{U}$ be Haar on $\mathcal{O}(d)$, and set $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$. The positive variance ensures $\lambda_{\max}(\mathbf{A}) > 0$. For statements involving a grid target, fix $\mathbf{w}^\star \in \mathcal{W}^\circ$ independently of $\mathbf{U}$. We first bound $\|\mathbf{K}\mathbf{z}\|_1$ for a fixed vector, then obtain a uniform bound over vectors with small support, and finally use that bound to control the conditional Hamming-distance expectation. The following constants are used throughout this subsection:

$$q \triangleq \frac{1}{256}\min\left\{1, \left(\frac{v}{b-a}\right)^4\right\}, \qquad u \triangleq \frac{Sv}{128b}, \tag{B.1}$$

$$\delta \triangleq \frac{q^2}{16e(1+4/u)}, \qquad \theta \triangleq \min\{\delta/2, u^2/9\}. \tag{B.2}$$

For a fixed nonzero vector $\mathbf{z}$, we first determine the conditional distribution of the component of $\mathbf{A}\mathbf{z}$ orthogonal to $\mathbf{z}$. This determines $(\mathbf{K}\mathbf{z})_i$ at coordinates with $z_i = 0$.

Lemma 28. *Let $d \ge 2$, let $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$ be deterministic and diagonal, and let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{U} \sim \mathrm{Unif}(\mathcal{O}(d))$. For a deterministic unit vector $\mathbf{e}$, set*

$$\mathbf{q} = \mathbf{U}^\top\mathbf{e}, \qquad r_\perp = \|(\mathbf{I}_d - \mathbf{e}\mathbf{e}^\top)\mathbf{A}\mathbf{e}\|_2.$$

Then

$$r_\perp^2 = \mathbf{q}^\top\mathbf{\Lambda}^2\mathbf{q} - (\mathbf{q}^\top\mathbf{\Lambda}\mathbf{q})^2.$$

Conditional on $\mathbf{q}$ with $r_\perp > 0$,

$$\begin{aligned} \mathbf{A}\mathbf{e} &= (\mathbf{q}^\top\mathbf{\Lambda}\mathbf{q})\mathbf{e} + r_\perp\boldsymbol{\xi}, \\ \boldsymbol{\xi} &\sim \mathrm{Unif}\{\mathbf{x} \in \mathbf{e}^\perp : \|\mathbf{x}\|_2 = 1\}. \end{aligned} \tag{B.3}$$

Proof. Since $\mathbf{U}\mathbf{q} = \mathbf{e}$ and $\mathbf{e}^\top\mathbf{A}\mathbf{e} = \mathbf{q}^\top\mathbf{\Lambda}\mathbf{q}$,

$$(\mathbf{I}_d - \mathbf{e}\mathbf{e}^\top)\mathbf{A}\mathbf{e} = \mathbf{U}\big(\mathbf{\Lambda}\mathbf{q} - (\mathbf{q}^\top\mathbf{\Lambda}\mathbf{q})\mathbf{q}\big).$$

Taking squared norms proves the formula for $r_\perp$. For every fixed $\mathbf{R} \in \mathcal{O}(d)$ with $\mathbf{R}\mathbf{e} = \mathbf{e}$, $\mathbf{R}\mathbf{U}$ has the same law as $\mathbf{U}$ and $(\mathbf{R}\mathbf{U})^\top\mathbf{e} = \mathbf{q}$. Replacing $\mathbf{U}$ by $\mathbf{R}\mathbf{U}$ therefore preserves the conditioning on $\mathbf{q}$ and rotates the displayed vector by $\mathbf{R}$. Its conditional law is invariant under all orthogonal transformations of $\mathbf{e}^\perp$. Since its length is $r_\perp$, its normalized direction has the asserted uniform distribution whenever $r_\perp > 0$. □

On $\{r_\perp > 0\}$, for $\mathbf{e} = \mathbf{z}/\|\mathbf{z}\|_2$ and any coordinate with $z_i = 0$, Eq. (B.3) gives

$$(\mathbf{K}\mathbf{z})_i = -\eta(\mathbf{A}\mathbf{z})_i = -\eta\|\mathbf{z}\|_2 r_\perp \xi_i.$$

We next bound $r_\perp$ and $\sum_{i:z_i=0}|\xi_i|$ from below.

B.1.1 AN ℓ_1 LOWER BOUND FOR A FIXED VECTOR

Lemma 29. *Under the assumptions of Appendix B.1, every deterministic nonzero $\mathbf{z} \in \mathbb{R}^d$ with $\|\mathbf{z}\|_0 \le d/2$ satisfies*

$$\Pr\Big(\|\mathbf{K}\mathbf{z}\|_1 < 2u\sqrt{d}\,\|\mathbf{z}\|_2\Big) \le 9e^{-qd}. \tag{B.4}$$

Proof. The eigenvalue variance gives $r_\perp \ge v/2$ with probability at least $1 - 6e^{-qd}$. Conditional on $\mathbf{q}$, the uniform direction in Lemma 28 satisfies $\sum_{i:z_i=0}|\xi_i| \ge \sqrt{d}/32$ with probability at least $1 - 3e^{-d/32}$. We prove these two bounds below. For independent standard Gaussians g_i and real t_i, not all zero, Bechar (2009) gives

$$\Pr\left(\left|\sum_i t_i(g_i^2 - 1)\right| \ge x\right) \le 2\exp\left[-\frac{1}{16}\min\left\{\frac{x^2}{\sum_i t_i^2}, \frac{x}{\max_i|t_i|}\right\}\right] \quad (x > 0), \tag{B.5}$$

since $s = \frac{1}{16}\min\{x^2/\sum_i t_i^2, x/\max_i |t_i|\}$ satisfies $2\sqrt{s\sum_i t_i^2} + 2s\max_i |t_i| \le 5x/8 < x$. If all coefficients vanish, the deviation is zero.

Set $\mathbf{e} = \mathbf{z}/\|\mathbf{z}\|_2$, $\mathbf{q} = \mathbf{U}^\top \mathbf{e}$, $\mathbf{B} = \mathbf{\Lambda} - \bar{\lambda}\mathbf{I}_d$, $\ell = b - a$, and $\sigma^2 = d^{-1}\operatorname{tr}(\mathbf{B}^2) \ge v^2$. Represent $\mathbf{q} = \mathbf{g}/\|\mathbf{g}\|_2$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and define

$$\mathcal{E} \triangleq \left\{ \frac{d}{2} \le \|\mathbf{g}\|_2^2 \le \frac{3d}{2}, \quad |\mathbf{g}^\top \mathbf{B}\mathbf{g}| \le \frac{dv}{4}, \quad \mathbf{g}^\top \mathbf{B}^2 \mathbf{g} \ge \frac{3d\sigma^2}{4} \right\}. \tag{B.6}$$

Apply Eq. (B.5) to coefficients 1, B_{ii}, and B_{ii}^2. The respective exponents are at least

$$\frac{d}{64}, \qquad d\min\left\{\frac{v^2}{256\ell^2}, \frac{v}{64\ell}\right\}, \qquad d\min\left\{\frac{v^4}{256\ell^4}, \frac{v^2}{64\ell^2}\right\},$$

each at least qd. Hence $\Pr(\mathcal{E}^c) \le 6e^{-qd}$. On $\mathcal{E}$,

$$\begin{aligned} r_\perp^2 &= \mathbf{q}^\top \mathbf{\Lambda}^2 \mathbf{q} - (\mathbf{q}^\top \mathbf{\Lambda}\mathbf{q})^2 \\ &= \mathbf{q}^\top \mathbf{B}^2 \mathbf{q} - (\mathbf{q}^\top \mathbf{B}\mathbf{q})^2 \ge \sigma^2/2 - v^2/4 \ge v^2/4. \end{aligned} \tag{B.7}$$

On $\mathcal{E}$, Lemma 28 applies with $r_\perp \ge v/2$.

Let $s = d - \|\mathbf{z}\|_0 \ge d/2$. For an independent standard Gaussian $\mathbf{g}'$, represent $\boldsymbol{\xi} = (\mathbf{I}_d - \mathbf{e}\mathbf{e}^\top)\mathbf{g}'/\|(\mathbf{I}_d - \mathbf{e}\mathbf{e}^\top)\mathbf{g}'\|_2$. The denominator is positive almost surely since $d \ge 2$. Define

$$\mathcal{E}_1 \triangleq \left\{ \sum_{i:z_i=0} |g_i'| \ge s/8, \quad \|\mathbf{g}'\|_2 \le 2\sqrt{d} \right\}.$$

Since $\mathbf{g}'$ is independent of $\mathbf{q}$, the coordinates indexed by $\{i : z_i = 0\}$ remain independent standard Gaussians conditional on $\mathbf{q}$. Apply Lemma 21 to this conditional law. Together with Eq. (B.5), it gives $\Pr(\mathcal{E}_1^c \mid \mathbf{q}) \le e^{-d/32} + 2e^{-3d/16} \le 3e^{-d/32}$. If $z_i = 0$, then $e_i = 0$, so $[(\mathbf{I}_d - \mathbf{e}\mathbf{e}^\top)\mathbf{g}']_i = g_i'$. Thus projection preserves the numerator at these coordinates and cannot increase the denominator. On $\mathcal{E}_1$, this gives

$$\sum_{i:z_i=0} |\xi_i| \ge \frac{s/8}{2\sqrt{d}} \ge \frac{\sqrt{d}}{32}. \tag{B.8}$$

For i with $z_i = 0$, $(\mathbf{K}\mathbf{z})_i = -\eta(\mathbf{A}\mathbf{z})_i$. Combining the last three displays and $\eta \ge S/b$ gives

$$\|\mathbf{K}\mathbf{z}\|_1 \ge \frac{S}{b}\|\mathbf{z}\|_2 r_\perp \sum_{i:z_i=0} |\xi_i| \ge \frac{Sv}{64b}\sqrt{d}\,\|\mathbf{z}\|_2 = 2u\sqrt{d}\,\|\mathbf{z}\|_2.$$

The failure probability is at most $6e^{-qd} + 3e^{-d/32} \le 9e^{-qd}$. □

B.1.2 A UNIFORM ℓ_1 LOWER BOUND UNDER $\|\mathbf{z}\|_0 \le \delta d$

The preceding lemma treats each fixed vector. The next lemma gives one event on which the lower bound holds for every vector with at most δd nonzero coordinates, as required by Proposition 3.

Lemma 30. *Under the assumptions of Appendix B.1, if $d \ge \max\{2/\delta, 4\log 9/q\}$, then, with probability at least $1 - e^{-qd/2}$ over $\mathbf{U}$,*

$$\|\mathbf{K}\mathbf{z}\|_1 \ge u\sqrt{d}\,\|\mathbf{z}\|_2 \quad \textit{for every } \mathbf{z} \in \mathbb{R}^d \textit{ with } 1 \le \|\mathbf{z}\|_0 \le \delta d. \tag{B.9}$$

Proof. We apply Lemma 29 to finite nets in all relevant coordinate subspaces. A union bound gives failure probability at most $9e^{-3qd/4} \le e^{-qd/2}$, and $\|\mathbf{K}\|_2 \le S$ extends the estimate from the nets to every vector with the prescribed support size. Put $r = \lfloor \delta d \rfloor$, $N = 1 + 4/u$, and $p = r/d \le \delta < 1/2$. For each $\mathcal{I} \subseteq [d]$ with $1 \le |\mathcal{I}| \le r$, choose a maximal deterministic subset $\mathcal{N}_\mathcal{I}$ of the unit sphere in that coordinate subspace such that distinct points have Euclidean distance greater than $u/2$. Maximality means that every unit vector supported on $\mathcal{I}$ lies within distance $u/2$ of a point of $\mathcal{N}_\mathcal{I}$; this is the defining property of a $u/2$-net. The open balls of radius $u/4$ centered at its points are disjoint and lie in the ball of radius $1 + u/4$ in the same subspace. Comparing their $|\mathcal{I}|$-dimensional volumes gives

$$|\mathcal{N}_\mathcal{I}|(u/4)^{|\mathcal{I}|} \le (1 + u/4)^{|\mathcal{I}|}, \qquad |\mathcal{N}_\mathcal{I}| \le (1 + 4/u)^{|\mathcal{I}|} = N^{|\mathcal{I}|}.$$

Define

$$\mathcal{E} \triangleq \bigcap_{1\le|\mathcal{I}|\le r} \bigcap_{\mathbf{z}_0\in\mathcal{N}_{\mathcal{I}}} \{\|\mathbf{K}\mathbf{z}_0\|_1 \ge 2u\sqrt{d}\}.$$

Since $p \le 1/2$, each term with $j \le r$ satisfies $p^j(1-p)^{d-j} \ge p^r(1-p)^{d-r}$. Therefore the binomial identity gives

$$1 = \sum_{j=0}^{d} \binom{d}{j} p^j(1-p)^{d-j} \ge p^r(1-p)^{d-r} \sum_{j=0}^{r} \binom{d}{j},$$

and hence

$$\sum_{j=0}^{r} \binom{d}{j} \le e^{d[-p\log p-(1-p)\log(1-p)]}. \tag{B.10}$$

Since $-p\log p-(1-p)\log(1-p) \le p\log(e/p)$ and $\log x \le \sqrt{x}$,

$$-p\log p-(1-p)\log(1-p)+p\log N \le \delta\log(eN/\delta) \le \sqrt{eN\delta} = q/4.$$

Thus Lemma 29 and a union bound give $\Pr(\mathcal{E}^c) \le 9e^{-3qd/4} \le e^{-qd/2}$ by the dimension hypothesis. To extend this estimate beyond the nets, note that both $\mathbf{A}$ and $\operatorname{diag}(\mathbf{A})$ lie between $\mathbf{0}$ and $\lambda_{\max}(\mathbf{A})\mathbf{I}_d$ in the positive semidefinite order. Their difference therefore has spectral norm at most $\lambda_{\max}(\mathbf{A})$, which gives $\|\mathbf{K}\|_2 = \eta\|\operatorname{diag}(\mathbf{A})-\mathbf{A}\|_2 \le S$. On $\mathcal{E}$, choose $\mathbf{z}_0 \in \mathcal{N}_{\mathcal{I}}$ within distance $u/2$ of any unit vector $\mathbf{z}$ supported on $\mathcal{I}$. Then

$$\|\mathbf{K}\mathbf{z}\|_1 \ge 2u\sqrt{d} - \sqrt{d}\|\mathbf{K}\|_2\|\mathbf{z}-\mathbf{z}_0\|_2 \ge u\sqrt{d}.$$

Homogeneity gives $\|\mathbf{K}\mathbf{z}\|_1 \ge u\sqrt{d}\,\|\mathbf{z}\|_2$ for $1 \le \|\mathbf{z}\|_0 \le \delta d$. □

B.1.3 Proof of Theorem 1

Recall the constants used in the uniform Haar estimate:

$$q = \frac{1}{256}\min\left\{1, \left(\frac{v}{b-a}\right)^4\right\}, \qquad u = \frac{Sv}{128b},$$
$$\delta = \frac{q^2}{16e(1+4/u)}, \qquad \theta = \min\{\delta/2, u^2/9\}.$$

The theorem holds with

$$\rho \triangleq \frac{1-S^2}{288M^2}\min\{u^2, \delta/2\}, \qquad c \triangleq \min\{q/8, \sqrt{\rho}/32\},$$
$$d_0 \triangleq \left\lceil \max\left\{e^8, \frac{4}{\theta}, \frac{1}{\rho}, \left(\frac{\log 2}{c}\right)^2\right\}\right\rceil. \tag{B.11}$$

For fixed S, a, b, v, these constants satisfy $\rho^{-1}, c^{-2}, d_0 = O(M^2)$.

Remark. Part (i) of Proposition 48 gives alternative admissible choices of the theorem constants satisfying $\rho^{-1}, c^{-2}, d_0 = O(\log(8M))$. The proof in Appendix F.2.3 combines the sparse-displacement estimate with a lattice bound for the remaining displacements to establish the conditional-mean conclusion of Proposition 3 at a larger threshold, then applies Lemma 16.

Proof of Theorem 1. As in the proof sketch, we combine Proposition 3 with Lemma 16 to prove the Hamming-distance recurrence. Specifically, Lemma 30 verifies the spreading assumption of Proposition 3 with $m = u\sqrt{d}$ and $r = \lfloor \delta d \rfloor$. The resulting bound $\bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}}$ allows us to apply Lemma 16 over the full time interval. We now verify the threshold and probability bounds for $k = \rho d$. Fix $d \ge d_0$, so $k \ge 1$. Put $r = \lfloor \delta d \rfloor \ge \delta d/2 \ge 1$. Define

$$\mathcal{E}_0 \triangleq \left\{\|\mathbf{K}\mathbf{z}\|_1 \ge u\sqrt{d}\|\mathbf{z}\|_2 \text{ for all } \mathbf{z}\in\mathbb{R}^d,\ 1 \le \|\mathbf{z}\|_0 \le r\right\}.$$

The dimension choice satisfies

$$\frac{4}{\theta} \ge \frac{2}{\delta}, \qquad \frac{4}{\theta} \ge \frac{128e}{q^2} \ge \frac{4\log 9}{q}.$$

Thus Lemma 30 gives $\Pr(\mathcal{E}_0^c) \le e^{-qd/2}$. On $\mathcal{E}_0$, the bound holds in particular for the grid differences in Proposition 3, with $m = u\sqrt{d}$ and the displayed r. The threshold supplied by that proposition is at least k, because

$$\frac{1-S^2}{144M^2}\min\{u^2 d, r\} \ge \frac{1-S^2}{144M^2}\min\{u^2, \delta/2\}d = 2k.$$

For each fixed $h \ge 0$, the function $x \mapsto x\min\{h, x\}$ is nondecreasing. We may therefore reduce the threshold in Proposition 3 to k, obtaining

$$\bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}).$$

Define

$$\mathcal{E}_k \triangleq \bigcap_{\substack{t\in\mathbb{Z}_{\ge 0}\\ t<e^k}} \left\{h^{(t+1)} \ge \min\{k, 2\sqrt{kh^{(t)}}\}\right\}.$$

Since $a > 0$, the target is the unique minimizer and $h(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|_0$. Conditional on $\mathbf{U}$, we may thus apply Lemma 16 on $\mathcal{E}_0$. The lemma gives $\Pr(\mathcal{E}_k^c \mid \mathbf{U}) \le e^{-\sqrt{k}/16}$ on $\mathcal{E}_0$. On $\mathcal{E}_k$, positivity of the distances gives $\tau_{\mathcal{M}} > e^k$. Finally, $\mathbf{A} \succeq a\mathbf{I}_d$ gives, for every $\mathbf{w} \in \mathcal{W}$,

$$f(\mathbf{w}) = \tfrac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{A}(\mathbf{w}-\mathbf{w}^\star) \ge \tfrac{a}{2}\|\mathbf{w}-\mathbf{w}^\star\|_2^2 \ge \tfrac{a}{2}\Delta^2 h(\mathbf{w}).$$

Consequently,

$$\Pr(\mathcal{E}_k^c) \le e^{-qd/2} + e^{-\sqrt{k}/16} \le 2e^{-2c\sqrt{d}} \le e^{-c\sqrt{d}},$$

where $\sqrt{d} \ge \log 2/c$. □

B.2 Universal exponential upper bound

B.2.1 Hitting-time bounds for rounded contractions

Proposition 31. *Let $\mathbf{w}^\star \in \mathcal{W}$ and consider SR-GD with any fixed step size $\eta > 0$. Suppose that, for some $\gamma \in (0,1)$,*

$$\|\mathbf{J}\|_2 \le \gamma. \tag{B.12}$$

Then $\mathcal{M} = \{\mathbf{w}^\star\}$ and, for every initializer, integer $j \ge 0$, and $\zeta \in (0,1)$,

$$\Pr(\tau_{\mathcal{M}} > jB_2 \mid \mathbf{A}, \mathbf{w}^{(0)}) \le (1-p_d)^j, \qquad \mathbb{E}[\tau_{\mathcal{M}} \mid \mathbf{A}, \mathbf{w}^{(0)}] \le B_2/p_d,$$
$$\Pr\Big(\tau_{\mathcal{M}} \le B_2\left\lceil p_d^{-1}\log(1/\zeta)\right\rceil \,\Big|\, \mathbf{A}, \mathbf{w}^{(0)}\Big) \ge 1-\zeta,$$

where

$$\varepsilon \triangleq (1-\gamma)/2, \qquad q \triangleq (1+\varepsilon)\gamma, \qquad C_\varepsilon \triangleq \max\{2, 2\log(2/\varepsilon)\},$$
$$B_2 \triangleq \left\lceil \frac{\log(2M\sqrt{d})}{\log(1/q)} \right\rceil + 1, \qquad p_d \triangleq \exp\left(-\frac{2MC_\varepsilon\gamma}{1-q}d\right). \tag{B.13}$$

Proof. The norm bound excludes a zero eigenvalue of $\mathbf{A}$, so $\mathbf{A} \succ \mathbf{0}$ and $\mathcal{M} = \{\mathbf{w}^\star\}$. We construct a sequence of possible SR-GD transitions that reaches $\mathbf{w}^\star$ within B_2 steps with probability at least p_d, uniformly over the starting grid point. The construction requires a scalar rounding outcome whose distance and probability can both be bounded.

For $u \in \mathbb{R}$, $g \in \mathcal{G}$, and $x = |\Pi(u) - g|/\Delta$, there is a rounding outcome r satisfying

$$\frac{|r-g|}{\Delta} \le (1+\varepsilon)x, \qquad \Pr(\mathrm{SR}(u) = r) \ge e^{-C_\varepsilon x}. \tag{B.14}$$

If x is an integer, $\Pi(u)$ is a grid point, so take $r = \Pi(u)$. Otherwise the two rounding outcomes have distances $\lfloor x\rfloor\Delta$ and $\lceil x\rceil\Delta$ from g, with probabilities $\lceil x\rceil - x$ and $x - \lfloor x\rfloor$. For $0 < x < 1/2$, choose the nearer outcome. Its probability satisfies $1 - x \ge e^{-2x} \ge e^{-C_\varepsilon x}$. For $x \ge 1/2$ and $\lceil x\rceil \le (1+\varepsilon)x$, both outcomes satisfy the distance bound, so choose the one with probability at least $1/2 \ge e^{-C_\varepsilon x}$. In the remaining case, the nearer outcome satisfies

$$\Pr(\mathrm{SR}(u) = r) = \lceil x\rceil - x > \varepsilon x \ge \varepsilon/2 \ge e^{-C_\varepsilon/2} \ge e^{-C_\varepsilon x}.$$

This proves Eq. (B.14) in every case.

Apply this scalar choice to each coordinate of the update from $\mathbf{w} \neq \mathbf{w}^\star$, with $g = w_i^\star$, and denote the resulting grid point by $\mathbf{r}(\mathbf{w})$. Projection fixes $\mathbf{w}^\star$, so

$$\|\Pi(\mathbf{w}^\star + \mathbf{J}(\mathbf{w} - \mathbf{w}^\star)) - \mathbf{w}^\star\|_2 \leq \|\mathbf{J}(\mathbf{w} - \mathbf{w}^\star)\|_2 \leq \gamma \|\mathbf{w} - \mathbf{w}^\star\|_2.$$

Consequently, the selected outcome satisfies $\|\mathbf{r}(\mathbf{w}) - \mathbf{w}^\star\|_2 \leq q\|\mathbf{w} - \mathbf{w}^\star\|_2$. Independent coordinatewise rounding and $\|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$ also give

$$\begin{aligned} &\Pr(\mathbf{w}^{(1)} = \mathbf{r}(\mathbf{w}) \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}) \\ &\geq \exp\left(-\frac{C_\varepsilon}{\Delta}\|\Pi(\mathbf{w}^\star + \mathbf{J}(\mathbf{w} - \mathbf{w}^\star)) - \mathbf{w}^\star\|_1\right) \\ &\geq \exp\left(-C_\varepsilon \gamma \sqrt{d}\frac{\|\mathbf{w} - \mathbf{w}^\star\|_2}{\Delta}\right). \end{aligned} \tag{B.15}$$

Starting from any $\mathbf{w} \neq \mathbf{w}^\star$, repeat this selection until the first visit to $\mathbf{w}^\star$. After j transitions, the distance is at most $q^j\|\mathbf{w} - \mathbf{w}^\star\|_2 \leq 2M\Delta\sqrt{d}\, q^j$. Since $2M\sqrt{d}\, q^{B_2} < 1$ and distinct grid points have distance at least Δ, the selected transitions reach $\mathbf{w}^\star$ within B_2 steps. Their probability is the product of the transition probabilities. By Eq. (B.15), it is at least

$$\exp\left(-C_\varepsilon \gamma \sqrt{d}\sum_{j=0}^{\infty} 2M\sqrt{d}\, q^j\right) = \exp\left(-\frac{2MC_\varepsilon\gamma}{1-q}d\right) = p_d.$$

Thus the probability of reaching $\mathbf{w}^\star$ within the next B_2 steps is at least p_d from every grid point outside $\mathcal{M}$. Applying Lemma 18 with $B = B_2$ and $p = p_d$ proves the stated bounds by conditioning at successive multiples of B_2. □

B.2.2 Proof of Theorem 2

For the condition-number bound $K \geq 1$, define

$$\gamma \triangleq 1 - \frac{S}{K}, \qquad \varepsilon \triangleq \frac{1-\gamma}{2}, \qquad q \triangleq (1+\varepsilon)\gamma, \qquad C_\varepsilon \triangleq \max\{2, 2\log(2/\varepsilon)\}. \tag{B.16}$$

Then $0 < q < 1$. Set

$$C \triangleq \frac{2MC_\varepsilon\gamma}{1-q}, \qquad B \triangleq \frac{\log(2M) + 1/2}{\log(1/q)} + 2. \tag{B.17}$$

An explicit choice in Theorem 2 is

$$\beta \triangleq C + 2 + \log(2B). \tag{B.18}$$

Proof of Theorem 2. We apply Proposition 31 with $\gamma = 1 - S/K$ and show that its hitting-time bound is at most $e^{\beta d}$ when the failure probability is e^{-d}. The spectral theorem and $\kappa(\mathbf{A}) \leq K$ give

$$\|\mathbf{J}\|_2 = 1 - \eta\lambda_{\min}(\mathbf{A}) = 1 - \frac{S}{\kappa(\mathbf{A})} \leq 1 - \frac{S}{K} = \gamma. \tag{B.19}$$

Since $\mathbf{w}^\star \in \mathcal{W}$, Proposition 31 applies to the stated SR-GD iteration with this γ. The resulting block length and lower bound on the probability of reaching $\mathbf{w}^\star$ within that block are

$$b_d = B_2 = \left\lceil \frac{\log(2M\sqrt{d})}{\log(1/q)} \right\rceil + 1, \qquad p_d = e^{-Cd}.$$

Since $(1-p_d)^j \leq e^{-jp_d}$, the block bound gives $\Pr(\tau_{\mathcal{M}} > H_d \mid \mathbf{A}, \mathbf{w}^{(0)}) \leq e^{-d}$ for

$$H_d \triangleq b_d\lceil d\, p_d^{-1}\rceil.$$

For $d \geq 1$,

$$b_d \leq \frac{\log(2M) + (\log d)/2}{\log(1/q)} + 2 \leq Bd.$$

Since $d\, p_d^{-1} \geq 1$, $\lceil d\, p_d^{-1}\rceil \leq 2d\, p_d^{-1}$. Therefore,

$$H_d \leq 2Bd^2 e^{Cd} \leq \exp\big[\big(C + 2 + \log(2B)\big)d\big] = e^{\beta d}, \tag{B.20}$$

since $\log d \leq d$ and $2B > 1$. Thus $\Pr(\tau_{\mathcal{M}} > e^{\beta d} \mid \mathbf{A}, \mathbf{w}^{(0)}) \leq e^{-d}$. □

B.3 Proof of Proposition 3

Take

$$\rho \triangleq \frac{1-S^2}{144M^2}. \tag{B.21}$$

Proof. By hypothesis, $\mathbf{w}^\star \in \mathcal{W}^\circ$ and

$$\|\mathbf{K}\mathbf{z}\|_1 \geq m\|\mathbf{z}\|_2 \quad \text{for every } \mathbf{z} \in \mathcal{W} - \mathbf{w}^\star \text{ with } 1 \leq \|\mathbf{z}\|_0 \leq r.$$

This is the assumption of Lemma 15. Its constant $\beta = (1-S^2)/(4M^2)$ satisfies

$$k = \rho \min\{m^2, r\} = \frac{\beta}{36}\min\{m^2, r\} \geq 1.$$

Consequently, part (ii) of that lemma applies with exactly this k and gives

$$\bar{h}(\mathbf{w}) \geq 4\sqrt{k \min\{\|\mathbf{w}-\mathbf{w}^\star\|_0, k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}),$$

which is the desired conclusion. □

B.4 Gaussian least squares

B.4.1 A uniform ℓ_1 lower bound under $\|\mathbf{z}\|_0 \leq r$

The estimate below is uniform over $\mathbf{z} \in \mathcal{W} - \mathbf{w}^\star$ satisfying $1 \leq \|\mathbf{z}\|_0 \leq r$. Throughout this subsection, let $\mathbf{X} \in \mathbb{R}^{n\times d}$ have independent standard Gaussian entries and set $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$. Then $\lambda_{\max}(\mathbf{A}) > 0$ almost surely. For an interior target fixed independently of $\mathbf{X}$, Lemma 25 gives $\mathcal{M} = \{\mathbf{w}^\star\}$ almost surely, and $\bar{h}(\mathbf{w}) = \mathbb{E}[h(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}, \mathbf{X}]$.

For $R > 1$, define

$$B_R \triangleq (2+\sqrt{R})^2, \qquad q_R \triangleq \min\{1/(16R), 1/32\}, \qquad \delta_R \triangleq \frac{q_R^2}{16(2+\log(2M))^2}. \tag{B.22}$$

Lemma 32. *Let $R > 1$, with B_R, q_R, δ_R as above. Fix $\mathbf{w}^\star \in \mathcal{W}$ independently of $\mathbf{X}$. Suppose $n \geq d/R$ and $d \geq 2/\delta_R$, and put $r = \lfloor \delta_R d \rfloor$ and $m = Sd/(16\sqrt{2}B_R\sqrt{n})$. With probability at least $1 - 2e^{-3q_R d/4} - e^{-n/2}$ over $\mathbf{X}$,*

$$\lambda_{\max}(\mathbf{A}) \leq B_R, \qquad \|\mathbf{K}\mathbf{z}\|_1 \geq m\|\mathbf{z}\|_2 \quad \textit{for all } \mathbf{z} \in \mathcal{W} - \mathbf{w}^\star \textit{ with } 1 \leq \|\mathbf{z}\|_0 \leq r.$$

Proof. We first lower-bound $\sum_{i:z_i=0} |(\mathbf{A}\mathbf{z})_i|$ for one fixed grid difference $\mathbf{z}$. By Lemma 24, this sum is a product of the square root of a chi-square variable and a sum of independent absolute Gaussian variables, with a deterministic scale factor. We then count the grid differences and take a union bound. Finally, a bound on $\lambda_{\max}(\mathbf{A})$ converts the estimate for $\mathbf{A}$ into the required estimate for $\mathbf{K}$. Fix $\mathbf{z} \neq \mathbf{0}$ with $h = \|\mathbf{z}\|_0 \leq r$, and write $\mathcal{T} = \{i \in [d] : z_i \neq 0\}$ and $s = d - h \geq d/2$. Apply Lemma 24 to this $\mathbf{z}$, with $Z = \|\mathbf{X}\mathbf{z}\|_2^2/(n\|\mathbf{z}\|_2^2)$. Since $z_i = 0$ for $i \in \mathcal{T}^c$, the term $Z\mathbf{z}$ vanishes on these coordinates, and the covariance $\mathbf{I}_d - \mathbf{z}\mathbf{z}^\top/\|\mathbf{z}\|_2^2$ restricts to $\mathbf{I}_s$. The lemma therefore gives

$$(\mathbf{A}\mathbf{z})_{\mathcal{T}^c} = \|\mathbf{z}\|_2\sqrt{Z/n}\,\mathbf{g}_{\mathcal{T}^c}, \qquad Y = nZ \sim \chi_n^2,$$

where $\mathbf{g}_{\mathcal{T}^c}$ is standard Gaussian and independent of Z. Since $\mathbb{E}[e^{-Y/2}] = 2^{-n/2}$, exponential Markov's inequality gives

$$\Pr(Y < n/2) \leq e^{-(n/2)(\log 2 - 1/2)} \leq e^{-n/16}. \tag{B.23}$$

For $\mathcal{E}_\mathbf{z} = \{Y \geq n/2,\ \|\mathbf{g}_{\mathcal{T}^c}\|_1 \geq s/8\}$, Lemma 21 gives $\Pr(\mathcal{E}_\mathbf{z}^c) \leq e^{-n/16} + e^{-s/16} \leq 2e^{-q_R d}$. On $\mathcal{E}_\mathbf{z}$,

$$\sum_{i\notin\mathcal{T}} |(\mathbf{A}\mathbf{z})_i| \geq \frac{s}{8\sqrt{2n}}\|\mathbf{z}\|_2 \geq \frac{d}{16\sqrt{2n}}\|\mathbf{z}\|_2. \tag{B.24}$$

Let $\mathcal{D}_r = \{\mathbf{z} \in \mathcal{W} - \mathbf{w}^\star : 1 \leq \|\mathbf{z}\|_0 \leq r\}$. For a support of size h, each coordinate has at most $2M$ nonzero values, so there are at most $\binom{d}{h}(2M)^h$ such differences. Since $r/d \leq \delta_R < 1/2$, the binomial identity gives

$$\sum_{h=0}^{r}\binom{d}{h} \leq (d/r)^r\big(d/(d-r)\big)^{d-r} \leq (ed/r)^r,$$

where the last inequality uses $\log(1+x) \le x$. The function $x \mapsto x\log(2eM/x)$ is increasing for $0 < x \le 1$. Using this fact, $r/d \le \delta_R$, and $u\log(e/u) \le 2\sqrt{u}$ for $0 < u < 1$, we obtain

$$|\mathcal{D}_r| \le \sum_{h=1}^{r} \binom{d}{h}(2M)^h \le e^{d\delta_R \log(2eM/\delta_R)} \le e^{d(2+\log(2M))\sqrt{\delta_R}} = e^{q_R d/4}.$$

Define $\mathcal{E} = \bigcap_{\mathbf{z}\in\mathcal{D}_r} \mathcal{E}_{\mathbf{z}}$ and $\mathcal{E}_0 = \{\|\mathbf{X}\|_2 \le 2\sqrt{n} + \sqrt{d}\}$. A union bound and Lemma 22 yield $\Pr(\mathcal{E}^c) \le 2e^{-3q_R d/4}$ and $\Pr(\mathcal{E}_0^c) \le e^{-n/2}$. On $\mathcal{E}_0$, $\lambda_{\max}(\mathbf{A}) \le B_R$ and $\eta \ge S/B_R$. Since $(\mathbf{Kz})_i = -\eta(\mathbf{Az})_i$ off the support, Eq. (B.24) proves $\|\mathbf{Kz}\|_1 \ge m\|\mathbf{z}\|_2$ simultaneously on $\mathcal{E}\cap\mathcal{E}_0$. □

B.4.2 Proof of Theorem 4

Put $\widehat{R} = \max\{2, R\}$. Recall the spectral bound, deviation exponent, and support-size coefficient from the uniform Gaussian estimate:

$$B_{\widehat{R}} = (2+\sqrt{\widehat{R}})^2, \qquad q_{\widehat{R}} = \min\{1/(16\widehat{R}), 1/32\},$$
$$\delta_{\widehat{R}} = \frac{q_{\widehat{R}}^2}{16(2+\log(2M))^2}.$$

For this proof, take

$$a = e^{-1}[e(2M+1)]^{-2\widehat{R}}, \qquad \rho = \min\left\{\frac{1}{256}, \frac{1-S^2}{144M^2}\min\left\{\frac{S^2}{512B_{\widehat{R}}^2}, \frac{\delta_{\widehat{R}}}{2}\right\}\right\},$$
$$c = 3q_{\widehat{R}}/8, \qquad d_0 = \left\lceil \max\left\{\frac{2}{\delta_{\widehat{R}}}, \frac{1}{\rho}, \frac{\log 4}{c}\right\}\right\rceil. \tag{B.25}$$

For fixed S, R, these constants give $\rho^{-1}, d_0 = O(M^2(1+\log M)^2)$.

Remark. Part (ii) of Proposition 48 gives alternative admissible choices of the theorem constants satisfying $\rho^{-1}, d_0 = O(\log(8M))$, while c remains bounded away from zero uniformly in M. The refined curvature coefficient a is specified separately in Eq. (F.16). The proof in Appendix F.3.4 strengthens the conditional-mean conclusion of Proposition 3 using Gaussian probability bounds summed over displacements grouped by Euclidean norm if $n \ge d$, and the refined Haar estimate if $n < d$. Applying Lemma 16 again yields alternative admissible choices of the theorem constants.

Proof of Theorem 4. We use the same argument as in the proof sketch: verify the assumption of Proposition 3, obtain the conditional-mean bound, and apply Lemma 16. Here Lemma 32 supplies the required uniform estimate with m proportional to $d/\sqrt{n}$ and r proportional to d, leading to the threshold $k = \rho\min\{d, d^2/n\}$. We also apply Lemma 23 to bound the restricted quadratic form from below. This verifies the theorem's restricted eigenvalue conclusion and converts the Hamming-distance bound into an objective lower bound. Fix $d \ge d_0$ and $d/R \le n \le \rho d^2$, so $k = \rho\min\{d, d^2/n\} \ge 1$, and define

$$\mathcal{E}_A(a,k) \triangleq \left\{\lambda_{\min}(\mathbf{A}; \mathcal{W}-\mathbf{w}^\star) \ge a, \quad \bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}} \quad (\mathbf{w}\in\mathcal{W}\setminus\{\mathbf{w}^\star\})\right\}.$$

Since $n \ge d/R \ge d/\widehat{R}$ and $d \ge 2/\delta_{\widehat{R}}$, Lemma 32 gives

$$m = \frac{Sd}{16\sqrt{2}B_{\widehat{R}}\sqrt{n}}, \qquad r = \lfloor \delta_{\widehat{R}} d\rfloor \ge \delta_{\widehat{R}} d/2,$$
$$\|\mathbf{Kz}\|_1 \ge m\|\mathbf{z}\|_2 \quad (\mathbf{z}\in\mathcal{W}-\mathbf{w}^\star,\ 1 \le \|\mathbf{z}\|_0 \le r).$$

On this event, Proposition 3 yields $\bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}}$ for $1 \le h(\mathbf{w}) \le d$, because

$$\frac{1-S^2}{144M^2}\min\{m^2, r\} \ge \frac{1-S^2}{144M^2}\min\left\{\frac{S^2}{512B_{\widehat{R}}^2}\frac{d^2}{n}, \frac{\delta_{\widehat{R}}}{2}d\right\}$$
$$\ge \rho\min\{d, d^2/n\} = k.$$

Lemma 23 gives $\Pr(\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) < a) \le e^{-d}$. Hence, since $q_{\widehat{R}} \le 1/(16\widehat{R})$,

$$\Pr(\mathcal{E}_A(a,k)^c) \le 2e^{-3q_{\widehat{R}}d/4} + e^{-n/2} + e^{-d} \le 4e^{-2cd} \le e^{-cd}.$$

We next combine the matrix probability with the conditional trajectory probability. Keeping the factor $99/100$ in Lemma 16 gives the stated bound $1 - e^{-c\sqrt{k}}$ also when k is small. The remaining argument applies whenever $\Pr(\mathcal{E}_A(a,k)^c) \le e^{-cd}$, $0 < c \le 1/16$, $\rho < 1/128$, and $d_0 \ge 1/\rho$. Define

$$\mathcal{E}_1 \triangleq \bigcap_{\substack{t \in \mathbb{Z}_{\ge 0} \\ t < e^k}} \left\{ h^{(t+1)} \ge \min\{k, 2\sqrt{k h^{(t)}}\} \right\}.$$

Put $b_0 = 1/16$, $x = \sqrt{k}$, and $\zeta = 99/100$. On $\mathcal{E}_A(a,k)$, $\mathcal{M} = \{\mathbf{w}^\star\}$. Conditional on $\mathbf{A}$, Lemma 16 gives $\Pr(\mathcal{E}_1^c \mid \mathbf{A}) \le \zeta e^{-b_0 x}$ on $\mathcal{E}_A(a,k)$. Since $x^2 = k \le \rho d$ and $d \ge 1/\rho$, $x \le \rho d$, $d - x \ge \rho^{-1} - 1 > 127$, and $e^{-b_0(d-x)} < 1 - \zeta$. Therefore

$$(1 - e^{-b_0 d})(1 - \zeta e^{-b_0 x}) \ge 1 - e^{-b_0 x}.$$

For $s > 0$, $F(s) = (1 - e^{-sx})/(1 - e^{-sd})$ is increasing, since $0 < x < d$ and

$$\frac{d}{ds} \log F(s) = \frac{x}{e^{sx} - 1} - \frac{d}{e^{sd} - 1} \ge 0.$$

Thus

$$\frac{1 - e^{-cx}}{1 - e^{-cd}} \le F(b_0) \le 1 - \zeta e^{-b_0 x},$$

and

$$\Pr(\mathcal{E}_A(a,k) \cap \mathcal{E}_1) \ge (1 - e^{-cd})(1 - \zeta e^{-b_0 x}) \ge 1 - e^{-cx} = 1 - e^{-c\sqrt{k}}.$$

Every nonzero coordinate of a grid difference has absolute value at least Δ. Thus, on $\mathcal{E}_A(a,k)$,

$$\begin{aligned} f(\mathbf{w}^{(t)}) &\ge \frac{1}{2}\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star)\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|_2^2 \\ &\ge \frac{\Delta^2}{2}\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) h^{(t)} \ge \frac{a}{2}\Delta^2 h^{(t)} \qquad (t \in \mathbb{Z}_{\ge 0}). \end{aligned}$$

On $\mathcal{E}_1$, Eq. (A.26) gives $h^{(t)} \ge k$ for every integer t satisfying

$$\log_2\left(1 + \tfrac{1}{2}\log_2 k\right) \le t \le e^k.$$

Thus $f(\mathbf{w}^{(t)}) \ge (a/2)\Delta^2 k$ throughout this interval. All distances through time e^k are positive, so $\tau_{\mathcal{M}} > e^k$. □

C Logarithmic hitting times

C.1 Proof of Proposition 5

Proof. Fix the Hessian and initializer. The assumed one-step inequality and conditional expectation give

$$\mathbb{E}[\psi(\mathbf{w}^{(t+1)}) \mid \mathcal{F}_t] \le q\psi(\mathbf{w}^{(t)}).$$

Taking expectations and iterating yields $\mathbb{E}[\psi(\mathbf{w}^{(t)}) \mid \mathbf{w}^{(0)}] \le q^t \psi(\mathbf{w}^{(0)})$. If $\tau_{\mathcal{M}} > t$, then $\mathbf{w}^{(t)} \ne \mathbf{w}^\star$ and $\psi(\mathbf{w}^{(t)}) \ge 1$. Thus Markov's inequality gives

$$\Pr(\tau_{\mathcal{M}} > t \mid \mathbf{w}^{(0)}) \le \Pr(\psi(\mathbf{w}^{(t)}) \ge 1 \mid \mathbf{w}^{(0)}) \le \mathbb{E}[\psi(\mathbf{w}^{(t)}) \mid \mathbf{w}^{(0)}] \le q^t \psi(\mathbf{w}^{(0)}).$$

□

C.2 Contracting potentials from local drift and hitting probabilities

We construct a potential that contracts at every grid point from estimates on the absolute distance to the target. The first lemma establishes contraction when this distance exceeds a threshold and a positive probability of hitting the target when it does not. The second lemma uses these estimates to bound the exponential moment of the hitting time.

Fix $\mathbf{w}^\star \in \mathcal{W}$ and assume $\ker(\mathbf{A}) \cap (\mathcal{W} - \mathbf{w}^\star) = \{\mathbf{0}\}$, so $\mathcal{M} = \{\mathbf{w}^\star\}$. Define

$$V(\mathbf{w}) \triangleq \|\mathbf{w} - \mathbf{w}^\star\|_1/\Delta \in \{0, \dots, 2Md\}.$$

Absorption gives $\{\tau_{\mathcal{M}} > t\} = \{\mathbf{w}^{(t)} \ne \mathbf{w}^\star\}$.

Local drift and hitting probabilities.

Lemma 33. *Under the assumptions of Appendix C.2, suppose $b \geq 0$, $g \in (0,1)$, $J_{ii} \in [0, 1-g]$, and $\sqrt{d}\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 \leq b$. Let*

$$r \triangleq \left\lceil \max\{1, 8Mb^2/g^2\} \right\rceil, \qquad q \triangleq 1 - g/2,$$
$$j \triangleq \left\lceil \frac{\log(2M+1)}{-\log q} \right\rceil, \qquad p \triangleq (g/2)^r e^{-4b\sqrt{2Mr}/g}.$$

Set $\mathcal{C} \triangleq \{\mathbf{w} \in \mathcal{W} : V(\mathbf{w}) \leq r\}$. If $d \geq r^2$, the following hold:

$$\begin{aligned} \mathbb{E}[V(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] &\leq qV(\mathbf{w}) \quad (\mathbf{w} \notin \mathcal{C}), \\ \Pr(\tau_{\mathcal{M}} \leq j \mid \mathbf{w}^{(0)} = \mathbf{w}) &\geq p \quad (\mathbf{w} \in \mathcal{C}), \\ \mathbb{E}[V(\mathbf{w}^{(j)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] &\leq r \quad (\mathbf{w} \in \mathcal{C}). \end{aligned} \tag{C.1}$$

The last bound holds at every deterministic integer time $t \geq 0$.

Proof. We first bound the expected absolute distance. The same bound gives contraction when $V(\mathbf{w}) > r$ and controls the expected distance at later times when $V(\mathbf{w}) \leq r$. To prove the hitting-probability bound from $\mathcal{C}$, we then require every nonzero coordinate to decrease in absolute value while coordinates already equal to their targets remain fixed. Fix $\mathbf{w}^{(0)} = \mathbf{w}$ and write $\widetilde{\mathbf{z}} \triangleq (\mathbf{w} - \mathbf{w}^\star)/\Delta$. Then $\|\widetilde{\mathbf{z}}\|_1 = V(\mathbf{w})$ and $\|\widetilde{\mathbf{z}}\|_2^2 \leq 2MV(\mathbf{w})$. The scalar rounding bound gives

$$\begin{aligned} \mathbb{E}[V(\mathbf{w}^{(1)})] &\leq \|\mathbf{J}\widetilde{\mathbf{z}}\|_1 \leq (1-g)V(\mathbf{w}) + \|\mathbf{K}\widetilde{\mathbf{z}}\|_1 \\ &\leq (1-g)V(\mathbf{w}) + b\sqrt{2MV(\mathbf{w})}. \end{aligned} \tag{C.2}$$

Here $\|\mathbf{K}\widetilde{\mathbf{z}}\|_1 \leq \sqrt{d}\|\mathbf{K}\|_2\|\widetilde{\mathbf{z}}\|_2$. For $v \geq r$, the inequality $r \geq 8Mb^2/g^2$ gives $b\sqrt{2Mv} \leq gv/2$. Substituting this into Eq. (C.2) proves the drift bound outside $\mathcal{C}$. Since $F(v) = (1-g)v + b\sqrt{2Mv}$ is increasing and concave with $F(r) \leq r$, Jensen's inequality gives, whenever $\mathbb{E}[V(\mathbf{w}^{(t)})] \leq r$,

$$\mathbb{E}[V(\mathbf{w}^{(t+1)})] \leq \mathbb{E}[F(V(\mathbf{w}^{(t)}))] \leq F(\mathbb{E}[V(\mathbf{w}^{(t)})]) \leq F(r) \leq r.$$

Starting with $V(\mathbf{w}^{(0)}) \leq r$, induction proves the last bound in Eq. (C.1) at every deterministic time. The argument requires only the expectation bound at the preceding time.

We next prove the probability of reaching $\mathbf{w}^\star$ from $\mathcal{C}$. For $V(\mathbf{w}) \leq r$, we also have

$$\|\mathbf{K}\widetilde{\mathbf{z}}\|_\infty \leq \frac{b}{\sqrt{d}}\sqrt{2Mr} \leq \frac{gr}{2\sqrt{d}} \leq g/2. \tag{C.3}$$

If $\widetilde{z}_i = \ell \geq 1$, then

$$J_{ii}\ell + (\mathbf{K}\widetilde{\mathbf{z}})_i \in [-g/2, (1-g)\ell + g/2].$$

Set $a_* = g/2$, so $q = 1 - a_*$, and let $\widetilde{\mathbf{z}}' \triangleq (\mathbf{w}^{(1)} - \mathbf{w}^\star)/\Delta$. Rounding gives, with probability at least a_*,

$$|\widetilde{z}_i'| \leq q\ell. \tag{C.4}$$

For $\ell = 1$, the input belongs to $[-a_*, 1-a_*]$. Its distance to zero is at most $1 - a_*$, so rounding gives zero with probability at least a_*. For $\ell \geq 2$, put $L_i = \lfloor q\ell \rfloor \geq 1$. The negative endpoint rounds to 0 or -1, while

$$(1-g)\ell + g/2 = q\ell - a_*(\ell - 1) \leq L_i + 1 - a_*.$$

Rounding to a value greater than L_i therefore has probability at most $1 - a_*$, and a value below $-L_i$ is impossible. Thus Eq. (C.4) holds with probability at least a_*. Clipping cannot decrease this probability; sign reversal handles $\widetilde{z}_i < 0$.

If $\widetilde{z}_i = 0$, then $\Pr(\widetilde{z}_i' \neq 0) \leq |(\mathbf{K}\widetilde{\mathbf{z}})_i| \leq g/2 < 1/2$. Independence and $\log(1-u) \geq -2u$ for $u \in [0, 1/2]$ give

$$\Pr(\widetilde{z}_i' = 0 \text{ for all } i \text{ with } \widetilde{z}_i = 0 \mid \mathbf{w}^{(0)} = \mathbf{w}) \geq e^{-2\|\mathbf{K}\widetilde{\mathbf{z}}\|_1} \geq e^{-2b\|\widetilde{\mathbf{z}}\|_2}.$$

For $\widetilde{\mathbf{z}}_t \triangleq (\mathbf{w}^{(t)} - \mathbf{w}^\star)/\Delta$, define

$$\mathcal{E}_t \triangleq \bigcap_{s=0}^{t-1} \bigcap_{i=1}^{d} \{|\widetilde{z}_{s+1,i}| \le q|\widetilde{z}_{s,i}|\}, \qquad \mathcal{E}_0 \triangleq \Omega.$$

On $\mathcal{E}_t$,

$$\|\widetilde{\mathbf{z}}_t\|_2 \le q^t \|\widetilde{\mathbf{z}}_0\|_2 \le q^t\sqrt{2Mr}, \qquad \|\widetilde{\mathbf{z}}_t\|_\infty \le 2Mq^t.$$

Also $V(\mathbf{w}^{(t)}) \le q^t V(\mathbf{w}^{(0)}) \le r$, so the bounds remain applicable. Since $2Mq^j \le 2M/(2M+1) < 1$, $V(\mathbf{w}^{(j)}) = 0$ on $\mathcal{E}_j$. Set $V_t \triangleq V(\mathbf{w}^{(t)})$, $m_t \triangleq \|\widetilde{\mathbf{z}}_t\|_0$, and $s_t \triangleq \exp\{-2bq^t\sqrt{2Mr}\}$. On $\mathcal{E}_t$, conditional independence gives $\Pr(\mathcal{E}_{t+1} \mid \mathcal{F}_t) \ge a_*^{m_t} s_t$. On $\mathcal{E}_{t+1}$, $V_{t+1} \le V_t - m_t$, because each nonzero integer coordinate decreases in absolute value by at least one when its new absolute value is at most $q\ell < \ell$. Since $0 < a_* < 1$, weighting by $a_*^{V_{t+1}}$ eliminates the factor $a_*^{m_t}$ from the conditional lower bound:

$$\mathbb{E}[\mathbb{1}\{\mathcal{E}_{t+1}\} a_*^{V_{t+1}} \mid \mathcal{F}_t] \ge \mathbb{1}\{\mathcal{E}_t\}\, a_*^{m_t} s_t a_*^{V_t - m_t} = s_t \mathbb{1}\{\mathcal{E}_t\}\, a_*^{V_t}.$$

Iteration and $V_j = 0$ on $\mathcal{E}_j$ yield

$$\Pr(\mathcal{E}_j) \ge a_*^{V_0} \prod_{t=0}^{j-1} s_t \ge (g/2)^r \exp\left\{-2b\sqrt{2Mr}\sum_{t=0}^{\infty} q^t\right\} = (g/2)^r e^{-4b\sqrt{2Mr}/g} = p.$$

Absorption includes paths hitting before j. □

Exponential moments and a contracting potential.

Lemma 34. *Under the assumptions of Appendix C.2, let $r, j \ge 1$ be integers, $q \in (0,1)$, and $p \in (0,1]$. Set $\mathcal{C} \triangleq \{\mathbf{w} \in \mathcal{W} : V(\mathbf{w}) \le r\}$ and suppose that*

$$\begin{aligned} \mathbb{E}[V(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] &\le qV(\mathbf{w}) \quad (\mathbf{w} \notin \mathcal{C}), \\ \Pr(\tau_{\mathcal{M}} \le j \mid \mathbf{w}^{(0)} = \mathbf{w}) &\ge p \qquad (\mathbf{w} \in \mathcal{C}), \\ \mathbb{E}[V(\mathbf{w}^{(j)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] &\le r \qquad (\mathbf{w} \in \mathcal{C}). \end{aligned}$$

Define

$$\theta \triangleq \frac{p}{12(j + 1/(1-q))}, \qquad c_\theta \triangleq \frac{e^\theta - 1}{1 - qe^\theta}. \tag{C.5}$$

Then $qe^\theta < 1$, $c_\theta \le p/5$, and

$$\mathbb{E}[e^{\theta\tau_{\mathcal{M}}} \mid \mathbf{w}^{(0)} = \mathbf{w}] \le \frac{2e^{j\theta}}{p}\left(1 + c_\theta \frac{V(\mathbf{w})}{r+1}\right) < \infty. \tag{C.6}$$

Define also

$$t_0(\mathbf{w}) \triangleq \left\lceil \frac{\log \max\{1, V(\mathbf{w})/(r+1)\}}{-\log q} \right\rceil. \tag{C.7}$$

The sharper bound

$$\mathbb{E}[e^{\theta\tau_{\mathcal{M}}} \mid \mathbf{w}^{(0)} = \mathbf{w}] \le \frac{3}{p} e^{\theta t_0(\mathbf{w})} \tag{C.8}$$

also holds. Define

$$\psi(\mathbf{w}^\star) \triangleq 0, \qquad \psi(\mathbf{w}) \triangleq \mathbb{E}[e^{\theta\tau_{\mathcal{M}}} \mid \mathbf{w}^{(0)} = \mathbf{w}] \quad (\mathbf{w} \ne \mathbf{w}^\star). \tag{C.9}$$

Then $\psi(\mathbf{w}) \ge 1$ for $\mathbf{w} \ne \mathbf{w}^\star$ and $\mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] \le e^{-\theta}\psi(\mathbf{w})$.

Proof. The main task is to prove that the exponential moment of $\tau_{\mathcal{M}}$ is finite with the stated bounds. First-step conditioning will then give contraction of the potential. We first bound the time to enter $\mathcal{C}$. For an initial state in $\mathcal{C}$, we combine the probability of hitting $\mathbf{w}^\star$ within j steps with the time to return to $\mathcal{C}$ when $\tau_{\mathcal{M}} > j$. Fix $\mathbf{w}^{(0)} = \mathbf{w}$ and set $V_t \triangleq V(\mathbf{w}^{(t)})$ and $\tau_{\mathcal{C}} \triangleq \inf\{t \ge 0 : \mathbf{w}^{(t)} \in \mathcal{C}\}$. Since V_t is integer-valued, $V_t \ge r+1$ before $\tau_{\mathcal{C}}$. Apply Lemma 17 with $Z_t = V_t$, $\tau = \tau_{\mathcal{C}}$, and $a = r+1$ to obtain

$$\Pr(\tau_{\mathcal{C}} > t) \le q^t V(\mathbf{w})/(r+1). \tag{C.10}$$

In particular, $\tau_{\mathcal{C}} < \infty$ almost surely.

The choice of θ controls both the first j steps and the return time. Since $j\theta \le p/12$, $\theta \le p(1-q)/12 \le 1/12$, and $e^u \le 1 + 2u$ for $u \in [0, 1/2]$,

$$1 - qe^{\theta} \ge \tfrac{5}{6}(1-q) > 0, \qquad c_\theta \le \frac{2\theta}{(5/6)(1-q)} \le p/5, \qquad e^{j\theta} \le 1 + p/6. \tag{C.11}$$

For this integer-valued entry time, Tonelli's theorem gives

$$\mathbb{E}[e^{\theta\tau_{\mathcal{C}}}] = 1 + (e^{\theta} - 1)\sum_{t=0}^{\infty} e^{\theta t}\Pr(\tau_{\mathcal{C}} > t).$$

The identity holds before finiteness of the expectation has been established. Using Eq. (C.10) and $qe^{\theta} < 1$ in the series yields

$$\mathbb{E}[e^{\theta\tau_{\mathcal{C}}}] \le 1 + c_\theta \frac{V(\mathbf{w})}{r+1} < \infty. \tag{C.12}$$

For the second bound, write $t_0 = t_0(\mathbf{w})$ and split the same series at t_0. Use $\Pr(\tau_{\mathcal{C}} > t) \le 1$ before t_0 and Eq. (C.10) afterwards. Since $q^{t_0}V(\mathbf{w})/(r+1) \le 1$, this gives

$$\begin{aligned}\mathbb{E}[e^{\theta\tau_{\mathcal{C}}}] &\le e^{\theta t_0} + (e^{\theta} - 1)\frac{V(\mathbf{w})}{r+1}\frac{(qe^{\theta})^{t_0}}{1 - qe^{\theta}} \\ &\le e^{\theta t_0}(1 + c_\theta).\end{aligned} \tag{C.13}$$

For $\mathbf{w} \in \mathcal{C}$, define $\mathcal{E} \triangleq \{\tau_{\mathcal{M}} > j\}$ and $\nu \triangleq \inf\{t \ge j : \mathbf{w}^{(t)} \in \mathcal{C}\}$. By $\Pr(\tau_{\mathcal{C}} > t) \le q^t V(\mathbf{w})/(r+1)$ and the Markov property at j, $\nu < \infty$ almost surely. Since $\nu \ge j$, $\mathcal{E} \in \mathcal{F}_\nu$; since $\mathbf{w}^\star \in \mathcal{C}$, $\nu \le \tau_{\mathcal{M}}$ on $\mathcal{E}$. We now show that the exponential factor accumulated up to ν, restricted to $\mathcal{E}$, has expectation strictly less than one. This will bound the contribution from paths that have not reached $\mathbf{w}^\star$ during their first j steps. Using $\Pr(\mathcal{E}) \le 1 - p$ and Eq. (C.12) conditionally on $\mathcal{F}_j$, we obtain

$$\begin{aligned}\mathbb{E}[e^{\theta\nu}\mathbb{1}\{\mathcal{E}\}] &\le e^{j\theta}\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\left(1 + c_\theta\frac{V_j}{r+1}\right)\right] \\ &\le e^{j\theta}\left(1 - p + c_\theta\frac{r}{r+1}\right) \le (1 + p/6)(1 - p + p/5) \le 1 - p/2.\end{aligned}$$

The second line uses $\mathbb{E}[V_j] \le r$. We apply this estimate first to truncated hitting times, so that taking a supremum does not assume the finiteness we are proving. For each integer $N \ge 1$, set

$$F_N(\mathbf{v}) \triangleq \mathbb{E}[e^{\theta(\tau_{\mathcal{M}} \wedge N)} \mid \mathbf{w}^{(0)} = \mathbf{v}], \qquad H_N \triangleq \sup_{\mathbf{v}\in\mathcal{C}} F_N(\mathbf{v}) \le e^{\theta N} < \infty.$$

On $\mathcal{E}$, $\tau_{\mathcal{M}} \wedge N \le \nu + ((\tau_{\mathcal{M}} - \nu) \wedge N)$. On $\mathcal{E}^c$, the exponential factor is at most $e^{j\theta}$. On $\mathcal{E}$, the strong Markov property at ν bounds the remaining truncated moment by H_N, because $\mathbf{w}^{(\nu)} \in \mathcal{C}$. Consequently,

$$F_N(\mathbf{w}) \le e^{j\theta} + \mathbb{E}[e^{\theta\nu}\mathbb{1}\{\mathcal{E}\}F_N(\mathbf{w}^{(\nu)})] \le e^{j\theta} + (1 - p/2)H_N.$$

Taking the supremum over $\mathbf{w} \in \mathcal{C}$ gives

$$H_N \le e^{j\theta} + (1 - p/2)H_N, \qquad \frac{p}{2}H_N \le e^{j\theta}.$$

Thus $H_N \le 2e^{j\theta}/p$ uniformly in N. Monotone convergence as $N \to \infty$ yields

$$\sup_{\mathbf{w}\in\mathcal{C}} \mathbb{E}[e^{\theta\tau_{\mathcal{M}}} \mid \mathbf{w}^{(0)} = \mathbf{w}] \le \frac{2e^{j\theta}}{p} < \infty. \tag{C.14}$$

For an arbitrary initial state $\mathbf{w}$, the strong Markov property at $\tau_{\mathcal{C}}$ and Eq. (C.14) give

$$\mathbb{E}[e^{\theta\tau_{\mathcal{M}}}] \le \frac{2e^{j\theta}}{p}\mathbb{E}[e^{\theta\tau_{\mathcal{C}}}].$$

Substituting Eq. (C.12) proves Eq. (C.6). Substituting Eq. (C.13) gives

$$\mathbb{E}[e^{\theta\tau_{\mathcal{M}}}] \le \frac{2e^{j\theta}}{p} e^{\theta t_0}(1+c_\theta) \le \frac{3}{p} e^{\theta t_0},$$

since $j\theta \le 1/12$, $c_\theta \le 1/5$, and $2e^{1/12}(6/5) < 3$. Both moments are therefore finite.

Finiteness now justifies using ψ as a potential. For $\mathbf{w} \neq \mathbf{w}^\star$, a transition directly to $\mathbf{w}^\star$ has $\tau_{\mathcal{M}} = 1$ and contributes e^θ to the exponential moment. Every other transition contributes e^θ times the remaining moment. Since $\psi(\mathbf{w}^\star) = 0$, first-step conditioning gives

$$\psi(\mathbf{w}) = e^\theta \left(\mathbb{E}[\psi(\mathbf{w}^{(1)})] + \Pr(\mathbf{w}^{(1)} = \mathbf{w}^\star) \right). \tag{C.15}$$

Dropping the probability term proves contraction; at $\mathbf{w}^\star$, both sides vanish. Elsewhere, $\tau_{\mathcal{M}} \ge 1$ gives $\psi \ge e^\theta > 1$. □

C.3 Matrix families with logarithmic hitting times

We verify the contraction condition of Proposition 5 for three matrix families. Weighted absolute distances suffice if $\||\mathbf{J}|\|_2$ is bounded by a constant less than one; the preceding construction handles near-diagonal Hessians, and a separate construction handles bounded blocks.

C.3.1 Weighted absolute-distance contraction

Proposition 35. *Assume $\mathbf{w}^\star \in \mathcal{W}$ and $\||\mathbf{J}|\|_2 \le \gamma < 1$. Set $q = (1+\gamma)/2$ and define*

$$\mathbf{y} \triangleq \sum_{k=0}^{\infty} (|\mathbf{J}|/q)^k \mathbf{1}, \qquad \psi_{\mathbf{y}}(\mathbf{w}) \triangleq \langle \mathbf{y}, |\mathbf{w} - \mathbf{w}^\star| \rangle / \Delta.$$

The series converges, $y_i \ge 1$, and $\sum_i y_i \le d(1+\gamma)/(1-\gamma)$. Moreover, $\psi_{\mathbf{y}}(\mathbf{w}^\star) = 0$, $\psi_{\mathbf{y}}(\mathbf{w}) \ge 1$ for $\mathbf{w} \neq \mathbf{w}^\star$, $\psi_{\mathbf{y}} \le 2Md(1+\gamma)/(1-\gamma)$, and $\mathbb{E}[\psi_{\mathbf{y}}(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] \le q\psi_{\mathbf{y}}(\mathbf{w})$.

Proof. The weights satisfy $|\mathbf{J}|\mathbf{y} \le q\mathbf{y}$. This inequality will turn the coordinatewise rounding bound into contraction of $\psi_{\mathbf{y}}$. Since $\||\mathbf{J}|/q\|_2 \le \gamma/q < 1$, the series converges in operator norm and has nonnegative entries, so $\mathbf{1} \le \mathbf{y}$. Cauchy–Schwarz gives

$$\mathbf{1}^\top \mathbf{y} = \sum_{k \ge 0} \mathbf{1}^\top (|\mathbf{J}|/q)^k \mathbf{1} \le d \sum_{k \ge 0} (\gamma/q)^k = d\frac{1+\gamma}{1-\gamma}.$$

The series gives $|\mathbf{J}|\mathbf{y} = q(\mathbf{y} - \mathbf{1}) \le q\mathbf{y}$. For $\mathbf{z} = \mathbf{w} - \mathbf{w}^\star$, Lemma 13 gives $\mathbb{E}[|w_i^{(1)} - w_i^\star| \mid \mathbf{w}^{(0)} = \mathbf{w}] \le |(\mathbf{J}\mathbf{z})_i| \le \sum_k |J_{ik}|\,|z_k|$. Summing with weights y_i and using symmetry yields

$$\mathbb{E}[\psi_{\mathbf{y}}(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] \le \Delta^{-1} \mathbf{y}^\top |\mathbf{J}|\,|\mathbf{z}| = \Delta^{-1} (|\mathbf{J}|\mathbf{y})^\top |\mathbf{z}| \le q\psi_{\mathbf{y}}(\mathbf{w}).$$

Grid spacing gives $\psi_{\mathbf{y}}(\mathbf{w}) \ge 1$ for $\mathbf{w} \neq \mathbf{w}^\star$; $|z_i| \le 2M\Delta$ gives its upper bound.

□

Matrix families satisfying the contraction condition. If $\mathbf{A} \succ \mathbf{0}$ and $A_{ij} \le 0$ for $i \neq j$, then $J_{ii} \ge 0$ and $J_{ij} = -\eta A_{ij} \ge 0$, so

$$|\mathbf{J}| = \mathbf{J}, \qquad \||\mathbf{J}|\|_2 = 1 - S/\kappa(\mathbf{A}).$$

Thus bounded condition number suffices. For fixed $a, b > 0$ and $d \ge 2$, $\mathbf{A} = (a+b)\mathbf{I}_d - (b/d)\mathbf{1}\mathbf{1}^\top$ has eigenvalues $a, a+b$ and

$$\||\mathbf{J}|\|_2 = 1 - \frac{Sa}{a+b}.$$

If $A_{ii} - \sum_{j \neq i} |A_{ij}| \ge \delta \lambda_{\max}(\mathbf{A})$ for all i and fixed $\delta > 0$, then $J_{ii} = 1 - \eta A_{ii} \ge 0$ and symmetry give

$$\||\mathbf{J}|\|_2 \le \max_i \sum_j |J_{ij}| = 1 - \eta \min_i \left(A_{ii} - \sum_{j \neq i} |A_{ij}| \right) \le 1 - S\delta.$$

Spectral norm of the entrywise absolute update matrix. Under the assumptions of Theorem 1, for all sufficiently large d, with probability at least $1 - e^{-c_H d/2}$ over $\mathbf{U}$,

$$\|\mathbf{J}\|_2 \le 1 - Sa/b < 1, \qquad u\sqrt{d} \le \big\||\mathbf{J}|\big\|_2 \le \sqrt{d},$$
$$u = \frac{Sv}{128b}, \qquad c_H = \frac{1}{256}\min\{1, (v/(b-a))^4\}.$$

Indeed, Lemma 30 gives, with probability at least $1-e^{-c_H d/2}$, $\sum_{i\neq j}|J_{ij}| \ge u\sqrt{d}$ for every $j \in [d]$, so

$$\big\||\mathbf{J}|\big\|_2 \ge d^{-1}\mathbf{1}^\top|\mathbf{J}|\mathbf{1} \ge u\sqrt{d}.$$

Deterministically, $\big\||\mathbf{J}|\big\|_2 \le (\sum_{i,j} J_{ij}^2)^{1/2} = \sqrt{\mathrm{tr}(\mathbf{J}^2)} \le \sqrt{d}$ and $\|\mathbf{J}\|_2 = 1 - S\lambda_{\min}(\mathbf{A})/\lambda_{\max}(\mathbf{A}) \le 1 - Sa/b$.

C.3.2 Contraction under an off-diagonal norm bound

Proposition 36. *Under the assumptions of Appendix C.2, suppose $\|\mathbf{J}\|_2 \le \gamma$ for $\gamma \in (0,1)$ and $\sqrt{d}\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 \le \delta$. Set*

$$g \triangleq 1-\gamma, \qquad r \triangleq \lceil \max\{1, 8M\delta^2/g^2\} \rceil, \qquad q \triangleq 1 - g/2,$$
$$j \triangleq \left\lceil \frac{\log(2M+1)}{-\log q} \right\rceil, \qquad p \triangleq (g/2)^r e^{-4\delta\sqrt{2Mr}/g}, \tag{C.16}$$
$$\theta \triangleq \frac{p}{12(j+1/(1-q))}, \qquad c_\theta \triangleq \frac{e^\theta - 1}{1 - qe^\theta}.$$

If $d \ge r^2$, there is a potential satisfying

$$\mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] \le e^{-\theta}\psi(\mathbf{w}), \qquad \psi(\mathbf{w}) \le \xi d, \qquad \xi \triangleq \frac{2e^{j\theta}}{p}\left(1 + \frac{2Mc_\theta}{r+1}\right).$$

Proof. Since $J_{ii} \in [1-S, \gamma]$, Lemma 33 applies with $b = \delta$, $g = 1-\gamma$. By Lemma 34, $\psi(\mathbf{w}^\star) = 0$ and $\psi(\mathbf{w}) = \mathbb{E}[e^{\theta\tau_{\mathcal{M}}} \mid \mathbf{w}^{(0)} = \mathbf{w}]$ for $\mathbf{w} \neq \mathbf{w}^\star$ define a contracting potential satisfying

$$\psi(\mathbf{w}) \le \frac{2e^{j\theta}}{p}\left(1 + \frac{c_\theta V(\mathbf{w})}{r+1}\right) \le \xi d,$$

since $V(\mathbf{w}) \le 2Md$. □

A near-diagonal example with $\big\||\mathbf{J}|\big\|_2 > 1$. For $d = 2^k$, define the symmetric Sylvester matrix by $\mathbf{H}_1 = (1)$ and $\mathbf{H}_{2d} = \left(\begin{smallmatrix} \mathbf{H}_d & \mathbf{H}_d \\ \mathbf{H}_d & -\mathbf{H}_d \end{smallmatrix}\right)$. Then $\mathbf{H}_d^2 = d\mathbf{I}_d$, and for $d \ge 16$, $\mathbf{A} = \mathbf{I}_d - 2\mathbf{H}_d/d$ has spectrum in $[1/2, 3/2]$. Consequently, $\eta \le 2S$ and $\|\mathbf{J}\|_2 = 1 - S\lambda_{\min}(\mathbf{A})/\lambda_{\max}(\mathbf{A}) \le 1 - S/3$. Since every entry of $\mathbf{H}_d$ is 1 or -1, $\|\mathbf{H}_d - \mathrm{diag}(\mathbf{H}_d)\|_2 \le \sqrt{d} + 1$, so

$$\sqrt{d}\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 = \frac{2\eta}{\sqrt{d}}\|\mathbf{H}_d - \mathrm{diag}(\mathbf{H}_d)\|_2 \le 2\eta(1 + d^{-1/2}) \le 6S.$$

However, $J_{ii} \ge 0$ and $|H_{ij}| = 1$ give

$$\sum_j |J_{ij}| = 1 + \eta + \frac{2\eta}{d}(H_{ii} - 1) \ge 1 + \eta(1 - 4/d) > 1.$$

Thus $\big\||\mathbf{J}|\big\|_2 \ge d^{-1}\mathbf{1}^\top|\mathbf{J}|\mathbf{1} > 1$.

C.3.3 Block-diagonal matrices with block size at most B

Proposition 37. *Under the assumptions of Appendix C.2, suppose $\mathbf{A}$ is block diagonal with blocks of size at most $B \in \mathbb{N}$ and $\|\mathbf{J}\|_2 \le \gamma$ for $\gamma \in (0,1)$. There is a potential satisfying Proposition 5 with $\psi \le \xi d$, where ξ and $q < 1$ depend only on M, B, γ.*

Proof. Apply Proposition 31 to each block and convert the resulting contraction over finitely many steps into one-step contraction. Set

$$\varepsilon = (1-\gamma)/2, \qquad \zeta = (1+\varepsilon)\gamma < 1, \qquad C_\varepsilon = \max\{2, 2\log(2/\varepsilon)\},$$
$$j = \left\lceil \frac{\log(2M\sqrt{B})}{\log(1/\zeta)} \right\rceil + 1, \qquad p = \exp\left\{ -\frac{2MB\gamma C_\varepsilon}{1-\zeta} \right\} > 0.$$

For each block $\mathcal{I}$, the SR-GD coordinates evolve according to

$$\mathbf{w}_{\mathcal{I}}^{(t+1)} = \mathrm{SR}\big(\mathbf{w}_{\mathcal{I}}^\star + \mathbf{J}_{\mathcal{I}\mathcal{I}}(\mathbf{w}_{\mathcal{I}}^{(t)} - \mathbf{w}_{\mathcal{I}}^\star)\big),$$
$$\mathbf{J}_{\mathcal{I}\mathcal{I}} = \mathbf{I}_{|\mathcal{I}|} - \eta \mathbf{A}_{\mathcal{I}\mathcal{I}}.$$

This is the same update with the original step size η restricted to the block, and $\|\mathbf{J}_{\mathcal{I}\mathcal{I}}\|_2 \le \|\mathbf{J}\|_2 \le \gamma$. Apply Proposition 31 in dimension $|\mathcal{I}| \le B$. Its step bound is at most j and its probability bound is at least p, so the block reaches $\mathbf{w}_{\mathcal{I}}^\star$ within j steps with probability at least p. If $\mathbf{w}_{\mathcal{I}} = \mathbf{w}_{\mathcal{I}}^\star$ for a block $\mathcal{I}$, subsequent updates leave that block unchanged. For the block partition $\mathcal{B}$, put $V_B(\mathbf{w}) \triangleq \sum_{\mathcal{I}\in\mathcal{B}} \mathbb{1}\{\mathbf{w}_{\mathcal{I}} \neq \mathbf{w}_{\mathcal{I}}^\star\}$ and $P\varphi(\mathbf{w}) \triangleq \mathbb{E}[\varphi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}]$. Then $P^j V_B \le (1-p)V_B$ and $P^s V_B \le V_B \le d$ for $s \ge 0$. To obtain contraction after one step, combine the expectations over the first j steps with weights that shift by the same factor q:

$$q \triangleq (1-p)^{1/j}, \qquad \psi \triangleq \sum_{s=0}^{j-1} q^{-s} P^s V_B, \qquad \xi \triangleq j q^{-(j-1)}.$$

Then $\psi(\mathbf{w}^\star) = 0$, $1 \le \psi(\mathbf{w}) \le \xi d$ for $\mathbf{w} \neq \mathbf{w}^\star$, and

$$P\psi = \sum_{s=0}^{j-1} q^{-s} P^{s+1} V_B = q \sum_{s=1}^{j} q^{-s} P^s V_B$$
$$= q\big(\psi - V_B + q^{-j} P^j V_B\big) \le q\psi,$$

where $q^{-j} P^j V_B \le V_B$ follows from $q^j = 1 - p$. □

C.4 Proof of Theorem 6

The constants in Theorem 6 may be chosen as

$$\beta \triangleq \frac{1-S^2}{8M^2}, \qquad \lambda \triangleq \log(1/\beta), \qquad \alpha \triangleq \frac{1}{4\lambda}, \qquad c \triangleq \frac{\beta\sqrt{\omega}}{8},$$
$$d_0 \triangleq \left\lceil \max\{2, \omega^{-2}, c^{-4}\} \right\rceil. \tag{C.17}$$

Proof. Condition on the Hessian and initializer. Since $\mathcal{M} = \{\mathbf{w}^\star\}$, $h^{(t+1)}$ is a sum of independent Bernoulli variables conditional on $\mathcal{F}_t$, by independent coordinatewise rounding. By Lemma 14,

$$\mathbb{E}[h^{(t+1)} \mid \mathcal{F}_t] \ge \frac{1-S^2}{4M^2} h^{(t)} = 2\beta h^{(t)}.$$

The multiplicative Chernoff bound with relative deviation $1/2$ therefore gives, on $\{h^{(t)} > 0\}$,

$$\Pr(h^{(t+1)} < \beta h^{(t)} \mid \mathcal{F}_t) \le \exp\left\{ -\frac{\mathbb{E}[h^{(t+1)} \mid \mathcal{F}_t]}{8} \right\} \le e^{-\beta h^{(t)}/4}.$$

Set $T_* \triangleq \lfloor \log h^{(0)}/(2\lambda) \rfloor$. If $h^{(s+1)} \ge \beta h^{(s)}$ for $0 \le s < t$, then $h^{(t)} \ge \beta^t h^{(0)} \ge \sqrt{h^{(0)}}$ for $t \le T_*$. Before the first violation, the current distance is at least $\sqrt{h^{(0)}}$. Thus each conditional probability of a first violation is at most $e^{-\beta\sqrt{h^{(0)}}/4}$. There are T_* transitions to consider, so summing these probabilities gives

$$\Pr\big(h^{(t)} \ge \sqrt{h^{(0)}} \text{ for every } 0 \le t \le T_*\big) \ge 1 - T_* e^{-\beta\sqrt{h^{(0)}}/4}. \tag{C.18}$$

Since $h^{(0)} \geq \omega d$ and $d \geq \omega^{-2}$, $\log h^{(0)} \geq \log(\omega d) \geq \frac{1}{2}\log d$. On the event in Eq. (C.18),

$$\tau_{\mathcal{M}} \geq T_* + 1 > \frac{\log h^{(0)}}{2\lambda} \geq \alpha \log d.$$

Since $\beta < 1/8$ and $\log d \leq \sqrt{d}$, $T_* \leq \log d/(2\lambda) \leq \sqrt{d}$. For $T_* > 0$, the inequality $\sup_{x\geq 1} \log x/x^{1/4} = 4/e$ gives $\log T_* \leq \frac{1}{2}\log d \leq d^{1/4}$. Using $d \geq c^{-4}$ and $c = \beta\sqrt{\omega}/8$, the failure probability in Eq. (C.18) is at most

$$T_* e^{-\beta\sqrt{h^{(0)}}/4} \leq \exp\{c\sqrt{d} - 2c\sqrt{d}\} = e^{-c\sqrt{d}}.$$

For $T_* = 0$, the failure probability in Eq. (C.18) is zero. □

C.5 Gaussian upper bounds

Throughout this subsection, consider Gaussian least squares with an on-grid target $\mathbf{w}^\star \in \mathcal{W}$ fixed independently of $\mathbf{X}$.

C.5.1 Gaussian matrix estimates and the near-diagonal condition

We first verify the near-diagonal condition from the main text and record the matrix estimates used in the quantitative hitting-time bound. Suppose $n \geq 16d$ and set

$$u \triangleq \sqrt{d/n}, \qquad a \triangleq (1-2u)^2, \qquad B \triangleq (1+2u)^2, \qquad b \triangleq 8d/\sqrt{n}.$$

By Lemma 22 with deviation $\sqrt{d}$, the singular values of $\mathbf{X}/\sqrt{n}$ lie in $[1-2u, 1+2u]$ with probability at least $1 - 2e^{-d/2}$. Thus the event $\mathcal{E}_X \triangleq \{a\mathbf{I}_d \preceq \mathbf{A} \preceq B\mathbf{I}_d\}$ has at least this probability. Since $u \leq 1/4$, we have $a \geq 1/4$ and $B \leq 9/4$. Work on $\mathcal{E}_X$. The target is unique and absorbing, and $\|\mathbf{A} - \mathbf{I}_d\|_2 \leq 4u + 4u^2 \leq 5u$. Moreover, $A_{ii} \in [a, B]$ gives

$$-(B-a)\mathbf{I}_d \preceq \mathbf{A} - \mathrm{diag}(\mathbf{A}) \preceq (B-a)\mathbf{I}_d.$$

Since $B - a = 8u$, we obtain

$$a\mathbf{I}_d \preceq \mathbf{A} \preceq B\mathbf{I}_d, \qquad \sqrt{d}\|\mathbf{A} - \mathrm{diag}(\mathbf{A})\|_2 \leq b. \tag{C.19}$$

The step size satisfies $S/B \leq \eta \leq S/a \leq 4S$, so

$$\|\mathbf{J}\|_2 = 1 - S\lambda_{\min}(\mathbf{A})/\lambda_{\max}(\mathbf{A}) \leq 1 - Sa/B \leq 1 - S/9,$$
$$\sqrt{d}\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 \leq \eta b \leq 32Sd/\sqrt{n}.$$

For fixed $R > 0$, if $n \geq Rd^2$, the last bound is at most $32S/\sqrt{R}$. Therefore, for sufficiently large d, Proposition 36 applies with $\gamma = 1 - S/9$ and $\delta = 32S/\sqrt{R}$. It gives a potential bounded by a constant times d and a contraction factor strictly below one, with both constants independent of d. Together with Proposition 5, this gives logarithmic hitting times. For smaller n, the same matrix estimates depend on d^2/n; the next lemma tracks this dependence through the local drift and hitting-probability estimates.

C.5.2 A finite Gaussian hitting-time bound

For the Gaussian upper bounds, C_1, C_2, C_3 depend only on M, S and are defined by

$$C_1 \triangleq 2 + 8192M, \qquad C_2 \triangleq C_1 \log(18/S) + 128\sqrt{2MC_1},$$
$$C_3 \triangleq 12\left(\left\lceil \frac{18}{S}\log(2M+1)\right\rceil + \frac{18}{S}\right). \tag{C.20}$$

Lemma 38. *Consider Gaussian least squares with an on-grid target fixed independently of* $\mathbf{X}$. *Put* $x = d^2/n$ *and suppose* $n \geq 16d$ *and* $d \geq C_1^2(1+x)^2$. *Define*

$$H_x(\varepsilon) \triangleq 2 + \frac{18}{S}\log(2Md) + C_3 e^{C_2(1+x)}[\log 3 + C_2(1+x) + \log(1/\varepsilon)]. \tag{C.21}$$

With probability at least $1 - 2e^{-d/2}$ *over* $\mathbf{X}$, *for every initializer* $\mathbf{w}^{(0)} \in \mathcal{W}$ *and every* $\varepsilon \in (0,1)$,

$$\Pr(\tau_{\mathcal{M}} > H_x(\varepsilon) \mid \mathbf{X}, \mathbf{w}^{(0)}) \leq \varepsilon.$$

Proof. We construct a finite contracting potential and track its constants in terms of $x = d^2/n$. The matrix estimates from the preceding subsection will give the three hypotheses of Lemma 34 through Lemma 33. The resulting exponential moment gives the potential, and Proposition 5 gives the tail bound. Use u, a, B, b and the event $\mathcal{E}_X$ from the preceding matrix estimates. This event has probability at least $1 - 2e^{-d/2}$; on it, the target is unique and absorbing. Work on $\mathcal{E}_X$.

The following parameters give bounds that are uniform over every fixed realization of $\mathbf{X}$ for which $\mathcal{E}_X$ occurs:

$$\begin{aligned} r &\triangleq \lceil \max\{1, 8Mb^2/a^2\} \rceil, \qquad q \triangleq 1 - \frac{Sa}{2B}, \\ j &\triangleq \left\lceil \frac{\log(2M+1)}{-\log q} \right\rceil, \qquad p \triangleq (1-q)^r \exp\left\{ -\frac{4b\sqrt{2Mr}}{a} \right\}, \\ \theta &\triangleq \frac{p}{12(j + 1/(1-q))}, \qquad t_0 \triangleq \left\lceil \frac{\log \max\{1, 2Md/(r+1)\}}{-\log q} \right\rceil. \end{aligned} \tag{C.22}$$

They depend only on M, S, n, d. In particular, the exponential parameter θ is chosen independently of the initializer and of ε. Since $u \le 1/4$, we have $a \ge 1/4$, $B \le 9/4$, and $b = 8\sqrt{x}$. Consequently,

$$r \le 2 + 8192Mx \le C_1(1+x), \qquad 1 - q = Sa/(2B) \ge S/18, \qquad j \le \left\lceil \frac{18}{S} \log(2M+1) \right\rceil. \tag{C.23}$$

Here $\lceil \max\{1, y\} \rceil \le 2 + y$. The dimension assumption therefore gives $d \ge r^2$.

The update matrix satisfies $J_{ii} \in [1 - S, 1 - \eta a]$ and $\sqrt{d}\|\mathbf{J} - \mathrm{diag}(\mathbf{J})\|_2 \le \eta b$. Apply Lemma 33 with $g' = \eta a$, $b' = \eta b$. Since $b'/g' = b/a$, the threshold in that lemma equals r, and $1 - \eta a/2 \le 1 - Sa/(2B) = q$. For $V(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|_1/\Delta \le r$, its lower bound on $\Pr(\tau_{\mathcal{M}} \le j \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w})$ is

$$(\eta a/2)^r e^{-4b\sqrt{2Mr}/a} \ge (Sa/(2B))^r \exp\{-4b\sqrt{2Mr}/a\} = p,$$

because $\eta \ge S/B$. The number of updates in the lemma is at most $\lceil \log(2M+1)/[-\log(1 - \eta a/2)] \rceil \le j$. The larger time j preserves the hitting-probability bound because $\mathbf{w}^\star$ is absorbing. The expectation bound in Lemma 33 holds at every deterministic time. We therefore obtain all three hypotheses required by Lemma 34:

$$\begin{aligned} \mathbb{E}[V(\mathbf{w}^{(1)}) \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w}] &\le qV(\mathbf{w}) \quad (V(\mathbf{w}) > r), \\ \Pr(\tau_{\mathcal{M}} \le j \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w}) &\ge p \qquad (V(\mathbf{w}) \le r), \\ \mathbb{E}[V(\mathbf{w}^{(j)}) \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w}] &\le r \qquad (V(\mathbf{w}) \le r). \end{aligned}$$

Condition on a realization of $\mathbf{X}$ in $\mathcal{E}_X$. Lemma 34 first establishes finiteness of the exponential moment and then proves that

$$\psi(\mathbf{w}^\star) = 0, \qquad \psi(\mathbf{w}) = \mathbb{E}[e^{\theta \tau_{\mathcal{M}}} \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w}] \quad (\mathbf{w} \ne \mathbf{w}^\star)$$

defines a finite potential satisfying $\mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{X}, \mathbf{w}^{(0)} = \mathbf{w}] \le e^{-\theta}\psi(\mathbf{w})$. The potential depends on $\mathbf{X}$ through the hitting-time distribution. Since $V(\mathbf{w}) \le 2Md$, replacing $V(\mathbf{w})$ by $2Md$ in the entry-time bound of Lemma 34 gives $t_0(\mathbf{w}) \le t_0$ for every initial state, with t_0 as displayed above. Its refined moment estimate and Proposition 5 therefore give

$$\psi(\mathbf{w}) \le \frac{3}{p} e^{\theta t_0}, \qquad \Pr(\tau_{\mathcal{M}} > t \mid \mathbf{X}, \mathbf{w}^{(0)}) \le \frac{3}{p} e^{-\theta(t - t_0)} \quad (t \in \mathbb{Z}_{\ge 0}). \tag{C.24}$$

Thus the integer $T(\varepsilon) \triangleq \lceil t_0 + \theta^{-1} \log(3/(p\varepsilon)) \rceil$ satisfies $\Pr(\tau_{\mathcal{M}} > T(\varepsilon) \mid \mathbf{X}, \mathbf{w}^{(0)}) \le \varepsilon$.

To bound $T(\varepsilon)$, use Eq. (C.23) to obtain

$$\begin{aligned} \log(1/p) &= r \log(1/(1-q)) + \frac{4b\sqrt{2Mr}}{a} \\ &\le C_1(1+x) \log(18/S) + 128\sqrt{2MC_1}\sqrt{x(1+x)} \\ &\le C_2(1+x). \end{aligned}$$

Hence

$$p \geq e^{-C_2(1+x)}, \qquad \theta^{-1} \leq C_3 e^{C_2(1+x)}, \qquad t_0 \leq 1 + (18/S)\log(2Md).$$

Here $-\log q \geq 1 - q$ and $\max\{1, 2Md/(r+1)\} \leq 2Md$. Substituting these bounds and $\log(3/(p\varepsilon)) \leq \log 3 + C_2(1+x) + \log(1/\varepsilon)$ into $T(\varepsilon)$, with at most two from the ceilings, proves $T(\varepsilon) \leq H_x(\varepsilon)$, with H_x defined in Eq. (C.21). This proves the asserted tail bound. All estimates hold on the same event $\mathcal{E}_X$, uniformly over the initializer and $\varepsilon \in (0,1)$. □

C.5.3 Proof of Theorem 7

For fixed M, S, R, use C_1, C_2, C_3 from Eq. (C.20). The theorem holds with

$$\begin{aligned} R_0 &\triangleq 2C_2, \qquad p \triangleq C_2/R, \\ \alpha &\triangleq 2 + \frac{18}{S}(1 + \log(2M)) + C_3 e^{C_2(1+1/R)}[1 + \log 3 + C_2(1 + 1/R)], \end{aligned} \tag{C.25}$$

and the dimension threshold

$$Q \triangleq \max\{1/R_0, 1/R\}, \qquad C \triangleq C_1(1+Q), \qquad d_0 \triangleq \left\lceil 16C^2 \log^2(4C) \right\rceil. \tag{C.26}$$

Here p is the exponent in Case (iii) of Theorem 7; R_0 depends only on M, S, and α, p, d_0 only on M, S, R. For fixed S, R, $R_0, p = O(M)$, $\log \alpha = O(M)$, and $d_0 = O(M^2 \log^2(2M))$.

Proof. We prove the three conclusions by evaluating the horizon from Lemma 38 under the three assumptions on n. With $x = d^2/n$, recall that this horizon is

$$\begin{aligned} H_x(\varepsilon) = 2 &+ \frac{18}{S}\log(2Md) \\ &+ C_3 e^{C_2(1+x)}[\log 3 + C_2(1+x) + \log(1/\varepsilon)]. \end{aligned}$$

The dependence on n enters through x and the factor $e^{C_2 x}$. We first verify that the stated d_0 ensures $n \geq 16d$ and $d \geq C_1^2(1+x)^2$, as required by that lemma. Set $L_0 = \log(4C) \geq 1$ and $y_0 = 4CL_0$. Then

$$y_0 - C(1 + 2\log y_0) = C(2L_0 - 1 - 2\log L_0) \geq 0.$$

Since $d \geq d_0$, we have $\sqrt{d} \geq y_0$. Apply monotonicity of $y - C(1 + 2\log y)$ for $y \geq y_0$ at $y = \sqrt{d}$ to obtain

$$\sqrt{d} \geq C(1 + \log d), \qquad \log d \geq 1. \tag{C.27}$$

In each applicable case, $x \leq Q \log d$ and $C \geq 2\sqrt{Q}$. Using $(1+z)^2 \geq 4z$ with $z = \log d$ gives

$$d \geq C^2(1 + \log d)^2 \geq 16Q \log d \geq 16x,$$

Thus $n \geq 16d$, and

$$C_1(1+x) \leq C_1(1 + Q\log d) \leq C(1 + \log d) \leq \sqrt{d}.$$

This verifies $d \geq C_1^2(1+x)^2$ in Lemma 38. Its hitting-time bound therefore holds with probability at least $1 - 2e^{-d/2}$ over $\mathbf{X}$, with the horizon recalled above, for every initializer and every $\varepsilon \in (0,1)$.

In Case (i), $x \leq 1/R$. Using $\log(d/\varepsilon) \geq \max\{1, \log d, \log(1/\varepsilon)\}$ gives

$$\begin{aligned} 2 + \frac{18}{S}\log(2Md) &\leq \left[2 + \frac{18}{S}(1 + \log(2M))\right]\log(d/\varepsilon), \\ \log 3 + C_2(1+x) + \log(1/\varepsilon) &\leq [1 + \log 3 + C_2(1 + 1/R)]\log(d/\varepsilon). \end{aligned}$$

Together with $e^{C_2(1+x)} \leq e^{C_2(1+1/R)}$ and the stated definition of α, these inequalities yield

$$H_x(\varepsilon) \leq \alpha \log(d/\varepsilon).$$

In Case (ii), $x \leq \log\log d/R_0$. Take $\varepsilon = 1/\log d$ and $y = \log d \geq 1$. Since $C_2 x \leq \frac{1}{2}\log y$,

$$H_x(1/\log d) \leq 2 + \frac{18}{S}(y + \log(2M)) + C_3 e^{C_2}\sqrt{y}\,[\log 3 + C_2 + \tfrac{3}{2}\log y].$$

For $A = \log 3 + C_2 \geq 3$, $\log y \leq 2(\sqrt{y} - 1)$ gives

$$\sqrt{y}\,[A + \tfrac{3}{2}\log y] \leq \sqrt{y}\,[A + 3(\sqrt{y} - 1)] \leq Ay.$$

Therefore

$$H_x(1/\log d) \leq \left[2 + \frac{18}{S}(1 + \log(2M)) + C_3 e^{C_2}(\log 3 + C_2)\right] \log d \leq \alpha \log d.$$

In Case (iii), $x \leq \log d/R$ gives $e^{C_2 x} \leq d^p$ and $C_2 x \leq p \log d$. Thus

$$\begin{aligned} H_x(\varepsilon) &\leq 2 + \frac{18}{S}\log(2Md) + C_3 e^{C_2} d^p [\log 3 + C_2 + p\log d + \log(1/\varepsilon)] \\ &\leq \left[2 + \frac{18}{S}(1 + \log(2M)) + C_3 e^{C_2}(1 + \log 3 + C_2 + p)\right] d^p \log(d/\varepsilon) \\ &\leq \alpha d^p \log(d/\varepsilon). \end{aligned}$$

Here $p = C_2/R$. □

C.5.4 Comparison with the Gaussian lower bounds

For fixed $\omega > 0$ and $h^{(0)} \geq \omega d$, averaging Theorem 6 over $\mathbf{X}$ gives failure at most $\exp\{-(1 - S^2)\sqrt{\omega d}/(64M^2)\}$. Together with Case (ii), this yields $\tau_{\mathcal{M}} = \Theta(\log d)$ with joint failure at most

$$2e^{-d/2} + 1/\log d + e^{-(1-S^2)\sqrt{\omega d}/(64M^2)} = O(1/\log d).$$

For fixed $C > 0$, Case (i) with $\varepsilon = d^{-C}$ gives a horizon $O(\log d)$ and joint failure

$$2e^{-d/2} + d^{-C} + e^{-(1-S^2)\sqrt{\omega d}/(64M^2)} = O(d^{-C}),$$

where the constant in the horizon may depend on C.

For the polynomial comparison, fix an interior on-grid target $\mathbf{w}^\star \in \mathcal{W}^\circ$, as required by Theorem 4. For fixed $0 < R_- \leq R_+ < \infty$, suppose

$$R_- \frac{d^2}{\log d} \leq n \leq R_+ \frac{d^2}{\log d}.$$

For the Gaussian lower bound in Theorem 4 with $R = 1$, recall

$$\rho = \min\left\{\frac{1}{256}, \frac{1 - S^2}{144M^2} \min\left\{\frac{S^2}{512(2 + \sqrt{2})^4}, \frac{1}{32768(2 + \log(2M))^2}\right\}\right\},$$
$$c = \frac{3}{256}.$$

For the upper bound in Case (iii) of Theorem 7 with $R = R_-$, the exponent is $p = C_2/R_-$, with C_2 from Eq. (C.20). For sufficiently large d, $\log d \geq R_+/\rho$ and $d \leq n \leq \rho d^2$. Thus Theorem 4 applies with $R = 1$ and

$$\frac{\rho}{R_+}\log d \leq k = \frac{\rho d^2}{n} \leq \frac{\rho}{R_-}\log d.$$

For every fixed nonoptimal initializer, it gives $\tau_{\mathcal{M}} > d^{\rho/R_+}$ with failure at most $e^{-cd} + e^{-c\sqrt{k}}$. Case (iii) with $R = R_-$ and $\varepsilon = d^{-1}$ gives a horizon $2\alpha d^p \log d \leq d^{p+1}$ for sufficiently large d, with failure at most $2e^{-d/2} + d^{-1}$. A union bound gives

$$\Pr\left(d^{\rho/R_+} < \tau_{\mathcal{M}} \leq d^{p+1}\right) \geq 1 - e^{-cd} - e^{-c\sqrt{k}} - 2e^{-d/2} - d^{-1} \geq 1 - e^{-c'\sqrt{\log d}}$$

with $c' = \frac{c}{2}\sqrt{\rho/R_+} > 0$ for sufficiently large d, over $\mathbf{X}$ and rounding.

D RANK-DEFICIENT HESSIANS

D.1 MATRIX ESTIMATES FOR A HAAR RANDOM SUBSPACE

The following two estimates are used in Lemma 40. The bound on $\sum_i A_{ii}^2$ controls the expected increase in the objective caused by rounding. The lower bound on the sum of absolute coordinates over any set of at least $d/4$ indices provides a lower bound on the decrease in the objective after clipping.

Lemma 39. *Fix $0 < a \le b$ and integers $1 \le r \le 2^{-13}d$. Let $\mathbf{V}$ consist of the first r columns of $\mathbf{U} \sim \mathrm{Unif}(\mathcal{O}(d))$. With probability at least $1 - 52\sqrt{r/d}$ over $\mathbf{V}$, every matrix*

$$\mathbf{A} = \frac{d}{r}\mathbf{V}\mathbf{D}\mathbf{V}^\top, \qquad a\mathbf{I}_r \preceq \mathbf{D} \preceq b\mathbf{I}_r, \tag{D.1}$$

satisfies

$$\sum_{i=1}^{d} A_{ii}^2 \le 24b^2 d, \qquad \sum_{i\in\mathcal{I}} |v_i| \ge 2^{-8}\sqrt{d}\,\|\mathbf{v}\|_2$$

for every $\mathbf{v} \in \mathrm{range}(\mathbf{A})$ and $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| \ge d/4$.

Proof. We construct one event determined by the random subspace on which both asserted inequalities hold for every allowed $\mathbf{D}$. For the coordinate-sum inequality, we first obtain a bound on a finite net of the unit sphere and then extend it to every unit vector. The diagonal-square bound follows from a fourth-moment bound for a Gaussian representation of the same subspace.

Let $\mathbf{G} \in \mathbb{R}^{r\times d}$ have independent standard Gaussian entries and set $\mathbf{H} = \mathbf{G}\mathbf{G}^\top$. Almost surely, $\mathbf{H}$ is positive definite and $(\mathbf{G}^\top\mathbf{H}^{-1/2})^\top(\mathbf{G}^\top\mathbf{H}^{-1/2}) = \mathbf{I}_r$. For every $\mathbf{R} \in \mathcal{O}(d)$, $\mathbf{G}\mathbf{R}$ has the same distribution as $\mathbf{G}$, and its normalized transpose is $\mathbf{R}^\top\mathbf{G}^\top\mathbf{H}^{-1/2}$. Orthogonal invariance therefore implies that $\mathbf{G}^\top\mathbf{H}^{-1/2}$ has the distribution of the first r columns of a Haar orthogonal matrix (Meckes, 2019). We may consequently take $\mathbf{V} = \mathbf{G}^\top\mathbf{H}^{-1/2}$. Set $\beta \triangleq 2^{-8}$ and $\rho \triangleq 2^{-13}$, suppose $r \le \rho d$, and define

$$a_0 \triangleq \frac{\sqrt{\pi/2}}{32}, \qquad \varepsilon_0 \triangleq \frac{\sqrt{\pi}}{128\sqrt{6}}, \qquad u_+ \triangleq \sqrt{3/2} - 1, \qquad u_- \triangleq 1 - 1/\sqrt{2}.$$

The choice $\rho = 2^{-13}$ satisfies

$$\rho \le \min\{u_+^2/4, u_-^2/4\}, \qquad \rho\log(1 + 2/\varepsilon_0) \le 1/1024.$$

Apply Lemma 22 to $\mathbf{G}^\top \in \mathbb{R}^{d\times r}$: the row and column dimensions denoted by n, d in that lemma are d, r here. Since $d \ge r$, both singular-value bounds apply, and

$$\|\mathbf{G}\|_2 = s_{\max}(\mathbf{G}^\top), \qquad \lambda_{\min}(\mathbf{H}) = s_{\min}(\mathbf{G}^\top)^2.$$

The bound $r/d \le \rho \le \min\{u_+^2/4, u_-^2/4\}$ gives $u_+\sqrt{d} - \sqrt{r} \ge u_+\sqrt{d}/2 > 0$ and $u_-\sqrt{d} - \sqrt{r} \ge u_-\sqrt{d}/2 > 0$. The corresponding thresholds are

$$\sqrt{d} + \sqrt{r} + (u_+\sqrt{d} - \sqrt{r}) = \sqrt{3d/2},$$
$$\sqrt{d} - \sqrt{r} - (u_-\sqrt{d} - \sqrt{r}) = \sqrt{d/2}.$$

Thus the upper and lower singular-value bounds give, respectively,

$$\Pr\big(\|\mathbf{G}\|_2^2 > 3d/2\big) \le \exp\left(-\frac{(u_+\sqrt{d} - \sqrt{r})^2}{2}\right) \le e^{-du_+^2/8},$$
$$\Pr(\lambda_{\min}(\mathbf{H}) < d/2) \le \exp\left(-\frac{(u_-\sqrt{d} - \sqrt{r})^2}{2}\right) \le e^{-du_-^2/8}.$$

A union bound therefore gives

$$\|\mathbf{G}\|_2^2 \le 3d/2, \qquad \lambda_{\min}(\mathbf{H}) \ge d/2 \tag{D.2}$$

with failure at most $e^{-du_+^2/8}+e^{-du_-^2/8}$. Let $\mathbf{g}_i \in \mathbb{R}^r$ denote the i-th column of $\mathbf{G}$. The independent $Y_i \triangleq \|\mathbf{g}_i\|_2^2 \sim \chi_r^2$ satisfy

$$\mathbb{E}[Y_i^2] = r(r+2) \le 3r^2, \qquad \mathrm{Var}(Y_i^2) = 8r(r+2)(r+3) \le 96r^3.$$

Chebyshev's inequality therefore gives

$$\Pr\left(\sum_i \|\mathbf{g}_i\|_2^4 > 6dr^2\right) \le \frac{32}{3dr}. \tag{D.3}$$

Take an ε_0-net $\mathcal{N}$ of the unit sphere in $\mathbb{R}^r$, with $|\mathcal{N}| \le (1+2/\varepsilon_0)^r$. For each $\mathbf{u}_0 \in \mathcal{N}$, the $\mathbf{g}_i^\top \mathbf{u}_0$ are independent standard normals, so

$$\Pr(|\mathbf{g}_i^\top \mathbf{u}_0| < 2a_0) \le \frac{4a_0}{\sqrt{2\pi}} = \frac{1}{16}.$$

Hoeffding's inequality and a union bound give

$$\Pr\big(\exists \mathbf{u}_0 \in \mathcal{N} : \big|\{i : |\mathbf{g}_i^\top \mathbf{u}_0| < 2a_0\}\big| > 3d/32\big) \le |\mathcal{N}| e^{-d/512} \le e^{-d/1024}.$$

On the complementary event and $\|\mathbf{G}\|_2^2 \le 3d/2$, for each unit $\mathbf{u} \in \mathbb{R}^r$ choose $\mathbf{u}_0 \in \mathcal{N}$ with $\|\mathbf{u}-\mathbf{u}_0\|_2 \le \varepsilon_0$. Then

$$\big|\{i : |\mathbf{g}_i^\top (\mathbf{u}-\mathbf{u}_0)| \ge a_0\}\big| \le \frac{\|\mathbf{G}^\top(\mathbf{u}-\mathbf{u}_0)\|_2^2}{a_0^2} \le \frac{3d\varepsilon_0^2}{2a_0^2} = \frac{d}{32}.$$

The triangle inequality consequently gives

$$\sup_{\|\mathbf{u}\|_2=1} \big|\{i : |\mathbf{g}_i^\top \mathbf{u}| < a_0\}\big| \le d/8. \tag{D.4}$$

Since $\mathbf{D}$ is positive definite and $\mathbf{H}^{-1/2}$ is invertible,

$$\mathrm{range}(\mathbf{A}) = \mathrm{range}(\mathbf{V}) = \mathrm{range}(\mathbf{G}^\top).$$

Every nonzero $\mathbf{v} \in \mathrm{range}(\mathbf{A})$ can therefore be written as $\mathbf{v} = \mathbf{G}^\top \mathbf{u}$ for some nonzero $\mathbf{u} \in \mathbb{R}^r$. Apply Eq. (D.4) to $\mathbf{u}/\|\mathbf{u}\|_2$. For every $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| \ge d/4$, this gives

$$\big|\{i \in \mathcal{I} : |\mathbf{g}_i^\top \mathbf{u}| \ge a_0 \|\mathbf{u}\|_2\}\big| \ge |\mathcal{I}| - d/8 \ge d/8.$$

Thus

$$\sum_{i\in\mathcal{I}} |v_i| \ge \frac{a_0 d}{8}\|\mathbf{u}\|_2 \ge \frac{a_0}{8\sqrt{3/2}}\sqrt{d}\,\|\mathbf{v}\|_2 = \frac{\sqrt{\pi/3}}{256}\sqrt{d}\,\|\mathbf{v}\|_2 \ge \beta\sqrt{d}\,\|\mathbf{v}\|_2.$$

For $\mathbf{v} = \mathbf{0}$ the inequality is immediate. On $\{\lambda_{\min}(\mathbf{H}) \ge d/2,\ \sum_i \|\mathbf{g}_i\|_2^4 \le 6dr^2\}$, the projector $\mathbf{P} \triangleq \mathbf{V}\mathbf{V}^\top$ satisfies

$$P_{ii} = \mathbf{g}_i^\top \mathbf{H}^{-1} \mathbf{g}_i \le \frac{2}{d}\|\mathbf{g}_i\|_2^2, \qquad \sum_i P_{ii}^2 \le \frac{24r^2}{d}.$$

For every $\mathbf{D}$ in Eq. (D.1), $A_{ii} \le (bd/r)P_{ii}$, proving $\sum_i A_{ii}^2 \le 24b^2 d$.

Set

$$\mathcal{E} \triangleq \big\{ \textstyle\sum_i P_{ii}^2 \le 24r^2/d, \quad \sum_{i\in\mathcal{I}} |v_i| \ge \beta\sqrt{d}\|\mathbf{v}\|_2 \\ \text{for all } \mathbf{v} \in \mathrm{range}(\mathbf{V}) \text{ and } \mathcal{I} \subseteq [d] \text{ with } |\mathcal{I}| \ge d/4\big\}.$$

The event $\mathcal{E}$ depends only on $\mathbf{V}$. By a union bound and $u_+^2/8, u_-^2/8 \ge 1/1024$,

$$\Pr(\mathcal{E}^c) \le 3e^{-d/1024} + \frac{32}{3dr} \le \left(3\sqrt{512/e} + \frac{32}{3}\right)\sqrt{\frac{r}{d}} \le 52\sqrt{\frac{r}{d}}. \tag{D.5}$$

Here $\sqrt{d}\,e^{-d/1024} \le \sqrt{512/e}$ and $1/(dr) \le \sqrt{r/d}$ for $d, r \ge 1$.

□

D.2 Objective drift and stopped trajectories

The next two lemmas control objective decrease and changes in Hamming distance for a fixed Hessian. They are used in Appendix D.3.

D.2.1 Objective drift on the matrix range

Lemma 40. *Set $\beta \triangleq 2^{-8}$ and $\chi \triangleq \beta^2/(1+\sqrt{2}\,S\beta)$. Suppose that*

$$\sum_{i\in\mathcal{I}} |v_i| \geq \beta\sqrt{d}\,\|\mathbf{v}\|_2 \quad \textit{for every } \mathbf{v} \in \mathrm{range}(\mathbf{A}) \textit{ and } \mathcal{I} \subseteq [d] \textit{ with } |\mathcal{I}| \geq d/4. \tag{D.6}$$

Assume that at least $d/2$ coordinates of $\mathbf{w}^{(0)}$ belong to $\mathcal{G}^\circ$. Define the cumulative number of coordinate changes by

$$C_t \triangleq \sum_{j=0}^{t-1} \|\mathbf{w}^{(j+1)} - \mathbf{w}^{(j)}\|_0 \tag{D.7}$$

and the stopping time by

$$\sigma \triangleq \inf\left\{t \geq 0 : C_t \geq d/4 \textit{ or } f(\mathbf{w}^{(t)}) > \Delta^2\lambda_{\max}(\mathbf{A})d\right\}. \tag{D.8}$$

On $\{t < \sigma\}$,

$$\mathbb{E}[f(\mathbf{w}^{(t+1)}) \mid \mathcal{F}_t] \leq q_1 f(\mathbf{w}^{(t)}) + A_1, \qquad \mathbb{E}[\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_0 \mid \mathcal{F}_t] \leq B_1\sqrt{f(\mathbf{w}^{(t)})},$$

where

$$q_1 = 1 - \frac{\chi\eta(2-S)\lambda_{\min}(\mathbf{A};\mathrm{range}(\mathbf{A}))}{2}, \qquad A_1 = \frac{\eta\Delta^2\sum_i A_{ii}^2}{4(2-S)}, \qquad B_1 = \frac{S}{\Delta}\sqrt{\frac{2d}{\lambda_{\max}(\mathbf{A})}}. \tag{D.9}$$

These coefficients satisfy $q_1 \in (0,1)$ and $A_1, B_1 \geq 0$.

Proof. We first show that the clipped gradient step decreases the objective by at least a fixed multiple of $\eta\|\nabla f(\mathbf{w}^{(t)})\|_2^2$. We then bound the increase caused by rounding and the expected number of coordinates changed by this step. The condition $t < \sigma$ ensures that enough initially interior coordinates remain unchanged and that the objective is below the threshold used to control clipping.

Condition on a history with $t < \sigma$ and write $\mathbf{w} \triangleq \mathbf{w}^{(t)}$, $\mathbf{g} \triangleq \mathbf{A}(\mathbf{w} - \mathbf{w}^\star)$, and $F \triangleq f(\mathbf{w})$. If $\mathbf{g} = \mathbf{0}$, then $F = 0$ and the claim is immediate. Assume $\mathbf{g} \neq \mathbf{0}$. At most $C_t < d/4$ coordinates have changed by time t. At least $d/2$ coordinates were initially interior, so more than $d/4$ of them are still unchanged and interior. Let $\mathcal{I}$ be this set and put $s \triangleq |\mathcal{I}|$; then $d/4 < s \leq d$. Since $\mathbf{g} \in \mathrm{range}(\mathbf{A})$, Eq. (D.6) gives $\sum_{i\in\mathcal{I}} |g_i| \geq \beta\sqrt{d}\,\|\mathbf{g}\|_2$.

Write the clipped point as $\bar{\mathbf{w}} \triangleq \Pi(\mathbf{w} - \eta\mathbf{g}) = \mathbf{w} - \eta\mathbf{D}_t\mathbf{g}$, where $\mathbf{D}_t \triangleq \mathrm{diag}(d_1,\dots,d_d)$, $0 \leq d_i \leq 1$, and $d_i = 1$ when $g_i = 0$. Set $R_t \triangleq \mathbf{g}^\top\mathbf{D}_t\mathbf{g}$. The quadratic expansion, $\mathbf{A} \preceq \lambda_{\max}(\mathbf{A})\mathbf{I}_d$, and $\mathbf{D}_t^2 \preceq \mathbf{D}_t$ give

$$\begin{aligned} f(\bar{\mathbf{w}}) &= F - \eta R_t + \frac{\eta^2}{2}(\mathbf{D}_t\mathbf{g})^\top\mathbf{A}(\mathbf{D}_t\mathbf{g}) \\ &\leq F - \eta R_t + \frac{\eta^2\lambda_{\max}(\mathbf{A})}{2}\|\mathbf{D}_t\mathbf{g}\|_2^2 \\ &\leq F - \frac{\eta(2-S)}{2}R_t. \end{aligned}$$

Thus it suffices to bound R_t below in terms of $\|\mathbf{g}\|_2^2$. Each coordinate in $\mathcal{I}$ is at least Δ from either endpoint of the box. Hence, for $i \in \mathcal{I}$ and $g_i \neq 0$,

$$d_i \geq \min\left\{1, \frac{\Delta}{\eta|g_i|}\right\}, \qquad d_i^{-1} \leq 1 + \frac{\eta|g_i|}{\Delta}.$$

With $u \triangleq \sum_{i\in\mathcal{I}} |g_i|$, Cauchy–Schwarz gives

$$u^2 \le \left(\sum_{i\in\mathcal{I}} d_i g_i^2\right)\left(\sum_{\substack{i\in\mathcal{I}\\ g_i\neq 0}} d_i^{-1}\right) \le R_t(s+\eta u/\Delta).$$

The function $u^2/(s+\eta u/\Delta)$ is increasing for $u \ge 0$. Also, $F \le \Delta^2\lambda_{\max}(\mathbf{A})d$ before σ, and thus $\|\mathbf{g}\|_2 \le \sqrt{2\lambda_{\max}(\mathbf{A})F} \le \sqrt{2}\,\Delta\lambda_{\max}(\mathbf{A})\sqrt{d}$. Using Eq. (D.6) yields

$$R_t \ge \frac{\beta^2 d\|\mathbf{g}\|_2^2}{s+(\eta/\Delta)\beta\sqrt{d}\|\mathbf{g}\|_2} \ge \frac{\beta^2}{1+\sqrt{2}\,S\beta}\|\mathbf{g}\|_2^2 = \chi\|\mathbf{g}\|_2^2. \tag{D.10}$$

Here the second inequality uses $s \le d$ and $\eta\lambda_{\max}(\mathbf{A}) = S$.

Let $\boldsymbol{\xi} \triangleq \mathbf{w}^{(t+1)} - \bar{\mathbf{w}}$. Conditional on the history, its coordinates are independent and centered. The scalar rounding bounds in Lemma 13 imply

$$\mathbb{E}[\xi_i^2 \mid \mathcal{F}_t] \le \Delta|\bar{w}_i - w_i| = \Delta\eta d_i|g_i|.$$

Cauchy–Schwarz gives

$$\sum_i A_{ii} d_i|g_i| \le \sqrt{\left(\sum_i A_{ii}^2 d_i\right) R_t} \le \sqrt{\left(\sum_i A_{ii}^2\right) R_t}.$$

The linear term in the quadratic expansion has conditional expectation zero, and the mixed second moments vanish by independence. Consequently,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{w}^{(t+1)}) \mid \mathcal{F}_t] &= f(\bar{\mathbf{w}}) + \frac{1}{2}\sum_i A_{ii}\mathbb{E}[\xi_i^2 \mid \mathcal{F}_t] \\
&\le F - \frac{\eta(2-S)}{2}R_t + \frac{\eta\Delta}{2}\sqrt{\left(\sum_i A_{ii}^2\right) R_t} \\
&\le F - \frac{\eta(2-S)}{4}R_t + \frac{\eta\Delta^2\sum_i A_{ii}^2}{4(2-S)} \\
&\le \left(1 - \frac{\chi\eta(2-S)\lambda_{\min}(\mathbf{A};\mathrm{range}(\mathbf{A}))}{2}\right)F + \frac{\eta\Delta^2\sum_i A_{ii}^2}{4(2-S)}.
\end{aligned}$$

The third line follows by completing the square in $\sqrt{R_t}$ and bounds the rounding term by $\eta(2-S)R_t/4 + \eta\Delta^2\sum_i A_{ii}^2/(4(2-S))$. For the last line, Eq. (D.10) and the spectral decomposition give

$$R_t \ge \chi\|\mathbf{g}\|_2^2 \ge 2\chi\lambda_{\min}(\mathbf{A};\mathrm{range}(\mathbf{A}))F.$$

The latter inequality remains valid for singular $\mathbf{A}$ because both F and $\|\mathbf{g}\|_2^2$ depend only on the component of $\mathbf{w}-\mathbf{w}^\star$ in $\mathrm{range}(\mathbf{A})$. Moreover, $0 < \lambda_{\min}(\mathbf{A};\mathrm{range}(\mathbf{A}))/\lambda_{\max}(\mathbf{A}) \le 1$, $0 < S(2-S) < 1$, and $0 < \chi < 1$ imply $q_1 \in (0,1)$. For the coordinate-change bound, the current coordinate w_i is a grid point. The scalar rounding formula gives $\Pr(w_i^{(t+1)} \neq w_i \mid \mathcal{F}_t) \le |\bar{w}_i - w_i|/\Delta$. Summing over the coordinates and using $\|\mathbf{g}\|_2^2 \le 2\lambda_{\max}(\mathbf{A})F$ yields

$$\mathbb{E}[\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_0 \mid \mathcal{F}_t] \le \frac{1}{\Delta}\sum_i |\bar{w}_i - w_i| \le \frac{\eta}{\Delta}\|\mathbf{g}\|_1 \le \frac{S}{\Delta}\sqrt{\frac{2d}{\lambda_{\max}(\mathbf{A})}}\sqrt{F}.$$

□

D.2.2 Objective decrease and Hamming-distance control

Lemma 41. *Fix $\mathbf{A}$ and $\mathbf{w}^{(0)}$, and let σ be the stopping time in Eq. (D.8). All probabilities and expectations in this lemma are over rounding. Suppose that $q_1 \in (0,1)$ and $A_1, B_1 \ge 0$ satisfy, on $\{t < \sigma\}$,*

$$\mathbb{E}[f(\mathbf{w}^{(t+1)}) \mid \mathcal{F}_t] \le q_1 f(\mathbf{w}^{(t)}) + A_1, \qquad \mathbb{E}[\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_0 \mid \mathcal{F}_t] \le B_1\sqrt{f(\mathbf{w}^{(t)})}. \tag{D.11}$$

For $0 < m < d/4$, $\varepsilon > 0$ with $\varepsilon \geq 2A_1/(1-q_1)$, and integers $0 \leq s \leq T$,

$$\begin{aligned}\Pr\left(\max_{0\leq t\leq T}|h^{(t)}-h^{(0)}|\leq m,\quad \min_{0\leq t\leq s} f(\mathbf{w}^{(t)})\leq\varepsilon\right)\\ \geq 1-\frac{B_1}{m}\left(\frac{\sqrt{f(\mathbf{w}^{(0)})}}{1-\sqrt{q_1}}+T\sqrt{\frac{A_1}{1-q_1}}\right)\\ -\frac{f(\mathbf{w}^{(0)})+A_1T}{\Delta^2\lambda_{\max}(\mathbf{A})d}-\frac{f(\mathbf{w}^{(0)})}{\varepsilon}\left(\frac{1+q_1}{2}\right)^s.\end{aligned} \tag{D.12}$$

Proof. The three subtracted terms bound, in order, the probabilities of more than m coordinate changes by time T before stopping, an objective value above the stopping threshold by time T, and remaining above objective value ε through time s without stopping. Excluding the first two events ensures that the drift assumptions apply throughout $0,\dots,T$; excluding the third then gives the required objective value.

Write $F_t \triangleq f(\mathbf{w}^{(t)})$ and $F_{\max} \triangleq \Delta^2\lambda_{\max}(\mathbf{A})d$, and define

$$N_t \triangleq \|\mathbf{w}^{(t+1)}-\mathbf{w}^{(t)}\|_0,\qquad \nu \triangleq \inf\{t\geq 0 : C_t \geq d/4\},\qquad \zeta \triangleq \inf\{t\geq 0 : F_t > F_{\max}\}.$$

Then $C_t = \sum_{j=0}^{t-1} N_j$ and $\sigma = \nu\wedge\zeta$. If $F_0 > F_{\max}$, the right-hand side of Eq. (D.12) is negative and the claim is immediate. We may therefore assume $F_0 \leq F_{\max}$. The triangle inequality for Hamming distance gives $|h(\mathbf{u})-h(\mathbf{w})| \leq \|\mathbf{u}-\mathbf{w}\|_0$, and hence

$$|h^{(t)}-h^{(0)}| \leq C_t. \tag{D.13}$$

Define

$$\tau_\varepsilon \triangleq \inf\{t\geq 0 : F_t \leq \varepsilon\},\qquad Q_1 \triangleq (1+q_1)/2,$$
$$\mathcal{E}_1 \triangleq \{C_{T\wedge\sigma} > m\},\qquad \mathcal{E}_2 \triangleq \{\zeta\leq T,\ \zeta\leq\nu\},\qquad \mathcal{E}_3 \triangleq \{\sigma > s,\ \tau_\varepsilon > s\}.$$

Since $m < d/4$, $\{\nu\leq T, \nu<\zeta\}\subseteq\mathcal{E}_1$, so $\{\sigma\leq T\}\subseteq\mathcal{E}_1\cup\mathcal{E}_2$. Outside $\mathcal{E}_1\cup\mathcal{E}_2$, we have $\sigma > T$ and $C_T = C_{T\wedge\sigma} \leq m$. Thus Eq. (D.13) bounds the Hamming-distance change at every time up to T. Since $s\leq T$, the additional exclusion of $\mathcal{E}_3$ implies $\tau_\varepsilon \leq s$. It remains to bound the probabilities of these three events.

First, nonnegativity gives

$$\begin{aligned}\mathbb{E}[F_{t+1}\mathbb{1}\{\sigma > t+1\}\mid\mathcal{F}_t] &\leq \mathbb{1}\{\sigma>t\}\mathbb{E}[F_{t+1}\mid\mathcal{F}_t]\\ &\leq q_1F_t\mathbb{1}\{\sigma>t\}+A_1\mathbb{1}\{\sigma>t\}.\end{aligned}$$

Taking expectations and iterating the resulting scalar recurrence yields

$$\mathbb{E}[F_t\mathbb{1}\{\sigma>t\}] \leq q_1^tF_0+\frac{A_1}{1-q_1}. \tag{D.14}$$

By the definition of C_t,

$$C_{T\wedge\sigma} = \sum_{j=0}^{T-1}\mathbb{1}\{j<\sigma\}N_j.$$

Thus Cauchy–Schwarz and Eq. (D.11) give

$$\begin{aligned}\mathbb{E}[C_{T\wedge\sigma}] &\leq B_1\sum_{j<T}\mathbb{E}[\mathbb{1}\{j<\sigma\}\sqrt{F_j}]\\ &\leq B_1\sum_{j<T}\sqrt{\mathbb{E}[F_j\mathbb{1}\{j<\sigma\}]}\\ &\leq B_1\left(\frac{\sqrt{F_0}}{1-\sqrt{q_1}}+T\sqrt{\frac{A_1}{1-q_1}}\right).\end{aligned}$$

Here we used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and summed $\sum_{j<T} q_1^{j/2} \le (1-\sqrt{q_1})^{-1}$. Markov's inequality gives $\Pr(\mathcal{E}_1) \le \mathbb{E}[C_{T\wedge\sigma}]/m$.

Second, the drift inequality and $q_1 < 1$ imply $\mathbb{E}[F_{t+1} - F_t \mid \mathcal{F}_t] \le A_1$ on $\{t < \sigma\}$. For the stopped process,

$$F_{(t+1)\wedge\sigma} - F_{t\wedge\sigma} = \mathbb{1}\{t < \sigma\}(F_{t+1} - F_t), \qquad \mathbb{E}[F_{T\wedge\sigma}] \le F_0 + A_1 T.$$

On $\mathcal{E}_2$, $F_{T\wedge\sigma} > F_{\max}$, also if $\zeta = 0$ or $\zeta = \nu$. Therefore

$$\Pr(\mathcal{E}_2) \le \frac{F_0 + A_1 T}{F_{\max}}.$$

Finally, before $\sigma \wedge \tau_\varepsilon$ we have $F_t > \varepsilon \ge 2A_1/(1-q_1)$, so

$$q_1 F_t + A_1 \le q_1 F_t + \frac{1-q_1}{2} F_t = Q_1 F_t.$$

Thus the additive drift bound becomes a multiplicative contraction until the objective reaches ε or the trajectory stops. Apply Lemma 17 with $Z_t = F_t$, $\tau = \sigma \wedge \tau_\varepsilon$, $a = \varepsilon$, and $q = Q_1$ to obtain

$$\mathbb{E}[F_s \mathbb{1}\{\sigma > s,\ \tau_\varepsilon > s\}] \le Q_1^s F_0, \qquad \Pr(\mathcal{E}_3) \le \frac{F_0}{\varepsilon} Q_1^s.$$

A union bound proves Eq. (D.12). □

D.3 Proof of Theorem 8

We will show that, with the probability stated in the theorem, the objective reaches $\ell\Delta^2 r$ by time $t_\star$ while the Hamming distance changes by at most $\min\{\omega/2, 1/8\}d$ through time T. Lemma 39 supplies the matrix estimates needed by Lemma 40 to control the objective and the number of coordinate changes in one step. We apply Lemma 41 conditionally on $\mathbf{U}$ and average over $\mathbf{U}$ to obtain the joint probability bound. The initial separation then gives the Hamming-distance conclusion of the theorem.

For $M \ge 2$, $0 < a \le b$, and $\omega \in (0, 1]$, define

$$\begin{aligned} &\rho \triangleq 2^{-13}, \qquad \chi \triangleq \frac{1}{2^{16} + 2^8\sqrt{2}\,S}, \qquad \alpha \triangleq \frac{8b}{\chi S(2-S)a}, \qquad \ell \triangleq \frac{24b^2}{\chi(2-S)^2 a}, \\ &c \triangleq 52 + \max\left\{\frac{2}{\omega}, 8\right\}\left(\frac{8M(b/a)}{\chi(2-S)} + \frac{2\sqrt{6}\,S(b/a)}{\sqrt{\chi}(2-S)}\right) + M\left(\frac{2b}{a} + \frac{b}{2\ell}\right) + \frac{6S(b/a)^2}{2-S}. \end{aligned} \tag{D.15}$$

Here α, ℓ depend only on S, a, b, and $c = O(M)$ for fixed S, a, b, ω.

Proof of Theorem 8. Fix the on-grid target and initializer as in the theorem before drawing $\mathbf{U}$. By Lemma 25, $\ker(\mathbf{A}) \cap (\mathcal{W} - \mathcal{W}) = \{\mathbf{0}\}$ almost surely. Hence $\mathcal{M} = \{\mathbf{w}^\star\}$ and $h^{(0)} = \|\mathbf{w}^{(0)} - \mathbf{w}^\star\|_0 \ge \omega d$ almost surely; in particular, the initial Hamming distance equals a fixed value almost surely. Put

$$\delta = \min\{\omega/2, 1/8\}, \qquad t_\star = \left\lceil \alpha \log \frac{2d}{r} \right\rceil.$$

Write $F_t = f(\mathbf{w}^{(t)})$. Let $\mathbf{V}$ be the first r columns of $\mathbf{U}$ and set $\mathbf{D} \triangleq \operatorname{diag}(r\lambda_1/d, \dots, r\lambda_r/d)$. Then $\mathbf{A} = (d/r)\mathbf{V}\mathbf{D}\mathbf{V}^\top$ and $a\mathbf{I}_r \preceq \mathbf{D} \preceq b\mathbf{I}_r$. Define

$$\begin{aligned} \mathcal{E} \triangleq \{ &\sum_i A_{ii}^2 \le 24b^2 d, \quad \textstyle\sum_{i\in\mathcal{I}} |v_i| \ge 2^{-8}\sqrt{d}\,\|\mathbf{v}\|_2 \\ &\text{for all } \mathbf{v} \in \operatorname{range}(\mathbf{A}),\ \mathcal{I} \subseteq [d] \text{ with } |\mathcal{I}| \ge d/4\}. \end{aligned}$$

By Lemma 39, $\Pr(\mathcal{E}^c) \le 52\sqrt{r/d}$. Rotational invariance and Cauchy–Schwarz give

$$\begin{aligned} \mathbb{E}[\mathbf{A}] = \frac{\operatorname{tr}(\mathbf{\Lambda})}{d}\mathbf{I}_d \preceq b\mathbf{I}_d, \qquad \mathbb{E}[F_0] \le \frac{b}{2}\|\mathbf{w}^{(0)} - \mathbf{w}^\star\|_2^2 \le 2bM^2\Delta^2 d, \\ \mathbb{E}[\sqrt{F_0}] \le \sqrt{\mathbb{E}[F_0]} \le \sqrt{2b}\,M\Delta\sqrt{d}. \end{aligned} \tag{D.16}$$

On $\mathcal{E}$, Lemma 40 gives the conditional bounds required by Lemma 41. Write $F_{\max} = \Delta^2 \lambda_{\max}(\mathbf{A}) d$ and apply the latter lemma with

$$m = \delta d, \qquad \varepsilon = \ell \Delta^2 r, \qquad s = t_\star = \left\lceil \alpha \log \frac{2d}{r} \right\rceil.$$

The definition of δ gives $0 < m < d/4$, and the theorem assumes $s \le T$. The coefficients supplied by Lemma 40 are

$$q_1 = 1 - \frac{\chi \eta (2 - S) \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))}{2},$$
$$A_1 = \frac{\eta \Delta^2 \sum_i A_{ii}^2}{4(2 - S)}, \qquad B_1 = \frac{S}{\Delta} \sqrt{\frac{2d}{\lambda_{\max}(\mathbf{A})}}.$$

The required inequality $\varepsilon \ge 2A_1/(1 - q_1)$ follows from

$$\frac{2A_1}{1 - q_1} = \frac{\Delta^2 \sum_i A_{ii}^2}{\chi (2 - S)^2 \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))} \le \frac{24 b^2 \Delta^2 r}{\chi (2 - S)^2 a} = \ell \Delta^2 r.$$

Moreover, with $Q_1 \triangleq (1 + q_1)/2$,

$$1 - Q_1 = \frac{\chi S (2 - S) \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))}{4 \lambda_{\max}(\mathbf{A})} \ge \frac{\chi S (2 - S) a}{4b}, \qquad Q_1^{t_\star} \le \left(\frac{r}{2d} \right)^2. \tag{D.17}$$

Indeed, $Q_1^{t_\star} \le e^{-(1 - Q_1) t_\star}$, and the choice of α makes $(1 - Q_1) t_\star \ge 2 \log(2d/r)$.

The coefficients q_1, B_1 are deterministic and A_1 is uniformly bounded on $\mathcal{E}$. For $Z \in \{F_0, \sqrt{F_0}\}$, $\mathbb{E}[\mathbb{1}\{\mathcal{E}\} Z] \le \mathbb{E}[Z]$. Write $x \triangleq r/d$. Using $1/(1 - \sqrt{q_1}) \le 2/(1 - q_1)$ and Eq. (D.16), the terms involving B_1 in Eq. (D.12) satisfy

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\} \frac{B_1 \sqrt{F_0}}{\delta d (1 - \sqrt{q_1})} \right] \le \frac{8M \sqrt{b \lambda_{\max}(\mathbf{A})}}{\delta \chi (2 - S) \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))} \le \frac{8M (b/a)}{\delta \chi (2 - S)} \sqrt{x},$$
$$\frac{B_1 T}{\delta d} \sqrt{\frac{A_1}{1 - q_1}} = \frac{ST}{\delta d} \sqrt{\frac{d \sum_i A_{ii}^2}{\chi (2 - S)^2 \lambda_{\max}(\mathbf{A}) \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A}))}} \le \frac{2\sqrt{6}\, S (b/a)}{\delta \sqrt{\chi} (2 - S)} xT.$$

The remaining terms satisfy

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\} \frac{F_0 + A_1 T}{F_{\max}} \right] \le 2(b/a) M^2 x + \frac{6 S (b/a)^2}{2 - S} x^2 T,$$
$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\} \frac{F_0}{\ell \Delta^2 r} Q_1^{t_\star} \right] \le \frac{bM^2}{2\ell} x.$$

Here $\lambda_{\max}(\mathbf{A}), \lambda_{\min}(\mathbf{A}; \mathrm{range}(\mathbf{A})) \ge ad/r$, $\lambda_{\max}(\mathbf{A}) \le bd/r$, and $\sum_i A_{ii}^2 \le 24 b^2 d$ on $\mathcal{E}$. Adding $\Pr(\mathcal{E}^c)$ to Eq. (D.12) gives

$$\Pr\left(\max_{0 \le t \le T} |h^{(t)} - h^{(0)}| \le \delta d, \quad \min_{0 \le t \le t_\star} F_t \le \ell \Delta^2 r \right)$$
$$\ge 1 - 52\sqrt{x} - \frac{8M (b/a)}{\delta \chi (2 - S)} \sqrt{x} - \frac{2\sqrt{6}\, S (b/a)}{\delta \sqrt{\chi} (2 - S)} xT$$
$$- \left(2(b/a) + \frac{b}{2\ell} \right) M^2 x - \frac{6 S (b/a)^2}{2 - S} x^2 T. \tag{D.18}$$

If $M\sqrt{x} < 1$, then $M^2 x \le M\sqrt{x}$ and $x^2 T \le xT$, so the failure is at most $c(\sqrt{x} + xT)$ with Eq. (D.15). If $M\sqrt{x} \ge 1$, the claimed probability lower bound is nonpositive because $c \ge M$, and hence holds as well. Thus the event in Eq. (D.18) has probability at least $1 - c(\sqrt{r/d} + rT/d)$. On the event just bounded,

$$\min_{0 \le t \le t_\star} f(\mathbf{w}^{(t)}) \le \ell \Delta^2 r, \qquad h^{(t)} \ge h^{(0)} - \delta d \ge h^{(0)}/2 > 0 \quad (0 \le t \le T).$$

This proves the theorem and also gives $\tau_{\mathcal{M}} > T$. □

D.4 Proof of Corollary 9

Constants. For this Gaussian specialization, define

$$\rho \triangleq 2^{-13}, \qquad \chi \triangleq \frac{1}{2^{16} + 2^8\sqrt{2}\,S}, \qquad \alpha \triangleq \frac{72}{\chi S(2-S)}, \qquad \ell \triangleq \frac{486}{\chi(2-S)^2},$$
$$c \triangleq 57 + \max\left\{\frac{2}{\omega}, 8\right\}\left(\frac{72M}{\chi(2-S)} + \frac{18\sqrt{6}\,S}{\sqrt{\chi}(2-S)}\right) + M\left(18 + \frac{9}{8\ell}\right) + \frac{486S}{2-S}. \tag{D.19}$$

Here α, ℓ depend only on S, and $c = O(M)$ for fixed S, ω.

Set $t_\star = \lceil \alpha \log(2d/n) \rceil$.

Proof of Corollary 9. By Lemma 22, the n nonzero eigenvalues belong to $[d/(4n), 9d/(4n)]$ except on an event of probability at most $2e^{-d/32}$. For every $\mathbf{Q} \in \mathcal{O}(d)$, $\mathbf{XQ} \overset{\mathrm{d}}{=} \mathbf{X}$, so the distribution of $\mathbf{A} = \mathbf{X}^\top\mathbf{X}/n$ is invariant under orthogonal conjugation. Consequently, $\mathbf{A} \overset{\mathrm{d}}{=} \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda}$ lists the n positive eigenvalues followed by zeros, and $\mathbf{U}$ is Haar distributed independently of $\mathbf{\Lambda}$ (Meckes, 2019). Conditional on $\mathbf{\Lambda}$, $\mathbf{U}$ therefore remains Haar distributed. For every spectrum in the stated interval, apply Theorem 8 with $r = n$, $a = 1/4$, and $b = 9/4$. Substituting these values into Eq. (D.15) gives the time and objective coefficients in Eq. (D.19), and the failure coefficient $c - 5$. The bound is uniform over the allowed spectra. Averaging the conditional probability and adding the spectral failure probability yields

$$\Pr\left(\min_{0\le t\le t_\star} f(\mathbf{w}^{(t)}) \le \ell\Delta^2 n, \quad \min_{0\le t\le T} h^{(t)} \ge \frac{h^{(0)}}{2}\right)$$
$$\ge 1 - (c-5)\left(\sqrt{\frac{n}{d}} + \frac{nT}{d}\right) - 2e^{-d/32}. \tag{D.20}$$

Since $n \ge 1$ and

$$2e^{-d/32} \le \frac{8}{\sqrt{e}}\sqrt{\frac{n}{d}} \le 5\sqrt{\frac{n}{d}},$$

the total failure probability is at most $c(\sqrt{n/d} + nT/d)$. Since $h^{(0)} \ge \omega d > 0$, the same event gives $h^{(t)} \ge h^{(0)}/2 > 0$ for every integer $0 \le t \le T$, and hence $\tau_{\mathcal{M}} > T$. □

E Off-grid targets

For any $\mathbf{w}, \mathbf{v} \in \mathcal{W}$, the displacement from $\mathbf{v}$ before rounding is

$$\mathbf{w} - \eta\mathbf{A}(\mathbf{w} - \mathbf{w}^\star) - \mathbf{v} = \mathbf{J}(\mathbf{w} - \mathbf{v}) - \eta\mathbf{A}(\mathbf{v} - \mathbf{w}^\star). \tag{E.1}$$

Applying Eq. (A.3) coordinatewise gives

$$\bar{h}_{\mathbf{v}}(\mathbf{w}) \triangleq \mathbb{E}[\|\mathbf{w}^{(1)} - \mathbf{v}\|_0 \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}]$$
$$= \sum_{i=1}^{d} \min\left\{1, \frac{|\Pi(w_i - \eta[\mathbf{A}(\mathbf{w} - \mathbf{w}^\star)]_i) - v_i|}{\Delta}\right\}. \tag{E.2}$$

Conditional on $\mathcal{F}_t$, $\|\mathbf{w}^{(t+1)} - \mathbf{v}\|_0$ is a sum of independent Bernoulli variables with mean $\bar{h}_{\mathbf{v}}(\mathbf{w}^{(t)})$. Conditioning on $\mathbf{w}^{(0)} = \mathbf{w}$ denotes the transition law from $\mathbf{w}$. Moreover, for $u \in \mathbb{R}$ and $g \in \mathcal{G}$, $\mathrm{SR}(u) = g$ almost surely if and only if $\Pi(u) = g$. Thus $\mathbf{v}$ is absorbing if and only if

$$\Pi\big(\mathbf{v} - \eta\mathbf{A}(\mathbf{v} - \mathbf{w}^\star)\big) = \mathbf{v}. \tag{E.3}$$

If $\mathbf{v} \in \mathcal{W}^\circ$, this is equivalent to $\mathbf{A}(\mathbf{v} - \mathbf{w}^\star) = \mathbf{0}$.

E.1 Grid minimizers with few boundary coordinates

E.1.1 A conditional-mean bound for the set of grid minimizers

Lemma 42. *Fix* $\mathbf{A}$ *and an initializer* $\mathbf{w}^{(0)} \in \mathcal{W} \setminus \mathcal{M}$. *Suppose that, for some real* $k \geq 1$, *the conditional means in Eq. (E.2) satisfy*

$$\bar{h}_{\mathbf{v}}(\mathbf{w}) \geq 4\sqrt{k \min\{h(\mathbf{w}), k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \mathcal{M},\ \mathbf{v} \in \mathcal{M}). \tag{E.4}$$

With probability at least $1 - \frac{99}{100}|\mathcal{M}|e^{-\sqrt{k}/16}$ *over rounding,*

$$h^{(t+1)} \geq \min\{k, 2\sqrt{kh^{(t)}}\} \quad (0 \leq t < e^k), \qquad \tau_{\mathcal{M}} > e^k. \tag{E.5}$$

All time indices are integers.

Proof. We first bound the probability that one transition violates the claimed recurrence. For each $\mathbf{v} \in \mathcal{M}$, conditional on $\mathcal{F}_t$, the distance $\|\mathbf{w}^{(t+1)} - \mathbf{v}\|_0$ is a sum of independent Bernoulli variables with mean $\bar{h}_{\mathbf{v}}(\mathbf{w}^{(t)})$. For a Bernoulli sum X of mean μ, the Chernoff inequalities

$$\Pr(X < \mu/2) \leq e^{-\mu/8}, \qquad \Pr(X < k) \leq \exp\big(2(\log 2)k - 3\mu/4\big)$$

follow from exponential Markov's inequality; the second uses the exponential parameter $2\log 2$. Substituting $\mu \geq 4\sqrt{k \min\{h^{(t)}, k\}}$ and taking a union bound over $\mathcal{M}$ gives

$$\Pr(h^{(t+1)} < 2\sqrt{kh^{(t)}} \mid \mathcal{F}_t) \leq |\mathcal{M}|e^{-\sqrt{kh^{(t)}}/2} \qquad (1 \leq h^{(t)} < k), \tag{E.6}$$

$$\Pr(h^{(t+1)} < k \mid \mathcal{F}_t) \leq |\mathcal{M}|e^{-(3-2\log 2)k} \qquad (h^{(t)} \geq k). \tag{E.7}$$

This argument requires independence of the coordinate indicators for each fixed minimizer, but no independence between distances to different minimizers. Define the event that the next transition violates its required bound by

$$\mathcal{B}_t \triangleq \{1 \leq h^{(t)} < k,\ h^{(t+1)} < 2\sqrt{kh^{(t)}}\} \cup \{h^{(t)} \geq k,\ h^{(t+1)} < k\}, \qquad \mathcal{E}_t \triangleq \bigcap_{s=0}^{t-1} \mathcal{B}_s^c.$$

Here $\mathcal{E}_0 = \Omega$, and $\mathcal{E}_t$ is the event that none of the first t transitions violates its bound. Put $T = \lceil e^k \rceil$. For $k = 1$ or $k \geq 2$, the same bound on the number of transitions before reaching k and the same sum over the first failed transition as in Lemma 16 apply. Each conditional failure bound acquires the factor $|\mathcal{M}|$. Thus Eq. (A.24) gives

$$\Pr(\mathcal{E}_T^c) \leq \frac{99}{100}|\mathcal{M}|e^{-\sqrt{k}/16}. \tag{E.8}$$

For $1 < k < 2$ and $|\mathcal{M}| = 1$, the two cases according to $h^{(0)} \geq k$ or $h^{(0)} = 1$ in that proof apply without change: they use only the displayed conditional bounds and exponential Markov's inequality. For $1 < k < 2$ and $|\mathcal{M}| \geq 2$, the right-hand side of Eq. (E.8) exceeds one, so the inequality also holds. On $\mathcal{E}_T$, the recurrence holds and all distances through time T are positive. Hence $\tau_{\mathcal{M}} > T \geq e^k$. □

E.1.2 A small-ball bound for a fixed grid pair

At every interior coordinate of $\mathbf{v}$, the corresponding summand in $\bar{h}_{\mathbf{v}}(\mathbf{w})$ (Eq. (E.4)) equals the truncated displacement before clipping. The next lemma ensures that this displacement is large on many coordinates. We apply it to all relevant deterministic grid pairs before restricting to the realized minimizers and discarding their boundary coordinates.

Lemma 43. *Let* $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ *and* $\mathbf{w}^\star \notin \mathcal{W}$ *be as in Theorem 10. Fix distinct* $\mathbf{w}, \mathbf{v} \in \mathcal{W}$ *before drawing* $\mathbf{U}$, *and write* $\mathbf{z} = \mathbf{w} - \mathbf{v}$ *and* $\mathbf{u} = \mathbf{w} - \mathbf{w}^\star$. *For* $\mathbf{u} \neq \mathbf{0}$, *define*

$$R(\mathbf{u}) \triangleq \frac{1}{\|\mathbf{u}\|_2}\left\|\mathbf{A}\mathbf{u} - \frac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\|\mathbf{u}\|_2^2}\mathbf{u}\right\|_2.$$

Put $Q = 2M + 1$ *and*

$$r_0 = \lceil d/8 \rceil, \quad p = d - r_0 - 2, \quad \chi = [2Q^2(8e)^{1/8}]^{-8/7}.$$

Suppose $d \geq 64$, $R_0 > 0$, $G \geq 1$, *and* $0 < b_0 < 1$ *satisfy*

$$\|\mathbf{u}\|_2 \geq \frac{\Delta\sqrt{\|\mathbf{w}-\mathbf{v}\|_0}}{G}, \qquad b_0 \leq \chi\sqrt{\frac{\pi}{2}}\frac{SR_0}{bG}.$$

The step size $\eta = S/\lambda_{\max}(\mathbf{\Lambda})$ *is deterministic. Define*

$$\mathcal{B}_{\mathbf{w},\mathbf{v}} \triangleq \left\{\left|\left\{i \in [d] : |z_i - \eta(\mathbf{A}\mathbf{u})_i| > b_0\Delta\sqrt{\|\mathbf{w}-\mathbf{v}\|_0/d}\right\}\right| \leq r_0\right\}. \tag{E.9}$$

Then

$$\Pr\big(\{R(\mathbf{u}) \geq R_0\} \cap \mathcal{B}_{\mathbf{w},\mathbf{v}}\big) \leq \binom{d}{r_0} d\chi^p. \tag{E.10}$$

Proof. Since $d \geq 64$, we have $p \geq 1$. On $\mathcal{B}_{\mathbf{w},\mathbf{v}}$, at least $d - r_0$ coordinates satisfy $|z_i - \eta(\mathbf{A}\mathbf{u})_i| \leq b_0\Delta\sqrt{\|\mathbf{w}-\mathbf{v}\|_0/d}$. Choose a set $\mathcal{J}$ of exactly $d - r_0$ such coordinates. For each such deterministic set $\mathcal{J}$, choose $\mathbf{P} \in \mathbb{R}^{p\times|\mathcal{J}|}$ with orthonormal rows satisfying $\mathbf{P}\mathbf{z}_{\mathcal{J}} = \mathbf{P}\mathbf{u}_{\mathcal{J}} = \mathbf{0}$. This is possible because the orthogonal complement of $\operatorname{span}\{\mathbf{z}_{\mathcal{J}}, \mathbf{u}_{\mathcal{J}}\}$ has dimension at least $|\mathcal{J}| - 2 = p$. Applying $\mathbf{P}$ removes $\mathbf{z}_{\mathcal{J}}$. The symmetry of the cube therefore gives

$$\mathcal{B}_{\mathbf{w},\mathbf{v},\mathcal{J}} \triangleq \{\mathbf{P}(\mathbf{A}\mathbf{u})_{\mathcal{J}} \in \mathbf{P}[-a_0, a_0]^{|\mathcal{J}|}\}, \qquad a_0 = \frac{b_0\Delta}{\eta}\sqrt{\|\mathbf{w}-\mathbf{v}\|_0/d}. \tag{E.11}$$

Thus $\mathcal{B}_{\mathbf{w},\mathbf{v}} \subseteq \bigcup_{|\mathcal{J}|=d-r_0} \mathcal{B}_{\mathbf{w},\mathbf{v},\mathcal{J}}$.

We next bound the conditional density in this containing event. Set $\mathbf{e} = \mathbf{u}/\|\mathbf{u}\|_2$ and $q = \mathbf{e}^\top\mathbf{A}\mathbf{e}$. By Lemma 28, conditional on $\mathbf{U}^\top\mathbf{e}$, the component of $\mathbf{A}\mathbf{e}$ orthogonal to $\mathbf{e}$ has uniform direction. Both q and $R(\mathbf{u})$ are functions of $\mathbf{U}^\top\mathbf{e}$, so the tower property gives

$$\mathbf{A}\mathbf{u} = q\mathbf{u} + R(\mathbf{u})\|\mathbf{u}\|_2\mathbf{y}, \qquad \mathbf{y} \mid (q, R(\mathbf{u})) \sim \operatorname{Unif}\{\mathbf{x} \in \mathbf{u}^\perp : \|\mathbf{x}\|_2 = 1\} \quad (R(\mathbf{u}) > 0). \tag{E.12}$$

Extending the rows of $\mathbf{P}$ by zeros outside $\mathcal{J}$ gives p orthonormal vectors in $\mathbf{u}^\perp$. Consequently, with $D = d - 1$, the conditional distribution of $\mathbf{P}\mathbf{y}_{\mathcal{J}}$ is that of the first p coordinates of a uniform point on the unit sphere in $\mathbb{R}^D$.

To compute its density, write that point as $\mathbf{g}/\|\mathbf{g}\|_2$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$. The variables $X = \sum_{i=1}^p g_i^2$ and $Y = \sum_{i=p+1}^D g_i^2$ are independent, with distributions χ_p^2 and χ_{D-p}^2. Changing variables from (X, Y) to $(X/(X+Y), X+Y)$ in their joint density gives, with Γ denoting the gamma function, the density

$$\frac{\Gamma(D/2)}{\Gamma(p/2)\Gamma((D-p)/2)} s^{p/2-1}(1-s)^{(D-p)/2-1} \qquad (0 < s < 1)$$

for the squared norm of the first p coordinates. Their direction is independent and uniform. Writing $s = r^2$ and dividing the radial density by $[2\pi^{p/2}/\Gamma(p/2)]r^{p-1}$ gives the Lebesgue density

$$\frac{\Gamma(D/2)}{\pi^{p/2}\Gamma((D-p)/2)}(1 - \|\mathbf{x}\|_2^2)^{(D-p-2)/2}\mathbb{1}\{\|\mathbf{x}\|_2 < 1\}.$$

Here $D - p = r_0 + 1 \geq 2$, so the last power is nonnegative. The density is at most its constant factor. For a gamma random variable Z with shape $x > 0$ and unit scale, differentiating the defining gamma integral and applying Jensen's inequality give

$$(\log\Gamma)'(x) = \mathbb{E}[\log Z] \leq \log\mathbb{E}[Z] = \log x.$$

It follows that

$$\log\frac{\Gamma(D/2)}{\Gamma((D-p)/2)} = \int_{(D-p)/2}^{D/2} (\log\Gamma)'(x)\,dx \leq \frac{p}{2}\log(D/2).$$

Since $\mathbf{P}\mathbf{u}_{\mathcal{J}} = \mathbf{0}$, Eq. (E.12) gives $\mathbf{P}(\mathbf{A}\mathbf{u})_{\mathcal{J}} = R(\mathbf{u})\|\mathbf{u}\|_2\mathbf{P}\mathbf{y}_{\mathcal{J}}$. On $\{R(\mathbf{u}) \geq R_0\}$, its conditional density given $(q, R(\mathbf{u}))$ is therefore at most

$$\left(\frac{\sqrt{d/(2\pi)}}{R_0\|\mathbf{u}\|_2}\right)^p \leq \left(\frac{\chi}{2a_0}\right)^p. \tag{E.13}$$

Indeed, the assumptions on $\|\mathbf{u}\|_2$ and b_0, together with $\eta \geq S/b$, give

$$2a_0 \frac{\sqrt{d/(2\pi)}}{R_0\|\mathbf{u}\|_2} \leq \sqrt{\frac{2}{\pi}} \frac{b_0 G}{\eta R_0} \leq \sqrt{\frac{2}{\pi}} \frac{b_0 G b}{S R_0} \leq \chi.$$

It remains to bound the volume of the projected cube. The zonotope volume formula (Gover, 2014), Cauchy–Schwarz, and Cauchy–Binet give

$$\begin{aligned}
\mathrm{vol}_p(\mathbf{P}[-a_0, a_0]^{|\mathcal{J}|}) &= (2a_0)^p \sum_{\mathcal{S}\subseteq\mathcal{J},\,|\mathcal{S}|=p} |\det \mathbf{P}_{\mathcal{S}}| \\
&\leq (2a_0)^p \binom{|\mathcal{J}|}{p}^{1/2} \left(\sum_{\mathcal{S}\subseteq\mathcal{J},\,|\mathcal{S}|=p} \det(\mathbf{P}_{\mathcal{S}})^2 \right)^{1/2} \\
&= (2a_0)^p \binom{p+2}{2}^{1/2} \det(\mathbf{P}\mathbf{P}^\top)^{1/2} \leq d(2a_0)^p,
\end{aligned}$$

where $\mathbf{P}_{\mathcal{S}}$ denotes the columns indexed by $\mathcal{S}$. On $\{R(\mathbf{u}) \geq R_0\}$, multiplying this volume by Eq. (E.13) gives

$$\Pr(\mathcal{B}_{\mathbf{w},\mathbf{v},\mathcal{J}} \mid q, R(\mathbf{u})) \leq d(2a_0)^p \left(\frac{\chi}{2a_0}\right)^p = d\chi^p.$$

Integrating this conditional bound yields

$$\Pr\big(\{R(\mathbf{u}) \geq R_0\} \cap \mathcal{B}_{\mathbf{w},\mathbf{v},\mathcal{J}}\big) = \mathbb{E}[\mathbb{1}\{R(\mathbf{u}) \geq R_0\} \Pr(\mathcal{B}_{\mathbf{w},\mathbf{v},\mathcal{J}} \mid q, R(\mathbf{u}))] \leq d\chi^p.$$

A union bound over the $\binom{d}{r_0}$ possible sets $\mathcal{J}$ proves the lemma. □

E.1.3 Proof of Theorem 10

We prove a conditional-mean lower bound for every nonoptimal grid point and every grid minimizer with at most θd boundary coordinates, then apply Lemma 42 to obtain the claimed trajectory. The constants below control the rotation event and the Hamming-distance threshold.

Under the assumptions of Theorem 10, define

$$\begin{aligned}
Q &\triangleq 2M+1, \quad B \triangleq b-a, \quad \alpha \triangleq \frac{v^2}{4B^2}, \quad g \triangleq \frac{v^2}{2B}, \\
L &\triangleq 4\log Q + 10, \qquad C \triangleq 4(L+1), \\
s_0 &\triangleq \frac{\alpha}{eC} e^{-2L/\alpha}, \qquad R_0 \triangleq g s_0.
\end{aligned} \tag{E.14}$$

$$\begin{aligned}
\chi &\triangleq [2Q^2(8e)^{1/8}]^{-8/7}, \qquad G \triangleq 1 + \sqrt{b/a}, \\
b_0 &\triangleq \min\left\{\frac{1}{2}, \chi\sqrt{\frac{\pi}{2}}\frac{SR_0}{bG}\right\}, \qquad \rho \triangleq (1-8\theta)^2 \min\{1/64, b_0^2/1024\}.
\end{aligned} \tag{E.15}$$

Define

$$p_d \triangleq 3Q^d e^{-Ld} + 8e\chi^{-3} d\, 2^{-d}. \tag{E.16}$$

Set

$$c \triangleq \min\{(\log 2)/4, \sqrt{\rho}/32\}.$$

Choose $d_0 \in \mathbb{N}$ so that $p_d \leq e^{-cd}$ for every $d \geq d_0$ and

$$d_0 \geq \max\left\{\frac{1}{\rho}, \left(\frac{\log(99/50)}{c}\right)^2\right\}.$$

Such a choice exists because $L - \log Q = 3\log Q + 10 > 2c$ and $\log 2 > 2c$. All these constants depend only on M, S, a, b, v, θ.

Proof of Theorem 10. Fix $d \geq d_0$ and put $k = \rho d \geq 1$. We first construct a rotation event $\mathcal{E}$ with $\Pr(\mathcal{E}^c) \leq p_d$ on which $|\mathcal{M}| \leq 2$ and, if every $\mathbf{v} \in \mathcal{M}$ has at most θd boundary coordinates,

$$\bar{h}_{\mathbf{v}}(\mathbf{w}) \geq 4\sqrt{k \min\{h(\mathbf{w}), k\}} \qquad (\mathbf{w} \in \mathcal{W} \setminus \mathcal{M},\ \mathbf{v} \in \mathcal{M}). \tag{E.17}$$

Once Eq. (E.17) holds, the trajectory conclusion follows from Lemma 42; the bound $|\mathcal{M}| \leq 2$ controls its union over minimizers. Because the minimizers depend on $\mathbf{U}$, we apply Lemma 43 to deterministic grid pairs and take a union bound before restricting to the realized minimizers. To apply that lemma, we first bound $R(\mathbf{w} - \mathbf{w}^\star) = \|\mathbf{A}\mathbf{e} - (\mathbf{e}^\top \mathbf{A}\mathbf{e})\mathbf{e}\|_2$ from below for $\mathbf{e} = (\mathbf{w} - \mathbf{w}^\star)/\|\mathbf{w} - \mathbf{w}^\star\|_2$, $\mathbf{w} \in \mathcal{W}$. This lower bound ensures that the random component orthogonal to $\mathbf{w} - \mathbf{w}^\star$ has a sufficiently large radius. Write $x_i \triangleq \lambda_i - \bar{\lambda}$. Since $|x_i| \leq B$, $\sum_i x_i = 0$ and $d^{-1}\sum_i x_i^2 \geq v^2$,

$$\frac{1}{d}\sum_i \max\{x_i, 0\} = \frac{1}{d}\sum_i \max\{-x_i, 0\} \geq \frac{v^2}{2B}.$$

For $s \triangleq v^2/(4B)$, define $\mathcal{S}_\pm \triangleq \{i : \pm x_i \geq s\}$. The bound $d^{-1}\sum_i \max\{x_i, 0\} \leq s + B|\mathcal{S}_+|/d$ and its negative counterpart give $|\mathcal{S}_\pm| \geq \alpha d$ and $\lambda_i - \lambda_j \geq g = 2s$ for $i \in \mathcal{S}_+$, $j \in \mathcal{S}_-$.

For a fixed unit $\mathbf{e}$, the vector $\mathbf{U}^\top \mathbf{e}$ has the same distribution as $\mathbf{g}/\|\mathbf{g}\|_2$ for $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. In this representation, set $N \triangleq \|\mathbf{g}\|_2^2$, $W_\pm \triangleq \sum_{i \in \mathcal{S}_\pm} g_i^2/N$. The chi-square tails give

$$\Pr(N > Cd) \leq e^{-d(C-1-\log C)/2} \leq e^{-Ld},$$

$$\Pr\left(\sum_{i \in \mathcal{S}_\pm} g_i^2 < s_0 C d\right) \leq \left(\frac{e s_0 C d}{|\mathcal{S}_\pm|}\right)^{|\mathcal{S}_\pm|/2} \leq e^{-Ld}.$$

Here $C \geq 4$ and $\log C \leq C/2$. Thus $\Pr(W_+ < s_0 \text{ or } W_- < s_0) \leq 3e^{-Ld}$. The definition of R gives the weighted-variance identity

$$\begin{aligned} R(\mathbf{e})^2 &= \sum_i (\mathbf{U}^\top \mathbf{e})_i^2 \lambda_i^2 - \left(\sum_i (\mathbf{U}^\top \mathbf{e})_i^2 \lambda_i\right)^2 \\ &= \sum_{i<j} (\mathbf{U}^\top \mathbf{e})_i^2 (\mathbf{U}^\top \mathbf{e})_j^2 (\lambda_i - \lambda_j)^2 \geq g^2 W_+ W_-. \end{aligned}$$

Consequently, $R(\mathbf{e}) \geq g s_0 = R_0$ whenever both W_+ and W_- are at least s_0. Define

$$\mathcal{E}_1 \triangleq \bigcap_{\mathbf{w} \in \mathcal{W}} \{R(\mathbf{w} - \mathbf{w}^\star) \geq R_0\}.$$

The target is off-grid, so all these directions are nonzero, and $\Pr(\mathcal{E}_1^c) \leq 3Q^d e^{-Ld}$.

For $\mathbf{w} \neq \mathbf{v}$, put $\mathbf{z} \triangleq \mathbf{w} - \mathbf{v}$ and $\mathbf{u} \triangleq \mathbf{w} - \mathbf{w}^\star$. Let $\mathcal{I}$ be the set of such deterministic grid pairs satisfying

$$\|\mathbf{u}\|_2 \geq \Delta\sqrt{\|\mathbf{w} - \mathbf{v}\|_0}/G. \tag{E.18}$$

This collection is determined before drawing $\mathbf{U}$ and contains every pair with $\mathbf{v} \in \mathcal{M}$ and $\mathbf{w} \neq \mathbf{v}$. Indeed, minimality and the spectral bounds give

$$\sqrt{a}\|\mathbf{v} - \mathbf{w}^\star\|_2 \leq \sqrt{2f(\mathbf{v})} \leq \sqrt{2f(\mathbf{w})} \leq \sqrt{b}\|\mathbf{w} - \mathbf{w}^\star\|_2.$$

The triangle inequality and grid spacing therefore give

$$\Delta\sqrt{\|\mathbf{w} - \mathbf{v}\|_0} \leq \|\mathbf{w} - \mathbf{v}\|_2 \leq \|\mathbf{w} - \mathbf{w}^\star\|_2 + \|\mathbf{v} - \mathbf{w}^\star\|_2 \leq G\|\mathbf{w} - \mathbf{w}^\star\|_2,$$

which verifies Eq. (E.18).

Set $\mathcal{B} = \bigcup_{(\mathbf{w},\mathbf{v}) \in \mathcal{I}} \mathcal{B}_{\mathbf{w},\mathbf{v}}$, with the events in Eq. (E.9). By Eq. (E.15), $0 < b_0 \leq 1/2$, $b_0 \leq \chi\sqrt{\pi/2}\, S R_0/(bG)$, and $d \geq d_0 \geq 1/\rho \geq 64$. Thus Lemma 43 applies to each fixed pair in $\mathcal{I}$. With $r_0 = \lceil d/8 \rceil$ and $p = d - r_0 - 2$, it gives

$$\Pr\big(\{R(\mathbf{u}) \geq R_0\} \cap \mathcal{B}_{\mathbf{w},\mathbf{v}}\big) \leq \binom{d}{r_0} d\chi^p.$$

Since $\mathcal{E}_1 \subseteq \{R(\mathbf{u}) \geq R_0\}$ for each pair, a union bound over at most Q^{2d} pairs gives

$$\begin{aligned}\Pr(\mathcal{E}_1 \cap \mathcal{B}) &\leq Q^{2d}\binom{d}{r_0} d\chi^p \\ &\leq 8e\chi^{-3}d[Q^2(8e)^{1/8}\chi^{7/8}]^d = 8e\chi^{-3}d\,2^{-d}.\end{aligned} \tag{E.19}$$

The second inequality uses $p \geq 7d/8 - 3$, $0 < \chi < 1$, and $\binom{d}{r_0} \leq (8e)^{d/8+1}$.

Define $\mathcal{E} \triangleq (\mathcal{E}_1 \setminus \mathcal{B}) \cap \{|\mathcal{M}| \leq 2\}$. By Lemma 26 and the two preceding bounds, $\Pr(\mathcal{E}^c) \leq p_d$. Suppose now that $\max_{\mathbf{v}\in\mathcal{M}} |\{i \in [d] : |v_i| = M\Delta\}| \leq \theta d$. Fix $\mathbf{w} \notin \mathcal{M}$ and $\mathbf{v} \in \mathcal{M}$, and put $H = \|\mathbf{w} - \mathbf{v}\|_0 \geq h(\mathbf{w}) \geq 1$. On $\mathcal{E}$, more than $r_0 \geq d/8$ coordinates satisfy

$$|z_i - \eta(\mathbf{A}\mathbf{u})_i| > b_0\Delta\sqrt{H/d}.$$

At most θd of these coordinates have $|v_i| = M\Delta$. For each remaining coordinate, the interior-coordinate identity in Lemma 13 gives mismatch probability $\min\{1, |z_i - \eta(\mathbf{A}\mathbf{u})_i|/\Delta\}$. Since $b_0\sqrt{H/d} \leq b_0 \leq 1/2$, summing over these coordinates yields

$$\begin{aligned}\bar{h}_{\mathbf{v}}(\mathbf{w}) &\geq (1/8 - \theta)b_0\sqrt{dH} = \frac{(1-8\theta)b_0}{8}\sqrt{dH} \\ &\geq 4\sqrt{kH} \geq 4\sqrt{k\min\{h(\mathbf{w}), k\}},\end{aligned}$$

where $k = \rho d$ and $\rho \leq (1-8\theta)^2 b_0^2/1024$. This proves Eq. (E.17). The rotation event constructed above satisfies

$$\Pr(\mathcal{E}^c) \leq p_d = 3e^{-(L-\log Q)d} + 8e\chi^{-3}d\,e^{-(\log 2)d} \leq e^{-cd}$$

by the choice of d_0.

Fix a realization of $\mathbf{U}$ for which $\mathcal{E}$ and Eq. (7) hold. For every initializer outside $\mathcal{M}$, Lemma 42 applies with $k = \rho d$. Since $|\mathcal{M}| \leq 2$, its rounding failure probability satisfies

$$\frac{99}{100}|\mathcal{M}|e^{-\sqrt{\rho d}/16} \leq \frac{99}{50}e^{-2c\sqrt{d}} \leq e^{-c\sqrt{d}}.$$

Thus the claimed recurrence holds for every integer $0 \leq t < e^{\rho d}$. This proves Theorem 10 with the stated constants.

The probability bound is for the rotation event $\mathcal{E}$; the boundary-coordinate condition is imposed separately. The sufficient conditions in Proposition 11 ensure it for every rotation with $\theta = 1/16$. The rounding conclusion holds for each initializer outside the realized set $\mathcal{M}$. □

E.2 Objective gaps and localization

E.2.1 Lower bounds on the objective gap

Lemma 44. *Suppose $\mu \triangleq \lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathcal{W}) > 0$ and $k \geq 1$. If $f^\star_{\mathcal{W}} \leq \mu\Delta^2 k/32$, then*

$$f(\mathbf{w}) - f^\star_{\mathcal{W}} \geq \mu\Delta^2 k/4 \qquad (\mathbf{w} \in \mathcal{W},\ h(\mathbf{w}) \geq k).$$

Proof. Fix $\mathbf{v} \in \mathcal{M}$ and $\mathbf{w} \in \mathcal{W}$ with $h(\mathbf{w}) \geq k$. Put $\mathbf{z} \triangleq \mathbf{w} - \mathbf{v}$ and $Q \triangleq \mathbf{z}^\top\mathbf{A}\mathbf{z} \geq \mu\Delta^2 k$. The quadratic expansion gives

$$f(\mathbf{w}) - f(\mathbf{v}) = \tfrac{1}{2}Q + \langle \mathbf{A}(\mathbf{v} - \mathbf{w}^\star), \mathbf{z}\rangle. \tag{E.20}$$

Cauchy–Schwarz in the quadratic seminorm gives

$$|\langle \mathbf{A}(\mathbf{v} - \mathbf{w}^\star), \mathbf{z}\rangle| \leq \sqrt{2f^\star_{\mathcal{W}}Q} \leq Q/4.$$

Substitution into Eq. (E.20) yields $f(\mathbf{w}) - f^\star_{\mathcal{W}} \geq Q/4 \geq \mu\Delta^2 k/4$. □

E.2.2 Proof of Proposition 11

Proof. To prove the boundary-coordinate bound, fix $\mathbf{v} \in \mathcal{M}$ and set $m = |\{i \in [d] : |v_i| = M\Delta\}|$. For every comparison point $\mathbf{u} \in \mathcal{W}$, minimality gives

$$\|\mathbf{v} - \mathbf{w}^\star\|_2^2 \le \frac{2f(\mathbf{v})}{\lambda_{\min}(\mathbf{A})} \le \frac{2f(\mathbf{u})}{\lambda_{\min}(\mathbf{A})} \le K\|\mathbf{u} - \mathbf{w}^\star\|_2^2. \tag{E.21}$$

Under (i), the target lies in $[-M\Delta, M\Delta]^d$. Choose $\mathbf{u}$ by rounding each coordinate of $\mathbf{w}^\star$ to a nearest grid value. Then $\|\mathbf{u} - \mathbf{w}^\star\|_2^2 \le d\Delta^2/4$. Every boundary coordinate of $\mathbf{v}$ satisfies $|v_i - w_i^\star| \ge M\Delta - |w_i^\star| \ge \frac{1}{2}\sqrt{K/\theta}\Delta$. Consequently,

$$m\frac{K\Delta^2}{4\theta} \le \|\mathbf{v} - \mathbf{w}^\star\|_2^2 \le \frac{Kd\Delta^2}{4},$$

so $m \le \theta d$.

Under (ii), put $\xi = (1 + \sqrt{K/\theta})^{-1}$ and use the specified $\mathbf{u} \in \mathcal{W}^\circ$ in Eq. (E.21). Then $\|\mathbf{v} - \mathbf{w}^\star\|_2^2 \le K\xi^2\Delta^2 d$. If $|v_i| = M\Delta$, interiority of u_i gives

$$|v_i - w_i^\star| \ge |v_i - u_i| - |u_i - w_i^\star| \ge (1 - \xi)\Delta.$$

It follows that

$$m(1 - \xi)^2\Delta^2 \le K\xi^2\Delta^2 d.$$

Namely, $m \le dK\xi^2/(1 - \xi)^2 = \theta d$. Note that both arguments apply to every $\mathbf{v} \in \mathcal{M}$ and $K = b/a \ge \kappa(\mathbf{A})$ is unchanged by orthogonal conjugation. □

E.3 Upper bound for general off-grid targets

The probabilities in the proposition are over rounding conditional on the Hessian and target.

Fix $K \ge 1$. Let $\mathbf{A} \succ \mathbf{0}$ satisfy $\kappa(\mathbf{A}) \le K$, let $\mathbf{w}^\star \in [-M\Delta, M\Delta]^d$, and choose a grid minimizer $\mathbf{v} \in \mathcal{M}$. Set

$$\varepsilon \triangleq \frac{S}{136MK}, \quad c \triangleq \frac{S}{4MK}, \quad q \triangleq 1 - c, \quad H \triangleq 2M^2 + \frac{M}{S}. \tag{E.22}$$

These assumptions give $0 < c < 1$, $0 < q < 1$, and $2SHd > 1$. From every grid state, we will show that SR-GD reaches a grid minimizer within b_d iterations with probability at least p_d, where

$$b_d \triangleq \left\lceil \frac{\log(2SHd)}{\log(1/q)} \right\rceil + 1, \qquad p_d \triangleq \varepsilon^{db_d}. \tag{E.23}$$

Proposition 45. *Under the preceding assumptions, for every initializer and integer $\ell \ge 0$,*

$$\Pr(\tau_{\mathcal{M}} > \ell b_d \mid \mathbf{A}, \mathbf{w}^{(0)}) \le (1 - p_d)^\ell \le e^{-\ell p_d}. \tag{E.24}$$

In particular, there is $C > 0$, depending only on M, S, K, such that, for every initializer, with probability at least $1 - e^{-d}$,

$$\tau_{\mathcal{M}} \le \exp(Cd\log(d + 1)). \tag{E.25}$$

The grid minimizer need not be unique or absorbing.

Proof. It suffices to show that, from every state outside $\mathcal{M}$, SR-GD reaches the fixed minimizer $\mathbf{v}$ within b_d transitions with probability at least p_d. We construct a possible rounding outcome at each transition that decreases a nonnegative potential by the factor q and has probability at least ε^d. Rescale coordinates by Δ and divide the Hessian by $\lambda_{\max}(\mathbf{A})$. This gives unit grid spacing and spectral norm without changing the rounding probabilities. We may therefore assume $\Delta = \lambda_{\max}(\mathbf{A}) = 1$, so $\eta = S$ and $\mu \triangleq \lambda_{\min}(\mathbf{A}) \ge 1/K$. For a grid state $\mathbf{w}$, put $\mathbf{z} \triangleq \mathbf{w} - \mathbf{v}$, $\mathbf{u} \triangleq \mathbf{w} - S\nabla f(\mathbf{w})$, and

$$\Psi_{\mathbf{v}}(\mathbf{w}) \triangleq f(\mathbf{w}) - f(\mathbf{v}) + \frac{\|\mathbf{z}\|_1}{2S}, \qquad Q_{\mathbf{w}}(\mathbf{x}) \triangleq \|\mathbf{x} - \mathbf{u}\|_2^2 + \|\mathbf{x} - \mathbf{v}\|_1. \tag{E.26}$$

The purpose of $Q_{\mathbf{w}}$ is to bound the change of $\Psi_{\mathbf{v}}$. The objective is now 1-smooth, and $\mathbf{w} - \mathbf{u} = S\nabla f(\mathbf{w})$. Thus, for every $\mathbf{x} \in \mathcal{W}$, expanding the square gives

$$\begin{aligned}\Psi_{\mathbf{v}}(\mathbf{x}) - \Psi_{\mathbf{v}}(\mathbf{w}) &\le \nabla f(\mathbf{w})^\top(\mathbf{x} - \mathbf{w}) + \tfrac{1}{2}\|\mathbf{x} - \mathbf{w}\|_2^2 \\ &\quad + \frac{\|\mathbf{x} - \mathbf{v}\|_1 - \|\mathbf{w} - \mathbf{v}\|_1}{2S} \\ &= \frac{Q_{\mathbf{w}}(\mathbf{x}) - Q_{\mathbf{w}}(\mathbf{w})}{2S} - \frac{1 - S}{2S}\|\mathbf{x} - \mathbf{w}\|_2^2.\end{aligned}$$

It therefore suffices to select an outcome for which $Q_{\mathbf{w}}$ decreases enough to imply contraction of $\Psi_{\mathbf{v}}$. The objective gap is nonnegative because $\mathbf{v} \in \mathcal{M}$. The added ℓ_1 term is positive at every grid point other than $\mathbf{v}$, even if another grid point has the same objective value. For each coordinate, let r_i be the rounding endpoint nearest v_i (or $\Pi(u_i)$ if it is a grid point). By Lemma 19, the resulting vector $\mathbf{r} = (r_i)_{i=1}^d$ satisfies

$$Q_{\mathbf{w}}(\mathbf{r}) \le Q_{\mathbf{w}}(\mathbf{x}) \qquad (\mathbf{x} \in \mathcal{W}),$$

including when the update is clipped.

To compare values of $Q_{\mathbf{w}}$, introduce an auxiliary random vector $\boldsymbol{\delta}$; it is not an SR-GD update. Set $\theta \triangleq 1/(2M)$ and choose its coordinates independently, with

$$\Pr(\delta_i = -1) = \theta \max\{z_i, 0\}, \qquad \Pr(\delta_i = 1) = \theta \max\{-z_i, 0\}, \qquad \Pr(\delta_i = 0) = 1 - \theta|z_i|.$$

In the next two displays, expectations are over this auxiliary vector. Then $\mathbf{w} + \boldsymbol{\delta} \in \mathcal{W}$ and

$$\mathbb{E}[\boldsymbol{\delta}] = -\theta\mathbf{z}, \qquad \mathbb{E}[\|\boldsymbol{\delta}\|_2^2] = \theta\|\mathbf{z}\|_1, \qquad \mathbb{E}[\|\mathbf{z} + \boldsymbol{\delta}\|_1 - \|\mathbf{z}\|_1] = -\theta\|\mathbf{z}\|_1.$$

The two expected changes involving $\|\mathbf{z}\|_1$ cancel. Since $\mathbf{w} - \mathbf{u} = S\nabla f(\mathbf{w})$, expanding the square gives

$$\begin{aligned}\mathbb{E}[Q_{\mathbf{w}}(\mathbf{w} + \boldsymbol{\delta}) - Q_{\mathbf{w}}(\mathbf{w})] &= 2S\nabla f(\mathbf{w})^\top \mathbb{E}[\boldsymbol{\delta}] + \mathbb{E}[\|\boldsymbol{\delta}\|_2^2] \\ &\quad + \mathbb{E}[\|\mathbf{z} + \boldsymbol{\delta}\|_1 - \|\mathbf{z}\|_1] \\ &= -2S\theta\nabla f(\mathbf{w})^\top \mathbf{z}.\end{aligned}$$

Since $Q_{\mathbf{w}}(\mathbf{r})$ is no larger than $Q_{\mathbf{w}}(\mathbf{w} + \boldsymbol{\delta})$ for every realization of $\boldsymbol{\delta}$, and strong convexity gives $\nabla f(\mathbf{w})^\top \mathbf{z} \ge f(\mathbf{w}) - f(\mathbf{v}) + (\mu/2)\|\mathbf{z}\|_2^2$, we obtain

$$\begin{aligned}Q_{\mathbf{w}}(\mathbf{r}) - Q_{\mathbf{w}}(\mathbf{w}) &\le -2S\theta\nabla f(\mathbf{w})^\top \mathbf{z} \\ &\le -2S\theta\{f(\mathbf{w}) - f(\mathbf{v})\} - S\theta\mu\|\mathbf{z}\|_2^2. \qquad \text{(E.27)}\end{aligned}$$

The selected outcome $\mathbf{r}$ may have arbitrarily small probability. For each coordinate, let $p_i = \Pr(\mathrm{SR}(u_i) = r_i)$. To obtain a uniform lower bound, replace each selected endpoint for which $p_i < \varepsilon$ by the other endpoint, obtaining $\widetilde{\mathbf{r}}$. We next bound the increase of $Q_{\mathbf{w}}$ caused by these replacements. Deterministic coordinates are unchanged. If coordinate i is replaced, the two endpoints are adjacent and the selected one is nearest v_i. The scalar rounding probabilities therefore give

$$(\widetilde{r}_i - u_i)^2 + |\widetilde{r}_i - v_i| - \big((r_i - u_i)^2 + |r_i - v_i|\big) = 2p_i \le 2\varepsilon.$$

Thus each replacement increases $Q_{\mathbf{w}}$ by at most 2ε.

Since nonzero coordinates of $\mathbf{z}$ are integers, at most $\|\mathbf{z}\|_0 \le \|\mathbf{z}\|_2^2$ replacements have $z_i \ne 0$. If $z_i = 0$ is replaced, put $s = \widetilde{r}_i - r_i \in \{-1, 1\}$. Both endpoints are feasible, and the alternative endpoint is farther from $v_i = w_i$ in direction s. Hence $s(u_i - v_i) > 1 - \varepsilon$ and $\mathbf{v} + s\mathbf{e}_i \in \mathcal{W}$. Here $\mathbf{e}_i$ is the i-th standard basis vector. Minimality of $\mathbf{v}$ gives

$$0 \le f(\mathbf{v} + s\mathbf{e}_i) - f(\mathbf{v}) = s\nabla_i f(\mathbf{v}) + A_{ii}/2.$$

Since $u_i - v_i = -S\{\nabla_i f(\mathbf{v}) + (\mathbf{A}\mathbf{z})_i\}$,

$$Ss(\mathbf{A}\mathbf{z})_i = -s(u_i - v_i) - Ss\nabla_i f(\mathbf{v}) < -(1 - \varepsilon) + SA_{ii}/2 \le -(1 - \varepsilon - S/2).$$

Taking absolute values and using $\varepsilon \le 1/136$ and $S < 1$ gives

$$|S(\mathbf{A}\mathbf{z})_i| > 1 - \varepsilon - S/2 \ge 1/4.$$

The number of these replacements is at most $16S^2\|\mathbf{Az}\|_2^2 \le 16\|\mathbf{z}\|_2^2$, by summing the squares of the preceding coordinate bound and using $\|\mathbf{A}\|_2 = 1$. Together with the coordinates where $z_i \neq 0$, there are at most $17\|\mathbf{z}\|_2^2$ replacements. Using $34\varepsilon \le S\theta\mu/2$ gives

$$Q_{\mathbf{w}}(\widetilde{\mathbf{r}}) - Q_{\mathbf{w}}(\mathbf{w}) \le -2S\theta\{f(\mathbf{w}) - f(\mathbf{v})\} - \frac{S\theta\mu}{2}\|\mathbf{z}\|_2^2. \tag{E.28}$$

The comparison between $\Psi_{\mathbf{v}}$ and $Q_{\mathbf{w}}$ established above, together with Eq. (E.28), now gives

$$\begin{aligned}\Psi_{\mathbf{v}}(\widetilde{\mathbf{r}}) - \Psi_{\mathbf{v}}(\mathbf{w}) &\le \frac{Q_{\mathbf{w}}(\widetilde{\mathbf{r}}) - Q_{\mathbf{w}}(\mathbf{w})}{2S} \\ &\le -\theta\{f(\mathbf{w}) - f(\mathbf{v})\} - \frac{\theta\mu}{4}\|\mathbf{z}\|_2^2 \le -c\Psi_{\mathbf{v}}(\mathbf{w}).\end{aligned}$$

For the last inequality, use $f(\mathbf{w}) - f(\mathbf{v}) \ge 0$, $\|\mathbf{z}\|_1 \le \|\mathbf{z}\|_2^2$, and

$$\theta = \frac{1}{2M} \ge c, \qquad \frac{\theta\mu}{4} \ge \frac{1}{8MK} = \frac{c}{2S}.$$

By construction, every selected coordinate outcome has probability at least ε. Independent coordinatewise rounding therefore gives

$$\Pr(\mathbf{w}^{(1)} = \widetilde{\mathbf{r}} \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}) \ge \varepsilon^d. \tag{E.29}$$

Since $\mathbf{w}, \mathbf{v}, \mathbf{w}^\star \in [-M, M]^d$ and $f(\mathbf{v}) \ge 0$,

$$0 \le \Psi_{\mathbf{v}}(\mathbf{w}) \le \tfrac{1}{2}\|\mathbf{w} - \mathbf{w}^\star\|_2^2 + \frac{\|\mathbf{w} - \mathbf{v}\|_1}{2S} \le 2M^2 d + \frac{Md}{S} = Hd.$$

If $\mathbf{w} \neq \mathbf{v}$, the unit grid spacing gives $\|\mathbf{w} - \mathbf{v}\|_1 \ge 1$ and hence $\Psi_{\mathbf{v}}(\mathbf{w}) \ge 1/(2S)$. Following the selected outcomes for at most b_d transitions reaches $\mathbf{v}$: otherwise, geometric decrease along the selected path would give $\Psi_{\mathbf{v}}(\mathbf{w}^{(b_d)}) \le q^{b_d} Hd < 1/(2S)$ with $\mathbf{w}^{(b_d)} \neq \mathbf{v}$. By Eq. (E.29), this path has probability at least $\varepsilon^{db_d} = p_d$. The same argument applies from every state reached before the first visit to $\mathcal{M}$. Conditioning successively at the block endpoints therefore proves Eq. (E.24); the path only needs to reach $\mathbf{v}$, which need not be absorbing.

Taking $\ell = \lceil d/p_d \rceil$ in Eq. (E.24) gives $\tau_{\mathcal{M}} \le b_d\lceil d/p_d \rceil$ with probability at least $1 - e^{-d}$. Since $b_d = O(\log(d+1))$ and $p_d^{-1} = \exp(db_d \log(1/\varepsilon))$, this proves Eq. (E.25). □

E.4 Target-location separation

E.4.1 Coordinate probabilities for a displaced target

Lemma 46. *Suppose* $\|\mathbf{A} - \mathbf{I}_d\|_\infty \le C < 1/3$ *and* $\mathbf{w}^\star = \Delta\mathbf{1}/4$. *Set*

$$p \triangleq \frac{S(1-C)}{2(1+C)}, \qquad q_+ \triangleq \frac{S(1-3C)}{5(1+C)}.$$

For every $\mathbf{w} \in \{0, \Delta\}^d$, $\mathbf{w} - \eta\mathbf{A}(\mathbf{w} - \mathbf{w}^\star) \in (0, \Delta)^d$. *Consequently,* $\{0, \Delta\}^d$ *is invariant, and every* $i \in [d]$ *satisfies*

$$\Pr(w_i^{(1)} = 0 \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}) \ge p, \qquad \Pr(w_i^{(1)} = \Delta \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}) \ge q_+. \tag{E.30}$$

Proof. For $b_i \triangleq \sum_{j \neq i} |A_{ij}|$,

$$|A_{ii} - 1| + b_i \le C, \qquad 1 - C \le \lambda_{\max}(\mathbf{A}) \le 1 + C.$$

For every $\zeta \ge 1$,

$$\frac{A_{ii} + \zeta b_i}{\lambda_{\max}(\mathbf{A})} \le 1 + \zeta C. \tag{E.31}$$

If $\lambda_{\max}(\mathbf{A}) \ge 1$, this follows from $A_{ii} + \zeta b_i \le 1 + \zeta C$. If $\lambda_{\max}(\mathbf{A}) < 1$, then $A_{ii} \le \lambda_{\max}(\mathbf{A}) < 1$ and $b_i \le C - 1 + A_{ii}$, so

$$\frac{A_{ii} + \zeta b_i}{\lambda_{\max}(\mathbf{A})} \le \zeta + 1 - \frac{\zeta(1-C)}{\lambda_{\max}(\mathbf{A})} \le 1 + \zeta C.$$

Normalize $\mathbf{x} \triangleq \mathbf{w}/\Delta \in \{0,1\}^d$ and write $\mathbf{u} \triangleq \mathbf{x} - \eta\mathbf{A}(\mathbf{x} - \mathbf{1}/4)$. If $x_i = 0$, then

$$\frac{\eta}{4}(A_{ii} - 3b_i) \le u_i \le \frac{\eta}{4}(A_{ii} + 3b_i).$$

Since $A_{ii} - 3b_i \ge 1 - 3C$, Eq. (E.31) with $\zeta = 3$ gives

$$u_i \ge \frac{S(1-3C)}{4(1+C)} \ge q_+, \qquad u_i \le \frac{S(1+3C)}{4} < \frac{1}{2}.$$

If $x_i = 1$, then

$$1 - \frac{3\eta}{4}(A_{ii} + b_i) \le u_i \le 1 - \frac{3\eta}{4}(A_{ii} - b_i).$$

Using $A_{ii} - b_i \ge 1 - C$ and $\zeta = 1$ gives

$$u_i \ge 1 - \frac{3S(1+C)}{4} \ge \frac{1-3C}{4} \ge q_+, \qquad 1 - u_i \ge \frac{3S(1-C)}{4(1+C)} \ge p.$$

In both cases, $q_+ \le u_i \le 1 - p$; for $x_i = 0$, use $1 - u_i > 1/2 \ge p$. Clipping is inactive, and $\Pr(w_i^{(1)} = \Delta \mid \mathbf{A}, \mathbf{w}^{(0)} = \mathbf{w}) = u_i$. □

E.4.2 Proof of Theorem 12

The following choices are valid in Theorem 12:

$$q \triangleq 1 - \frac{S(1-C)}{2(1+C)}, \qquad \rho \triangleq \frac{S(1-3C)}{10(1+C)},$$
$$a \triangleq \frac{S(1-3C)^2}{40(1+C)}, \qquad c \triangleq \frac{S(1-3C)}{80(1+C)}. \tag{E.32}$$

For $C \in (0, 1/3)$, these constants satisfy $q, \rho \in (0,1)$, $a, c > 0$, and $a = (1-3C)\rho/4$.

Proof of Theorem 12. For (i), we use the potential $\psi(\mathbf{w}) \triangleq \|\mathbf{w}\|_1/\Delta$. The displacement of the target contributes an additive term to its one-step expectation. The assumption $\varepsilon d \le 1/8$ makes this term small enough to retain multiplicative contraction away from $\mathbf{0}$, so Lemma 17 gives the hitting-time tail. For (ii), Lemma 46 gives a positive lower bound on the probability of each coordinate being nonzero. A conditional Chernoff bound then keeps the Hamming distance proportional to d; a deterministic objective-gap bound converts this distance into the claimed optimization error. Fix $\mathbf{A}$ and $\mathbf{w}^{(0)} = \Delta\mathbf{1}$; probabilities are over rounding. The matrix assumption gives

$$(1-C)\mathbf{I}_d \preceq \mathbf{A} \preceq (1+C)\mathbf{I}_d, \qquad \|\mathbf{A}\mathbf{1}\|_\infty \le 1 + C.$$

For $\mathbf{w} \in \mathcal{W}$ and $0 \le \varepsilon \le 1/4$, $\mathbf{w}^\star = \varepsilon\Delta\mathbf{1}$ and $\Delta\|\mathbf{w}\|_1 \le \|\mathbf{w}\|_2^2$ give

$$\begin{aligned} f(\mathbf{w}) - f(\mathbf{0}) &= \tfrac{1}{2}\mathbf{w}^\top\mathbf{A}\mathbf{w} - \varepsilon\Delta\mathbf{w}^\top\mathbf{A}\mathbf{1} \\ &\ge \left(\frac{1-C}{2} - \varepsilon(1+C)\right)\|\mathbf{w}\|_2^2 \ge \frac{1-3C}{4}\|\mathbf{w}\|_2^2. \end{aligned} \tag{E.33}$$

Thus $\mathcal{M} = \{\mathbf{0}\}$ for both targets.

For (i), the diagonal entries of $\mathbf{J}$ are nonnegative. The row sums and symmetry therefore give

$$\big\||\mathbf{J}|\big\|_2 \le \|\mathbf{J}\|_1 = \|\mathbf{J}\|_\infty \le 1 - \eta(1-C) \le 1 - \frac{S(1-C)}{1+C} < 1.$$

The target displacement satisfies

$$\frac{\eta\|\mathbf{A}\mathbf{w}^\star\|_1}{\Delta} \le \eta\varepsilon d(1+C) \le \frac{\eta(1+C)}{8}.$$

For $\mathbf{w} \ne \mathbf{0}$, Lemma 13 therefore gives

$$\begin{aligned} \mathbb{E}[\psi(\mathbf{w}^{(1)}) \mid \mathbf{w}^{(0)} = \mathbf{w}] &\le \frac{\|\mathbf{J}\mathbf{w} + \eta\mathbf{A}\mathbf{w}^\star\|_1}{\Delta} \\ &\le [1 - \eta(1-C)]\psi(\mathbf{w}) + \frac{\eta(1+C)}{8} \\ &\le \left[1 - \frac{\eta(1-C)}{2}\right]\psi(\mathbf{w}) \le q\psi(\mathbf{w}). \end{aligned}$$

Here $\psi(\mathbf{w}) \geq 1$, $(1-C)-(1+C)/8 \geq (1-C)/2$, and $\eta \geq S/(1+C)$. Stopping at the first visit to $\mathbf{0}$ ensures that no contraction assumption is needed after that visit. Apply Lemma 17 with threshold 1, $Z_t = \psi(\mathbf{w}^{(t)})$, $\tau = \tau_{\mathcal{M}}$, and $Z_0 = d$ to obtain $\Pr(\tau_{\mathcal{M}} > t) \leq dq^t$.

For (ii), set

$$q_+ \triangleq \frac{S(1-3C)}{5(1+C)} = 2\rho, \qquad \mathcal{E} \triangleq \bigcap_{\substack{t\in\mathbb{Z}_{\geq 0}\\ t\leq e^{cd}}} \{h^{(t)} \geq q_+ d/2\}.$$

By Lemma 46, $h^{(t+1)}$ conditional on $\mathcal{F}_t$ is a Bernoulli sum with mean at least $q_+ d$. The Chernoff bound $\Pr(X < \mu/2) \leq e^{-\mu/8}$ for a Bernoulli sum of mean μ, followed by a union bound, gives

$$\Pr(h^{(t+1)} < q_+ d/2 \mid \mathcal{F}_t) \leq e^{-q_+ d/8} = e^{-2cd}, \qquad \Pr(\mathcal{E}^c) \leq e^{cd} e^{-2cd} \leq e^{-cd}.$$

Here $h^{(0)} = d$. On $\mathcal{E}$, Eq. (E.33) gives

$$h^{(t)} \geq \rho d > 0, \qquad f(\mathbf{w}^{(t)}) - f^\star_{\mathcal{W}} \geq \frac{1-3C}{4}\Delta^2 h^{(t)} \geq a\Delta^2 d \quad (0 \leq t \leq e^{cd},\ t \in \mathbb{Z}).$$

Since $\tau_{\mathcal{M}}$ is integer-valued, $\tau_{\mathcal{M}} > e^{cd}$ on $\mathcal{E}$. □

E.4.3 An upper bound for the displaced target

Proposition 47. *Suppose $\|\mathbf{A} - \mathbf{I}_d\|_\infty \leq C < 1/3$ and $\mathbf{w}^\star = \Delta\mathbf{1}/4$. Set $p \triangleq S(1-C)/(2(1+C))$. For every initializer $\mathbf{w}^{(0)} \in \{0, \Delta\}^d$ and integer $T \geq 0$,*

$$\Pr(\tau_{\mathcal{M}} > T \mid \mathbf{w}^{(0)}) \leq (1-p^d)^T \leq e^{-Tp^d}. \tag{E.34}$$

In particular, $\tau_{\mathcal{M}} \leq \lceil dp^{-d} \rceil$ with probability at least $1 - e^{-d}$.

Proof. By Lemma 46 and Eq. (E.33), the binary state space is invariant, $\mathcal{M} = \{\mathbf{0}\}$, and independent rounding gives $\Pr(\mathbf{w}^{(t+1)} = \mathbf{0} \mid \mathcal{F}_t) \geq p^d$. On $\{\tau_{\mathcal{M}} > t\}$, this gives $\Pr(\tau_{\mathcal{M}} > t+1 \mid \mathcal{F}_t) \leq 1-p^d$. Taking expectations and iterating yields

$$\Pr(\tau_{\mathcal{M}} > T \mid \mathbf{w}^{(0)}) \leq (1-p^d)^T \leq e^{-Tp^d},$$

which proves Eq. (E.34). This argument does not require $\mathbf{0}$ to be absorbing. Taking $T = \lceil dp^{-d} \rceil$ gives the stated probability bound. □

F Improved dependence on the grid range

In this section, we show that the constants in Theorems 1 and 4 can be chosen so that ρ^{-1} and d_0 depend at most logarithmically on the grid range M. Specifically, we improve the dependence on M of the thresholds in the conditional-mean conclusion of Proposition 3 for the random Hessians in Appendix B. The proofs strengthen the matrix estimates; once the conditional-mean bound is established, we apply Lemma 16 as in the baseline proofs.

Proposition 48. *The following statements hold.*

(i) Fix $S \in (0,1)$, $0 < a < b$, and $v > 0$. There exists $C_{\mathrm{H}} > 0$, depending only on S, a, b, v, such that, for every $M \in \mathbb{N}$, the constants ρ, c, d_0 in Theorem 1 can be chosen to satisfy

$$\max\{\rho^{-1}, c^{-2}, d_0\} \leq C_{\mathrm{H}} \log(8M).$$

(ii) Fix $S \in (0,1)$ and $R > 0$. There exists $C_{\mathrm{G}} > 0$, depending only on S, R, such that, for every $M \in \mathbb{N}$, the constants ρ, a, c, d_0 in Theorem 4 can be chosen to satisfy

$$\max\{\rho^{-1}, d_0\} \leq C_{\mathrm{G}} \log(8M), \qquad c \geq C_{\mathrm{G}}^{-1}.$$

The curvature constant a in part (ii) is specified separately in Eq. (F.16). We first establish the lattice estimates used in the proofs, then prove parts (i) and (ii) in Appendix F.2.3 and Appendix F.3.4, respectively.

F.1 LATTICE ESTIMATES

F.1.1 SPHERICAL APPROXIMATION BY LATTICE POINTS

Lemma 49. *Let $D \geq 2$ and let $1 \leq r \leq D$ be integers. Let $\mathbf{v}$ be uniform on the unit sphere in $\mathbb{R}^D$. For every deterministic $R \geq 0$ and $0 < \varepsilon \leq 2^{-12}$,*

$$\Pr\left(\exists \mathbf{z} \in \mathbb{Z}^D : \|\mathbf{z}\|_2 \geq \sqrt{r}, \quad \|R\mathbf{v} - \mathbf{z}\|_2 \leq \varepsilon\sqrt{r}\right) \leq e^{-D}.$$

The same conclusion holds for a random radius, provided that, conditional on the radius, the direction is uniform whenever the radius is positive.

Proof. We bound the probability for each lattice point by the area of a spherical cap and then sum over the lattice points within distance $\varepsilon\sqrt{r}$ of the sphere. For large R, the lattice count has an upper bound proportional to R^{D-1}, and the cap probability has an upper bound proportional to $R^{-(D-1)}$. Multiplying these bounds eliminates R. Set $s = \varepsilon\sqrt{r}$ and

$$\mathcal{E}_R \triangleq \{\exists \mathbf{z} \in \mathbb{Z}^D : \|\mathbf{z}\|_2 \geq \sqrt{r},\ \|R\mathbf{v} - \mathbf{z}\|_2 \leq s\}.$$

If $R < (1-\varepsilon)\sqrt{r}$, this event is empty. Otherwise $R \geq \sqrt{r}/2$ and $s/R \leq 2\varepsilon < 1/2$. For each lattice point $\mathbf{z}$ with $|\|\mathbf{z}\|_2 - R| \leq s$, the event $\mathcal{E}_{\mathbf{z}} \triangleq \{\|R\mathbf{v} - \mathbf{z}\|_2 \leq s\}$ is contained in a spherical cap of angular sine at most s/R. Writing V_j for the volume of the unit j-ball, projection onto the equatorial hyperplane gives

$$\Pr(\mathcal{E}_{\mathbf{z}}) \leq \frac{2V_{D-1}}{DV_D}(s/R)^{D-1} \leq (2s/R)^{D-1}.$$

The surface Jacobian is at most 2 because $s/R < 1/2$, and $DV_D \geq 2V_{D-1}$ follows by projecting the two hemispheres. The Gaussian integral gives $V_D \leq (\sqrt{2\pi e}/\sqrt{D})^D \leq (5/\sqrt{D})^D$: on the ball of radius $\sqrt{D}$, the integrand $e^{-\|\mathbf{x}\|_2^2/2}$ is at least $e^{-D/2}$.

If $R \leq \sqrt{D}$, the disjoint unit cubes centered at these lattice points lie in a ball of radius $R + s + \sqrt{D}/2 \leq 2\sqrt{D}$. There are therefore at most 10^D such lattice points, and

$$\Pr(\mathcal{E}_R) \leq 10^D(2s/R)^{D-1} \leq 10(40\varepsilon)^{D-1} \leq e^{-D}.$$

If $R \geq \sqrt{D}$, put $w = s + \sqrt{D}/2 \leq \sqrt{D}$. The cubes lie in the annulus with radii $R - w$ and $R + w$, whose volume is

$$V_D\big((R+w)^D - (R-w)^D\big) \leq 2DwV_D(R+w)^{D-1} \leq 10D(10R/\sqrt{D})^{D-1}.$$

Thus

$$\Pr(\mathcal{E}_R) \leq 10D(10R/\sqrt{D})^{D-1}(2s/R)^{D-1} \leq 10D(20\varepsilon)^{D-1} \leq e^{-D}.$$

Both final inequalities hold at $D = 2$; their successive-dimension ratios are at most 40ε and 30ε, respectively, below e^{-1}. The bound is uniform in R, so conditioning on a random radius proves the last assertion; for $R = 0$, the event is empty. □

F.1.2 LATTICE POINTS IN EUCLIDEAN BALLS

Lemma 50. *For every $R > 0$,*

$$\left|\{\mathbf{z} \in \Delta\mathbb{Z}^d : \|\mathbf{z}\|_2 \leq R\Delta\sqrt{d}\}\right| \leq [\sqrt{2\pi e}(R + 1/2)]^d. \tag{F.1}$$

At radius $R = 1/4$, the sharper bound

$$\left|\{\mathbf{z} \in \Delta\mathbb{Z}^d : \|\mathbf{z}\|_2 \leq \Delta\sqrt{d}/4\}\right| \leq e^{d/3} \tag{F.2}$$

holds.

Proof. By rescaling, take $\Delta = 1$. The disjoint cubes $\mathbf{z} + [-1/2, 1/2)^d$ with $\|\mathbf{z}\|_2 \leq R\sqrt{d}$ lie in the ball of radius $(R + 1/2)\sqrt{d}$. For $R' > 0$,

$$\mathrm{vol}_d\{\|\mathbf{x}\|_2 \leq R'\sqrt{d}\} \leq e^{d/2} \int_{\mathbb{R}^d} e^{-\|\mathbf{x}\|_2^2/(2R'^2)}\, d\mathbf{x} = [\sqrt{2\pi e}\, R']^d,$$

using the integrand's lower bound $e^{-d/2}$ on the ball. Take $R' = R + 1/2$. For $\|\mathbf{z}\|_2^2 \le d/16$, $1 \le e^{d/4} e^{-4\|\mathbf{z}\|_2^2}$, so

$$\left|\{\mathbf{z} \in \mathbb{Z}^d : \|\mathbf{z}\|_2^2 \le d/16\}\right| \le e^{d/4} \left(\sum_{a \in \mathbb{Z}} e^{-4a^2}\right)^d \le \left[e^{1/4}\left(1 + \frac{2e^{-4}}{1 - e^{-12}}\right)\right]^d \le e^{d/3}.$$

Here $a^2 \ge 1 + 3(a-1)$ for integers $a \ge 1$, $\log(1+x) \le x$, and $2e^{-4}/(1-e^{-12}) < 1/12$. □

F.2 Haar eigenbases

Use the Haar model of Appendix B.1, with deterministic eigenvalues in $[a, b]$, $0 \le a < b$, and empirical variance at least $v^2 > 0$. The positive variance ensures that $\lambda_{\max}(\mathbf{A}) > 0$. For this subsection, use the constants from Eqs. (B.1) and (B.2):

$$q = \frac{1}{256} \min\left\{1, \left(\frac{v}{b-a}\right)^4\right\}, \qquad u = \frac{Sv}{128b}, \qquad \delta = \frac{q^2}{16e(1 + 4/u)}, \qquad \theta = \min\{\delta/2, u^2/9\}. \tag{F.3}$$

The bound in Lemma 30 controls displacements with at most θd nonzero coordinates. For the remaining displacements, the following lattice estimate gives a threshold proportional to $d/\log(8M)$. Define

$$\varepsilon \triangleq 2^{-12}, \qquad Q \triangleq 1 + \frac{2^{16}}{(1-S)\sqrt{\theta}}, \qquad \chi \triangleq \frac{1}{4096 Q^2}. \tag{F.4}$$

F.2.1 A lower bound under $\|\mathbf{z}\|_0 > \theta d$

Lemma 51. *Under the Haar model above, if $d \ge 4/\theta$ and $1 \le k \le \chi d/\log(8M)$ is real, then, with probability at least $1 - e^{-d/4}$ over $\mathbf{U}$,*

$$\sum_i \min\{1, |(\mathbf{J}\mathbf{z})_i|\} \ge 4k \quad \textit{for every } \mathbf{z} \in \mathbb{Z}^d \textit{ with } \|\mathbf{z}\|_\infty \le 2M, \quad \|\mathbf{z}\|_0 > \theta d.$$

Proof. We show that a sum below $4k$ in the statement forces an event in which a random vector is close to a nonzero lattice point. Such a sum forces $\mathbf{J}\mathbf{z}$ to be close to a vector supported on at most $\lfloor 4k \rfloor$ coordinates. For each deterministic net point $\mathbf{b}_0$ supported on $\mathcal{I}$, the subvector $(\mathbf{J}^{-1}\mathbf{b}_0)_{\mathcal{I}^c}$ has uniform direction conditional on its norm being fixed and positive. We first apply Lemma 49 and a union bound to exclude all such lattice approximations. We then show that a sum below $4k$ would force one of the excluded events. Put $s = \lfloor 4k \rfloor$ and $r_0 = \lfloor \theta d/2 \rfloor \ge \theta d/4$. Since $(1-S)\mathbf{I}_d \preceq \mathbf{J} \preceq \mathbf{I}_d$, we have $\|\mathbf{J}\|_2 \le 1$ and $\|\mathbf{J}^{-1}\|_2 \le (1-S)^{-1}$. For each $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| \le s$, choose a deterministic a_0-net $\mathcal{N}_\mathcal{I}$ of $\{\mathbf{b} : b_i = 0 \ (i \notin \mathcal{I}),\ \|\mathbf{b}\|_2 \le 2M\sqrt{d}\}$, where $a_0 = \varepsilon(1-S)\sqrt{r_0}/2$. The volume bound gives

$$|\mathcal{N}_\mathcal{I}| \le (1 + 4M\sqrt{d}/a_0)^{|\mathcal{I}|} \le \left(1 + \frac{16M}{\varepsilon(1-S)\sqrt{\theta}}\right)^{|\mathcal{I}|} \le (QM)^{|\mathcal{I}|}.$$

For each $\mathbf{b}_0 \in \mathcal{N}_\mathcal{I}$, define

$$\mathcal{B}_{\mathcal{I},\mathbf{b}_0} \triangleq \left\{\exists \mathbf{z}' \in \mathbb{Z}^{d-|\mathcal{I}|} : \|\mathbf{z}'\|_2 \ge \sqrt{r_0}, \quad \|(\mathbf{J}^{-1}\mathbf{b}_0)_{\mathcal{I}^c} - \mathbf{z}'\|_2 \le \varepsilon\sqrt{r_0}\right\}.$$

Conjugation by orthogonal matrices acting identically on $\mathcal{I}$ preserves the law of $\mathbf{J}$ and fixes $\mathbf{b}_0$. Thus, conditional on its norm, $(\mathbf{J}^{-1}\mathbf{b}_0)_{\mathcal{I}^c}$ has uniform direction whenever its norm is positive. Since $d - |\mathcal{I}| \ge d/2 \ge 2$ and $r_0 \le d - |\mathcal{I}|$, Lemma 49 gives $\Pr(\mathcal{B}_{\mathcal{I},\mathbf{b}_0}) \le e^{-(d-|\mathcal{I}|)} \le e^{-d/2}$.

Set $p = s/d \le 4\chi/\log(8M) < 1/2$. Since

$$\sum_{j=0}^{s} \binom{d}{j} \le \exp\{d[-p\log p - (1-p)\log(1-p)]\},$$

the number of pairs $(\mathcal{I}, \mathbf{b}_0)$ is at most

$$\exp\{d[-p\log p-(1-p)\log(1-p)+p\log(QM)]\}.$$

Monotonicity of $x\log(eQM/x)$ gives

$$\begin{aligned}
&-p\log p-(1-p)\log(1-p)+p\log(QM)\\
&\le \frac{4\chi}{\log(8M)}\log\frac{eQM\log(8M)}{4\chi}\\
&= \frac{\log(1024e)+3\log Q+\log M+\log\log(8M)}{1024Q^2\log(8M)}\\
&\le \frac{6+(3/2)Q}{1024Q^2} < \frac{1}{4}.
\end{aligned}$$

Here $\log(8M) > 2$, $\log M, \log\log(8M) \le \log(8M)$, $\log Q \le Q$, and $\log(1024e) < 8$. For $\mathcal{E}_2 \triangleq \bigcap_{|\mathcal{I}|\le s}\bigcap_{\mathbf{b}_0\in\mathcal{N}_\mathcal{I}} \mathcal{B}^c_{\mathcal{I},\mathbf{b}_0}$, the union bound gives $\Pr(\mathcal{E}_2^c) \le e^{-d/4}$.

On $\mathcal{E}_2$, suppose an admissible $\mathbf{z}$ satisfies $\|\mathbf{z}\|_0 > \theta d$ and $\mu \triangleq \sum_i \min\{1, |(\mathbf{J}\mathbf{z})_i|\} < 4k$. Set $\mathbf{x} = \mathbf{J}\mathbf{z}$, $\mathcal{I} = \{i : |x_i| \ge 1\}$, and $\mathbf{b} = \mathbf{x}_\mathcal{I}$ extended by zero outside $\mathcal{I}$. Then $|\mathcal{I}| \le \mu < 4k$, so $|\mathcal{I}| \le s$, and

$$\|\mathbf{x}-\mathbf{b}\|_2^2 = \sum_{i\notin\mathcal{I}} x_i^2 \le \sum_{i\notin\mathcal{I}} |x_i| \le \mu < 4k, \qquad \|\mathbf{b}\|_2 \le 2M\sqrt{d}.$$

Choose $\mathbf{b}_0 \in \mathcal{N}_\mathcal{I}$ with $\|\mathbf{b}-\mathbf{b}_0\|_2 \le a_0$. The parameter bounds

$$\chi \le \frac{\varepsilon^2(1-S)^2\theta}{2^{20}}, \qquad k \le \chi d/2, \qquad r_0 \ge \theta d/4$$

imply $k \le \varepsilon^2(1-S)^2 r_0/16$ and $4k \le \theta d/2$. Therefore

$$\|\mathbf{z}-\mathbf{J}^{-1}\mathbf{b}_0\|_2 \le \frac{2\sqrt{k}+a_0}{(1-S)} \le \varepsilon\sqrt{r_0}, \qquad \|\mathbf{z}_{\mathcal{I}^c}\|_2^2 \ge \|\mathbf{z}\|_0 - |\mathcal{I}| > \theta d - 4k \ge r_0.$$

This is the excluded event $\mathcal{B}_{\mathcal{I},\mathbf{b}_0}$, a contradiction. □

F.2.2 Conditional Hamming-distance expectation for a Haar eigenbasis

Combining the preceding estimate with Proposition 3 gives the larger of the two admissible thresholds. Set

$$\rho \triangleq \frac{1}{2}\max\left\{\frac{1-S^2}{144M^2}\min\{u^2, \delta/2\},\ \frac{1}{2^{12}(1+2^{16}/((1-S)\sqrt{\theta}))^2\log(8M)}\right\}. \tag{F.5}$$

Lemma 52. *Under the Haar model and constants above, fix $\mathbf{w}^\star \in \mathcal{W}^\circ$ independently of $\mathbf{U}$. If $d \ge 4/\theta$, then, with probability at least $1-e^{-qd/4}$ over $\mathbf{U}$,*

$$\bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}} \quad \textit{for every } \mathbf{w}\in\mathcal{W}\setminus\{\mathbf{w}^\star\} \textit{ and every real } 1\le k\le 2\rho d.$$

Moreover, $\mathcal{M} = \{\mathbf{w}^\star\}$ almost surely.

Proof. We verify the conditional-mean hypothesis of Lemma 16 at the largest permitted threshold $k = 2\rho d$. When the first term in Eq. (F.5) is maximal, Proposition 3 suffices. When the second term is larger, we use Lemma 15(i) for displacements with at most θd nonzero coordinates and Lemma 51 for the remaining displacements. The spectrum is positive semidefinite and nonzero, so Lemma 25 gives $\mathcal{M} = \{\mathbf{w}^\star\}$ almost surely, including when $a = 0$. On this probability-one event, since $\mathbf{w}^\star$ is interior, coordinatewise stochastic rounding gives

$$h(\mathbf{w}) = \|\mathbf{w}-\mathbf{w}^\star\|_0, \qquad \bar{h}(\mathbf{w}) = \sum_i \min\{1, |(\mathbf{J}(\mathbf{w}-\mathbf{w}^\star))_i|/\Delta\} \quad (\mathbf{w}\in\mathcal{W}).$$

Let ξ_1, ξ_2 denote, respectively, the two expressions in the maximum defining ρ above, so $2\rho = \max\{\xi_1, \xi_2\}$. The threshold range is empty if $2\rho d < 1$. Otherwise it suffices by monotonicity to take $k = 2\rho d$. The dimension hypothesis gives

$$d \ge \frac{4}{\theta} \ge \frac{2}{\delta}, \qquad \frac{4}{\theta} \ge \frac{128e}{q^2} \ge \frac{4\log 9}{q}.$$

Define

$$\mathcal{E}_1 \triangleq \left\{ \|\mathbf{K}\mathbf{z}\|_1 \geq u\sqrt{d}\,\|\mathbf{z}\|_2 \text{ for every } \mathbf{z} \in \mathbb{R}^d \text{ with } 1 \leq \|\mathbf{z}\|_0 \leq \delta d \right\},$$

and

$$\mathcal{E}_2(k) \triangleq \left\{ \sum_i \min\{1, |(\mathbf{J}\mathbf{z})_i|\} \geq 4k \text{ for every } \mathbf{z} \in \mathbb{Z}^d \text{ with } \|\mathbf{z}\|_\infty \leq 2M,\ \|\mathbf{z}\|_0 > \theta d \right\}.$$

On $\mathcal{E}_1$, use $m = u\sqrt{d}$ and $r = \lfloor \delta d \rfloor \geq \delta d/2$. If $\xi_1 \geq \xi_2$, then

$$\frac{1-S^2}{144M^2}\min\{m^2, r\} \geq \frac{1-S^2}{144M^2}\min\{u^2, \delta/2\}d = \xi_1 d = k.$$

Hence Proposition 3 gives $\bar{h}(\mathbf{w}) \geq 4\sqrt{k\min\{h(\mathbf{w}), k\}}$ for every $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$, with probability at least $1 - e^{-qd/2}$.

If $\xi_2 > \xi_1$, then $k = \chi d/\log(8M)$ by Eq. (F.4). By Lemmas 30 and 51,

$$\Pr\left((\mathcal{E}_1 \cap \mathcal{E}_2(k))^c\right) \leq e^{-qd/2} + e^{-d/4} \leq 2e^{-qd/2} \leq e^{-qd/4}.$$

For $1 \leq h \triangleq h(\mathbf{w}) \leq \theta d$, apply Lemma 15(i) with $m = u\sqrt{d}$ and $r = \lfloor \delta d \rfloor$. The event $\mathcal{E}_1$ verifies its spreading assumption, and $h \leq \theta d \leq \min\{r, m^2/9\}$. Thus

$$\bar{h}(\mathbf{w}) \geq \frac{u}{3}\sqrt{dh} \geq 4\sqrt{kh} \geq 4\sqrt{k\min\{h, k\}}.$$

Indeed, $\theta \leq \min\{\delta/2, u^2/9\}$ and

$$\chi \leq \frac{\varepsilon^2(1-S)^2\theta}{2^{20}}, \qquad \log(8M) > 2 \implies k \leq u^2 d/144.$$

For $h(\mathbf{w}) > \theta d$, apply $\mathcal{E}_2(k)$ to the integer vector $(\mathbf{w} - \mathbf{w}^\star)/\Delta$, whose coordinates have magnitude at most $2M$. This gives

$$\bar{h}(\mathbf{w}) = \sum_i \min\{1, |(\mathbf{J}(\mathbf{w} - \mathbf{w}^\star))_i|/\Delta\} \geq 4k = 4\sqrt{k\min\{h(\mathbf{w}), k\}},$$

since $k \leq \theta d$. Which term attains the maximum in Eq. (F.5) depends only on the fixed parameters, so no additional union bound is needed. Monotonicity gives every smaller real threshold on the same event. □

F.2.3 Proof of part (i) of Proposition 48

Use Eqs. (F.3) and (F.5). The threshold $k = \rho d$ determines the exponential time horizon e^k. Set

$$c \triangleq \min\{q/8, \sqrt{\rho}/32\}, \qquad d_0 \triangleq \left\lceil \max\left\{ e^8, \frac{4}{\theta}, \frac{1}{\rho}, \left(\frac{\log 2}{c}\right)^2 \right\} \right\rceil. \tag{F.6}$$

For fixed S, a, b, v, the second term in Eq. (F.5) yields

$$\rho^{-1} = O(\log(8M)), \qquad c^{-2} = O(\log(8M)), \qquad d_0 = O(\log(8M)).$$

These bounds hold uniformly for $M \geq 1$; they are $O(\log M)$ as $M \to \infty$.

Proof. As in the proof sketch of Theorem 1, we first lower-bound the conditional expectation of the next Hamming distance uniformly over the current iterate. Here Lemma 52 supplies this bound at the improved threshold $k = \rho d$. Conditional on this matrix event, Lemma 16 gives the recurrence and excludes hitting up to time e^k. Finally, $\mathbf{A} \succeq a\mathbf{I}_d$ converts the Hamming-distance lower bound into the claimed objective lower bound. Fix $d \geq d_0 \geq 4/\theta$ and set $k = \rho d \geq 1$. Let $\mathcal{B}_k$ be the event that, for every $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$,

$$\bar{h}(\mathbf{w}) \geq 4\sqrt{k\min\{h(\mathbf{w}), k\}}.$$

By Lemma 52, $\Pr(\mathcal{B}_k^c) \le e^{-qd/4}$. Define

$$\mathcal{E}_k \triangleq \bigcap_{\substack{t\in\mathbb{Z}_{\ge 0}\\ t<e^k}} \{h^{(t+1)} \ge \min\{k, 2\sqrt{kh^{(t)}}\}\}.$$

Apply Lemma 16 conditionally on $\mathbf{U}$. It gives $\Pr(\mathcal{E}_k^c \mid \mathbf{U}) \le e^{-\sqrt{k}/16}$ on $\mathcal{B}_k$. Hence

$$\Pr(\mathcal{E}_k^c) \le e^{-qd/4} + e^{-\sqrt{k}/16} \le 2e^{-2c\sqrt{d}} \le e^{-c\sqrt{d}},$$

since $\sqrt{d} \ge \log 2/c$. On $\mathcal{E}_k$, $\tau_{\mathcal{M}} > e^k$. Moreover, $\mathbf{A} \succeq a\mathbf{I}_d$ and Eq. (A.26) give

$$\begin{aligned} f(\mathbf{w}^{(t)}) &= \frac{1}{2}(\mathbf{w}^{(t)} - \mathbf{w}^\star)^\top \mathbf{A}(\mathbf{w}^{(t)} - \mathbf{w}^\star) \\ &\ge \frac{a}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|_2^2 \ge \frac{a}{2}\Delta^2 h^{(t)} \ge \frac{a}{2}\Delta^2 k, \end{aligned}$$

for every integer t satisfying $\log_2(1 + \frac{1}{2}\log_2 k) \le t \le e^k$. The middle inequality uses the fact that every nonzero coordinate of $\mathbf{w}^{(t)} - \mathbf{w}^\star$ has magnitude at least Δ. □

F.3 Gaussian designs

Throughout this subsection, use the Gaussian model of Theorem 4.

F.3.1 Variance of an underdetermined Gaussian spectrum

Lemma 53. *Let $d \ge 4$ and $1 \le n < d$. Write $\lambda_1, \dots, \lambda_d$ for the ordered eigenvalues of $\mathbf{A}$ and $\bar\lambda = d^{-1}\operatorname{tr}(\mathbf{A})$. Then, with probability at least $1 - 3e^{-d/1728}$,*

$$\frac{1}{d}\sum_{i=1}^{d}(\lambda_i - \bar\lambda)^2 \ge \frac{d}{12n} \ge \frac{1}{12}.$$

Proof. We lower-bound the spectral variance by the squared entries in two off-diagonal blocks of $\mathbf{A}$. Splitting the columns of $\mathbf{X}$ into two groups makes their total squared magnitude a Gaussian quadratic form conditional on the first group. A lower-tail bound for this quadratic form will give the required spectral variance. Write $\mathbf{X} = [\mathbf{Y}, \mathbf{Z}]$, where $\mathbf{Y}$ has $d_1 = \lfloor d/2 \rfloor \ge d/3$ columns and $\mathbf{Z}$ has $d_2 = d - d_1 \ge d/2$ columns. Put $\mathbf{B} = \mathbf{Y}\mathbf{Y}^\top$ and

$$\mathcal{E}_0 \triangleq \{\operatorname{tr}(\mathbf{B}) \ge nd_1/2, \quad \|\mathbf{Y}\|_2 \le \sqrt{n} + \sqrt{d_1} + \sqrt{d}\}.$$

The chi-square Laplace transform and Gaussian norm bound give

$$\Pr(\operatorname{tr}(\mathbf{B}) < nd_1/2) \le e^{-nd_1/16} \le e^{-nd/48}, \qquad \Pr(\|\mathbf{Y}\|_2 > \sqrt{n} + \sqrt{d_1} + \sqrt{d}) \le e^{-d/2}.$$

On $\mathcal{E}_0$, $\operatorname{tr}(\mathbf{B}) \ge nd/6 > 0$ and $\|\mathbf{B}\|_2 \le 9d$ because $n < d$. Conditional on $\mathbf{Y}$, write the columns of $\mathbf{Z}$ as $\mathbf{z}_j$ and set

$$Q = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}(\mathbf{Y}^\top\mathbf{Z})_{ij}^2 = \sum_{j=1}^{d_2}\mathbf{z}_j^\top\mathbf{B}\mathbf{z}_j, \qquad \mu = d_2\operatorname{tr}(\mathbf{B}).$$

For $s \ge 0$, Gaussian integration and $\log(1+x) \ge x - x^2/2$ give

$$\mathbb{E}[e^{-sQ} \mid \mathbf{Y}] = \det(\mathbf{I}_n + 2s\mathbf{B})^{-d_2/2} \le \exp\{-s\mu + s^2 d_2 \operatorname{tr}(\mathbf{B}^2)\}.$$

On $\mathcal{E}_0$, Markov's inequality with $s = \operatorname{tr}(\mathbf{B})/[4\operatorname{tr}(\mathbf{B}^2)]$ yields

$$\Pr(Q < \mu/2 \mid \mathbf{Y}) \le \exp\left\{-\frac{d_2\operatorname{tr}(\mathbf{B})^2}{16\operatorname{tr}(\mathbf{B}^2)}\right\} \le \exp\left\{-\frac{d_2\operatorname{tr}(\mathbf{B})}{16\|\mathbf{B}\|_2}\right\} \le e^{-nd/1728}.$$

Thus $\mathcal{E} = \mathcal{E}_0 \cap \{Q \ge \mu/2\}$ satisfies

$$\Pr(\mathcal{E}^c) \le e^{-nd/48} + e^{-d/2} + e^{-nd/1728} \le 3e^{-d/1728}.$$

On $\mathcal{E}$, $Q \ge nd^2/24$, and the two off-diagonal Gram blocks give

$$\frac{1}{d}\sum_i(\lambda_i - \bar\lambda)^2 = \frac{1}{d}\operatorname{tr}\big((\mathbf{A} - \bar\lambda\mathbf{I}_d)^2\big) \ge \frac{2Q}{dn^2} \ge \frac{d}{12n}.$$

□

F.3.2 Uniform conditional-mean bounds for Gaussian designs

We prove the conditional-mean lower bound in Lemma 55 uniformly over all grid displacements. We first bound the probability of failure for a fixed displacement in Lemma 54, then sum these bounds over displacements grouped by their Euclidean norms.

For the following two lemmas, assume $n \geq d$, define η_0, g, and choose a real k as follows:

$$\eta_0 \triangleq S/(1+2\sqrt{d/n})^2, \qquad g \triangleq \eta_0\sqrt{d/n}, \\ 1 \leq k \leq \frac{1}{2^{16}} \min\left\{g^2 d, \frac{d}{\log(8M)}\right\}. \tag{F.7}$$

Here $0 < g < 1$, $d \geq 2^{16}$, and $g^2 d = S^2 d^2/[n(1+2\sqrt{d/n})^4]$ is independent of M.

The actual step size $\eta = S/\lambda_{\max}(\mathbf{A})$ depends on the Gaussian design. We therefore bound the probability of failure simultaneously over all $\eta' \geq \eta_0$. For $\mathbf{z} \in \Delta\mathbb{Z}^d \setminus \{\mathbf{0}\}$, define

$$\mathcal{B}_{\mathbf{z}} \triangleq \left\{\exists \eta' \geq \eta_0 : \sum_i \min\{1, |((\mathbf{I}_d - \eta'\mathbf{A})\mathbf{z})_i|/\Delta\} < 4\sqrt{k \min\{\|\mathbf{z}\|_0, k\}}\right\}.$$

A fixed displacement.

Lemma 54. *Under Eq. (F.7), for a deterministic nonzero $\mathbf{z} \in \Delta\mathbb{Z}^d$, define*

$$r \triangleq \frac{\|\mathbf{z}\|_2}{\Delta\sqrt{d}}, \qquad u \triangleq 4\sqrt{k \min\{\|\mathbf{z}\|_0, k\}}, \qquad \ell \triangleq \lfloor 4k \rfloor, \qquad p \triangleq d - \ell - 1. \tag{F.8}$$

Then $p \geq 99d/100$ and

$$\Pr(\mathcal{B}_{\mathbf{z}}) \leq \binom{d}{\ell}\left(\frac{6u}{grp}\right)^p. \tag{F.9}$$

Proof of Lemma 54. We first show that $\mathcal{B}_{\mathbf{z}}$ forces a small ℓ_1 norm on some set of $p+1$ coordinates. On that set, projecting orthogonally to $\mathbf{z}$ removes its deterministic contribution for every $\eta' \geq \eta_0$ simultaneously. We can then apply the Gaussian density bound to a fixed projected ℓ_1 ball and sum over the choices of coordinates. Since $k \geq 1$ and $k \leq d/2^{16}$, $p = d - \ell - 1 \geq d - 5k \geq 99d/100$. On $\mathcal{B}_{\mathbf{z}}$, some $\eta' \geq \eta_0$ gives $\mathbf{u} = (\mathbf{I}_d - \eta'\mathbf{A})\mathbf{z}$ with at most ℓ coordinates satisfying $|u_i| \geq \Delta$, since $\sum_i \min\{1, |u_i|/\Delta\} < u \leq 4k$. Deleting these and additional coordinates leaves $|\mathcal{I}| = p+1$ with

$$\|\mathbf{u}_{\mathcal{I}}\|_1 < u\Delta. \tag{F.10}$$

Define

$$\mathcal{B}_{\mathcal{I}} \triangleq \{\exists \eta' \geq \eta_0 : \|((\mathbf{I}_d - \eta'\mathbf{A})\mathbf{z})_{\mathcal{I}}\|_1 < u\Delta\}.$$

Then $\mathcal{B}_{\mathbf{z}} \subseteq \bigcup_{|\mathcal{I}|=p+1} \mathcal{B}_{\mathcal{I}}$. For each $\mathcal{I}$, choose $\mathbf{P}_{\mathcal{I}}$ with p orthonormal rows in $\mathbf{z}_{\mathcal{I}}^{\perp}$ (any such rows if $\mathbf{z}_{\mathcal{I}} = \mathbf{0}$). Condition on $Z = \|\mathbf{X}\mathbf{z}\|_2^2/(n\|\mathbf{z}\|_2^2) \sim \chi_n^2/n$ in Lemma 24. With $\sigma = \|\mathbf{z}\|_2\sqrt{Z/n} > 0$ almost surely,

$$(\mathbf{A}\mathbf{z})_{\mathcal{I}} = Z\mathbf{z}_{\mathcal{I}} + \sigma\mathbf{g}_{\mathcal{I}}, \qquad \mathbf{P}_{\mathcal{I}}\mathbf{z}_{\mathcal{I}} = \mathbf{0}, \qquad \mathbf{P}_{\mathcal{I}}\mathbf{g}_{\mathcal{I}} \mid Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p).$$

Conditional on Z, the projection $\mathbf{P}_{\mathcal{I}}(\mathbf{A}\mathbf{z})_{\mathcal{I}}$ is Gaussian with covariance $\sigma^2\mathbf{I}_p$ and mean zero, so its density is at most $(2\pi\sigma^2)^{-p/2}$. Apply Lemma 27 to this conditional law, with $\mathbf{y} = (\mathbf{A}\mathbf{z})_{\mathcal{I}}$, $\mathbf{a} = \mathbf{z}_{\mathcal{I}}$, $\mathbf{P} = \mathbf{P}_{\mathcal{I}}$, and $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^{\mathcal{I}} : \|\mathbf{x}\|_1 \leq u\Delta\}$. In this application,

$$\mathbf{P}_{\mathcal{I}}((\mathbf{I}_d - \eta'\mathbf{A})\mathbf{z})_{\mathcal{I}} = -\eta'\mathbf{P}_{\mathcal{I}}(\mathbf{A}\mathbf{z})_{\mathcal{I}}, \qquad -\frac{1}{\eta'}\mathbf{P}_{\mathcal{I}}\mathcal{K} \subseteq \frac{1}{\eta_0}\mathbf{P}_{\mathcal{I}}\mathcal{K} \quad (\eta' \geq \eta_0).$$

The equality removes $\mathbf{z}_{\mathcal{I}}$, and the inclusion uses the convexity and symmetry of $\mathcal{K}$. Consequently, a single projected set bounds $\mathcal{B}_{\mathcal{I}}$, including the union over all $\eta' \geq \eta_0$. It remains to bound $\mathrm{vol}_p(\mathbf{P}_{\mathcal{I}}\mathcal{K})$. Let $\mathbf{v}$ be a unit vector spanning $\ker(\mathbf{P}_{\mathcal{I}})$. For any $\mathbf{x} \in \mathcal{K}$, varying $\mathbf{x} + t\mathbf{v}$ preserves its projection $\mathbf{P}_{\mathcal{I}}\mathbf{x}$. The function $t \mapsto \|\mathbf{x} + t\mathbf{v}\|_1$ is continuous and piecewise linear, and tends to infinity as $|t| \to \infty$. A minimum occurs at a point where at least one coordinate vanishes: choose

an endpoint of the interval of minimizers, where the slope must change. The minimizing vector still belongs to $\mathcal{K}$ and has the same projection as $\mathbf{x}$. Consequently,

$$\mathbf{P}_{\mathcal{I}}\mathcal{K} \subseteq \bigcup_{i\in\mathcal{I}} \mathbf{P}_{\mathcal{I}}\big(\mathcal{K} \cap \{x_i = 0\}\big).$$

Each coordinate section is a p-dimensional ℓ_1 ball of radius $u\Delta$, with volume $(2u\Delta)^p/p!$. Let $\mathbf{P}_{\mathcal{I},-i}$ be $\mathbf{P}_{\mathcal{I}}$ with the column indexed by i removed. The Cauchy–Binet identity gives

$$\sum_{i\in\mathcal{I}} \det(\mathbf{P}_{\mathcal{I},-i})^2 = \det(\mathbf{P}_{\mathcal{I}}\mathbf{P}_{\mathcal{I}}^\top) = 1.$$

Summing the projected volumes and applying Cauchy–Schwarz therefore gives

$$\begin{aligned}\mathrm{vol}_p(\mathbf{P}_{\mathcal{I}}\mathcal{K}) &\le \frac{(2u\Delta)^p}{p!}\sum_{i\in\mathcal{I}} |\det(\mathbf{P}_{\mathcal{I},-i})| \\ &\le \frac{(2u\Delta)^p\sqrt{p+1}}{p!}.\end{aligned} \tag{F.11}$$

Thus

$$\Pr(\mathcal{B}_{\mathcal{I}} \mid Z) \le (2\pi)^{-p/2}\frac{(2u\Delta)^p\sqrt{p+1}}{p!(\eta_0\sigma)^p}.$$

Since $\sigma = \|\mathbf{z}\|_2\sqrt{Z/n}$, averaging this conditional bound requires an upper bound on $\mathbb{E}[Z^{-p/2}]$. Since $n - p \ge d - p = \ell + 1 \ge 5$, chi-square integration gives, for integers $j \ge 1$ with $2j + 2 \le n$,

$$\mathbb{E}[Z^{-j}] = \prod_{a=1}^{j}\frac{n}{n-2a} \le \frac{(j+1)^j}{j!} \le e^{2j}.$$

Here each factor decreases with n, and $\log(j!) \ge \int_1^j \log x\,dx \ge j\log j - j$. Together with $\mathbb{E}[Z^0] = 1$, this gives $\mathbb{E}[Z^{-p/2}] \le e^p$ for even p. For $p = 2j + 1$, Cauchy–Schwarz gives

$$\mathbb{E}[Z^{-p/2}] \le \big(\mathbb{E}[Z^{-j}]\mathbb{E}[Z^{-(j+1)}]\big)^{1/2} \le e^p. \tag{F.12}$$

To absorb the factor $\sqrt{p+1}$, concavity of log on each interval $[j - 1/2, j + 1/2]$ gives

$$\begin{aligned}\log(p!) &\ge \int_{1/2}^{p+1/2}\log x\,dx \\ &= (p+1/2)\log(p+1/2) - p + \tfrac{1}{2}\log 2 \\ &\ge p\log p - p + \tfrac{1}{2}\log(p+1).\end{aligned}$$

Thus $p! \ge \sqrt{p+1}(p/e)^p$. Using this bound, $\mathbb{E}[Z^{-p/2}] \le e^p$, and $\eta_0\|\mathbf{z}\|_2/(\Delta\sqrt{n}) = gr$ gives

$$\Pr(\mathcal{B}_{\mathcal{I}}) \le \left(\frac{2e^2u}{\sqrt{2\pi}\,grp}\right)^p \le \left(\frac{6u}{grp}\right)^p.$$

Sum over the $\binom{d}{\ell}$ choices of $\mathcal{I}$. □

Uniformity over grid displacements. We now sum the fixed-displacement probability bounds to obtain the conditional-mean inequality simultaneously at every nonoptimal grid point.

Lemma 55. *Under Eq. (F.7), fix $\mathbf{w}^\star \in \mathcal{W}^\circ$ independently of $\mathbf{X}$. There is an event depending only on $\mathbf{X}$, with probability at least $1 - e^{-d/2} - 2e^{-d}$, on which $\lambda_{\max}(\mathbf{A}) \le (1 + 2\sqrt{d/n})^2$ and*

$$\bar{h}(\mathbf{w}) \ge 4\sqrt{k\min\{h(\mathbf{w}), k\}} \qquad (\mathbf{w} \in \mathcal{W}\setminus\{\mathbf{w}^\star\}). \tag{F.13}$$

Proof of Lemma 55. We first exclude $\mathcal{B}_{\mathbf{z}}$ simultaneously for every grid displacement, and then substitute the actual step size. For $\mathcal{D} = (\mathcal{W} - \mathbf{w}^\star)\setminus\{\mathbf{0}\}$, write $r(\mathbf{z}), u(\mathbf{z})$ for the radius and threshold in Eq. (F.8), and use ℓ, p from the same equation. For radii $r \ge 1/4$, the lattice count has an upper bound proportional to r^d, and using $u(\mathbf{z}) \le 4k$ in Lemma 54 gives a probability bound proportional

to r^{-p}. Their product therefore depends on the radius through $r^{d-p} = r^{\ell+1}$. Since $\ell + 1 \le 5k$ and every displacement has radius at most $2M$, this contributes at most $(2M)^{5k}$. The assumption $k \log(8M) \le d/2^{16}$ makes this contribution small enough for the union bound. The smaller radii are handled separately using Eq. (F.2). Define

$$\mathcal{B}(r_1, r_2) \triangleq \bigcup_{\mathbf{z}\in\mathcal{D}:\ r_1\le r(\mathbf{z})\le r_2} \mathcal{B}_{\mathbf{z}}.$$

Since $1 \le k \le d/2^{16}$, $4 \le \ell \le 4k$ and

$$p = d - \ell - 1 \ge d - 5k \ge \frac{99d}{100}, \qquad \binom{d}{\ell} \le e^{d/100}. \tag{F.14}$$

The second bound follows from $\binom{d}{\ell} \le (ed/\ell)^\ell$ and $x \log(e/x) < 1/100$ for $0 < x \le 4/2^{16}$. For each $\mathbf{z} \in \mathcal{D}$, write $r = r(\mathbf{z})$ and $u = u(\mathbf{z})$. Since $\|\mathbf{z}\|_0 \le r^2 d$, $u \le \min\{4\sqrt{k\|\mathbf{z}\|_0}, 4k\}$, and $p \ge d/2$,

$$\frac{6u}{grp} \le \frac{48\sqrt{k/d}}{g} \le \frac{3}{16}, \qquad \frac{6u}{grp} \le \frac{b}{r}, \quad b \triangleq \frac{48k}{gd} \le \frac{3}{4096}. \tag{F.15}$$

By Lemmas 50 and 54,

$$\Pr(\mathcal{B}(0, 1/4)) \le e^{d/3}\binom{d}{\ell}(3/16)^p \le \exp\left\{\frac{d}{3} + \frac{d}{100} - \frac{3}{2}\frac{99d}{100}\right\} \le e^{-d},$$

using $\log(3/16) < -3/2$. Set $r_j = 2^{j-2}$ and $J_M = \lfloor \log_2(8M) \rfloor$. Since $r(\mathbf{z}) \le 2M$, the intervals $[r_j, 2r_j]$, $0 \le j \le J_M$, cover the remaining radii. There are at most $8M$ such intervals, and $1/4 \le r_j \le 2M$. The lattice counts are bounded by

$$[\sqrt{2\pi e}(2r_j + 1/2)]^d \le (17 r_j)^d,$$

so

$$\Pr(\mathcal{B}(r_j, 2r_j)) \le \binom{d}{\ell} 17^d b^p r_j^{d-p} = \binom{d}{\ell} 17^d b^p r_j^{\ell+1}.$$

As $\ell + 1 \le 5k$, $r_j^{\ell+1} \le (2M)^{5k}$. The number of annuli and their radius factors satisfy

$$\log(8M) + 5k \log(2M) \le 6k \log(8M) \le \frac{6d}{2^{16}},$$

where the first inequality uses $k \ge 1$. For $\mathcal{B}_* \triangleq \bigcup_{j=0}^{J_M} \mathcal{B}(r_j, 2r_j)$,

$$\begin{aligned}\Pr(\mathcal{B}_*) &\le 8M\binom{d}{\ell} 17^d b^p (2M)^{5k} \\ &\le \exp\left\{\frac{d}{100} + 3d - 7\frac{99d}{100} + \frac{6d}{2^{16}}\right\} \le e^{-3d}.\end{aligned}$$

Here $k \log(8M) \le d/2^{16}$, $k \ge 1$, $\log 17 < 3$, and $\log(3/4096) < -7$. Thus $\Pr(\mathcal{B}(0, 1/4) \cup \mathcal{B}_*) \le e^{-d} + e^{-3d} \le 2e^{-d}$. By Lemma 22,

$$\Pr\Big(\lambda_{\max}(\mathbf{A}) > (1 + 2\sqrt{d/n})^2\Big) \le e^{-d/2}.$$

By Lemma 25, $\mathcal{M} = \{\mathbf{w}^\star\}$ almost surely. On the spectral event above, $\eta = S/\lambda_{\max}(\mathbf{A}) \ge \eta_0$. On its intersection with $(\mathcal{B}(0, 1/4) \cup \mathcal{B}_*)^c$, substituting this η into the inequalities defining $\mathcal{B}_{\mathbf{z}}^c$ gives, since $\mathbf{w}^\star$ is interior and rounding acts coordinatewise,

$$\begin{aligned}\bar{h}(\mathbf{w}) &= \sum_i \min\left\{1, \frac{|((\mathbf{I}_d - \eta\mathbf{A})(\mathbf{w} - \mathbf{w}^\star))_i|}{\Delta}\right\} \\ &\ge 4\sqrt{k \min\{h(\mathbf{w}), k\}} \quad (\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}).\end{aligned}$$

The intersection has probability at least $1 - e^{-d/2} - 2e^{-d}$. □

F.3.3 Simultaneous conditional-mean and quadratic-form bounds

We next construct the matrix event needed for the improved version of Theorem 4. On this event, the conditional mean satisfies the hypothesis of Lemma 16, and a restricted quadratic-form bound converts Hamming distance into objective error. The constants below specify the common probability bound and the largest admissible threshold in both ranges $n \ge d$ and $n < d$.

Constants. Fix $R > 0$ and use the Gaussian model above. For the restricted quadratic-form lower bound, set

$$a \triangleq \max\left\{e^{-1}[e(2M+1)]^{-2R},\ \frac{\max\{1-\sqrt{R},0\}^2}{4}\right\}. \tag{F.16}$$

For the range $n \ge d$, use

$$\begin{aligned}\rho_+ &\triangleq 2^{-16}\min\left\{\frac{S^2}{81}, \frac{1}{\log(8M)}\right\},\\ c_+ &\triangleq \begin{cases}1/16, & a = e^{-1}[e(2M+1)]^{-2R},\\ \min\{1/16, (1-\sqrt{R})^2/(16R)\}, & a > e^{-1}[e(2M+1)]^{-2R},\end{cases}\\ d_{0,+} &\triangleq \lceil \max\{4, \log 4/c_+, 1/\rho_+\}\rceil.\end{aligned} \tag{F.17}$$

The second case requires $R < 1$; the first includes $R > 1$.

For $R > 1$, recall the spectral upper bound, probability exponent, and support fraction from Eq. (B.22):

$$B_R = (2+\sqrt{R})^2, \qquad q_R = \min\{1/(16R), 1/32\}, \qquad \delta_R = \frac{q_R^2}{16(2+\log(2M))^2}.$$

Define

$$\delta_{1,R} \triangleq \frac{S}{2^{36} e B_R^8 (S + 2^{11} B_R)}, \qquad \theta_R \triangleq \min\left\{\frac{\delta_{1,R}}{2}, \frac{S^2}{9\cdot 2^{18} B_R^2}\right\}. \tag{F.18}$$

Set

$$\rho_- \triangleq \max\left\{\begin{array}{l}\dfrac{1-S^2}{144M^2}\min\left\{\dfrac{S^2}{512B_R^2}, \dfrac{\delta_R}{2}\right\},\\[2ex] \dfrac{1-S^2}{144M^2}\min\left\{\dfrac{S^2}{2^{18}B_R^2}, \dfrac{\delta_{1,R}}{2}\right\},\\[2ex] \dfrac{1}{2^{12}\left(1+2^{16}/((1-S)\sqrt{\theta_R})\right)^2\log(8M)}\end{array}\right\}. \tag{F.19}$$

The first term follows from Gaussian spreading in Lemma 32 and Proposition 3. The second and third terms come from the sparse-displacement and lattice bounds in Lemma 52, respectively, with spectral interval $[0, B_R]$ and variance parameter $1/4$. If the first term attains the maximum, including a tie, use the first case below.

$$(c_-, d_{0,-}) \triangleq \begin{cases}(3q_R/8,\ \lceil\max\{2/\delta_R, 1/\rho_+, 1/\rho_-\}\rceil), & \text{first term attains the maximum,}\\ (1/(2^{19}B_R^4),\ \lceil\max\{4/\theta_R, 1/\rho_+, 1/\rho_-\}\rceil), & \text{otherwise.}\end{cases} \tag{F.20}$$

For $n < d$, $a = e^{-1}[e(2M+1)]^{-2R}$.

For $d/R \le n \le \rho d^2$, take

$$(\rho, c, d_0) \triangleq \begin{cases}(\rho_+, c_+, d_{0,+}), & 0 < R \le 1,\\ (\min\{\rho_+, \rho_-\}, c_-, d_{0,-}), & R > 1,\end{cases} \tag{F.21}$$

with a from Eq. (F.16). For $R > 1$, $c_- \le c_+ = 1/16$ and $d_{0,-} \ge d_{0,+}$, so this also covers $n \ge d$. Here $d_0 \ge 1/\rho$; hence $d \ge d_0$, $n \le \rho d^2$ imply $\rho \min\{d, d^2/n\} \ge 1$. The constants depend only on M, S, R. For fixed S, R, they satisfy $\rho^{-1}, d_0 = O(\log(8M))$, and c is bounded away from zero uniformly in M. The curvature coefficient a has the separate dependence given in Eq. (F.16).

Lemma 56. *Fix $\mathbf{w}^\star \in \mathcal{W}^\circ$ independently of $\mathbf{X}$ and use the constants specified above. If $d \ge d_0$ and $d/R \le n \le \rho d^2$, then, with probability at least $1 - e^{-cd}$ over $\mathbf{X}$,*

$$\bar{h}(\mathbf{w}) \ge 4\sqrt{k \min\{h(\mathbf{w}), k\}}, \qquad \lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \ge a. \tag{F.22}$$

The bounds hold simultaneously for every $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}$ and every real k with $1 \le k \le \rho \min\{d, d^2/n\}$.

Proof of Lemma 56. We construct a single matrix event on which both bounds hold. First, we lower-bound the quadratic form either on grid displacements or on all vectors, according to the value of a. For the conditional mean, the preceding Gaussian estimate applies when $n \ge d$. When $n < d$, we use Gaussian spreading or condition on the spectrum and apply the Haar estimate, according to which term defines ρ_-. Define

$$\mathcal{E}_0 = \{\lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \ge a\}, \qquad \mathcal{E}(k) = \{\bar{h}(\mathbf{w}) \ge 4\sqrt{k \min\{h(\mathbf{w}), k\}}\ \forall \mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^\star\}\}.$$

If $a = e^{-1}[e(2M+1)]^{-2R}$, Lemma 23 gives $\Pr(\mathcal{E}_0^c) \le e^{-d} \le e^{-2c_+ d}$. Otherwise $R < 1$, and Lemma 22 with $u = (1 - \sqrt{R})\sqrt{n}/2$ gives

$$\Pr(\mathcal{E}_0^c) \le e^{-(1-\sqrt{R})^2 n/8} \le e^{-(1-\sqrt{R})^2 d/(8R)} \le e^{-2c_+ d}.$$

The stronger e^{-d} bound holds when $n < d$.

For $n \ge d$, take $k = \rho d^2/n \ge 1$. Set $\eta_0 = S/(1 + 2\sqrt{d/n})^2$ and $g = \eta_0\sqrt{d/n}$ as in Eq. (F.7). Then

$$g^2 d = \frac{S^2 d^2}{n(1 + 2\sqrt{d/n})^4} \ge \frac{S^2 d^2}{81 n}, \qquad k \le \rho_+ d^2/n \le 2^{-16} \min\{g^2 d, d/\log(8M)\}.$$

Thus Lemma 55, applied with this k, gives $\Pr(\mathcal{E}(k)^c) \le e^{-d/2} + 2e^{-d}$. Since $c \le c_+ \le 1/16$ and $d \ge d_{0,+} \ge \log 4/c_+$,

$$\Pr((\mathcal{E}_0 \cap \mathcal{E}(k))^c) \le e^{-d/2} + 2e^{-d} + e^{-2c_+ d} \le 4e^{-2c_+ d} \le e^{-c_+ d} \le e^{-cd}.$$

For $n < d$, $R > 1$, $k = \rho d$, and $c = c_-$. If $\rho_- = (1 - S^2)\min\{S^2/(512 B_R^2), \delta_R/2\}/(144 M^2)$, set $r = \lfloor \delta_R d \rfloor$ and $m = Sd/(16\sqrt{2} B_R \sqrt{n})$. Then Lemma 32 gives

$$r \ge \delta_R d/2, \qquad m^2 = \frac{S^2 d^2}{512 B_R^2 n} \ge \frac{S^2 d}{512 B_R^2}, \qquad \frac{1 - S^2}{144 M^2} \min\{m^2, r\} \ge \rho_- d \ge k.$$

Proposition 3 therefore gives $\mathcal{E}(k)$. Using $q_R \le 1/(16R)$ and $n \ge d/R$,

$$\Pr((\mathcal{E}_0 \cap \mathcal{E}(k))^c) \le 2e^{-3q_R d/4} + e^{-n/2} + e^{-d} \le 4e^{-2c_- d} \le e^{-c_- d},$$

because

$$d \ge \frac{2}{\delta_R} = \frac{32(2 + \log(2M))^2}{q_R^2} \ge \frac{8 \log 4}{3 q_R} = \frac{\log 4}{c_-}.$$

Otherwise, $\mathbf{X}\mathbf{Q} \overset{\mathrm{d}}{=} \mathbf{X}$ for every deterministic $\mathbf{Q} \in \mathcal{O}(d)$, so the law of $\mathbf{A}$ is invariant under orthogonal conjugation. It therefore admits the representation in distribution $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda}$ has the ordered eigenvalue distribution of $\mathbf{A}$ and $\mathbf{U}$ is independent of $\mathbf{\Lambda}$ and Haar distributed (Meckes, 2019). This representation includes repeated and zero eigenvalues. Write $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ and $\bar{\lambda} = d^{-1}\sum_{i=1}^d \lambda_i$. Define

$$\mathcal{E}_1 = \left\{0 \le \lambda_i \le B_R\ (i \in [d]), \quad d^{-1}\textstyle\sum_i (\lambda_i - \bar{\lambda})^2 \ge 1/12\right\}.$$

Lemmas 22 and 53 give $\Pr(\mathcal{E}_1^c) \le e^{-n/2} + 3e^{-d/1728}$. For each spectrum in $\mathcal{E}_1$, apply Lemma 52 with spectral interval $[0, B_R]$ and variance parameter $1/4$. Its constants are

$$q = \frac{1}{2^{16} B_R^4}, \qquad u = \frac{S}{512 B_R}, \qquad \delta = \delta_{1,R}, \qquad \theta = \theta_R.$$

With these constants, the coefficient of d in the lemma's largest threshold equals the maximum of the last two terms in Eq. (F.19). Since the first term is not maximal in the present case, this coefficient is ρ_-. For $d \geq 4/\theta_R$ and every $1 \leq k \leq \rho_- d$, the lemma gives $\Pr(\mathcal{E}(k)^c \mid \mathbf{\Lambda}) \leq e^{-qd/4}$ on $\mathcal{E}_1$. Conditioning on the spectrum gives $\Pr(\mathcal{E}(k)^c \cap \mathcal{E}_1) \leq e^{-qd/4}$. Since $q/4 \leq \min\{1/(2R), 1/1728\}$ and $c_- = q/8$,

$$\Pr((\mathcal{E}_0 \cap \mathcal{E}(k))^c) \leq e^{-qd/4} + e^{-n/2} + 3e^{-d/1728} + e^{-d} \leq 6e^{-2c_- d} \leq e^{-c_- d}.$$

Indeed, $\theta_R \leq q^2/(32e)$ gives

$$d \geq 4/\theta_R \geq 128e/q^2 \geq 8 \log 6/q = \log 6/c_-.$$

The maximizing term in Eq. (F.19) is deterministic. Since $k \mapsto 4\sqrt{k \min\{h(\mathbf{w}), k\}}$ is nondecreasing, $\mathcal{E}(\rho \min\{d, d^2/n\}) \subseteq \mathcal{E}(k)$ for $1 \leq k \leq \rho \min\{d, d^2/n\}$. □

F.3.4 Proof of part (ii) of Proposition 48

We prove the Hamming-distance recurrence and the hitting-time lower bound by applying Lemma 16 on the matrix event from Lemma 56. The same event gives the restricted quadratic-form bound that converts the recurrence into the claimed objective lower bound. Thus the argument follows the proof sketch of Theorem 4, with the improved constants recalled below.

Constants. Use the constants from Appendix F.3.3. In particular,

$$a = \max\left\{ e^{-1}[e(2M+1)]^{-2R}, \ \frac{\max\{1-\sqrt{R}, 0\}^2}{4} \right\},$$

$$(\rho, c, d_0) = \begin{cases} (\rho_+, c_+, d_{0,+}), & 0 < R \leq 1, \\ (\min\{\rho_+, \rho_-\}, c_-, d_{0,-}), & R > 1. \end{cases}$$

Here $(\rho_+, c_+, d_{0,+})$ are given in Eq. (F.17), and $(\rho_-, c_-, d_{0,-})$ are given in Eqs. (F.19) and (F.20). For fixed S, R, these constants satisfy $\rho^{-1}, d_0 = O(\log(8M))$, and c is bounded away from zero uniformly in M. Indeed, $\rho_+^{-1} = O(\log(8M))$ directly, and the last term in Eq. (F.19) gives $\rho_-^{-1} = O(\log(8M))$ if $R > 1$. In the first case of Eq. (F.20),

$$\rho_- \leq \frac{(1-S^2)\delta_R}{288M^2}, \qquad \frac{2}{\delta_R} \leq \frac{1}{\rho_-}.$$

In the other case, θ_R is independent of M. Thus the formulas for $d_{0,+}$ and $d_{0,-}$ give the asserted bound on d_0. Every possible value of c is a positive constant depending only on the fixed parameters. The curvature coefficient a follows the separate formula recalled above.

Proof. For $d \geq d_0$ and $d/R \leq n \leq \rho d^2$, set $k = \rho \min\{d, d^2/n\} \geq 1$ and define

$$\mathcal{E}_A(a, k) \triangleq \left\{ \begin{matrix} \lambda_{\min}(\mathbf{A}; \mathcal{W} - \mathbf{w}^\star) \geq a, \\ \bar{h}(\mathbf{w}) \geq 4\sqrt{k \min\{h(\mathbf{w}), k\}} \quad (\mathbf{w} \in \mathcal{W}, \ 1 \leq h(\mathbf{w}) \leq d) \end{matrix} \right\}.$$

Lemma 56 gives $\Pr(\mathcal{E}_A(a, k)) \geq 1 - e^{-cd}$. Conditional on $\mathbf{A}$ in this event, Lemma 16 gives the recurrence below with probability at least $1 - \frac{99}{100}e^{-\sqrt{k}/16}$. Since $c \leq 1/16$, $\rho \leq \rho_+ < 1/128$, and $d_0 \geq 1/\rho$, the probability comparison proved in Appendix B.4.2 gives

$$(1 - e^{-cd})(1 - \tfrac{99}{100}e^{-\sqrt{k}/16}) \geq 1 - e^{-c\sqrt{k}}.$$

Thus, with joint probability at least $1 - e^{-c\sqrt{k}}$,

$$h^{(t+1)} \geq \min\{k, 2\sqrt{k h^{(t)}}\} \qquad (0 \leq t < e^k, \ t \in \mathbb{Z}),$$

and $\tau_{\mathcal{M}} > e^k$. The restricted quadratic-form bound and Eq. (A.26) also give

$$\begin{aligned} f(\mathbf{w}^{(t)}) &= \frac{1}{2}(\mathbf{w}^{(t)} - \mathbf{w}^\star)^\top \mathbf{A}(\mathbf{w}^{(t)} - \mathbf{w}^\star) \\ &\geq \frac{a}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|_2^2 \geq \frac{a}{2}\Delta^2 h^{(t)} \geq \frac{a}{2}\Delta^2 k, \end{aligned}$$

for every integer t satisfying $\log_2(1 + \frac{1}{2}\log_2 k) \leq t \leq e^k$. The middle inequality uses the fact that every nonzero coordinate of $\mathbf{w}^{(t)} - \mathbf{w}^\star$ has magnitude at least Δ. □