LOGARITHMIC ACTIVATION CLOCKS IN DIAGONAL LINEAR NETWORKS: WHEN SGD DOES AND DOES NOT FOLLOW LASSO

Anonymous authors
Paper under double-blind review

ABSTRACT

Under a path-monotonicity condition, the from-start average of gradient flow in a diagonal linear network, rather than its raw output, recovers the Lasso regularization path. We ask what survives for stochastic gradient flow (SGF) and finite-step stochastic gradient descent (SGD). Integrating the square-factor log dynamics yields the Lasso KKT system plus coordinatewise Itô variance clocks and a self-normalized martingale. In an exactly solvable least-squares model, fixed SGF noise and a fixed SGD step produce different weighted-Lasso paths; ordinary Lasso therefore need not survive in either regime. In weak noise, only clock mass accumulated before future support motion enters the comparison. Under the stated nonexplosion, localization, and path-variation conditions, the averaged path returns to Lasso without requiring $\delta_\alpha \log(1/\alpha) \to 0$. For the standard signed product parametrization, an exact rotation produces two nonnegative channels driven by the same minibatch, with distinct finite-step clocks $R(\eta G)$ and $R(-\eta G)$. When the positive and negative Lasso channels have zero downward variation, the stopped, averaged actual-SGD path converges in prediction norm and signed-Lasso objective as $\alpha \downarrow 0$ and $\eta_\alpha \to 0$, again without an $\eta_\alpha \log(1/\alpha)$ condition. For correlated row sampling, our positive results retain explicit tail-control assumptions; a universal off-face excursion bound remains open.

1 INTRODUCTION

Early stopping is often described as implicit regularization, but pathwise versions of this principle are rare. Even for least squares, an optimization trajectory and an explicitly regularized solution path are generally different objects. A particularly clean benchmark is provided by Ali et al. (2020): in the direct linear parametrization, their moment-matched stochastic gradient flow (SGF) admits quantitative, whole-path comparisons with ridge regression at regularization level $1/t$. This raises a natural sparse analogue. If the predictor is instead represented by a two-layer diagonal linear network, $\beta = a \odot b$, and the factors are trained without an explicit penalty, does stochastic optimization trace the Lasso path?

The deterministic theory makes this question both compelling and subtle. The product parametrization turns Euclidean factor dynamics into a sparsity-favoring geometry; at convergence, small initialization is closely connected to ℓ_1 selection (Chou et al., 2023; Pesme et al., 2021). More recently, Berthier (2025) proved a finite-time Lasso-path correspondence. Under the path-monotonicity condition in that work, the relevant estimator is not the raw network output. If $L_\alpha = \log(1/\alpha)$ and $T_\alpha(s) = sL_\alpha/2$, it is the from-start average

$$\bar{\beta}^\alpha(s) = \frac{1}{T_\alpha(s)} \int_0^{T_\alpha(s)} \beta_t \, dt$$

that approaches Lasso at penalty $1/s$ under the stated small-initialization scaling. The distinction is substantive: the raw infinitesimal-initialization trajectory can instead exhibit saddle-to-saddle, incremental dynamics (Pesme & Flammarion, 2023; Zhao et al., 2026). Thus the stochastic problem is not whether noise perturbs a pointwise identity. It is whether noise preserves an *averaged* variational path on a logarithmically long time horizon.

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.

We show that the answer is controlled by coordinatewise *logarithmic activation clocks*. On the positive square slice, taking logarithms of the multiplicative dynamics produces the Lasso KKT system plus a cumulative variance clock and a self-normalized martingale. Fixed-noise clocks deform the path; when the clocks relevant to future support motion vanish, ordinary Lasso re-emerges under the stated localization and path-variation conditions. Finite-step SGD instead carries the exact logarithmic remainder of its multiplicative update.

Contributions. Our main results are:

1. **Fixed-noise separation.** In an exactly solvable axis sampler, averaged SGF and actual factor SGD converge to differently weighted positive- Lasso paths. Unequal weights rule out a single scalar time calibration and show that factor-space moment matching does not determine the path.
2. **Future-active weak-noise recovery.** A perturbed KKT comparison charges only clock mass created before future upward comparator motion. It yields uniform diagonal recovery under nonexplosion and a horizon-wide variance bound, and localized correlated-design results, without requiring $\delta_\alpha \log(1/\alpha) \to 0$.
3. **Finite-horizon signed-SGD comparison.** Building on the exact per-step logarithmic corrections identified by Even et al. (2023), the rotation $u^{\pm} = (a \pm b)/2$ produces two nonnegative channels with exact corrections $R(\eta G)$ and $R(-\eta G)$, where $R(q) = -\log(1-q) - q$, driven by one shared minibatch. Our contribution is to insert these corrections into the logarithmic-horizon, from-start averaged Lasso KKT system. Under stopped localization and zero channel deletion, the averaged actual-SGD path converges in prediction norm and signed-Lasso objective as $\alpha \downarrow 0$ and $\eta_\alpha \to 0$, with no $\eta_\alpha \log(1/\alpha) \to 0$ condition.

Scope and unresolved global step. Our correlated-design results are conditional. For SGF, recovery follows under an explicit tail-cover or localization condition; for actual SGD, it follows up to a stopping time whose failure probability vanishes. Signed paths that delete or reverse coordinates retain an explicit downward-variation term. Establishing a uniform off-face excursion bound for general correlated row sampling remains open.

2 Related work

Early stopping and explicit regularization paths. For direct least squares, Ali et al. (2019) compare gradient flow with ridge along $\lambda = 1/t$, and Ali et al. (2020) obtain whole-training risk and coefficient bounds for a moment-matched SGF. Matching the first two moments of minibatch SGD and forward-Euler SGF is not a pathwise coupling of actual SGD with continuous SGF. Moreover, the affine direct-parameter recursion and deliberately weighted averages used in ridge comparisons do not carry over directly through the nonlinear Hadamard map. Extended comparisons appear in Appendix A.1.

Deterministic sparse bias from Hadamard parametrization. Explicit quadratic factor penalties induce an exact ℓ_1 predictor penalty (Hoff, 2017; Tibshirani, 2021), while unpenalized diagonal dynamics have sparse-recovery or sparse-endpoint guarantees under additional conditions (Chou et al., 2023; Vaškevičius et al., 2019; Zhao et al., 2022; Li et al., 2021). These are not full Lasso-path theorems. At the trajectory level, the raw vanishing-initialization predictor can converge to a saddle-to-saddle, LARS-like jump process (Pesme & Flammarion, 2023; Zhao et al., 2026); in contrast, Berthier (2025) identifies the deterministic *from-start average* with the Lasso path under monotonicity, retaining a path-variation defect more generally. We extend this averaged object, not the raw path.

Stochasticity and finite steps in diagonal networks. Pesme et al. (2021) characterize the SGF endpoint through an evolving hypentropy geometry. Their trajectory-dependent effective initialization is the endpoint precursor of our Itô variance clock, not a finite-time averaged- path theorem; deliberately injected label noise can likewise select one fixed weighted-Lasso endpoint (Pillaud-Vivien et al., 2022). Most closely, Even et al. (2023) derive $q_+(x) = 2R(x)$ and $q_-(x) = 2R(-x)$ for $|x| < 1$ in their endpoint analysis. We use these exact remainders in a uniform finite-horizon KKT

comparison for the from-start average. We found no such theorem for ordinary minibatching; a detailed claim-by-claim comparison appears in Appendix A.1.

3 SETUP AND THE LOGARITHMIC KKT IDENTITY

Quadratic loss and stochastic gradients. We study the quadratic objective

$$\mathcal{L}(\beta) = \frac{1}{2}\beta^\top H\beta - r^\top\beta, \qquad H \in \mathbb{R}^{d\times d}, \quad H \succ 0, \tag{3.1}$$

up to an irrelevant additive constant. For every positive-definite matrix Q, write $\|v\|_Q = (v^\top Qv)^{1/2}$. A sampled gradient $G(\beta,\zeta)$ is unbiased and has conditional covariance

$$\mathbb{E}[G(\beta,\zeta)\mid\beta] = g(\beta) = H\beta - r, \qquad \Sigma(\beta) = \mathrm{Cov}(G(\beta,\zeta)\mid\beta), \qquad V_i(\beta) = \Sigma_{ii}(\beta). \tag{3.2}$$

For empirical least squares, $G(\beta,\zeta) = x_\zeta(x_\zeta^\top\beta - y_\zeta)$, and Σ is an affine-quadratic function of β.

Following the moment-matched continuous-time convention of Ali et al. (2020), we first work on the balanced positive slice of the Hadamard model. The factor-space stochastic gradient flow (SGF) is

$$\begin{aligned} du_t &= -v_t\odot g(\beta_t)\,dt + \sqrt{\delta}\ \mathrm{Diag}(v_t)\Sigma(\beta_t)^{1/2}dB_t,\\ dv_t &= -u_t\odot g(\beta_t)\,dt + \sqrt{\delta}\ \mathrm{Diag}(u_t)\Sigma(\beta_t)^{1/2}dB_t, \qquad \beta_t = u_t\odot v_t. \end{aligned} \tag{3.3}$$

The same Brownian motion appears in both equations because both factors see the same sampled gradient. The effective noise δ includes the batch-size scaling and may depend on α; we write δ_α in weak-noise statements. If $u_0 = v_0 > 0$, balance is invariant and $\beta = u^{\odot 2} > 0$, so coordinatewise logarithms are well defined. The logarithm is natural because the dynamics are multiplicative: a coordinate initialized at scale α must traverse roughly $L_\alpha = \log(1/\alpha)$ log-units before becoming order one. Itô's formula gives

$$d\beta_t = \{-2\,\mathrm{Diag}(\beta_t)g(\beta_t) + \delta\,\mathrm{Diag}(\beta_t)V(\beta_t)\}\,dt + 2\sqrt{\delta}\,\mathrm{Diag}(\beta_t)\Sigma(\beta_t)^{1/2}dB_t, \tag{3.4}$$

$$d\log\beta_{t,i} = -2g_i(\beta_t)\,dt - \delta V_i(\beta_t)\,dt + 2\sqrt{\delta}\,dM_{t,i}, \qquad d\langle M_i\rangle_t = V_i(\beta_t)\,dt. \tag{3.5}$$

Why the Itô correction appears. For a diffusion the ordinary chain rule is incomplete: Brownian increments are of order $\sqrt{dt}$, so their squares contribute at order dt. Accordingly, $d\log X = X^{-1}dX - \frac{1}{2}X^{-2}d\langle X\rangle$. Here $d\langle\beta_i\rangle_t = 4\delta\beta_i^2V_i\,dt$, so the second term contributes $-2\delta V_i\,dt$. It more than cancels the $+\delta V_i\,dt$ obtained from $d\beta_i/\beta_i$ in Equation (3.4), leaving $-\delta V_i\,dt$ in Equation (3.5). Omitting it would give the wrong logarithmic growth and support-entry times.

Small initialization and from-start averaging. Initialize

$$\beta_0 = \alpha a, \qquad a \in (0,\infty)^d, \qquad L_\alpha = \log(1/\alpha), \tag{3.6}$$

$$T_\alpha(s) = \frac{sL_\alpha}{2}, \qquad b^\alpha(s) = \beta_{T_\alpha(s)}, \qquad z^\alpha(s) = \int_0^s b^\alpha(u)\,du. \tag{3.7}$$

The object compared with Lasso is the average from initialization. The time change gives the first identity below, and affineness of the gradient in Equation (3.1) gives the second:

$$\overline{\beta}^\alpha(s) = \frac{1}{T_\alpha(s)}\int_0^{T_\alpha(s)}\beta_t\,dt = \frac{z^\alpha(s)}{s}. \tag{3.8}$$

$$\frac{1}{T}\int_0^T g(\beta_t)\,dt = g\left(\frac{1}{T}\int_0^T\beta_t\,dt\right) \tag{3.9}$$

For comparison, let

$$x_s = \arg\min_{x\ge 0}\left\{\frac{1}{2}x^\top Hx - r^\top x + \frac{1}{s}\mathbf{1}^\top x\right\}, \qquad z^*(s) = sx_s, \qquad w^*(s) = \mathbf{1} + Hz^*(s) - sr. \tag{3.10}$$

The Karush–Kuhn–Tucker (KKT) conditions are

$$z^*(s)\ge 0, \qquad w^*(s)\ge 0, \qquad z^*(s)\odot w^*(s) = 0. \tag{3.11}$$

Proposition 3.1 (Master logarithmic KKT identity)**.** *Suppose the positive solution of Equation* (3.4) *is nonexplosive through* $T_\alpha(S)$*. Define*

$$w_i^\alpha(s) = -\frac{\log b_i^\alpha(s)}{L_\alpha}, \qquad c_{\alpha,i} = -\frac{\log a_i}{L_\alpha},$$
$$A_{\alpha,i}(s) = \frac{\delta}{2}\int_0^s V_i(b^\alpha(u))\,du, \qquad N_{\alpha,i}(s) = \frac{2\sqrt{\delta}}{L_\alpha} M_i(T_\alpha(s)). \tag{3.12}$$

Then, pathwise for $0 \le s \le S$*,*

$$\boxed{w^\alpha(s) = \mathbf{1} + Hz^\alpha(s) - sr + c_\alpha + A_\alpha(s) - N_\alpha(s)}. \tag{3.13}$$

Moreover,

$$\langle N_{\alpha,i}\rangle_s = \frac{4}{L_\alpha} A_{\alpha,i}(s), \qquad \dot{z}^\alpha(s) = b^\alpha(s) > 0, \qquad \langle w^\alpha(s), \dot{z}^\alpha(s)\rangle \le \frac{d}{eL_\alpha}. \tag{3.14}$$

Why "activation clock"? The variable w_i^α is the remaining normalized logarithmic distance: it starts near one and becomes asymptotically zero once coordinate i reaches order-one scale. For an active coordinate, Equation (3.13) reads, after vanishing terms are removed, $sr_i - (Hz^\alpha)_i \approx 1 + A_{\alpha,i}$. Thus the accumulated signal must pay the initial unit barrier and the nondecreasing toll $A_{\alpha,i}$, which ticks at rate $\delta V_i/2$. Because this toll shifts support-entry time, we call it a *logarithmic activation clock*. Vanishing clocks recover ordinary Lasso; deterministic nonzero clocks become time-dependent Lasso weights. Actual SGD has a different clock given by its exact discrete logarithmic remainder, as Section 4 makes explicit.

The raw path is not the comparator. In one deterministic dimension, away from $s = 1/\mu$,

$$\beta_{T_\alpha(s)} \longrightarrow \mu\mathbf{1}\{s > 1/\mu\}, \qquad \overline{\beta}^\alpha(s) \longrightarrow (\mu - 1/s)_+. \tag{3.15}$$

Thus averaging turns the activation jump into the Lasso soft threshold.

The calculations behind Proposition 3.1 and this scalar limit are given in Appendix B.1.

4 Fixed SGF noise and fixed SGD steps: weighted-path separation

The master identity already suggests that persistent stochastic corrections need not preserve the ordinary Lasso path. We first make this failure exact in an orthogonal problem, so the deformation cannot be attributed to feature correlation.

4.1 An exactly solvable axis sampler

Let coordinate i be sampled with probability $p_i > 0$, and take

$$x_i = \frac{e_i}{\sqrt{p_i}}, \qquad y_i = \frac{\mu_i}{\sqrt{p_i}}, \qquad \mu_i > 0. \tag{4.1}$$

Then $H = I$, $r = \mu$, and the expected loss is $\frac{1}{2}\|\beta - \mu\|_2^2$. When sample i is selected, its gradient is $(\beta_i - \mu_i)e_i/p_i$. With $q_i = (1 - p_i)/p_i$,

$$V_i(\beta) = q_i(\beta_i - \mu_i)^2. \tag{4.2}$$

Theorem 4.1 (Exact fixed SGF-noise and SGD-step axis paths)**.** *Initialize* $\beta_{0,i} = \alpha a_i$ *with* $a_i > 0$*.*

SGF. *Fix* $\delta > 0$ *such that* $\delta q_i \mu_i < 2$ *for every* i*, and define*

$$\rho_i^{\mathrm{F}} = 2\mu_i - \delta q_i \mu_i^2, \qquad \omega_i^{\mathrm{F}} = \frac{2\mu_i}{\rho_i^{\mathrm{F}}} = \frac{1}{1 - \delta q_i \mu_i/2}, \tag{4.3}$$

$$\overline{\beta}_i^{\alpha,\mathrm{F}}(s) \xrightarrow[\alpha\downarrow 0]{\mathbb{P}} \left(\mu_i - \frac{\omega_i^{\mathrm{F}}}{s}\right)_+. \tag{4.4}$$

These limits hold for every $s > 0$*. Away from* $s = 2/\rho_i^{\mathrm{F}}$*, the raw coordinate converges in probability to* $\mu_i\mathbf{1}\{s > 2/\rho_i^{\mathrm{F}}\}$*.*

Actual SGD. *Fix a step size $\eta > 0$ such that $\eta\mu_i/p_i < 1/2$. Run balanced factor SGD: when coordinate i is selected, $\beta_{k+1,i} = \beta_{k,i}\{1 + \eta(\mu_i - \beta_{k,i})/p_i\}^2$, and otherwise leave it unchanged. For α small enough that $K_\alpha(s) = \lfloor sL_\alpha/(2\eta)\rfloor \geq 1$, let $\overline{\beta}^{\alpha,\mathrm{D}}(s) = K_\alpha(s)^{-1}\sum_{k<K_\alpha(s)}\beta_k$, and set*

$$\rho_i^{\mathrm{D}} = \frac{2p_i}{\eta}\log\left(1 + \frac{\eta\mu_i}{p_i}\right), \qquad \omega_i^{\mathrm{D}} = \frac{2\mu_i}{\rho_i^{\mathrm{D}}} = \frac{\eta\mu_i}{p_i\log(1+\eta\mu_i/p_i)}, \tag{4.5}$$

$$\overline{\beta}_i^{\alpha,\mathrm{D}}(s) \longrightarrow \left(\mu_i - \frac{\omega_i^{\mathrm{D}}}{s}\right)_+ \quad \textit{almost surely} \tag{4.6}$$

along every deterministic sequence $\alpha \downarrow 0$, under a common sample-stream coupling. Away from $s = 2/\rho_i^{\mathrm{D}}$, the raw coordinate converges almost surely to $\mu_i\mathbf{1}\{s > 2/\rho_i^{\mathrm{D}}\}$ under the same coupling.

In both cases the limiting vector is the unique minimizer of

$$\min_{x\geq 0}\left\{\frac{1}{2}\|x-\mu\|_2^2 + \frac{1}{s}\sum_i \omega_i x_i\right\}, \tag{4.7}$$

with $\omega = \omega^{\mathrm{F}}$ for SGF and $\omega = \omega^{\mathrm{D}}$ for SGD.

The two weights expose the limitation of moment matching after nonlinear reparametrization. For the two-coordinate uniform sampler $p_1 = p_2 = 1/2$, at the matched nominal scale $\delta = \eta$,

$$\omega_i^{\mathrm{F}} = 1 + \frac{\eta\mu_i}{2} + O(\eta^2), \qquad \omega_i^{\mathrm{D}} = 1 + \eta\mu_i + O(\eta^2). \tag{4.8}$$

The SGF correction uses the conditional variance; the discrete logarithm also sees the squared mean gradient. If two active coordinates have unequal weights, their deficits $\mu_i - x_i = \omega_i/s$ are unequal. Hence no scalar reparametrization of time can turn that limiting curve into ordinary Lasso. The proof of Theorem 4.1 is in Appendix B.2.

Correlated clocks. For correlated designs, the existence of a deterministic clock limit is a probabilistic question. Under the joint convergence and complementarity hypotheses of Theorem A.1, $x(s) = z(s)/s$ minimizes

$$\frac{1}{2}x^\top Hx - r^\top x + \frac{1}{s}\{\mathbf{1} + A(s)\}^\top x, \qquad x \geq 0.$$

Continued clock growth on an active coordinate rules out one fixed weight vector. A structured four-observation example instead reduces to scalar Lasso under the nonlinear calibration $\lambda(s) = 1/s + \delta\sigma^2/2$, so this symmetric clock remains compatible with a scalar path after time recalibration. The full comparator, criterion, construction, and proof are in Appendices A.2 and B.3.

5 Weak noise and future-active clocks

At fixed SGF noise, Theorem 4.1 gives a genuine deformation of the Lasso path. In the diagonal class below, the deformation disappears at the sharp weak-noise scale. Crucially, one need not couple the noisy and deterministic trajectories over their full $O(L_\alpha)$ horizon. Throughout this section, the noise in Equation (3.3) is $\delta = \delta_\alpha$.

5.1 A sharp diagonal theorem

Theorem 5.1 (Sharp diagonal weak-noise recovery)**.** *Let $H = \mathrm{Diag}(h_1, \ldots, h_d)$, with $h_i > 0$, and let $r_i > 0$. Suppose $\beta_0 = \alpha a$, the positive SGF is nonexplosive through $T_\alpha(S)$, and, for a deterministic $\bar{V} < \infty$, almost surely,*

$$\max_i \sup_{t\leq T_\alpha(S)} V_i(\beta_t) \leq \bar{V}. \tag{5.1}$$

Let $x_{s,i} = (r_i - 1/s)_+/h_i$. For every $0 < s_0 < S$, there is a constant $C < \infty$, independent of α and ρ, such that for all sufficiently small α, with probability at least $1-\rho$,

$$\sup_{s\in[s_0,S]}\left\|\overline{\beta}^\alpha(s) - x_s\right\|_\infty \leq C\left\{\frac{1}{L_\alpha} + \delta_\alpha + \sqrt{\frac{\delta_\alpha\log(2d/\rho)}{L_\alpha}}\right\}. \tag{5.2}$$

Consequently, for a fixed problem, $\delta_\alpha \to 0$ implies uniform convergence in probability on $[s_0, S]$. No condition of the form $\delta_\alpha L_\alpha \to 0$ is required.

The complete finite-α constants and proof are in Appendix B.4. The argument uses contraction of the scalar log dynamics and the exact averaged-gradient identity. The unscaled log paths can differ by $O(\delta_\alpha L_\alpha)$; after KKT normalization, the same discrepancy costs only $O(\delta_\alpha)$. If Equation (5.1) holds only before a stopping time, the same statement holds on the no-exit event, with its failure probability added to ρ.

5.2 Only variance before future support motion matters

For general $H \succ 0$, the scaled Lasso path in Equation (3.10) is Lipschitz. Write, almost everywhere,

$$\mathcal{V}_i^{\uparrow}(s) = \int_0^s (\dot{z}_i^*(u))_+ \, du, \qquad \mathcal{V}_i^{\downarrow}(s) = \int_0^s (\dot{z}_i^*(u))_- \, du. \tag{5.3}$$

Define the future-active clock and the integrated dual defect

$$\mathcal{K}_\alpha^{\mathrm{act}}(s) = \sum_i \int_0^s A_{\alpha,i}(u)(\dot{z}_i^*(u))_+ \, du, \tag{5.4}$$

$$J_\alpha^+(s) = \int_0^s \sum_i (w_i^\alpha(u))_-(\dot{z}_i^*(u))_+ \, du. \tag{5.5}$$

The future-active clock records variance accumulated before later upward motion of the comparator. The term J_α^+ measures violations of nonnegative log-dual feasibility along that same motion.

Lemma 5.2 (Clock-uniform martingale bound)**.** *For $0 < \theta \le 1$ and $0 < \rho < 1$, put*

$$\gamma_\alpha(\theta, \rho) = \frac{2\log(2d/\rho)}{\theta L_\alpha}. \tag{5.6}$$

With probability at least $1 - \rho$, simultaneously for all i and $s \le S$,

$$|N_{\alpha,i}(s)| \le \theta A_{\alpha,i}(s) + \gamma_\alpha(\theta, \rho). \tag{5.7}$$

Thus the martingale is controlled by its own Itô clock, even when the clock is not known to be small.

Theorem 5.3 (Future-active pathwise comparison)**.** *Let $H \succ 0$, suppose the positive SGF is nonexplosive through $T_\alpha(S)$, fix $0 < \theta \le 1$ and $0 < \rho < 1$, and work on the event in Lemma 5.2. Write $\gamma_\alpha = \gamma_\alpha(\theta, \rho)$, and put*

$$\begin{aligned} &A_a = \max_i |\log a_i|, \quad Z_\alpha = \|z^\alpha(S)\|_1, && h_+ = \max_{i,j}(H_{ij})_+, \\ &r^- = (-r)_+, && U_\alpha = 1 + h_+ Z_\alpha + S\left\|r^-\right\|_\infty, \end{aligned} \tag{5.8}$$

and

$$\begin{aligned} \mathcal{R}_\alpha^{\mathrm{act}}(s) = \frac{ds}{eL_\alpha} + \left(\frac{A_a}{L_\alpha} + \gamma_\alpha\right)\left\{\|z^\alpha(s)\|_1 + \sum_i \mathcal{V}_i^{\uparrow}(s)\right\} + J_\alpha^+(s) \\ + (1+\theta)\mathcal{K}_\alpha^{\mathrm{act}}(s) + U_\alpha \sum_i \mathcal{V}_i^{\downarrow}(s). \end{aligned} \tag{5.9}$$

Then, simultaneously for every $s \le S$,

$$\frac{1}{2}\|z^\alpha(s) - z^*(s)\|_H^2 \le \mathcal{R}_\alpha^{\mathrm{act}}(s). \tag{5.10}$$

If

$$\mathcal{G}_s(x) = \frac{1}{2}x^\top H x - r^\top x + \frac{1}{s}\mathbf{1}^\top x, \qquad x \ge 0, \tag{5.11}$$

then, for $s > 0$,

$$\left\|\overline{\beta}^\alpha(s) - x_s\right\|_H \le \frac{\sqrt{2\mathcal{R}_\alpha^{\mathrm{act}}(s)}}{s}, \tag{5.12}$$

$$0 \le \mathcal{G}_s(\overline{\beta}^\alpha(s)) - \mathcal{G}_s(x_s) \le \frac{\mathcal{R}_\alpha^{\mathrm{act}}(s)}{s^2} + \frac{\|r\|_{H^{-1}}\sqrt{2\mathcal{R}_\alpha^{\mathrm{act}}(s)}}{s}. \tag{5.13}$$

Here $\|r\|_{H^{-1}} = (r^\top H^{-1} r)^{1/2}$.

In particular, for fixed d, choose $\theta = 1$ and any $\rho_\alpha \downarrow 0$ such that $\log(1/\rho_\alpha) = o(L_\alpha)$. If z^* is coordinatewise nondecreasing and

$$Z_\alpha = O_{\mathbb{P}}(1), \qquad J_\alpha^+(S) = o_{\mathbb{P}}(1), \qquad \mathcal{K}_\alpha^{\mathrm{act}}(S) = o_{\mathbb{P}}(1), \tag{5.14}$$

then, on every $[s_0, S]$, $\sup_{s\in[s_0,S]} \left\|\overline{\beta}^\alpha(s) - x_s\right\|_H \xrightarrow{\mathbb{P}} 0$. For a nonmonotone comparator, the last term in Equation (5.9) is an explicit deletion/downward-variation defect; Theorem 5.3 does not turn that defect into an unconditional convergence theorem.

Fubini's theorem gives the revealing equivalent form

$$\mathcal{K}_\alpha^{\mathrm{act}}(s) = \sum_i \int_0^s \{\mathcal{V}_i^\uparrow(s) - \mathcal{V}_i^\uparrow(v)\}\, dA_{\alpha,i}(v). \tag{5.15}$$

Variance accumulated after a coordinate's last upward motion is absent; variance before a future entry can shift its activation time. Proofs of Lemma 5.2 and Theorem 5.3 are in Appendix B.5.

Checkable closure and the global boundary. When H is Stieltjes—positive definite with nonpositive off-diagonal entries—a transverse tail-cover criterion gives the full-clock rate $O(\delta_\alpha^{p/(p+2)})$, hence weak-noise recovery without coupling δ_α to L_α. A tail cover partitions large states into low-output, visible-output, and transverse-leakage regions. The criterion holds for coordinate-local covariance, one-dimensional finite data, and an exact two-coordinate cross-noise model, but is not automatic for every dataset. For conventional one-row sampling, the only exact leakage directions are coordinate axes, and transverse control holds locally near each axis. A global bound on repeated departures from and returns to these neighborhoods remains open. Precise statements and qualifications are in Appendices A.3 and A.4.

6 Signed product-factor SGD: two exact logarithmic clocks

We now return to the standard Hadamard parametrization $\beta = a \odot b$. This case is not obtained by applying the positive result twice: the two sign channels use the same minibatch, the lifted quadratic is singular, and the exact finite-step corrections differ beyond second order. Nevertheless, a linear rotation exposes enough structure to prove a signed, actual-SGD path theorem.

Model and rotation. Let

$$\mathcal{L}(\beta) = \frac{1}{2}\beta^\top H\beta - r^\top\beta, \qquad g(\beta) = H\beta - r, \qquad H \succ 0. \tag{6.1}$$

At iteration k, an unbiased stochastic gradient G_k satisfies

$$\mathbb{E}[G_k \mid \mathcal{F}_k] = g(\beta_k), \qquad \zeta_k = G_k - g(\beta_k). \tag{6.2}$$

The shared-minibatch product-factor recursion is

$$a_{k+1} = a_k - \eta b_k \odot G_k, \qquad b_{k+1} = b_k - \eta a_k \odot G_k. \tag{6.3}$$

Introduce the rotated factors and their nonnegative square channels

$$u_k^+ = \frac{a_k + b_k}{2}, \qquad u_k^- = \frac{a_k - b_k}{2}, \qquad q_k^\pm = (u_k^\pm)^{\odot 2}. \tag{6.4}$$

Then $\beta_k = q_k^+ - q_k^-$, and the recursion becomes exactly

$$u_{k+1}^+ = u_k^+ \odot (\mathbf{1} - \eta G_k), \qquad u_{k+1}^- = u_k^- \odot (\mathbf{1} + \eta G_k). \tag{6.5}$$

Thus the two channels are coupled through the same random gradient; replacing one occurrence of G_k by an independent copy would define a different algorithm.

Fix $0 < \alpha < 1$ and $\eta > 0$, and initialize

$$q_{0,i}^\pm = \alpha a_i^\pm, \qquad a_i^\pm > 0, \qquad L = \log(1/\alpha), \qquad h = \frac{2\eta}{L}, \qquad s_k = kh, \tag{6.6}$$

and work before the first loss of positive-step stability $|\eta G_{k,i}| < 1$. Define

$$R(q) = -\log(1-q) - q, \qquad z_k^{\pm} = h\sum_{j<k} q_j^{\pm}, \qquad z_k = z_k^{+} - z_k^{-}. \tag{6.7}$$

The same per-step logarithmic algebra appears in the endpoint analysis of Even et al. (2023): their channel corrections satisfy $q_+(x) = 2R(x)$ and $q_-(x) = 2R(-x)$. Here we retain the two corrections on a finite logarithmic horizon and insert them into an averaged-path KKT system. Define also

$$\begin{aligned} w_{k,i}^{\pm} &= -\frac{\log q_{k,i}^{\pm}}{L}, \qquad c_i^{\pm} = -\frac{\log a_i^{\pm}}{L}, \\ M_{k,i} &= h\sum_{j<k} \zeta_{j,i}, \quad \mathcal{A}_{k,i}^{\pm} = \frac{2}{L}\sum_{j<k} R(\pm\eta G_{j,i}). \end{aligned} \tag{6.8}$$

Here and below, the plus sign in $R(\pm\eta G)$ corresponds to the positive channel.

Lemma 6.1 (Exact two-clock system)**.** *For every stable iterate,*

$$w_k^{+} = \mathbf{1} + Hz_k - s_k r + c^{+} + M_k + \mathcal{A}_k^{+}, \tag{6.9}$$

$$w_k^{-} = \mathbf{1} - Hz_k + s_k r + c^{-} - M_k + \mathcal{A}_k^{-}. \tag{6.10}$$

The common channel obeys the exact contraction and predictor control

$$q_{k,i}^{+} q_{k,i}^{-} = \alpha^2 a_i^{+} a_i^{-} e^{-L(\mathcal{A}_{k,i}^{+} + \mathcal{A}_{k,i}^{-})} \le \alpha^2 a_i^{+} a_i^{-}, \qquad q_{k,i}^{+} + q_{k,i}^{-} \le |\beta_{k,i}| + 2\alpha\sqrt{a_i^{+} a_i^{-}}. \tag{6.11}$$

If $k \ge 1$ and $|\eta G_{j,i}| \le \kappa_{\mathrm{step}} < 1$ for $j < k$, each exact clock is comparable to $s_k\eta$ times the empirical average of $G_{j,i}^2$. The detailed identities and two-sided bounds are given in Appendix C.1.

The overlap $q_k^{+} \odot q_k^{-}$ is the redundant common direction of the two-channel lift: both channels can grow along it without changing their difference. Its exact contraction prevents such hidden channel inflation, so a predictor bound also controls both nonnegative channels.

Signed Lasso in prediction geometry. For $s > 0$, let

$$\beta_s = \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \left\{ \frac{1}{2}\beta^{\top} H\beta - r^{\top}\beta + \frac{1}{s}\|\beta\|_1 \right\}, \qquad y^*(s) = s\beta_s, \tag{6.12}$$

and set $y^*(0) = 0$. Write

$$p^* = (y^*)_+, \quad n^* = (-y^*)_+, \quad \widetilde{z}^* = (p^*, n^*), \quad \nu^* = sr - Hy^* \in \partial\|y^*\|_1. \tag{6.13}$$

Linearly interpolate $z_k^{\pm}$ on the mesh $\{s_k\}$, and set $\widetilde{z} = (z^{+}, z^{-})$ and $z = z^{+} - z^{-}$. With

$$P = (I, -I), \quad \widetilde{H} = P^{\top}HP, \quad \widetilde{r} = P^{\top}r, \quad \widetilde{w}^* = (\mathbf{1} - \nu^*, \mathbf{1} + \nu^*), \tag{6.14}$$

we have $\widetilde{H} \succeq 0$, but $\ker(\widetilde{H}) = \{(v, v) : v \in \mathbb{R}^d\}$. We therefore compare only in the quotient, through

$$\frac{1}{2}\|P(\widetilde{z} - \widetilde{z}^*)\|_H^2 = \frac{1}{2}\|z - y^*\|_H^2, \tag{6.15}$$

and never invoke $\widetilde{H}^{-1}$.

The mesh-complementarity construction and the full pathwise comparison are stated in Appendix A.6. Their remainder keeps separate the mesh error, the shared martingale, the two future-active exact clocks, and the channel downward variation. In particular, no monotonicity is needed for the pathwise bound; deletion and sign reversal remain as an explicit defect. Write

$$\mathcal{V}^{\downarrow}(s) := \sum_i \int_0^s \{(\dot{p}_i^*(v))_- + (\dot{n}_i^*(v))_-\}\, dv \tag{6.16}$$

for the total downward variation of the positive and negative comparator channels.

For $s > 0$, define the from-start discrete average $\overline{\beta}^{\eta,\alpha}(s) = z(s)/s$.

Corollary 6.2 (Localized weak-step signed SGD). *Fix $d, H, r, a^{\pm}$ and $0 < s_0 < S$. Consider a triangular array with $L_\alpha \to \infty$ and $\eta_\alpha \to 0$. Suppose there are fixed $B, G_*, \overline{V} < \infty$ and $0 \le \kappa_{\mathrm{step}} < 1$, and, for each α, an integer-valued stopping time τ_α. With $h_\alpha = 2\eta_\alpha/L_\alpha$ and $N_\alpha = \lceil S/h_\alpha \rceil$, suppose almost surely that*

$$\max_i |\beta_{k,i}| \le B \qquad (0 \le k \le N_\alpha,\ k < \tau_\alpha),$$

$$\max_i |G_{k,i}| \le G_*, \qquad \max_i |\eta_\alpha G_{k,i}| \le \kappa_{\mathrm{step}} \qquad (0 \le k < N_\alpha,\ k < \tau_\alpha), \tag{6.17}$$

$$\max_i \mathbb{E}[\zeta_{k,i}^2 \mid \mathcal{F}_k] \le \overline{V} \qquad (k < N_\alpha,\ k < \tau_\alpha), \tag{6.18}$$

and $\mathbb{P}(\tau_\alpha \le N_\alpha) \to 0$. If $\mathcal{V}^{\downarrow}(S) = 0$, then

$$\sup_{s \in [s_0, S]} \left\| \overline{\beta}^{\eta_\alpha, \alpha}(s) - \beta_s \right\|_H \xrightarrow{\mathbb{P}} 0, \tag{6.19}$$

and the signed-Lasso objective gap converges uniformly in probability to zero on $[s_0, S]$. At grid times with $k \ge 1$,

$$\overline{\beta}^{\eta,\alpha}(s_k) = \frac{1}{k} \sum_{j<k} \beta_j. \tag{6.20}$$

No condition of the form $\eta_\alpha L_\alpha \to 0$ is required.

The zero-downward-variation condition is used only to turn the pathwise estimate into convergence. Without it, Theorem A.8 remains valid with the explicit channel-deletion term displayed in its remainder.

Boundary of the result. Corollary 6.2 is an actual-SGD theorem for the standard signed product factors, but it does not establish localization for every finite dataset or identify the correlated fixed-step limit. It controls prediction and the ordinary signed-Lasso objective rather than the redundant channel norm. Removing the stopping assumption requires uniform occupation and clock bounds for both channels; treating sign reversals without a downward-variation penalty requires a sharper quotient-geometry comparison.

7 Numerical checks of the clock mechanism

An exactly normalized two-dimensional experiment isolates from-start averaging, variance-clock deformation, and the extra finite-step correction; its design and diagnostics are in Appendix A.5. Here we report a correlated signed-SGD pilot. Both experiments diagnose the clock mechanism on fixed instances; the universal correlated-design question remains open.

In this scoped CPU pilot, we simulated actual signed Hadamard SGD on 15 Toeplitz–Gaussian datasets generated without post hoc selection, with $d = 32, n = 128$ and $\rho \in \{0.25, 0.60, 0.90\}$. All 150 frozen dataset–setting combinations passed the numerical audits. Using integrated mean pathwise H-error, averaging improved on the raw iterate in every combination, with median averaged/raw ratio 0.159. At the predeclared setting $(\alpha, \eta, b) = (10^{-6}, 0.02, 1)$, the trajectory-specific exact- R comparator had lower ensemble-mean systematic H-error than scalar Lasso in 14 of 15 datasets. Along $\eta_\alpha = 0.05/\sqrt{L_\alpha}$, ensemble-mean scalar-Lasso error decreased while $\eta_\alpha L_\alpha$ increased. The exact-R comparator did not improve overall on its conditional-second-moment Taylor approximation; holding η/b fixed produced mixed changes, and exact signed-support recovery remained weak. Complete specifications, values, and diagnostics appear in Appendix A.7.

8 Discussion and limitations

The averaged Hadamard path is governed by a logarithmic budget rather than by stochasticity alone. At fixed SGF noise or fixed SGD step size, activation clocks remain visible and produce coordinate-dependent shrinkage; in weak-noise and weak-step regimes, only clock mass accumulated before future support motion affects the comparison. This separates deterministic flow, moment-matched SGF, and actual SGD even when they share the same population loss. The remaining limitations are structural: the signed theorem is localized and retains a deletion defect, while general correlated one-row SGF still requires a global off-face excursion bound.

REPRODUCIBILITY STATEMENT

All assumptions and complete proofs appear in Appendices B and C. The experiments in Appendices A.5 and A.7 state the data-generating mechanisms, initialization, time scaling, sampling scheme, trajectory counts, and comparator definitions. The retained experiment records include frozen configurations, trajectory-level data, solver and KKT diagnostics, exact logarithmic-identity checks, and output hashes. A separate reconstruction from all 150 stored trajectory archives and a clean rerun reproduced every deterministic output byte for byte. The accompanying anonymous code package contains both runners, version-pinned dependencies, frozen configurations, compact reference outputs, and integrity checks; it downloads no data. The full pilot regenerates the omitted trajectory archives, and the packaged figure builder reconstructs both manuscript figures.

AI USE STATEMENT

OpenAI's GPT-5.6 Sol was used to survey and summarize literature; stress-test and audit mathematical claims; assist with exposition; and design and implement numerical experiments. The author provided the main ideas and formulated the major results and proofs through conversation with GPT-5.6 Sol. The author takes responsibility for the final content of this work, including all text, claims, proofs, code, and artifacts produced with generative-AI assistance.

REFERENCES

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. URL https://proceedings.mlr.press/v89/ali19a.html.

Alnur Ali, Edgar Dobriban, and Ryan J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 233–244. PMLR, 2020. URL https://proceedings.mlr.press/v119/ali20a.html.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 468–477. PMLR, 2021. URL https://proceedings.mlr.press/v139/azulay21a.html.

Raphaël Berthier. Diagonal linear networks and the Lasso regularization path, 2025. URL https://arxiv.org/abs/2509.18766. arXiv:2509.18766, version 3, revised 2026.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3):1437–1460, 2023. doi: 10.1093/imaiai/iaad012. URL https://doi.org/10.1093/imaiai/iaad012.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b117cd5a0af6f5ecfdac7b47b893b96f-Abstract-Conference.html.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29406–29448. Curran Associates, Inc., 2023. doi: 10.52202/075280-1279. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5da6ce80e97671b70c01a2e703b868b3-Abstract-Conference.html.

Begoña García Malaxechebarría, Courtney Paquette, Maryam Fazel, and Dmitriy Drusvyatskiy. High-dimensional limit of SGD for diagonal linear networks, 2026. URL https://arxiv.org/abs/2605.17177.

Peter D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017. doi: 10.1016/j.csda.2017.06.007. URL https://doi.org/10.1016/j.csda.2017.06.007.

Chendi Huang and Yuan Yao. A unified dynamic approach to sparse model selection. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 2047–2055. PMLR, 2018. URL https://proceedings.mlr.press/v84/huang18a.html.

Chris Kolb, Christian L. Müller, Bernd Bischl, and David Rügamer. Smoothing the edges: Smooth optimization for sparse regularization using Hadamard overparametrization. *Machine Learning*, 115:87, 2026. doi: 10.1007/s10994-026-06997-0. URL https://doi.org/10.1007/s10994-026-06997-0.

Jiangyuan Li, Thanh V. Nguyen, Chinmay Hegde, and Raymond K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28298–28309. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ee188463935a061dee6df8bf449cb882-Abstract.html.

Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34626–34640. Curran Associates, Inc., 2022. doi: 10.52202/068431-2509. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfa1106ea7065899b13f2be9da04efb4-Abstract-Conference.html.

Ziyin Liu and Zihao Wang. spred: Solving L1 penalty with SGD. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43407–43422. PMLR, 2023. URL https://proceedings.mlr.press/v202/ziyin23a.html.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16270–16295. PMLR, 2022. URL https://proceedings.mlr.press/v162/nacson22a.html.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 3222–3242. PMLR, 2018. URL https://proceedings.mlr.press/v75/neu18a.html.

Hristo Papazov, Scott Pesme, and Nicolas Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3556–3564. PMLR, 2024. URL https://proceedings.mlr.press/v238/papazov24a.html.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7475–7505. Curran Associates, Inc., 2023. doi: 10.52202/075280-0329. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/17a9ab4190289f0e1504bbb98d1d111a-Abstract-Conference.html.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f4661398cb1a3abd3ffe58600bf11322-Abstract.html.

Loucas Pillaud-Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of Thirty*

Fifth Conference on Learning Theory, volume 178 of *Proceedings of Machine Learning Research*, pp. 2127–2159. PMLR, 2022. URL https://proceedings.mlr.press/v178/vivien22a.html.

Tom Sander, Maxime Sylvestre, and Alain Durmus. Implicit bias in noisy-SGD: With applications to differentially private training. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3295–3303. PMLR, 2024. URL https://proceedings.mlr.press/v238/sander24a.html.

Arun Sai Suggala, Adarsh Prasad, and Pradeep K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10608–10619. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6459257ddab7b85bf4b57845e875e4d4-Abstract.html.

Ryan J. Tibshirani. Equivalences between sparse models and neural networks. Manuscript, 2021. URL https://www.stat.berkeley.edu/~ryantibs/papers/sparsitynn.pdf.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2972–2983. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5cf21ce30208cfffaa832c6e44bb567d-Abstract.html.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Jingfeng Wu, Vladimir Braverman, and Lin Yang. Obtaining adjustable regularization for free via iterate averaging. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10344–10354. PMLR, 2020. URL https://proceedings.mlr.press/v119/wu20a.html.

Lei Wu and Weijie J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37656–37684. PMLR, 2023. URL https://proceedings.mlr.press/v202/wu23r.html.

Jiajie Zhao, Jianxing Wang, Junjie Yang, Zhiwei Bai, and Yaoyu Zhang. Gradient flow dynamics and implicit bias of diagonal linear networks under infinitesimal initialization. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2607.12332.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 109(4):1033–1046, 2022. doi: 10.1093/biomet/asac010. URL https://doi.org/10.1093/biomet/asac010.

A Extended context, theory, and numerical diagnostics

A.1 Extended related work

The phrase "implicit regularization" covers several inequivalent claims. The taxonomy in Table 1 separates the objects most easily conflated with our question. The central distinctions are static penalty versus unregularized dynamics, endpoint versus path, raw iterate versus from-start average, and continuous SGF versus the exact discrete update.

Direct least squares and ridge. Ali et al. (2019) compare direct-parameter least-squares gradient flow with ridge along $\lambda = 1/t$, principally at the risk level. Ali et al. (2020) construct an SGF whose forward-Euler increment matches the first two conditional moments of minibatch SGD, and compare its risk and coefficients with ridge throughout training. The comparison is quantitative, not an equality: exponential flow filters and ridge resolvents are different. Moreover, matching the first two moments of discrete SGD and *Euler-discretized* SGF does not give equality in distribution, a pathwise coupling, or a bound from actual SGD to continuous SGF.

There are exact ridge results for deliberately designed averages. Neu & Rosasco (2018) use geometrically weighted streaming-SGD iterates and obtain an asymptotic ridge equivalence together with a finite-time expectation identity involving explicitly regularized SGD. Wu et al. (2020) generalize the construction to adjustable regularization and broader smooth, strongly convex objectives. By contrast, the average here is the unweighted from-start average forced by the logarithmic product dynamics. Even in direct least squares, the mean raw-SGD path equals GD and hence an exact *matrix*-ridge path; a single scalar ridge parameter does not generally match all Hessian eigendirections.

Suggala et al. (2018) give a broader deterministic raw-path comparison: gradient flow, gradient descent, and mirror descent stay close to corresponding regularization paths under smooth strong convexity. They also show that this comparison can fail for general convex objectives. Their assumptions therefore do not cover the singular, underdetermined, small-initialization Hadamard limit, and their object is not a stochastic from-start average.

Table 1: Closest literature organized by the mathematical object proved and the additional ingredient needed for the present comparison.

Representative work	Object	Established conclusion	Additional ingredient needed here
Hoff (2017); Tibshirani (2021); Liu & Wang (2023); Kolb et al. (2026)	Explicit factor penalty	Exact or optimization-preserving ℓ_1 reformulation	Unregularized training path
Chou et al. (2023); Vaškevičius et al. (2019); Zhao et al. (2022); Li et al. (2021)	Endpoint or stopped estimator	Sparse interpolant bias or statistical recovery	Full finite-time Lasso path
Pesme & Flammarion (2023); Zhao et al. (2026)	Raw infinitesimal-init path	Saddle-to-saddle/LARS-like dynamics and endpoint bias	From-start average and stochasticity
Berthier (2025)	Deterministic from-start average	Lasso correspondence under monotonicity; path-variation defect in general	Quadratic variation and finite steps
Ali et al. (2019; 2020)	Direct linear GF/SGF	Whole-training ridge risk and coefficient comparisons	Product geometry and Lasso
Pesme et al. (2021); Pillaud-Vivien et al. (2022)	Stochastic endpoint	Evolving hypentropy bias or one weighted-Lasso target	Ordinary-minibatch averaged path
Nacson et al. (2022); Even et al. (2023); Papazov et al. (2024)	Actual steps or momentum	Endpoint regularizers; exact finite-step gain and log remainders	Uniform finite-horizon averaged-path KKT comparison
García Malaxechebarría et al. (2026)	High-dimensional SGD limit	SDE/PDE approximation for broad observables	Fixed-design Lasso comparator

Static Hadamard equivalence. Hoff (2017) proves that minimizing over two product factors with quadratic factor penalties induces an exact ℓ_1 predictor penalty; deeper products generate fractional quasi-norms. Tibshirani (2021) develops this sparse-model/neural-network correspondence for general losses. The same variational identity underlies the exact smooth solver of Liu & Wang (2023), the optimization-transfer framework of Kolb et al. (2026), and the high-dimensional explicit-weight-decay analysis of Defilippis et al. (2026). These are exact and important, but the explicit penalty or weight decay changes the objective. None says that unpenalized training follows the penalized solution path.

Deterministic implicit sparse regularization. For unpenalized regression, Vaškevičius et al. (2019), Zhao et al. (2022), and Li et al. (2021) prove sparse-recovery guarantees for early-stopped product or deeper diagonal gradient descent under stated initialization, step size, noise, and restricted-design conditions. They establish statistically strong stopped estimators rather than equality with every Lasso solution along training. Chou et al. (2023) instead study the interpolating endpoint of small-initialization gradient flow: it is a Bregman-selected solution and approaches weighted minimum ℓ_1 as the initialization vanishes, but is not exactly Lasso at a fixed nonzero initialization.

The selected geometry itself depends on initialization. Woodworth et al. (2020) show how initialization scale separates kernel and rich regimes, and Azulay et al. (2021) show that relative initialization shape also changes the implicit regularizer. Li et al. (2022) characterize when a reparametrized gradient flow is equivalent to a continuous mirror flow through the commuting- parametrization property. These results explain why diagonal networks favor sparse solutions, but they do not add ordinary minibatch noise to an averaged Lasso-path theorem.

Raw and averaged deterministic paths. Pesme & Flammarion (2023) characterize the raw vanishing-initialization trajectory of a two-layer diagonal network as a sequence of saddle-to-saddle jumps described by a LARS-like recursion. Zhao et al. (2026) extend the raw-trajectory description

to deeper and more general two-layer diagonal networks and identify a modified-ℓ_1 endpoint bias. A jump process reminiscent of LARS is not the continuous Lasso path.

The closest deterministic result is Berthier (2025). On the logarithmic time scale set by initialization, the *from-start average*, not the raw predictor, has the Lasso variational interpretation. Under a coordinatewise monotonicity condition, the limiting objective correspondence is exact; the general comparison retains an explicit path-variation error. This paper studies how stochastic quadratic variation and exact finite steps modify that averaged correspondence.

Stochastic endpoints. Pesme et al. (2021) analyze diagonal-network SGF and characterize its endpoint through a trajectory-dependent hypentropy geometry, showing that structured noise can strengthen sparse bias. More precisely, their effective initialization contains a coordinatewise accumulated loss/variance term, so it is an endpoint precursor of the Itô clock used here. It is integrated to infinite time to characterize the selected interpolant, rather than retained on a finite logarithmic horizon in an averaged Lasso KKT system. The closest related endpoint result is Pillaud-Vivien et al. (2022): persistent, deliberately injected label noise drives a quadratic parametrization to one fixed weighted-Lasso solution, with convergence and support-recovery guarantees. In that result, time approaches one penalized endpoint and the noise covariance is designed as part of the algorithm. It does not index the ordinary Lasso path generated by finite-data minibatch sampling.

Continuous-time noisy SGD with additional isotropic Gaussian noise also changes the endpoint bias in linear and diagonal-linear models (Sander et al., 2024). Stability-based analyses relate large-step SGD minima to sharpness and parameter norms (Wu & Su, 2023). These are endpoint or stable-minimum statements, not averaged-path comparisons.

Finite steps, momentum, and high-dimensional limits. Nacson et al. (2022) show that a discrete GD step size can change a diagonal network's endpoint bias, including a transition from kernel-like to sparsity-favoring behavior. Even et al. (2023) characterize endpoint implicit regularizers for actual GD and SGD with macroscopic step sizes, including edge-of-stability effects. Crucially, their exact quantities

$$q_+(x) = -2x - 2\log(1-x) = 2R(x), \qquad q_-(x) = 2x - 2\log(1+x) = 2R(-x), \qquad |x| < 1,$$

are the same signed finite-step logarithmic remainders used here, up to the factor-of-two convention from squared channel amplitudes. They also identify competition between accumulated gain and the initialization logarithm in the endpoint bias. Papazov et al. (2024) derive a momentum-dependent implicit regularizer in a continuous-time scaling and give weaker stochastic-momentum results. These results demonstrate that the optimizer matters. Our discrete contribution is different: we retain those known exact remainders inside a uniform, logarithmic-horizon comparison of the from-start average with the Lasso KKT system and its localization and path-variation defects.

García Malaxechebarría et al. (2026) prove that, in a proportional high-dimensional regime, streaming diagonal-network SGD is approximated by an SDE and that a deterministic PDE propagates broad observables such as risk and curvature. This supplies rigorous continuous-dynamics machinery for actual SGD in its asymptotic regime, but does not identify a finite-dimensional Lasso-path comparator. Thus factor-space SGF, its Euler discretization, and the exact factor-SGD product update remain distinct mathematical objects.

Other dynamic sparse paths. Linearized Bregman iteration and sparse mirror descent provide deliberately constructed dynamic regularization paths. For example, Huang & Yao (2018) prove path-consistency and early-stopping guarantees under restricted strong convexity and an irrepresentability condition. Such methods, like classical LARS/homotopy, are useful comparators but do not imply that ordinary Hadamard minibatch dynamics follows the same path.

Across the reviewed literature, no located result combines the unweighted from-start average, ordinary minibatch SGF, actual finite-step factor SGD, and the fixed-noise/weak-noise Lasso phase diagram. This is a search-based novelty boundary, not a claim that no overlapping later preprint can exist.

A.2 Correlated clocks and time-dependent weights

For a general fixed-noise problem, identifying the limiting clock is a probabilistic question. Conditional on such a limit, however, its variational meaning is exact.

Theorem A.1 (Deterministic clock comparator)**.** *Fix a compact interval $I \subset (0, \infty)$. Suppose, along a sequence $\alpha \downarrow 0$, that*

$$z^\alpha \longrightarrow z, \qquad A_\alpha \longrightarrow A \quad \textit{uniformly on } I \textit{ in probability}, \tag{A.1}$$

where z, A are deterministic and continuous. Assume that

$$w(s) = \mathbf{1} + Hz(s) - sr + A(s), \qquad z(s) \geq 0, \quad w(s) \geq 0, \quad z(s) \odot w(s) = 0. \tag{A.2}$$

Then $x(s) = z(s)/s$ uniquely minimizes

$$\frac{1}{2} x^\top H x - r^\top x + \frac{1}{s}\{\mathbf{1} + A(s)\}^\top x \quad \textit{over } x \geq 0. \tag{A.3}$$

A fixed weight vector $\omega \geq 0$ generates the same path at nominal level $1/s$ if and only if, for every i and $s \in I$,

$$\begin{cases} \omega_i = 1 + A_i(s), & z_i(s) > 0, \\ w_i(s) + \omega_i - 1 - A_i(s) \geq 0, & z_i(s) = 0. \end{cases} \tag{A.4}$$

In particular, continued clock growth on a coordinate active at two times rules out a single fixed weight vector.

The hypotheses in Equations (A.1) and (A.2) are conditional inputs. Verifying them for arbitrary correlated finite data at fixed noise is the unresolved averaging problem on interacting stochastic faces.

Proposition A.2 (A correlated finite-data clock deformation)**.** *Let*

$$H = \begin{pmatrix} h & c \\ c & h \end{pmatrix} \succ 0, \qquad c \neq 0, \qquad K = h + c, \tag{A.5}$$

and consider an ordinary four-observation least-squares problem grouped into two equiprobable blocks whose stochastic block gradients are

$$G_\zeta(\beta) = H\beta - r_0 \mathbf{1} + \zeta\sigma\mathbf{1}, \qquad \zeta \in \{-1, 1\}, \tag{A.6}$$

where $r_0 > 0$ and $\sigma > 0$. Thus $\Sigma = \sigma^2 \mathbf{1}\mathbf{1}^\top$. Initialize $\beta_0 = \alpha\mathbf{1}$, fix $\delta > 0$, and suppose

$$a_\delta = r_0 - \frac{\delta\sigma^2}{2} > 0. \tag{A.7}$$

Then, for every $s > 0$,

$$\overline{\beta}^\alpha(s) \xrightarrow[\alpha \downarrow 0]{\mathbb{P}} \frac{1}{K}\left(a_\delta - \frac{1}{s}\right)_+ \mathbf{1}. \tag{A.8}$$

Relative to the original loss, this is scalar Lasso at $\lambda(s) = 1/s + \delta\sigma^2/2$, or time-dependent weighted Lasso at the nominal level $1/s$, with weights $(1 + s\delta\sigma^2/2)\mathbf{1}$. It cannot be represented by one fixed weight vector on an interval containing two active times. This symmetric example therefore remains compatible with scalar Lasso after nonlinear time recalibration; more general deterministic clocks may instead produce time-dependent coordinate weights.

Proofs of Theorem A.1 and Proposition A.2 appear in Appendix B.3.

A.3 When does the clock vanish?

The comparison theorem reduces the stochastic problem to a clock estimate. We record a checkable sufficient condition that permits cross-coordinate variance leakage. Assume in this subsection that H is Stieltjes (positive definite with nonpositive off-diagonal entries), $r > 0$, and $\mu = H^{-1}r > 0$. Fix $p \in (0, 1/2)$, and write

$$\mathsf{S}(\beta) = 1 + \|\beta\|_1, \qquad V_\Sigma(\beta) = \sum_i V_i(\beta), \qquad W_p(\beta) = \sum_i \mu_i(\beta_i/\mu_i)^p V_i(\beta). \tag{A.9}$$

Lemma A.3 (Fractional Stieltjes occupation budgets)**.** *For ordinary finite-data SGF under the assumptions above, the positive process is nonexplosive. For every fixed $S > 0$, there are constants $C_p, B_p, C_V < \infty$, independent of all sufficiently small α, such that, with $T_\alpha = SL_\alpha/2$,*

$$\frac{1}{T_\alpha}\mathbb{E}\int_0^{T_\alpha} \mathsf{S}(\beta_t)^{p+1} dt \leq C_p, \qquad \frac{\delta_\alpha}{T_\alpha}\mathbb{E}\int_0^{T_\alpha} W_p(\beta_t) dt \leq B_p, \qquad V_\Sigma(\beta) \leq C_V \mathsf{S}(\beta)^2. \tag{A.10}$$

Proposition A.4 (Transverse tail-cover criterion)**.** *Suppose that, outside a fixed compact set, every state belongs to at least one of the following regions:*

$$\begin{aligned} V_\Sigma(\beta) &\le C_{\mathrm{low}}\mathsf{S}(\beta)^{p+1}, && \textit{(low output)},\\ W_p(\beta) &\ge c_{\mathrm{vis}}\mathsf{S}(\beta)^p V_\Sigma(\beta), && \textit{(visible output)},\\ \beta &\in \mathscr{L}, && \textit{(transverse leakage)}. \end{aligned} \tag{A.11}$$

Here $\mathscr{A}_\delta$ denotes the generator of Equation (3.4)*. Assume there is a nonnegative C^2 function Ψ for which stopped Dynkin formulas are valid, $\Psi(\beta_0)/T_\alpha = O(1)$, and, for all sufficiently small δ,*

$$\begin{aligned} \mathscr{A}_\delta\Psi(\beta) \le\ & -c_\perp \mathsf{S}(\beta)V_\Sigma(\beta)\mathbf{1}\{\beta\in\mathscr{L}\} + C_\perp\{1+\mathsf{S}(\beta)^{p+1}\}\\ & + C_\perp\delta W_p(\beta) + C_\perp\delta V_\Sigma(\beta). \end{aligned} \tag{A.12}$$

Then, whenever $\delta_\alpha \to 0$,

$$\frac{\delta_\alpha}{T_\alpha}\mathbb{E}\int_0^{T_\alpha} V_\Sigma(\beta_t)dt = O\Big(\delta_\alpha^{p/(p+2)}\Big). \tag{A.13}$$

Consequently $\max_i A_{\alpha,i}(S) \to 0$ in L^1, and hence the future-active clock vanishes. No relation between δ_α and L_α is imposed.

The criterion covers models with coordinate-local covariance and one-dimensional finite datasets as simpler special cases. It also closes an exact two-coordinate cross-noise design using $\Psi = (\beta_1-\beta_2)^2$. It is a sufficient condition, not an automatic theorem for every finite dataset. Its short threshold proof is given in Appendix B.6.

A.4 The precise global boundary

For a fixed finite dataset with conventional one-row sampling, write

$$H = \mathbb{E}[x_A x_A^\top] \succ 0, \qquad Z_A = x_A x_A^\top - H, \qquad Q_i(v) = \mathbb{E}(e_i^\top Z_A v)^2. \tag{A.14}$$

A boundary direction $v \ge 0$ is an exact leakage direction if $Q_i(v) = 0$ for $i \in \mathrm{supp}(v)$, while $\sum_i Q_i(v) > 0$.

Theorem A.5 (Axis rigidity and local dissipation)**.** *Every exact leakage direction in Equation* (A.14) *is a coordinate axis. Axis e_j has zero within-axis variance exactly when*

$$x_{Aj}^2 = H_{jj} \quad \textit{almost surely}. \tag{A.15}$$

For every genuinely leaking such axis, there are $\varepsilon_j, c_j, C_j, \delta_j > 0$ such that the cut

$$D_j(\beta) = \beta_j - \sum_{i\ne j}\beta_i \tag{A.16}$$

satisfies

$$\mathscr{A}_\delta D_j(\beta)^2 \le -c_j\mathsf{S}(\beta)V_\Sigma(\beta) + C_j \tag{A.17}$$

whenever $\sum_{i\ne j}\beta_i \le \varepsilon_j\beta_j$ and $0 < \delta \le \delta_j$.

Thus the obstruction is not local geometry at the exact leakage face. In fact, on the invariant singleton face $\beta = Xe_j$, the active coordinate is a logistic diffusion. For every fixed $\delta_0 < \infty$, uniformly over $0 < \delta \le \delta_0$ and uniformly bounded initial values X_0,

$$\frac{\delta}{T}\mathbb{E}\int_0^T V_i(X_t e_j)dt \le C_i\delta\left(1 + \frac{X_0}{T}\right) \tag{A.18}$$

for every output coordinate i. Hence the ideal face itself has an $O(\delta)$ leaked clock. What remains is the cumulative cost from repeated departures from and returns to the narrow axis cones.

That residual cannot be dismissed by standard smooth patching. For the admissible Rademacher rows $x_\varepsilon = \varepsilon \in \{-1,1\}^d$ with labels $y_\varepsilon = \varepsilon^\top\mu$, one has $H = I$ and every axis leaks. If Ψ is two-homogeneous, denote by $\mathfrak{S}_\Psi(v)$ the coefficient of $\delta\varrho^4$ in $\mathscr{A}_\delta\Psi(\varrho v)$.

Proposition A.6 (No-go boundary for global patches)**.** *In the Rademacher model with $d \ge 3$:*

(i) for $v_{jk} = (e_j + e_k)/2$, every C^2 two-homogeneous Ψ obeys

$$\mathfrak{S}_\Psi(v_{jk}) = \frac{3}{2}\Psi(v_{jk}); \tag{A.19}$$

(ii) if $\Psi_M(\beta) = \beta^\top M\beta$, $M \succeq 0$, and $\min_j M_{jj} \geq m > 0$, then some pair satisfies

$$\mathfrak{S}_{\Psi_M}(v_{jk}) \geq \frac{3(d-2)}{4(d-1)}m > 0; \tag{A.20}$$

(iii) every cutoff of the form $a(\beta)[\ell(\beta)_+]^m$, where ℓ is linear, $m \geq 2$, and the positive C^2 prefactor is homogeneous of degree $2 - m$, has positive leading symbol along every fixed transverse line just inside a noncharacteristic noisy boundary;

(iv) every nonconstant, nonnegative, nonincreasing C^2 separable soft selector $F(x, u) = x\phi(u)$ has a ray, already on a two-coordinate face, on which $\mathscr{A}_\delta F(x, u) \sim c\delta x^3 > 0$.

These conclusions obstruct the indicated proof strategies; they do not produce a trajectory with a macroscopic variance clock.

The rank-one proof of axis rigidity and the key no-go calculations appear in Appendix B.7. Taken together, the results locate the remaining fixed-dimensional Stieltjes row-sampling problem precisely: a theorem covering every such finite dataset still requires a genuinely nonseparable cancellation or a quantitative off-face excursion/local-time estimate.

A.5 Exact two-dimensional clock diagnostic

We use a two-dimensional least-squares instance whose population Hessian is the identity but whose row-gradient covariance is anisotropic. The diagnostic distinguishes the raw trajectory from its from-start average, scalar Lasso from the variance-clock comparator, and the finite-step SGD correction from the canonical SGF correction; it does not test the unresolved universal correlated-design question.

The design has three rows and parameter $q = 0.95$:

$$X = \sqrt{3}\begin{pmatrix} \sqrt{q} & 0 \\ \sqrt{(1-q)/2} & 1/\sqrt{2} \\ \sqrt{(1-q)/2} & -1/\sqrt{2} \end{pmatrix}, \qquad \beta^\star = (1, 1), \qquad y = X\beta^\star. \tag{A.21}$$

Then $X^\top X/3 = I_2$, and the scalar Lasso path is $\beta^{\mathrm{las}}(s) = (1 - 1/s)_+(1, 1)$. Nevertheless, the two row-gradient variances differ by a factor of approximately 3.11 at initialization. There is no response noise. We initialize balanced positive factors at predictor level $\alpha = 10^{-6}$, use batch-one sampling and $\eta = 0.16$, and average 4096 independent paths. Rows are sampled uniformly with replacement; each sampled row drives both factors. The base seed is 20260902, with offsets 101 and 202 for SGF and SGD, respectively.

For SGF we set $\delta = \eta = 0.16$ and use factor-coordinate Euler–Maruyama with microstep $dt = 0.002$, sharing the Gaussian covariance increment between the factors. Both SGF and SGD are recorded at

$$t_k = k\eta, \qquad s_k = \frac{2k\eta}{L_\alpha}, \qquad 1 \leq k \leq \left\lfloor \frac{4L_\alpha}{2\eta} \right\rfloor = 172,$$

so s_k ranges from 0.0231624 to 3.98393. Stochastic averages and clock integrals use pre-update/left endpoints; reported errors use trapezoidal integration over the recorded points with $s \geq 1.25$.

For completeness, let $C_i(\beta) = \mathrm{Var}_I(G_i(\beta, I) \mid \beta)$ and $M_i(\beta) = \mathbb{E}_I[G_i(\beta, I)^2 \mid \beta]$. The weighted positive-Lasso comparators soft-threshold coordinate i at

$$\lambda_i^{\mathrm{F}}(t) = \frac{1}{s} + \frac{\delta}{2t}\int_0^t C_i(\beta_u)\, du,$$

$$\lambda_{k,i}^{\mathrm{D,var}} = \frac{1}{s_k} + \frac{\eta}{2k}\sum_{j<k} C_i(\beta_j), \qquad \lambda_{k,i}^{\mathrm{D,m2}} = \frac{1}{s_k} + \frac{\eta}{2k}\sum_{j<k} M_i(\beta_j),$$

$$\lambda_{k,i}^{\mathrm{D,R}} = \frac{1}{s_k} + \frac{1}{\eta k}\sum_{j<k} R(\eta G_{j,i}), \qquad R(q) = -\log(1-q) - q.$$

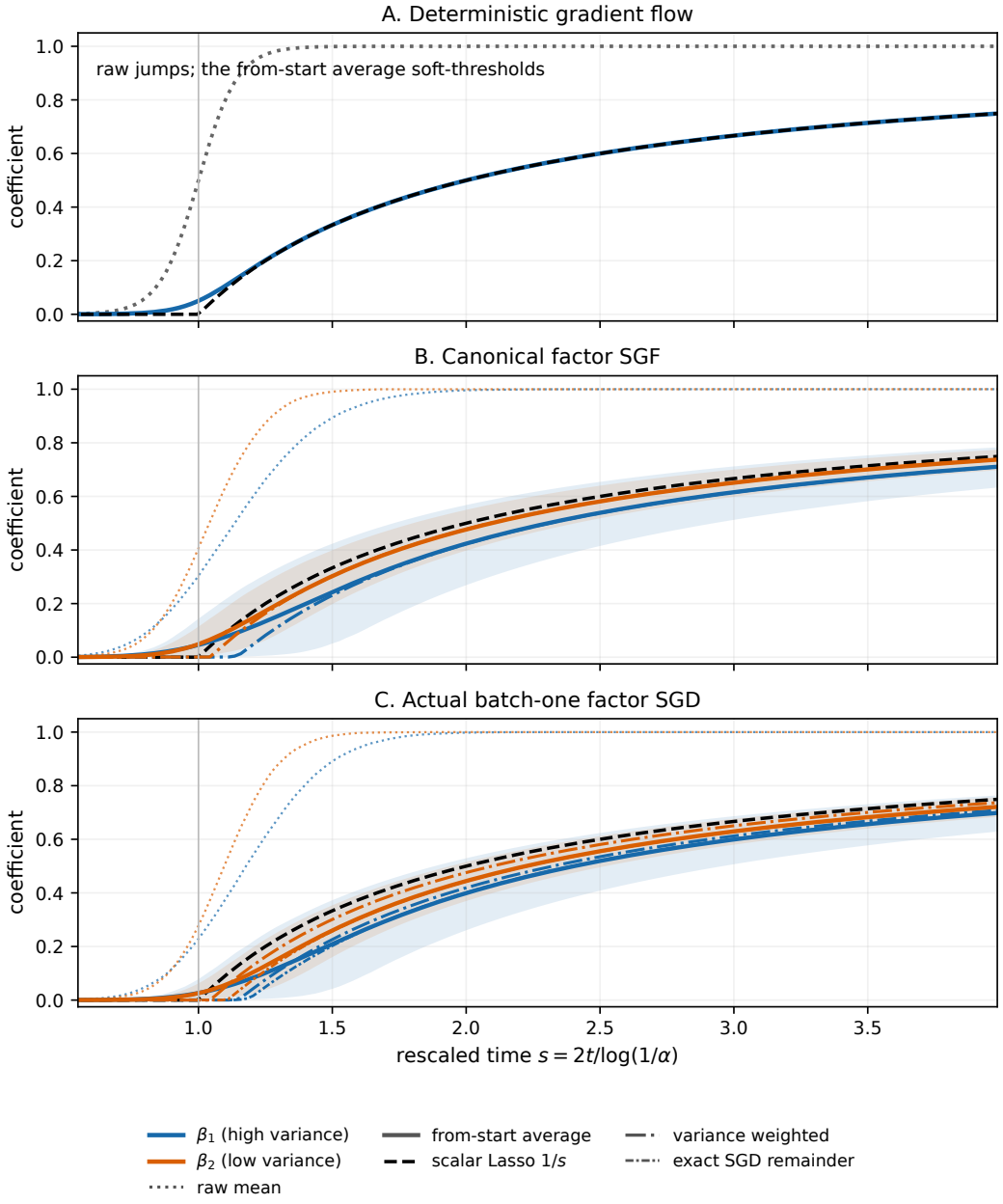


Figure 1: Raw predictors (dotted) and from-start averages (solid) in the exactly normalized two-dimensional experiment. Deterministic averaging recovers scalar Lasso (top). Fixed SGF noise breaks the scalar path's coordinate symmetry and is captured by a variance-weighted comparator (middle). Actual SGD follows a different finite-step correction (bottom). Shaded bands are pointwise 10th–90th percentiles across the 4096 stochastic trajectories.

The stochastic comparators are trajectory-specific and are averaged only after they are solved.

The deterministic integrated path error over $s \geq 1.25$ falls from 1.635 for the raw path to 1.76×10^{-4} after averaging. For SGF, the integrated ensemble-mean error is 0.176 against scalar Lasso and 0.0105 against the variance-clock comparator. For actual SGD, the corresponding scalar error is 0.259; the variance-only correction leaves error 0.0824, whereas the exact R-clock reduces it to 0.0111. The summed discrete logarithmic identity closes numerically to 9.99×10^{-16}. These diagnostics support

the algebraic distinction between fixed SGF noise and a fixed SGD step; the theoretical results do not rely on the simulation.

A.6 SIGNED FINITE-STEP PATHWISE MACHINERY

Set $\widetilde{z} = (z^+, z^-)$ and $\widetilde{w} = (w^+, w^-)$. Linearly interpolate $z_k^\pm, w_k^\pm, M_k, \mathcal{A}_k^\pm$ on the mesh $\{s_k\}$. Thus $\dot{z}^\pm = q_k^\pm$ in the interior of the k-th cell. Define the mesh-complementarity defect

$$\mathcal{J}_h^{\mathrm{sgn}}(s) = \sum_k \int_{[s_k, s_{k+1}]\cap[0,s]} \left\langle \widetilde{w}(v) - \widetilde{w}_k, (q_k^+, q_k^-) \right\rangle_+ dv. \tag{A.22}$$

Lemma A.7 (Mesh complementarity)**.** *Suppose, for $N = \lceil S/h \rceil$, that*

$$|\beta_{k,i}| \le B \quad (0 \le k \le N), \qquad |G_{k,i}| \le G_*, \quad |\eta G_{k,i}| \le \kappa_{\mathrm{step}} < 1 \quad (0 \le k < N). \tag{A.23}$$

Let

$$a_\times = \max_i \sqrt{a_i^+ a_i^-}, \qquad B_{\mathrm{ch}} = B + 2\alpha a_\times. \tag{A.24}$$

Then, for every $s \le S$,

$$\int_0^s \left\langle \widetilde{w}(v), \dot{\widetilde{z}}(v) \right\rangle dv \le \frac{2ds}{eL} + \mathcal{J}_h^{\mathrm{sgn}}(s), \tag{A.25}$$

and

$$\mathcal{J}_h^{\mathrm{sgn}}(S) \le \frac{(S+h)h\, dB_{\mathrm{ch}} G_*}{2(1-\kappa_{\mathrm{step}})} = \frac{(S+h)\eta d B_{\mathrm{ch}} G_*}{(1-\kappa_{\mathrm{step}})L}. \tag{A.26}$$

Define the comparator-channel variations

$$\mathcal{V}^\uparrow(s) = \sum_i \int_0^s \{(\dot{p}_i^*(v))_+ + (\dot{n}_i^*(v))_+\}\, dv, \tag{A.27}$$

$$\mathcal{V}^\downarrow(s) = \sum_i \int_0^s \{(\dot{p}_i^*(v))_- + (\dot{n}_i^*(v))_-\}\, dv, \tag{A.28}$$

the channel occupation and future-active clock

$$Z_{\mathrm{ch}}(s) = \left\|z^+(s)\right\|_1 + \left\|z^-(s)\right\|_1, \tag{A.29}$$

$$\mathcal{K}_{\eta,\alpha}^{\mathrm{sgn}}(s) = \sum_i \int_0^s \left\{\mathcal{A}_i^+(v)(\dot{p}_i^*(v))_+ + \mathcal{A}_i^-(v)(\dot{n}_i^*(v))_+\right\} dv. \tag{A.30}$$

Set

$$\varepsilon_0 = \max_{i,\pm} \sup_{v \le S} |c_i^\pm \pm M_i(v)|, \qquad \kappa_{\mathrm{ch}} = \frac{(\log B_{\mathrm{ch}})_+}{L}, \tag{A.31}$$

where the upper sign accompanies the positive channel, and put

$$h_{\mathrm{abs}} = \max_{i,j} |H_{ij}|, \qquad U = 1 + h_{\mathrm{abs}} Z_{\mathrm{ch}}(S) + S\left\|r\right\|_\infty. \tag{A.32}$$

Theorem A.8 (Pathwise signed perturbed-complementarity bound)**.** *Under Equation* (A.23)*, define*

$$\begin{aligned}\mathcal{R}_{\eta,\alpha}^{\mathrm{sgn}}(s) = \frac{2ds}{eL} + \mathcal{J}_h^{\mathrm{sgn}}(s) + \varepsilon_0\{Z_{\mathrm{ch}}(s) + \mathcal{V}^\uparrow(s)\} \\ + \kappa_{\mathrm{ch}}\mathcal{V}^\uparrow(s) + \mathcal{K}_{\eta,\alpha}^{\mathrm{sgn}}(s) + U\mathcal{V}^\downarrow(s).\end{aligned} \tag{A.33}$$

Then, simultaneously for all $s \in [0, S]$,

$$\frac{1}{2}\left\|z(s) - y^*(s)\right\|_H^2 \le \mathcal{R}_{\eta,\alpha}^{\mathrm{sgn}}(s). \tag{A.34}$$

For

$$\mathcal{G}_s(\beta) = \frac{1}{2}\beta^\top H\beta - r^\top\beta + \frac{1}{s}\left\|\beta\right\|_1, \qquad \overline{\beta}^{\eta,\alpha}(s) = \frac{z(s)}{s}, \tag{A.35}$$

and every $s > 0$,

$$0 \le \mathcal{G}_s(\overline{\beta}^{\eta,\alpha}(s)) - \mathcal{G}_s(\beta_s) \le \frac{\mathcal{R}^{\mathrm{sgn}}_{\eta,\alpha}(s)}{s^2} + \frac{\|r\|_{H^{-1}}\sqrt{2\mathcal{R}^{\mathrm{sgn}}_{\eta,\alpha}(s)}}{s}, \tag{A.36}$$

and

$$\left\|\overline{\beta}^{\eta,\alpha}(s) - \beta_s\right\|_H \le \frac{\sqrt{2\mathcal{R}^{\mathrm{sgn}}_{\eta,\alpha}(s)}}{s}. \tag{A.37}$$

No monotonicity assumption is needed for these pathwise inequalities. In particular, $U\mathcal{V}^{\downarrow}$ remains visible when the signed Lasso path deletes magnitude or reverses sign.

The future-active clock has the equivalent Stieltjes form

$$\mathcal{K}^{\mathrm{sgn}}_{\eta,\alpha}(s) = \sum_i \int_{[0,s]} \{\mathcal{V}^{\uparrow}_{i,+}(s) - \mathcal{V}^{\uparrow}_{i,+}(v)\}\, d\mathcal{A}^{+}_i(v) + \sum_i \int_{[0,s]} \{\mathcal{V}^{\uparrow}_{i,-}(s) - \mathcal{V}^{\uparrow}_{i,-}(v)\}\, d\mathcal{A}^{-}_i(v), \tag{A.38}$$

where $\mathcal{V}^{\uparrow}_{i,+}(s) = \int_0^s (\dot{p}^*_i(v))_+\, dv$ and $\mathcal{V}^{\uparrow}_{i,-}(s) = \int_0^s (\dot{n}^*_i(v))_+\, dv$. Hence a correction accumulated after a channel's final upward comparator motion does not enter the comparison. The coarser bound

$$\mathcal{K}^{\mathrm{sgn}}_{\eta,\alpha}(S) \le \max_{i,\pm} \mathcal{A}^{\pm}_i(S)\, \mathcal{V}^{\uparrow}(S) \tag{A.39}$$

is sufficient for Corollary 6.2.

A.7 Moderate-dimensional signed-SGD pilot

This experiment has no SGF arm. For each $\rho \in \{0.25, 0.60, 0.90\}$, we generated five datasets without post hoc selection. Before empirical column standardization, $X_i \sim N(0, \Sigma_\rho)$, where $(\Sigma_\rho)_{jk} = \rho^{|j-k|}$, with $d = 32$ and $n = 128$. Each column is then empirically centered and divided by its root mean square, so the empirical Hessian has unit diagonal. Four support coordinates were sampled uniformly without replacement and assigned amplitudes $(3/2, -7/6, 5/6, -1/2)$. With $\epsilon \sim N(0, I_n)$, we formed

$$y_0 = X\beta^0 + \tfrac{1}{2}\epsilon, \qquad c = \frac{3/2}{\|X^\top y_0/n\|_\infty}, \qquad y = cy_0, \qquad \beta^\star = c\beta^0.$$

Thus the pre-rescaling response-noise standard deviation is $1/2$; the final noise standard deviation and signal are both multiplied by c.

Each dataset–setting pair uses 16 independent trajectories. The ten frozen settings are

$$\begin{aligned}
&\text{fixed:} \quad \alpha \in \{10^{-2}, 10^{-4}, 10^{-8}\}, \quad \eta = 0.02, \quad b = 1;\\
&\text{weak-step:} \quad \alpha \in \{10^{-2}, 10^{-4}, 10^{-8}\}, \quad \eta = 0.05/\sqrt{L_\alpha}, \quad b = 1;\\
&\text{matched } \eta/b: \quad \alpha = 10^{-6}, \quad (\eta, b) \in \{(0.01, 2), (0.02, 4)\};\\
&\text{batch contrast:} \quad \alpha = 10^{-6}, \quad \eta = 0.02, \quad b \in \{1, 16\}.
\end{aligned}$$

Here b is the minibatch size. We take $a^+_i = a^-_i = 1$, so $q^+_{0,i} = q^-_{0,i} = \alpha$ and $\beta_0 = 0$. At every step and for every trajectory, the b row indices are sampled independently and uniformly with replacement; their gradients are averaged and the same minibatch drives both signed channels. No clipping is used, and all archived runs satisfy $\min_{k,i}(1 - |\eta G_{k,i}|) \ge 0.1$.

The target grid consists of 120 linearly spaced points in $[0.5, 12]$. For each setting, a target s is replaced by

$$k = \max\left\{1, \mathrm{round}\left(\frac{sL_\alpha}{2\eta}\right)\right\}, \qquad s_k = \frac{2\eta k}{L_\alpha};$$

the resulting indices are distinct. The average is the pre-update quantity $k^{-1}\sum_{j<k}\beta_j$, and all integrated errors use the trapezoidal rule on the 120 realized points.

The comparators are trajectory-specific weighted signed Lassos. For channel penalties $\lambda^{\pm}$, they minimize

$$\frac{1}{2}(p-n)^\top H(p-n) - r^\top(p-n) + (\lambda^+)^\top p + (\lambda^-)^\top n, \qquad p, n \geq 0.$$

Nominal Lasso uses $\lambda^+ = \lambda^- = \mathbf{1}/s_k$. Writing

$$v_{j,i} = \mathrm{Var}(G_{j,i} \mid \mathcal{F}_j), \qquad m_{j,i} = \mathbb{E}[G_{j,i}^2 \mid \mathcal{F}_j] = g_i(\beta_j)^2 + v_{j,i},$$

the variance and conditional-second-moment clocks are

$$A_{k,i}^{\mathrm{var}} = \frac{\eta^2}{L_\alpha} \sum_{j<k} v_{j,i}, \qquad A_{k,i}^{\mathrm{m2}} = \frac{\eta^2}{L_\alpha} \sum_{j<k} m_{j,i},$$

and use the symmetric penalties $(1 + A_{k,i}^{\mathrm{var}})/s_k$ and $(1 + A_{k,i}^{\mathrm{m2}})/s_k$. The exact finite-step comparator uses

$$A_{k,i}^{\pm} = \frac{2}{L_\alpha} \sum_{j<k} R(\pm \eta G_{j,i}), \qquad \lambda_{k,i}^{\pm} = \frac{1 + A_{k,i}^{\pm}}{s_k}, \qquad R(q) = -\log(1-q) - q.$$

Systematic error is the trapezoidal integral of the H-distance between the trajectory-ensemble mean of the averaged predictor and the corresponding ensemble-mean comparator. Mean pathwise error instead integrates the trajectory average of the H-distance between each averaged predictor and its trajectory-specific comparator.

For support diagnostics, an estimated sign is set to zero when $|\bar{\beta}_i| < 0.05$; the reference is the unthresholded nominal-Lasso sign. Precision and recall ignore signs, while exact signed support requires equality of all 32 signs. Grid points within 0.08 of a nominal-Lasso sign-change midpoint are omitted.

All 150 frozen dataset–setting experiments passed the multiplier, exact-log, product, and Lasso-solver audits. From-start averaging reduced mean pathwise nominal-Lasso error relative to the raw iterates in every experiment, with median integrated error ratio 0.159. At the predeclared setting $(\alpha, \eta, b) = (10^{-6}, 0.02, 1)$, where b is the minibatch size, the post-hoc exact two-channel R-clock comparator reduced systematic error in 14 of 15 datasets; the median per-dataset exact/scalar ratios were 0.683, 0.646, and 0.699 as ρ increased. Along $\eta_\alpha = 0.05/\sqrt{L_\alpha}$, nominal-Lasso systematic error decreased across all three tested initializations for every dataset even as $\eta_\alpha L_\alpha$ increased from 0.107 to 0.215. This finite-grid evidence is consistent with the weak-step theorem but does not establish it.

The negative diagnostics are equally important. Exact R did not improve on the conditional second-moment comparator overall, matched-effective-noise effects were mixed, and signed exact-support recovery was weak. Thus the pilot supports averaging and the clock mechanism on these frozen instances, but supports neither a universal finite-step ranking nor a support-recovery claim.

B Proofs for positive stochastic gradient flow

B.1 Factor algebra, logarithmic rescaling, and averaging

Proof of Proposition 3.1. For completeness, first apply Itô's formula to the unbalanced factors in Equation (3.3). Coordinatewise,

$$d(u_i^2 - v_i^2) = -\delta V_i(\beta)(u_i^2 - v_i^2)\, dt. \tag{B.1}$$

Thus $u_0 = v_0$ implies $u_t = v_t$, up to the lifetime of the solution. Applying the product rule to $u_i v_i$, including the common noise cross-variation, gives Equation (3.4). The quadratic variation of its ith coordinate is $4\delta\beta_i^2 V_i\, dt$; a second application of Itô's formula gives Equation (3.5).

Integrating Equation (3.5) to $T_\alpha(s)$ and using $\log \beta_{0,i} = -L_\alpha + \log a_i$ yields

$$-\frac{\log b_i^\alpha(s)}{L_\alpha} = 1 - \frac{\log a_i}{L_\alpha} + \frac{2}{L_\alpha}\int_0^{T_\alpha(s)} g_i(\beta_t)\, dt + \frac{\delta}{L_\alpha}\int_0^{T_\alpha(s)} V_i(\beta_t)\, dt - \frac{2\sqrt{\delta}}{L_\alpha} M_i(T_\alpha(s)). \tag{B.2}$$

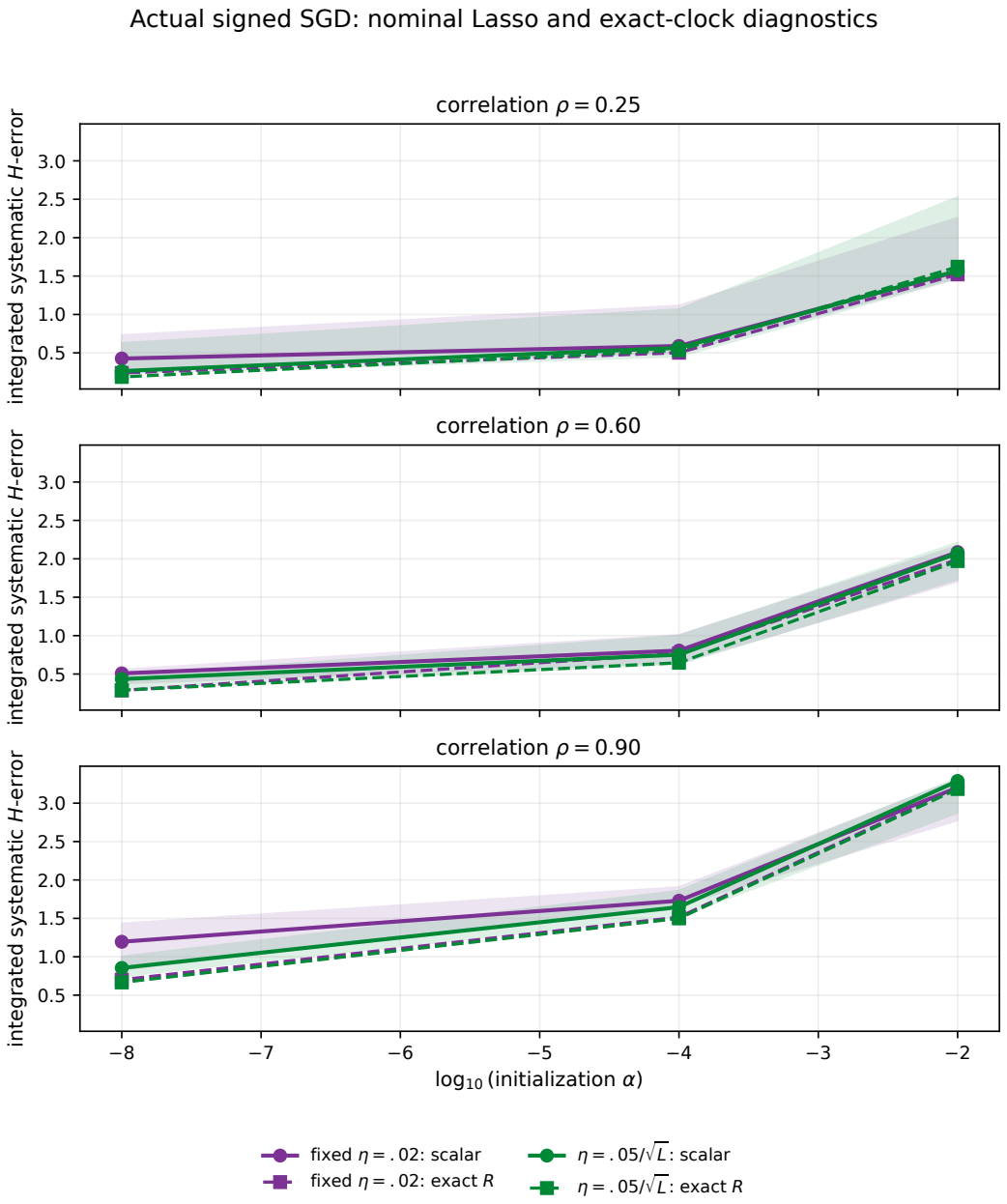


Figure 2: Unselected $d = 32, n = 128$ actual shared-minibatch signed-SGD pilot. Curves show medians over five frozen datasets per ρ, using 16-trajectory ensemble estimates and integrated error on $s \simeq [0.5, 12]$. Solid nominal-Lasso curves have empirical 5th–95th percentile bands; dashed post-hoc exact-R curves show medians only. Canonical SGF and the full benchmark are not included; the figure does not establish exact-R superiority over the second-moment diagnostic.

Affineness of g, the change of variables $t = T_\alpha(u)$, and $z^\alpha = s\overline{\beta}^\alpha$ turn the third term into $(Hz^\alpha - sr)_i$ and the fourth into $A_{\alpha,i}$. This proves Equation (3.13). The same time change gives

$$\langle N_{\alpha,i}\rangle_s = \frac{4\delta}{L_\alpha^2}\int_0^{T_\alpha(s)} V_i(\beta_t)\,dt = \frac{4}{L_\alpha}A_{\alpha,i}(s). \tag{B.3}$$

Finally, $\dot{z}^\alpha = b^\alpha$, and

$$\langle w^\alpha, \dot{z}^\alpha\rangle = -\frac{1}{L_\alpha}\sum_i b_i^\alpha \log b_i^\alpha \le \frac{d}{eL_\alpha}, \tag{B.4}$$

because $-x\log x \le 1/e$ for $x > 0$. □

Proof of the scalar raw-versus-average limit in Section 3. The deterministic predictor solves

$$\dot{\beta} = 2\beta(\mu - \beta), \qquad \beta_t = \frac{\mu}{1 + (\mu/(\alpha a) - 1)e^{-2\mu t}}. \tag{B.5}$$

At $t = T_\alpha(s)$, the denominator's second term is asymptotic to $(\mu/a)\alpha^{\mu s - 1}$, which proves the raw jump away from its critical time. Integrating the log equation gives

$$\frac{1}{T_\alpha(s)}\int_0^{T_\alpha(s)} \beta_t\,dt = \mu - \frac{\log\beta_{T_\alpha(s)} - \log(\alpha a)}{2T_\alpha(s)}. \tag{B.6}$$

Substituting the logarithmic asymptotics from Equation (B.5) proves the soft-threshold limit, including at the critical value. □

B.2 The exact axis-sampled fixed SGF-noise and SGD-step theorem

Proof of Theorem 4.1. We prove the two parts separately.

SGF. For the sampler in Equation (4.1), direct centering gives $V_i = q_i(\beta_i - \mu_i)^2$. Although different coordinates of the gradient noise are correlated, the marginal coordinate diffusion may be represented as

$$d\beta_i = \{2\beta_i(\mu_i - \beta_i) + \delta q_i\beta_i(\mu_i - \beta_i)^2\}\,dt + 2\sqrt{\delta q_i}\,\beta_i(\mu_i - \beta_i)\,dW_{t,i}. \tag{B.7}$$

Let

$$Z_i = \log\frac{\beta_i}{\mu_i - \beta_i}. \tag{B.8}$$

An exact Itô calculation yields

$$dZ_i = \{2\mu_i - \delta q_i\mu_i^2 + 3\delta q_i\mu_i\beta_i\}\,dt + 2\sqrt{\delta q_i}\mu_i\,dW_{t,i}. \tag{B.9}$$

This equation is global and maps back to $0 < \beta_i < \mu_i$. Its drift is bounded below by $\rho_i^{\mathrm{F}} > 0$; hence $Z_i(t) \to \infty$ and $\beta_i(t) \to \mu_i$ almost surely.

To locate the escape time, set $\varepsilon_\alpha = e^{-\sqrt{L_\alpha}}$ and let τ_i^α be the first time $\beta_i = \varepsilon_\alpha$. The logit distance to be crossed is $L_\alpha - \sqrt{L_\alpha} + O(1)$. Before this hit, the drift in Equation (B.9) lies between ρ_i^{F} and $\rho_i^{\mathrm{F}} + 3\delta q_i\mu_i\varepsilon_\alpha$, while the diffusion coefficient is constant. Comparison with Brownian motions having these two constant drifts sandwiches the hitting time. A Brownian motion with drift $c > 0$, diffusion σ, and target distance D has hitting-time mean D/c and variance $\sigma^2 D/c^3$. Therefore

$$\frac{\tau_i^\alpha}{L_\alpha} \xrightarrow{\mathbb{P}} \frac{1}{\rho_i^{\mathrm{F}}}. \tag{B.10}$$

The additional passage from ε_α to any fixed interior level is $o_{\mathbb{P}}(L_\alpha)$. After that passage, almost-sure convergence to μ_i, the strong Markov property, and the almost-surely finite last exit from a fixed neighborhood of μ_i show that the remaining settling time is $O_{\mathbb{P}}(1)$. This proves the claimed raw jump.

The coordinate is bounded by μ_i. Pointwise convergence at every noncritical rescaled time, followed by Fubini and dominated convergence, therefore gives

$$\overline{\beta}_i^{\alpha,\mathrm{F}}(s) \xrightarrow{\mathbb{P}} \frac{2}{s}\int_0^{s/2} \mu_i\mathbf{1}\{u > 1/\rho_i^{\mathrm{F}}\}\,du = \mu_i\left(1 - \frac{2}{s\rho_i^{\mathrm{F}}}\right)_+. \tag{B.11}$$

Since $\omega_i^{\mathrm{F}} = 2\mu_i/\rho_i^{\mathrm{F}}$, this is the coordinatewise solution of Equation (4.7).

Actual SGD. When coordinate i is selected, balanced factor SGD gives

$$\beta_{k+1,i} = \beta_{k,i}\left\{1 + \frac{\eta}{p_i}(\mu_i - \beta_{k,i})\right\}^2; \tag{B.12}$$

otherwise it is unchanged. Under $\eta\mu_i/p_i < 1/2$, the selected map sends $(0, \mu_i)$ monotonically into itself and its iterates converge to μ_i.

Until the coordinate reaches ε_α, every selected update adds between

$$2\log\left\{1 + \frac{\eta}{p_i}(\mu_i - \varepsilon_\alpha)\right\} \quad \text{and} \quad 2\log\left(1 + \frac{\eta\mu_i}{p_i}\right) \tag{B.13}$$

to $\log\beta_i$. Hence the required number of selections, divided by L_α, tends to $[2\log(1 + \eta\mu_i/p_i)]^{-1}$. If $N_i(k)$ counts selections up to step k, the uniform strong law on windows of order L_α gives $N_i(k) = p_i k + o(L_\alpha)$ almost surely. In optimization time $k\eta$, the resulting hitting constant is $1/\rho_i^{\mathrm{D}}$. Passage to a fixed level takes $o(L_\alpha)$ more updates, and monotone convergence of the selected map settles the coordinate thereafter. Thus the raw transition occurs at $s = 2/\rho_i^{\mathrm{D}}$.

The rescaled mesh is $2\eta/L_\alpha \to 0$. A bounded Riemann-sum argument now gives

$$\overline{\beta}_i^{\alpha,\mathrm{D}}(s) \longrightarrow \mu_i\left(1 - \frac{2}{s\rho_i^{\mathrm{D}}}\right)_+ = \left(\mu_i - \frac{\omega_i^{\mathrm{D}}}{s}\right)_+ \tag{B.14}$$

almost surely, including at the single critical time. This is again the coordinatewise KKT solution of Equation (4.7). □

B.3 Deterministic clock limits and the correlated example

Proof of Theorem A.1. Uniform convergence of A_α in probability makes $A_{\alpha,i}(\sup I)$ tight. On the event that this terminal clock is at most R, Equation (3.14) and optional stopping give

$$\mathbb{P}\left\{\sup_{s\in I}|N_{\alpha,i}(s)| > \varepsilon,\ A_{\alpha,i}(\sup I) \le R\right\} \le \frac{4R}{\varepsilon^2 L_\alpha}. \tag{B.15}$$

First let $\alpha \downarrow 0$, then $R \to \infty$. A finite union bound shows $N_\alpha \to 0$ uniformly on I in probability. Since $c_\alpha \to 0$, the master identity converges to Equation (A.2). Dividing that identity by s shows that its assumed inequalities are precisely the KKT conditions for Equation (A.3); strict convexity gives uniqueness.

For a candidate fixed weight ω, its scaled KKT dual is

$$w^\omega(s) = \omega + Hz(s) - sr = w(s) + \omega - \mathbf{1} - A(s). \tag{B.16}$$

The conditions $w^\omega \ge 0$ and $z \odot w^\omega = 0$ are exactly Equation (A.4). If one active coordinate has different clock values at two times, the active equality demands two different values of ω_i, which is impossible. □

Proof of Proposition A.2. We first verify that Equation (A.6) comes from a finite least-squares problem. Let $X = \sqrt{2}H^{1/2}$, and for $\zeta \in \{-1, 1\}$ set

$$y^{(\zeta)} = 2X^{-\top}(r_0\mathbf{1} - \zeta\sigma\mathbf{1}). \tag{B.17}$$

The two rows of X, paired with $y^{(+1)}$, form one block, and the same rows paired with $y^{(-1)}$ form the other. The average gradient in block ζ is

$$\frac{1}{2}X^\top(X\beta - y^{(\zeta)}) = H\beta - r_0\mathbf{1} + \zeta\sigma\mathbf{1}, \tag{B.18}$$

so its conditional covariance is $\sigma^2\mathbf{1}\mathbf{1}^\top$.

The shared rank-one Brownian noise and exchangeable initialization preserve the diagonal $\beta_1 = \beta_2 = X_t$. Since $H\mathbf{1} = K\mathbf{1}$, Equation (3.4) reduces to

$$dX_t = X_t(2r_0 + \delta\sigma^2 - 2KX_t)\,dt + 2\sigma\sqrt{\delta}X_t\,dW_t, \tag{B.19}$$

and hence

$$d \log X_t = 2(a_\delta - K X_t)\, dt + 2\sigma\sqrt{\delta}\, dW_t. \tag{B.20}$$

Its explicit positive solution is

$$X_t = \frac{\alpha e^{2a_\delta t + 2\sigma\sqrt{\delta} W_t}}{1 + 2K\alpha \int_0^t e^{2a_\delta u + 2\sigma\sqrt{\delta} W_u}\, du}. \tag{B.21}$$

For $t_\alpha = T_\alpha(s)$, Brownian scaling gives $L_\alpha^{-1} \sup_{u \le t_\alpha} |W_u| \to 0$ in probability. Upper-bounding the integral in Equation (B.21) by its largest exponential and lower-bounding it on $[(1-\varepsilon)t_\alpha, t_\alpha]$, then sending $\varepsilon \downarrow 0$, gives

$$\frac{1}{L_\alpha} \log \int_0^{t_\alpha} e^{2a_\delta u + 2\sigma\sqrt{\delta} W_u}\, du \xrightarrow{\mathbb{P}} a_\delta s. \tag{B.22}$$

It follows that

$$\frac{\log X_{t_\alpha}}{L_\alpha} \xrightarrow{\mathbb{P}} \min\{a_\delta s - 1, 0\}. \tag{B.23}$$

Integrating Equation (B.20) exactly gives

$$\frac{1}{t_\alpha} \int_0^{t_\alpha} X_t\, dt = \frac{a_\delta}{K} + \frac{\log \alpha - \log X_{t_\alpha}}{2K t_\alpha} + \frac{\sigma\sqrt{\delta}}{K} \frac{W_{t_\alpha}}{t_\alpha}. \tag{B.24}$$

The last term vanishes in probability, and Equation (B.23) proves Equation (A.8), including at activation.

Finally, write $\kappa = \delta\sigma^2/2$. On every active time, the nominal weighted-Lasso KKT equation forces $\omega_i = 1 + \kappa s$. Two different active times therefore rule out a fixed ω, whereas $a_\delta - K x_i = 1/s$ shows that the same curve is scalar Lasso at $\lambda(s) = 1/s + \kappa$. □

B.4 The sharp diagonal weak-noise bound

For reference, set

$$h_{\min} = \min_i h_i, \qquad c_i = \frac{h_i a_i}{r_i}, \qquad C_{\det} = \max_i \frac{\log 2 + |\log c_i| + \log(1 + c_i)}{h_i s_0}. \tag{B.25}$$

The sharper finite-α form of Equation (5.2) is

$$\sup_{s \in [s_0, S]} \left\| \overline{\beta}^{\alpha}(s) - x_s \right\|_\infty \le \frac{C_{\det}}{L_\alpha} + \frac{1}{h_{\min}} \left[\bar{V} \delta_\alpha \left(\frac{S}{s_0} + \frac{1}{2} \right) + \frac{6}{s_0} \sqrt{\frac{\delta_\alpha \bar{V} S \log(2d/\rho)}{L_\alpha}} \right]. \tag{B.26}$$

Lemma B.1 (Contraction under a decreasing drift)**.** *Let $b : \mathbb{R} \to \mathbb{R}$ be nonincreasing, and suppose continuous paths satisfy*

$$y(t) = y(0) + \int_0^t b(y(u))\, du + f(t), \qquad y^0(t) = y(0) + \int_0^t b(y^0(u))\, du, \qquad f(0) = 0. \tag{B.27}$$

Then

$$\sup_{t \le T} \left| y(t) - y^0(t) \right| \le 2 \sup_{t \le T} |f(t)|. \tag{B.28}$$

Proof. Let $D = y - y^0$. If $D(t) > 0$, let $\tau \le t$ be the last zero of D before t. On $(\tau, t]$, monotonicity gives $b(y) - b(y^0) \le 0$, and therefore $D(t) \le f(t) - f(\tau) \le 2 \|f\|_{\infty,[0,T]}$. Apply the same argument on an interval where $D < 0$. □

Proof of Theorem 5.1. Put $Y_i = \log \beta_i$, and let Y_i^0 solve

$$\dot{Y}_i^0 = 2(r_i - h_i e^{Y_i^0}), \qquad Y_i^0(0) = \log(\alpha a_i). \tag{B.29}$$

Its drift is decreasing. Apply Lemma B.1 to Equation (3.5) with

$$F_i(t) = -\delta_\alpha \int_0^t V_i(\beta_u)\, du + 2\sqrt{\delta_\alpha} M_i(t). \tag{B.30}$$

The martingale bracket is at most $\bar{V} S L_\alpha / 2$. The two-sided maximal inequality and a union bound show that, with probability at least $1 - \rho$,

$$\max_i \sup_{t \le T_\alpha(S)} |M_i(t)| \le \sqrt{\bar{V} S L_\alpha \log(2d/\rho)}. \tag{B.31}$$

On this event,

$$\max_i \sup_{t \le T_\alpha(S)} \left|Y_i(t) - Y_i^0(t)\right| \le \delta_\alpha \bar{V} S L_\alpha + 4\sqrt{\delta_\alpha \bar{V} S L_\alpha \log(2d/\rho)}. \tag{B.32}$$

Integrating the stochastic and deterministic log equations to $T = T_\alpha(s)$, then subtracting, gives

$$h_i \left|\overline{\beta}_i^{\alpha}(s) - \overline{\beta}_i^{0,\alpha}(s)\right| \le \frac{\left|Y_i(T) - Y_i^0(T)\right|}{s L_\alpha} + \frac{\delta_\alpha \bar{V}}{2} + \frac{\sqrt{\delta_\alpha}}{T} \left|M_i(T)\right|. \tag{B.33}$$

Uniformly over $s \in [s_0, S]$, Equations (B.31) and (B.32) bound the right side by

$$\bar{V}\delta_\alpha \left(\frac{S}{s_0} + \frac{1}{2}\right) + \frac{6}{s_0}\sqrt{\frac{\delta_\alpha \bar{V} S \log(2d/\rho)}{L_\alpha}}. \tag{B.34}$$

The deterministic logistic solution integrates explicitly:

$$\overline{\beta}_i^{0,\alpha}(s) = \frac{1}{h_i s L_\alpha} \log\left[1 + c_i\{\alpha^{1 - r_i s} - \alpha\}\right]. \tag{B.35}$$

The softplus inequality $|\log(1 + e^q) - q_+| \le \log 2$, together with the fixed multiplicative shift c_i, gives

$$\sup_{s \in [s_0, S]} \left|\overline{\beta}_i^{0,\alpha}(s) - (r_i - 1/s)_+ / h_i\right| \le \frac{C_{\mathrm{det}}}{L_\alpha}. \tag{B.36}$$

Combining Equations (B.34) and (B.36) and dividing by $h_{\min}$ proves Equation (B.26). □

B.5 Future-active comparison

Lemma B.2 (Regularity of the scaled Lasso path)**.** *For $H \succ 0$, the pair (z^*, w^*) in Equation (3.10) satisfies, for all $s, t \ge 0$,*

$$\|z^*(t) - z^*(s)\|_H^2 + \|w^*(t) - w^*(s)\|_{H^{-1}}^2 \le |t - s|^2 \|r\|_{H^{-1}}^2. \tag{B.37}$$

Consequently both paths are absolutely continuous, $\|\dot{w}^*\|_{H^{-1}} \le \|r\|_{H^{-1}}$ almost everywhere, and

$$\langle w^*(s), \dot{z}^*(s)\rangle = 0 \quad \textit{almost everywhere}. \tag{B.38}$$

Proof. Let $\Delta z = z^*(t) - z^*(s)$ and $\Delta w = w^*(t) - w^*(s)$. Cross-complementarity gives $\langle \Delta z, \Delta w\rangle \le 0$. Since $(t - s)r = H\Delta z - \Delta w$, expansion in the H^{-1} norm yields

$$|t - s|^2 \|r\|_{H^{-1}}^2 = \|\Delta z\|_H^2 + \|\Delta w\|_{H^{-1}}^2 - 2\langle \Delta z, \Delta w\rangle, \tag{B.39}$$

which proves the estimate. Absolute continuity follows. Coordinatewise complementarity, continuity, and absolute continuity then imply Equation (B.38). □

Proof of Lemma 5.2. For a continuous local martingale N, $\exp\{\lambda N_s - \lambda^2 \langle N\rangle_s / 2\}$ is a nonnegative local martingale and hence, after localization, a supermartingale. Take $\lambda = \theta L_\alpha / 2$ and use $\langle N_{\alpha,i}\rangle = 4A_{\alpha,i}/L_\alpha$. Ville's inequality gives

$$\mathbb{P}\left\{\sup_{s \le S}\{N_{\alpha,i}(s) - \theta A_{\alpha,i}(s)\} > \gamma_\alpha\right\} \le \frac{\rho}{2d}. \tag{B.40}$$

Apply the same argument to $-N_{\alpha,i}$, then union bound over coordinates. This proves Equation (5.7). □

Proof of Theorem 5.3. On the event in Lemma 5.2, the perturbation $\xi^\alpha = c_\alpha + A_\alpha - N_\alpha$ in Equation (3.13) obeys, coordinatewise,

$$-\left(\frac{A_a}{L_\alpha} + \gamma_\alpha\right) \le \xi_i^\alpha(s) \le \frac{A_a}{L_\alpha} + (1+\theta)A_{\alpha,i}(s) + \gamma_\alpha. \tag{B.41}$$

Set

$$e = z^\alpha - z^*, \qquad \widetilde{w}^\alpha = w^\alpha - \xi^\alpha = \mathbf{1} + Hz^\alpha - sr, \qquad E = \frac{1}{2}\left\|e\right\|_H^2. \tag{B.42}$$

Using Equation (B.38), differentiation gives

$$E' = \langle w^\alpha, \dot{z}^\alpha\rangle - \langle \xi^\alpha, \dot{z}^\alpha\rangle - \langle \widetilde{w}^\alpha, \dot{z}^*\rangle - \langle w^*, \dot{z}^\alpha\rangle. \tag{B.43}$$

The first term is at most $d/(eL_\alpha)$ by Equation (3.14); the last is nonpositive. Because $\dot{z}^\alpha \ge 0$, the lower bound in Equation (B.41) gives

$$-\langle \xi^\alpha, \dot{z}^\alpha\rangle \le \left(\frac{A_a}{L_\alpha} + \gamma_\alpha\right)\left\|\dot{z}^\alpha\right\|_1. \tag{B.44}$$

On the upward part of $\dot{z}^*$, write $-\widetilde{w}^\alpha = -w^\alpha + \xi^\alpha$ and use the upper bound in Equation (B.41). Its contribution is bounded by

$$\sum_i (w_i^\alpha)_-(\dot{z}_i^*)_+ + \left(\frac{A_a}{L_\alpha} + \gamma_\alpha\right)\sum_i (\dot{z}_i^*)_+ + (1+\theta)\sum_i A_{\alpha,i}(\dot{z}_i^*)_+. \tag{B.45}$$

On the downward part,

$$\widetilde{w}_i^\alpha(u) = 1 + (Hz^\alpha(u))_i - ur_i \le 1 + h_+ Z_\alpha + S\left\|r^-\right\|_\infty = U_\alpha. \tag{B.46}$$

Integrating Equation (B.43) from zero proves Equation (5.10).

For the objective, direct expansion gives

$$s^2\{\mathcal{G}_s(\overline{\beta}^\alpha(s)) - \mathcal{G}_s(x_s)\} = E(s) + \langle w^*(s), e(s)\rangle. \tag{B.47}$$

The left side is nonnegative by optimality. Differentiating the right side cancels the final term in Equation (B.43); the only new term satisfies, by Lemma B.2,

$$\langle \dot{w}^*, e\rangle \le \left\|\dot{w}^*\right\|_{H^{-1}}\left\|e\right\|_H \le \left\|r\right\|_{H^{-1}}\sqrt{2E}. \tag{B.48}$$

Since $\mathcal{R}_\alpha^{\mathrm{act}}$ is nondecreasing, integration proves Equation (5.13); division of the energy estimate by s^2 proves Equation (5.12). □

To verify Equation (5.15), substitute $A_{\alpha,i}(u) = \int_0^u dA_{\alpha,i}(v)$ into Equation (5.4) and apply Tonelli's theorem to the nonnegative integrand.

B.6 Proof of the transverse tail-cover criterion

Proof of Lemma A.3. Put

$$q_i = \frac{\beta_i}{\mu_i}, \qquad A = \mathrm{Diag}(\mu)H\,\mathrm{Diag}(\mu), \qquad d_i = (A\mathbf{1})_i = \mu_i r_i, \qquad d_{\min} = \min_i d_i > 0.$$

Since H is Stieltjes,

$$A = L_c + \mathrm{Diag}(d), \qquad c_{ij} = -A_{ij} \ge 0,$$

where L_c is the weighted graph Laplacian. For

$$F_p(\beta) = \sum_i \mu_i q_i^p,$$

Itô's formula applied to Equation (3.4) gives the exact generator identity

$$\mathscr{A}_\delta F_p = -2p(q^{\odot p})^\top A(q - \mathbf{1}) - p(1-2p)\delta\sum_i \mu_i q_i^p V_i. \tag{B.49}$$

Before estimating any summand, the complete graph contribution regroups as

$$(q^{\odot p})^\top A(q-\mathbf{1}) = \sum_i d_i q_i^p (q_i - 1) + \sum_{i<j} c_{ij}(q_i^p - q_j^p)(q_i - q_j). \tag{B.50}$$

The second term is nonnegative. If

$$c_p^\star = \sup_{x\geq 0}\left(x^p - \frac{1}{2}x^{p+1}\right) = \frac{1}{p+1}\left(\frac{2p}{p+1}\right)^p,$$

then Equations (B.49) and (B.50) imply

$$\mathscr{A}_\delta F_p \leq 2pc_p^\star \sum_i d_i - pd_{\min}\sum_i q_i^{p+1} - p(1-2p)\delta W_p(\beta). \tag{B.51}$$

Stop first on compact positive boxes. Since $q_i^p \leq 1 + q_i^{p+1}$, Equation (B.51) also implies $\mathscr{A}_\delta F_p \leq C - cF_p$ for constants $C, c > 0$. Stopped Dynkin's formula and Gronwall's inequality therefore give nonexplosion and a uniform first-moment bound for F_p. Applying Dynkin's formula once more, retaining both nonnegative terms on the left of Equation (B.51), and then removing the stopping gives

$$\frac{1}{T}\mathbb{E}\int_0^T \sum_i q_i^{p+1} dt \leq \frac{F_p(\beta_0)}{pd_{\min}T} + \frac{2c_p^\star \sum_i d_i}{d_{\min}}, \tag{B.52}$$

$$\frac{\delta}{T}\mathbb{E}\int_0^T W_p(\beta_t)dt \leq \frac{F_p(\beta_0)}{p(1-2p)T} + \frac{2c_p^\star \sum_i d_i}{1-2p}. \tag{B.53}$$

Here

$$F_p(\alpha a) = \alpha^p \sum_i \mu_i^{1-p} a_i^p,$$

so $F_p(\beta_0)/T_\alpha \to 0$. Fixed-dimensional norm equivalence gives

$$\mathsf{S}(\beta)^{p+1} \leq C\left(1 + \sum_i q_i^{p+1}\right),$$

which proves the first two bounds in Equation (A.10). Finally, for a finite dataset the sample gradients are affine in β, and hence

$$V_\Sigma(\beta) = \operatorname{tr}\operatorname{Cov}(G(\beta,\zeta) \mid \beta) \leq \mathbb{E}\,\|G(\beta,\zeta)\|_2^2 \leq C\{1 + \|\beta\|_1^2\} \leq C_V \mathsf{S}(\beta)^2.$$

□

Proof of Proposition A.4. Stop the positive process on compact boxes at

$$\tau_N = \inf\left\{t \geq 0 : \max_i \beta_i(t) \geq N \text{ or } \min_i \beta_i(t) \leq N^{-1}\right\}, \qquad \inf \varnothing = \infty, \tag{B.54}$$

and write

$$X_{\alpha,N} = \frac{\delta_\alpha}{T_\alpha}\mathbb{E}\int_0^{T_\alpha \wedge \tau_N} V_\Sigma(\beta_t)\, dt, \tag{B.55}$$

$$Y_{\alpha,N} = \frac{1}{T_\alpha}\mathbb{E}\int_0^{T_\alpha \wedge \tau_N} \mathsf{S}(\beta_t) V_\Sigma(\beta_t) \mathbf{1}\{\beta_t \in \mathscr{L}\}\, dt. \tag{B.56}$$

Dynkin's formula for Equation (A.12), nonnegativity of the terminal value, and the budgets in Equation (A.10) give, uniformly in N,

$$Y_{\alpha,N} \leq C\{1 + X_{\alpha,N}\}. \tag{B.57}$$

Fix M larger than the compact-set threshold. On $\mathsf{S}(\beta) \leq M$, use $\delta V_\Sigma \leq C\delta M^2$. In the low-output region, use the first occupation budget; in the visible region with $\mathsf{S}(\beta) > M$, use $\delta V_\Sigma \leq c_{\text{vis}}^{-1} M^{-p}\delta W_p$; and in the leakage region with $\mathsf{S}(\beta) > M$, use $\delta V_\Sigma \leq (\delta/M)\mathsf{S}(\beta)V_\Sigma$. Because these regions cover the tail, summing the bounds and invoking Equation (B.57) yields

$$X_{\alpha,N} \leq C\{\delta_\alpha M^2 + \delta_\alpha + M^{-p}\} + \frac{C\delta_\alpha}{M}\{1 + X_{\alpha,N}\}. \tag{B.58}$$

The last multiple is absorbed for small α. Choosing $M = \delta_\alpha^{-1/(p+2)}$ balances $\delta_\alpha M^2 = M^{-p}$ and proves a bound uniform in N of order $\delta_\alpha^{p/(p+2)}$. Nonexplosion and monotone convergence remove the stopping.

Finally, the time change in Equation (3.7) gives

$$\mathbb{E}\max_i A_{\alpha,i}(S) \le \frac{\delta_\alpha}{L_\alpha}\mathbb{E}\int_0^{T_\alpha} V_\Sigma(\beta_t)\,dt = \frac{S}{2}\left\{\frac{\delta_\alpha}{T_\alpha}\mathbb{E}\int_0^{T_\alpha} V_\Sigma(\beta_t)\,dt\right\}. \tag{B.59}$$

This proves the clock conclusion and, by $\mathcal{K}_\alpha^{\mathrm{act}}(S) \le \max_i A_{\alpha,i}(S)\sum_i \mathcal{V}_i^{\uparrow}(S)$, the future-active-clock conclusion. □

B.7 Axis geometry, exact-face clocks, and patching no-go results

Proof of the rigidity claim in Theorem A.5. Let $F = \mathrm{supp}(v)$ and $v_F > 0$. Zero within-support variance means, for every sampled row,

$$x_{AF}(x_{AF}^\top v_F) = H_{FF}v_F. \tag{B.60}$$

The right side is nonzero because $H_{FF} \succ 0$. Hence every x_{AF} lies in the one-dimensional span of $H_{FF}v_F$. But then $H_{FF} = \mathbb{E}[x_{AF}x_{AF}^\top]$ has rank at most one. Positive definiteness forces $|F| = 1$. For $F = \{j\}$, Equation (B.60) is exactly $x_{Aj}^2 = H_{jj}$ almost surely. □

Proof of local dissipation in Theorem A.5. Place the leaking coordinate first, set $s = (1, -\mathbf{1})$, and write

$$\beta = (X, u), \qquad Z_A = \begin{pmatrix} 0 & a_A^\top \\ a_A & B_A \end{pmatrix}, \qquad D_j = s^\top\beta = X - \mathbf{1}^\top u. \tag{B.61}$$

For the empirical least-squares sampler, also put

$$c_A = x_A y_A - r, \qquad h_A(\beta) = Z_A\beta - c_A.$$

Symmetry gives the exact homogeneous-noise cancellation

$$s^\top \mathrm{Diag}(\beta)Z_A\beta = Xa_A^\top u - u^\top(Xa_A + B_A u) = -u^\top B_A u. \tag{B.62}$$

Consequently the homogeneous stochastic part of $\mathscr{A}_\delta D_j^2$ is $\delta\mathfrak{B}_j(X,u)$, where

$$\begin{aligned}\mathfrak{B}_j(X,u) = 2D_j\,\mathbb{E}\left[X(a_A^\top u)^2 - \sum_i u_i\{Xa_{A,i} + (B_A u)_i\}^2\right] \\ + 4\mathbb{E}(u^\top B_A u)^2.\end{aligned} \tag{B.63}$$

For $X > 0$, set

$$w = \frac{u}{X}, \qquad \rho = \mathbf{1}^\top w, \qquad I = \{i : \mathbb{E}a_{A,i}^2 > 0\}, \qquad I_0 = I^c, \qquad m = \sum_{i\in I} w_i.$$

Choose $K < \infty$ so that, for every sampled row,

$$\|a_A\|_\infty \le K, \qquad \|B_A\|_{1\to\infty} \le K.$$

Then $\|B_A w\|_\infty \le K\rho$. Define

$$\begin{aligned} T &= \mathbb{E}(a_A^\top w)^2, & C_I &= \mathbb{E}\sum_{i\in I} w_i\{a_{A,i} + (B_A w)_i\}^2, \\ C_0 &= \mathbb{E}\sum_{i\in I_0} w_i(B_A w)_i^2, & Q &= \mathbb{E}(w^\top B_A w)^2. \end{aligned} \tag{B.64}$$

Dividing Equation (B.63) by $2X^4$ gives the exact regrouping

$$\frac{\mathfrak{B}_j(X,u)}{2X^4} = (1-\rho)(T - C_I - C_0) + 2Q, \tag{B.65}$$

Suppose first that $I \neq \varnothing$, and let $\sigma = \min_{i \in I} \mathbb{E} a_{A,i}^2 > 0$. Weighted Cauchy–Schwarz, applied only after the exact regrouping above, yields

$$\begin{aligned} T &\leq \rho\, \mathbb{E} \sum_{i \in I} w_i a_{A,i}^2 \leq K^2 \rho m, \\ C_I &\geq \sum_{i \in I} w_i \left\{ \frac{1}{2} \mathbb{E} a_{A,i}^2 - \mathbb{E}(B_A w)_i^2 \right\} \geq \frac{\sigma}{2} m - K^2 \rho^2 m, \\ Q &\leq 2\mathbb{E} \left(\sum_{i \in I_0} w_i (B_A w)_i \right)^2 + 2\mathbb{E} \left(\sum_{i \in I} w_i (B_A w)_i \right)^2 \\ &\leq 2\rho C_0 + 2K^2 \rho^2 m^2. \end{aligned} \tag{B.66}$$

Substitution into Equation (B.65) gives

$$\begin{aligned} \frac{\mathfrak{B}_j(X,u)}{2X^4} \leq (-1+5\rho) C_0 - \frac{(1-\rho)\sigma}{2} m \\ + (1-\rho) K^2 (\rho + \rho^2) m + 4K^2 \rho^2 m^2. \end{aligned} \tag{B.67}$$

Because $m \leq \rho$, the right side is nonpositive for all sufficiently small fixed ρ. If $I = \varnothing$, then $m = T = C_I = 0$, while the same weighted Cauchy–Schwarz argument gives $Q \leq 2\rho C_0$; hence Equation (B.65) is at most $(-1+5\rho)C_0$. Thus there is an $\varepsilon_j > 0$ for which $\mathfrak{B}_j(X,u) \leq 0$ whenever $\mathbf{1}^\top u \leq \varepsilon_j X$.

It remains to retain the affine and deterministic terms. With $V = \operatorname{diag} \Sigma$, the full generator is

$$\mathscr{A}_\delta D_j^2 = -4 D_j\, s^\top \operatorname{Diag}(\beta)(H\beta - r) + 2\delta D_j\, s^\top \operatorname{Diag}(\beta) V(\beta) + 4\delta\, \mathbb{E}\{ s^\top \operatorname{Diag}(\beta) h_A(\beta) \}^2. \tag{B.68}$$

The complete deterministic quadratic regroups exactly as

$$s^\top \operatorname{Diag}(\beta) H \beta = H_{jj} X^2 - u^\top H_{-j,-j} u. \tag{B.69}$$

In a sufficiently narrow cone, $D_j \geq X/2$ and $H_{jj} X^2 - u^\top H_{-j,-j} u \geq H_{jj} X^2/2$. The first term of Equation (B.68) is therefore $-cX^3 + O(X^2)$, where the $O(X^2)$ term is the contribution of r.

For completeness, let

$$Q_i^h(\beta) = \mathbb{E}(Z_A \beta)_i^2, \qquad \Delta_i(\beta) = V_i(\beta) - Q_i^h(\beta) = -2\mathbb{E}[(Z_A\beta)_i c_{A,i}] + \mathbb{E} c_{A,i}^2.$$

On the fixed cone, finiteness of the row ensemble gives

$$\max_i |\Delta_i(\beta)| \leq C(1+X).$$

Moreover, if

$$U_A = s^\top \operatorname{Diag}(\beta) Z_A \beta, \qquad \Gamma_A = s^\top \operatorname{Diag}(\beta) c_A,$$

then Equation (B.62) gives $|U_A| \leq CX^2$, while $|\Gamma_A| \leq CX$. Hence the difference between the affine stochastic terms in Equation (B.68) and their homogeneous counterpart is exactly

$$2\delta D_j\, s^\top \operatorname{Diag}(\beta) \Delta(\beta) + 4\delta \mathbb{E}\{ -2U_A \Gamma_A + \Gamma_A^2 \},$$

whose absolute value is at most

$$C\delta(X^3 + X^2). \tag{B.70}$$

Since the complete homogeneous quartic is nonpositive, choosing the cone first and then $\delta_j > 0$ sufficiently small gives

$$\mathscr{A}_\delta D_j^2 \leq -c_0 X^3 + C_0 X^2, \qquad \mathbf{1}^\top u \leq \varepsilon_j X, \quad 0 < \delta \leq \delta_j. \tag{B.71}$$

Finally, finite-data affine covariance and the cone relation imply

$$\mathsf{S}(\beta) V_\Sigma(\beta) \leq C(1 + X^3).$$

Taking $c_j > 0$ sufficiently small in Equation (B.71) and enlarging C_j on the remaining compact set proves Equation (A.17). □

Exact singleton-face clock. On the invariant face $\beta = Xe_j$, condition Equation (A.15) makes the active centered-gradient variance a constant σ_j^2, while every output variance is a quadratic polynomial satisfying $V_i(Xe_j) \le C_i(1 + X^2)$. The active process is

$$dX = X\{2r_j + \delta\sigma_j^2 - 2H_{jj}X\}\,dt + 2\sigma_j\sqrt{\delta}X\,dW. \tag{B.72}$$

The required moment identities follow first by stopping. Indeed, for the generator of Equation (B.72),

$$\mathscr{A}_\delta X^2 = (4r_j + 6\delta\sigma_j^2)X^2 - 4H_{jj}X^3. \tag{B.73}$$

Stopping on compact intervals, dropping the negative cubic term, and applying Gronwall's inequality gives compact-time second-moment bounds, uniformly for $0 < \delta \le \delta_0$ and bounded X_0. Letting the stops tend to infinity therefore makes the stochastic integral in the first-moment identity a true martingale. That identity and scalar logistic comparison give $\sup_{t\ge 0}\mathbb{E}X_t \le C$, with the same uniformity. Consequently,

$$2H_{jj}\mathbb{E}\int_0^T X_t^2 dt = X_0 - \mathbb{E}X_T + (2r_j + \delta\sigma_j^2)\int_0^T \mathbb{E}X_t dt \le X_0 + CT. \tag{B.74}$$

This proves Equation (A.18).

Proof of Proposition A.6. For a two-homogeneous C^2 function, define the homogeneous stochastic symbol

$$\mathfrak{S}_\Psi(v) = \nabla\Psi(v)^\top \operatorname{Diag}(v)q(v) + 2\operatorname{tr}\big\{\operatorname{Diag}(v)\Sigma_h(v)\operatorname{Diag}(v)\nabla^2\Psi(v)\big\}, \tag{B.75}$$

where $q = \operatorname{diag}\Sigma_h$ and $\Sigma_h(v) = \mathbb{E}[(Z_A v)(Z_A v)^\top]$. This is the coefficient of $\delta\varrho^4$ along the ray ϱv.

For Rademacher rows and $v_{jk} = (e_j + e_k)/2$, the active noise is radial and

$$\operatorname{Diag}(v_{jk})q(v_{jk}) = \frac{1}{4}v_{jk}, \qquad \operatorname{Diag}(v_{jk})\Sigma_h(v_{jk})\operatorname{Diag}(v_{jk}) = \frac{1}{4}v_{jk}v_{jk}^\top. \tag{B.76}$$

Euler's identities $\nabla\Psi(v)^\top v = 2\Psi(v)$ and $v^\top\nabla^2\Psi(v)v = 2\Psi(v)$ give Equation (A.19).

For $\Psi_M(\beta) = \beta^\top M\beta$, put $q_{jk} = (e_j + e_k)^\top M(e_j + e_k)$. The exact sum identity

$$\sum_{j<k} q_{jk} = (d-2)\operatorname{tr}M + \mathbf{1}^\top M\mathbf{1} \tag{B.77}$$

and $M \succeq 0$, $\operatorname{tr}M \ge dm$, show that some pair has $q_{jk} \ge 2(d-2)m/(d-1)$. Since $\Psi_M(v_{jk}) = q_{jk}/4$, Equation (A.20) follows.

For the finite-order statement, let $\ell(\beta) = n^\top\beta$, let $\ell(v_0) = 0$, and suppose the boundary is noncharacteristic:

$$\kappa(v_0, n) = \mathbb{E}\{n^\top \operatorname{Diag}(v_0)Z_A v_0\}^2 > 0. \tag{B.78}$$

Along a fixed transverse line $v(t) = v_0 + tu$ with $n^\top u > 0$ and $t \downarrow 0$, take, on the side $\ell > 0$, $\Psi = a(\beta)\ell(\beta)^m$, where $m \ge 2$, $a > 0$, and a is homogeneous of degree $2 - m$. Expanding its Hessian in Equation (B.75) gives

$$\mathfrak{S}_\Psi(v) = 2m(m-1)a(v_0)\kappa(v_0, n)\ell(v)^{m-2} + O(\ell(v)^{m-1}) > 0, \tag{B.79}$$

immediately inside the boundary. This positive $\delta\varrho^4$ term cannot be charged to the available $O(\delta\varrho^{p+2})$ fractional budget.

Finally, restrict the Rademacher model to $d = 2$, write $\beta = (x, u)$, and set

$$e_1 = x - \mu_1, \qquad e_2 = u - \mu_2.$$

Direct centering of the sampled gradient gives

$$V_1 = e_2^2, \qquad V_2 = e_1^2, \qquad \Sigma_{12} = e_1e_2.$$

For $F(x, u) = x\phi(u)$, the predictor generator is therefore exactly

$$\begin{aligned}\mathscr{A}_\delta F(x,u) = {}& \phi(u)\{-2xe_1 + \delta xe_2^2\} + x\phi'(u)\{-2ue_2 + \delta ue_1^2\} \\ & + 4\delta xue_1e_2\phi'(u) + 2\delta xu^2e_1^2\phi''(u).\end{aligned} \tag{B.80}$$

Consequently, for every fixed $u > 0$,

$$\mathscr{A}_\delta F(x,u) = \delta x^3\{u\phi'(u) + 2u^2\phi''(u)\} + O_{u,\delta,\mu,\phi}(x^2), \qquad x \to \infty. \tag{B.81}$$

If the bracket were nonpositive for every u, then with $y = -\phi' \geq 0$,

$$\frac{d}{du}\{\sqrt{u}\, y(u)\} \geq 0. \tag{B.82}$$

If $y(u_0) > 0$, this implies $y(u) \geq y(u_0)\sqrt{u_0/u}$ for $u \geq u_0$, contradicting global nonnegativity of the decreasing ϕ. Hence ϕ would be constant. Every nonconstant admissible profile therefore has a ray with positive leading symbol, completing the proof. □

C Proofs for the signed discrete theory

This appendix proves the statements in Section 6. We keep the two nonnegative channels explicit throughout. In particular, every energy estimate is made after applying $P = (I, -I)$; no inverse of the singular matrix $P^\top HP$ is used.

C.1 Rotation and common-channel identities

Proof of Lemma 6.1. From Equations (6.3) and (6.4),

$$u^+_{k+1} = \frac{a_k + b_k - \eta(a_k + b_k) \odot G_k}{2} = u^+_k \odot (\mathbf{1} - \eta G_k),$$
$$u^-_{k+1} = \frac{a_k - b_k + \eta(a_k - b_k) \odot G_k}{2} = u^-_k \odot (\mathbf{1} + \eta G_k).$$

Squaring coordinatewise proves

$$q^+_{k+1,i} = q^+_{k,i}(1 - \eta G_{k,i})^2, \qquad q^-_{k+1,i} = q^-_{k,i}(1 + \eta G_{k,i})^2. \tag{C.1}$$

The positive-step assumption makes both logarithms below well defined. The identities

$$\log(1-q) = -q - R(q), \qquad \log(1+q) = q - R(-q) \tag{C.2}$$

therefore give

$$\log q^+_{k+1,i} - \log q^+_{k,i} = -2\eta G_{k,i} - 2R(\eta G_{k,i}), \tag{C.3}$$
$$\log q^-_{k+1,i} - \log q^-_{k,i} = +2\eta G_{k,i} - 2R(-\eta G_{k,i}). \tag{C.4}$$

Summing, using $q^\pm_{0,i} = \alpha a^\pm_i$, and recalling that $h = 2\eta/L$, yields

$$w^+_k = \mathbf{1} + c^+ + h\sum_{j<k} G_j + \mathcal{A}^+_k,$$
$$w^-_k = \mathbf{1} + c^- - h\sum_{j<k} G_j + \mathcal{A}^-_k.$$

Because $G_j = g(\beta_j) + \zeta_j$, the affine form of g gives

$$h\sum_{j<k} G_j = H\left(h\sum_{j<k}\beta_j\right) - khr + h\sum_{j<k}\zeta_j = Hz_k - s_k r + M_k. \tag{C.5}$$

This proves Equations (6.9) and (6.10), including both martingale signs.

Multiplying the two identities in Equation (C.1) gives

$$q^+_{k,i}q^-_{k,i} = \alpha^2 a^+_i a^-_i \prod_{j<k}(1-\eta^2 G^2_{j,i})^2$$
$$= \alpha^2 a^+_i a^-_i \exp\{-L(\mathcal{A}^+_{k,i} + \mathcal{A}^-_{k,i})\}, \tag{C.6}$$

where the second equality uses $R(q) + R(-q) = -\log(1-q^2)$. Finally,

$$(q_{k,i}^+ + q_{k,i}^-)^2 = (q_{k,i}^+ - q_{k,i}^-)^2 + 4q_{k,i}^+ q_{k,i}^-, \tag{C.7}$$

and $\sqrt{x^2+y^2} \le |x| + |y|$ imply, with $\vartheta_k = q_k^+ + q_k^-$,

$$\vartheta_{k,i} \le |\beta_{k,i}| + 2\alpha\sqrt{a_i^+ a_i^-}. \tag{C.8}$$

It remains to verify the clock comparison. For $0 \le q \le \kappa_{\mathrm{step}}$,

$$R(q) = \int_0^q \frac{t}{1-t}\,dt, \qquad \frac{q^2}{2} \le R(q) \le \frac{q^2}{2(1-\kappa_{\mathrm{step}})}.$$

For $q = -a$, $0 \le a \le \kappa_{\mathrm{step}}$,

$$R(-a) = \int_0^a \frac{t}{1+t}\,dt, \qquad \frac{a^2}{2(1+\kappa_{\mathrm{step}})} \le R(-a) \le \frac{a^2}{2}.$$

Consequently, for either sign and every $|q| \le \kappa_{\mathrm{step}}$,

$$\frac{q^2}{2(1+\kappa_{\mathrm{step}})^2} \le R(q) \le \frac{q^2}{2(1-\kappa_{\mathrm{step}})}. \tag{C.9}$$

Insert $q = \pm\eta G_{j,i}$, sum over $j < k$, and use $s_k = 2k\eta/L$ to obtain, for either channel,

$$\frac{s_k\eta}{2(1+\kappa_{\mathrm{step}})^2}\left(\frac{1}{k}\sum_{j<k} G_{j,i}^2\right) \le \mathcal{A}_{k,i}^{\pm} \le \frac{s_k\eta}{2(1-\kappa_{\mathrm{step}})}\left(\frac{1}{k}\sum_{j<k} G_{j,i}^2\right). \tag{C.10}$$

□

C.2 Signed-Lasso geometry and mesh complementarity

We first record the predictor-space regularity used by the energy argument.

Lemma C.1 (Joint predictor–dual regularity)**.** *Let $y^*(s)$ and $\nu^*(s)$ be defined by Equations* (6.12) *and* (6.13)*. For all $s, t \ge 0$,*

$$\|y^*(t) - y^*(s)\|_H^2 + \|\nu^*(t) - \nu^*(s)\|_{H^{-1}}^2 \le |t-s|^2 \|r\|_{H^{-1}}^2. \tag{C.11}$$

In particular, y^*, ν^*, p^*, n^* are absolutely continuous,

$$\|\dot\nu^*(s)\|_{H^{-1}} \le \|r\|_{H^{-1}} \quad \textit{for almost every } s, \tag{C.12}$$

and

$$\langle \mathbf{1} - \nu^*, \dot p^*\rangle + \langle \mathbf{1} + \nu^*, \dot n^*\rangle = 0 \quad \textit{for almost every } s. \tag{C.13}$$

Moreover, $\widetilde{z}^* = (p^*, n^*)$ is the unique optimizer of the lifted signed-Lasso problem.

Proof. The KKT condition for the scaled signed-Lasso objective is

$$Hy^*(s) - sr + \nu^*(s) = 0, \qquad \nu^*(s) \in \partial\|y^*(s)\|_1. \tag{C.14}$$

For $\Delta y = y^*(t) - y^*(s)$ and $\Delta\nu = \nu^*(t) - \nu^*(s)$, subtracting the two KKT equations gives $(t-s)r = H\Delta y + \Delta\nu$. Monotonicity of $\partial\|\cdot\|_1$ gives $\langle \Delta y, \Delta\nu\rangle \ge 0$, and hence

$$\begin{aligned} |t-s|^2\|r\|_{H^{-1}}^2 &= \|H\Delta y + \Delta\nu\|_{H^{-1}}^2 \\ &= \|\Delta y\|_H^2 + \|\Delta\nu\|_{H^{-1}}^2 + 2\langle \Delta y, \Delta\nu\rangle. \end{aligned}$$

This proves Equation (C.11), and Equation (C.12) follows by differentiation.

The subgradient condition gives $\mathbf{1} - \nu^* \ge 0$, $\mathbf{1} + \nu^* \ge 0$, and static complementarity with p^* and n^*, respectively. A nonnegative absolutely continuous function has derivative zero almost everywhere on its zero set; applying this coordinatewise proves Equation (C.13).

Finally, for every $x^+, x^- \ge 0$,

$$\mathbf{1}^\top(x^+ + x^-) \ge \|x^+ - x^-\|_1, \tag{C.15}$$

with equality only for a coordinatewise disjoint pair. The predictor objective is strictly convex because $H \succ 0$, so its optimizer is unique, and Equation (C.15) uniquely selects $((y^*)_+, (-y^*)_+)$ among its channel representations. □

Proof of Lemma A.7. At a mesh point,

$$\left\langle \widetilde{w}_k, (q_k^+, q_k^-) \right\rangle = -\frac{1}{L} \sum_i \{ q_{k,i}^+ \log q_{k,i}^+ + q_{k,i}^- \log q_{k,i}^- \} \le \frac{2d}{eL}, \tag{C.16}$$

where we used $-x \log x \le 1/e$ for $x > 0$. On each mesh cell, $\dot{\widetilde{z}} = (q_k^+, q_k^-)$, so adding the positive part of the interpolation error and integrating proves Equation (A.25).

By Equations (A.23) and (C.8),

$$\sum_i (q_{k,i}^+ + q_{k,i}^-) \le dB_{\mathrm{ch}}. \tag{C.17}$$

Also, for both signs,

$$|w_{k+1,i}^{\pm} - w_{k,i}^{\pm}| \le \frac{2\eta |G_{k,i}|}{L(1-\kappa_{\mathrm{step}})} \le \frac{hG_*}{1-\kappa_{\mathrm{step}}}, \tag{C.18}$$

because $|\log(1+x)| \le |x|/(1-\kappa_{\mathrm{step}})$ for $|x| \le \kappa_{\mathrm{step}}$. The interpolation error grows linearly from zero on a cell. Therefore a full cell contributes at most

$$\frac{h^2 d B_{\mathrm{ch}} G_*}{2(1-\kappa_{\mathrm{step}})}$$

to $\mathcal{J}_h^{\mathrm{sgn}}$; the same expression bounds a partial cell. Since $\lceil S/h \rceil h \le S + h$, summing gives exactly Equation (A.26). □

C.3 Prediction-energy and objective comparison

Proof of Theorem A.8. Write

$$\widetilde{z} = (z^+, z^-), \qquad \widetilde{e} = \widetilde{z} - \widetilde{z}^*, \qquad e = P\widetilde{e} = z - y^*, \tag{C.19}$$

and

$$\widetilde{\xi} = (c^+ + M + \mathcal{A}^+, c^- - M + \mathcal{A}^-), \qquad \widehat{\widetilde{w}} = \widetilde{w} - \widetilde{\xi}. \tag{C.20}$$

Linear interpolation preserves the exact affine identity, so

$$\widehat{\widetilde{w}} = \mathbf{1} + \widetilde{H}\widetilde{z} - s\widetilde{r}, \qquad \widehat{\widetilde{w}} - \widetilde{w}^* = \widetilde{H}\widetilde{e}. \tag{C.21}$$

Define the semidefinite energy

$$E(s) = \frac{1}{2} \widetilde{e}(s)^\top \widetilde{H} \widetilde{e}(s) = \frac{1}{2} \left\| e(s) \right\|_H^2. \tag{C.22}$$

Using Equation (C.13), differentiation gives, almost everywhere,

$$E' = \left\langle \widetilde{w}, \dot{\widetilde{z}} \right\rangle - \left\langle \widetilde{\xi}, \dot{\widetilde{z}} \right\rangle - \left\langle \widehat{\widetilde{w}}, \dot{\widetilde{z}}^* \right\rangle - \left\langle \widetilde{w}^*, \dot{\widetilde{z}} \right\rangle. \tag{C.23}$$

The last term is nonpositive because $\widetilde{w}^* \ge 0$ and $\dot{\widetilde{z}} \ge 0$. The time integral of the first term is bounded by Equation (A.25).

Decompose

$$\widetilde{\xi} = (c^+ + M, c^- - M) + (\mathcal{A}^+, \mathcal{A}^-).$$

The clocks and $\dot{\widetilde{z}}$ are coordinatewise nonnegative, so

$$-\left\langle \widetilde{\xi}, \dot{\widetilde{z}} \right\rangle \le \varepsilon_0 \left\| \dot{\widetilde{z}} \right\|_1. \tag{C.24}$$

At every mesh endpoint, the channel bound gives $q_{k,i}^{\pm} \le B_{\mathrm{ch}}$, and linear interpolation therefore preserves

$$w_i^{\pm} \ge -\kappa_{\mathrm{ch}}. \tag{C.25}$$

On an upward comparator channel this yields

$$-\widehat{w}_i^{\pm} = -w_i^{\pm} + (c_i^{\pm} \pm M_i) + \mathcal{A}_i^{\pm} \le \kappa_{\mathrm{ch}} + \varepsilon_0 + \mathcal{A}_i^{\pm}. \tag{C.26}$$

On a downward comparator channel, the unperturbed affine system gives

$$\widehat{w}_i^{\pm} \le 1 + h_{\mathrm{abs}} Z_{\mathrm{ch}}(S) + S \left\| r \right\|_\infty = U. \tag{C.27}$$

Integrate Equation (C.23), use $E(0) = 0$, and split the comparator velocities into their positive and negative parts. Apply Equation (C.24) to the primal term, Equation (C.26) to upward comparator motion, and Equation (C.27) to downward motion. The resulting right-hand side is precisely Equation (A.33); this proves Equation (A.34).

For the objective estimate, introduce the lifted objective

$$\widetilde{\mathcal{G}}_s(x^+, x^-) = \frac{1}{2}\left\|P(x^+, x^-)\right\|_H^2 - r^\top P(x^+, x^-) + \frac{1}{s}\mathbf{1}^\top(x^+ + x^-). \tag{C.28}$$

Expansion at the disjoint lifted optimum gives the exact identity

$$s^2\left\{\widetilde{\mathcal{G}}_s(\widetilde{z}/s) - \widetilde{\mathcal{G}}_s(\widetilde{z}^*/s)\right\} = E + \langle \widetilde{w}^*, \widetilde{e}\rangle =: Q(s). \tag{C.29}$$

Differentiating Q cancels the last term in Equation (C.23). The only new term is

$$\left\langle \dot{\widetilde{w}}^*, \widetilde{e}\right\rangle = -\langle \dot{\nu}^*, e\rangle \le \|\dot{\nu}^*\|_{H^{-1}}\|e\|_H \le \|r\|_{H^{-1}}\sqrt{2E}, \tag{C.30}$$

where the last inequality is Equation (C.12). The remainder in Equation (A.33) is nondecreasing. Hence

$$Q(s) \le \mathcal{R}^{\mathrm{sgn}}_{\eta,\alpha}(s) + s\|r\|_{H^{-1}}\sqrt{2\mathcal{R}^{\mathrm{sgn}}_{\eta,\alpha}(s)}. \tag{C.31}$$

For every nonnegative channel pair,

$$\mathcal{G}_s(P(x^+, x^-)) \le \widetilde{\mathcal{G}}_s(x^+, x^-), \tag{C.32}$$

with equality for the disjoint comparator. Combining Equations (C.29), (C.31) and (C.32) proves Equation (A.36); nonnegativity follows from optimality of β_s. Finally, Equation (A.37) follows directly from Equation (A.34) and $z = s\overline{\beta}^{\eta,\alpha}$, $y^* = s\beta_s$. □

For completeness, we verify the future-active representation. Let the clocks denote their continuous linear interpolants and let $d\mathcal{V}^{\uparrow}_{i,+} = (\dot{p}^*_i)_+ dv$, with the analogous definition for the negative channel. Each clock and each comparator variation starts at zero. Stieltjes integration by parts gives

$$\begin{aligned}\int_0^s \mathcal{A}^+_i(v)\, d\mathcal{V}^{\uparrow}_{i,+}(v) &= \mathcal{A}^+_i(s)\mathcal{V}^{\uparrow}_{i,+}(s) - \int_{[0,s]} \mathcal{V}^{\uparrow}_{i,+}(v)\, d\mathcal{A}^+_i(v)\\ &= \int_{[0,s]} \{\mathcal{V}^{\uparrow}_{i,+}(s) - \mathcal{V}^{\uparrow}_{i,+}(v)\}\, d\mathcal{A}^+_i(v).\end{aligned}$$

The same calculation for the negative channel proves Equation (A.38). Since both integrands are bounded by $\mathcal{V}^{\uparrow}(S)$, Equation (A.39) follows.

C.4 Stopped martingale control and weak-step convergence

The following estimate supplies all stochastic constants used below.

Lemma C.2 (Stopped shared-martingale estimate)**.** *Assume the stopped hypotheses of Corollary 6.2, fix $0 < \rho < 1$, and set*

$$A_a = \max_{i,\pm} |\log a^{\pm}_i|, \qquad x_\rho = \log(2d/\rho). \tag{C.33}$$

For the stopped martingale

$$M^{\tau}_{k,i} = h\sum_{j<k} \mathbf{1}_{\{j<\tau\}}\zeta_{j,i}, \tag{C.34}$$

with probability at least $1 - \rho$,

$$\max_{i,k\le N} |M^{\tau}_{k,i}| \le 2\sqrt{\frac{(S+h)\eta\overline{V}x_\rho}{L}} + \frac{8\eta G_* x_\rho}{3L}. \tag{C.35}$$

On $\{\tau > N\}$, it follows that

$$\varepsilon_0 \le \varepsilon^0_{\eta,\alpha}(\rho) := \frac{A_a}{L} + 2\sqrt{\frac{(S+h)\eta\overline{V}x_\rho}{L}} + \frac{8\eta G_* x_\rho}{3L}, \tag{C.36}$$

$$\max_{i,\pm}\sup_{s\le S} \mathcal{A}^{\pm}_i(s) \le \frac{(S+h)\eta G_*^2}{2(1-\kappa_{\mathrm{step}})}. \tag{C.37}$$

Proof. On $\{k < \tau\}$, conditional unbiasedness and $|G_{k,i}| \le G_*$ give $|g_i(\beta_k)| \le G_*$, and therefore $|\zeta_{k,i}| \le 2G_*$. Each stopped increment in Equation (C.34) is consequently bounded by

$$2hG_* = \frac{4\eta G_*}{L}. \tag{C.38}$$

Its predictable quadratic variation through $N = \lceil S/h \rceil$ is at most

$$Nh^2\overline{V} \le (S+h)h\overline{V} = \frac{2(S+h)\eta\overline{V}}{L}. \tag{C.39}$$

The two-sided maximal Freedman inequality, followed by a union bound over the d coordinates, gives Equation (C.35). There is one d-dimensional martingale shared by both channels; only its sign changes.

On $\{\tau > N\}$, the stopped and original martingales agree through the horizon. The initialization contributes at most A_a/L, proving Equation (C.36). Finally, Equation (C.9) and $s_N = Nh \le S + h$ give Equation (C.37); linear interpolation cannot increase the maximum of the clock endpoints. □

Proof of Corollary 6.2. Choose $\rho_\alpha \downarrow 0$ so that

$$\log(d/\rho_\alpha) = o(L_\alpha). \tag{C.40}$$

On $\{\tau_\alpha > N_\alpha\}$, the common-channel estimate Equation (C.8) and the predictor barrier imply

$$Z_{\mathrm{ch}}(S) \le dB_{\mathrm{ch}}S, \qquad U \le 1 + h_{\mathrm{abs}}dB_{\mathrm{ch}}S + S\,\|r\|_\infty. \tag{C.41}$$

The constants in Lemmas A.7 and C.2 then give

$$\mathcal{J}_h^{\mathrm{sgn}}(S) = o(1), \qquad \varepsilon_0 = o_{\mathbb{P}}(1), \qquad \max_{i,\pm} \mathcal{A}_i^{\pm}(S) = o(1), \qquad \kappa_{\mathrm{ch}} = o(1). \tag{C.42}$$

Indeed, these four terms have respective orders

$$O(\eta_\alpha/L_\alpha), \quad O(1/L_\alpha) + O_{\mathbb{P}}(\sqrt{\eta_\alpha/L_\alpha}), \quad O(\eta_\alpha), \quad O(1/L_\alpha), \tag{C.43}$$

up to the logarithmic confidence factor controlled by Equation (C.40). In particular, none contains the product $\eta_\alpha L_\alpha$.

The deterministic comparator is absolutely continuous on $[0, S]$, so its channel variations are finite. By Equation (A.39), the future-active clock tends to zero. The assumption $\mathcal{V}^{\downarrow}(S) = 0$ removes the sole deterministic deletion term. Therefore

$$\mathcal{R}^{\mathrm{sgn}}_{\eta_\alpha,\alpha}(S) \xrightarrow{\mathbb{P}} 0 \tag{C.44}$$

on events whose probability tends to one; their complement has probability at most $\rho_\alpha + \mathbb{P}(\tau_\alpha \le N_\alpha)$. Apply Equations (A.36) and (A.37) with $s \ge s_0$ and the nondecreasing terminal remainder to obtain uniform prediction and objective convergence. Finally, at a grid time,

$$\frac{z_k}{s_k} = \frac{h\sum_{j<k}(q_j^+ - q_j^-)}{kh} = \frac{1}{k}\sum_{j<k}\beta_j,$$

which proves Equation (6.20). □