ORDERING AS CONTROL: FASTER, SINGLE-CALL, AND BIAS-CANCELLING STOCHASTIC EXTRAGRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The order in which stochastic extragradient (SEG) visits the components of a finite sum changes where it settles under a constant step. We treat this order as a control input, for strongly monotone finite-sum variational inequalities. Expanding one epoch shows that the order acts through a single prefix-sum term, $X(\pi)$, beside an order-free correction and a higher-order remainder. Four actions follow. Herded orders *shrink* $X(\pi)$. Alternating an order with its reverse *cancels* its leading contribution across epochs (affine model). *Reusing* the records herding already keeps removes the second operator call: SC-HERD makes one call per step and, for a fixed order herding the ideal offsets, matches the two-call herded plateau up to a higher-order term. An order can even *null* the whole leading bias of single-call past extragradient on an explicit affine family; BC-SPEG searches for such orders with estimated Jacobians, under a conditional a-posteriori certificate. Against an instance-wise lower bound for random reshuffling, herding improves the worst case by $\sqrt{n}$ on rotation-dominated instances at the largest admissible step, up to conditioning and component-dominance factors; on generic instances in fixed dimension the leading coefficient improves by a factor n under per-node split accuracy, with finite-step gains still conditioning-dependent. Experiments cover affine, nonlinear, multi-player and real-data problems. At two frozen adversarial ResNet checkpoints, $\langle g, X(\pi)\rangle$ ranks the one-epoch outcomes of unseen random orders at partial Spearman 0.97; the size of the prefix sums alone does not.

1 INTRODUCTION

An epoch of stochastic extragradient (SEG) visits every component of a finite sum, but the visits need not commute. Each update changes where later components are evaluated, so the order can change the constant-step bias. We study finite-sum variational inequalities (VIs) in root-finding form

$$F(z^*) = 0, \qquad F(z) = \frac{1}{n}\sum_{i=1}^{n} F_i(z), \qquad z \in \mathbb{R}^d, \tag{1}$$

VIs arise in adversarial learning, games, and robust optimization; our guarantees assume strongly monotone components (Section 3), hence strongly monotone F. Each SEG step evaluates the same component before and after extrapolation.

Ordering already improves minimization through herding (Lu et al., 2022); finite-sum min–max analyses include adversarial and fixed orders for GDA and proximal-point methods (Das et al., 2022). SEG-FFA combines random flip-flop orders with iterate anchoring (Chae et al., 2024). We ask which part of SEG's bias the order controls, and how to exploit it.

Order as a control input. At z^*, an epoch expansion (Lemma 1) separates an order-free correction from the leading order-dependent term

$$X(\pi) \;=\; \sum_{t=1}^{n} M_{\pi_t} P_{t-1}, \qquad P_{t-1} = \sum_{s<t} F_{\pi_s}(z^*), \tag{2}$$

Here $M_i = DF_i(z^*)$: each prefix P_{t-1} records earlier components' cumulative contribution, and M_{π_t} describes the next component's response to that displacement. The expansion retains a higher-order remainder. This structure leads to four complementary actions:

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.

- **Shrink it.** Herded orders reduce the order-dependent plateau relative to an instance-wise reshuffling lower bound at the mean-Jacobian scale (Lemma 7). On rotation-dominated instances, the gain is $\widetilde{\Omega}(\sqrt{n}\,\mu/L)$ at the largest admissible step under Theorem 1's non-dominance and step conditions. For generic instances, per-node split accuracy yields a leading-coefficient gain of $\widetilde{\Omega}(n\sigma_\star/(L\sqrt{d}))$; finite-step gains remain conditioning-dependent (Proposition 1). GD-D certifies this accuracy for one generic split in near-linear time at fixed $d \geq 2$ (Theorem 2), not the whole recursion, which is checked empirically.
- **Cancel it.** Reversal preserves discrepancy and negates $X(\pi)$ up to an order-free term (Proposition 2). In the affine model, alternating a herded order with its reverse cancels the leading order-dependent bias, removing coherent accumulation under step decay and lowering the constant-step plateau below the worst-case herded scale without additional operator calls (Theorem 3). Smooth and single-call variants are empirical.
- **Reuse it.** SC-HERD uses the ordering records as a component-aligned lookahead, making one call per step. In the affine model, it matches the two-call herded plateau up to a higher-order term for a fixed or stabilized order herding the ideal offsets; switched orders have a boundedness guarantee (Theorem 4). Lemma 30 transfers a separately bounded record discrepancy at an augmented fixed point. On covtype (Blackard & Dean, 1999) ($p = 8192$, $n = 64$), SC-HERD matches the herded plateau with $1.86\times$ faster epochs than SEG-RR, including ordering overhead (Table 3).
- **Null it.** An explicit order cancels SPEG's entire leading bias on an affine family, letting single-call SPEG beat two-call SEG under every order within a specified step range (Theorem 5). BC-SPEG searches using trajectory-estimated Jacobians. Its affine plateau certificate requires the stated step cap and stabilized orders (Theorem 6); it becomes observable once each component has d linearly independent recorded increments (Lemma 32).

2 Related Work

Extragradient and single-call methods originate with Korpelevich (1976) and Popov (1980); stochastic and single-call variants are analysed by Gidel et al. (2020); Hsieh et al. (2019); Mishchenko et al. (2020b); Gorbunov et al. (2022). Stochastic past extragradient, SPEG (Choudhury et al., 2023), reuses the previous step's value; SC-HERD instead keeps component-aligned records, resembling SAGA-style memory (Defazio et al., 2014). Weak-Minty regimes (Diakonikolas et al., 2021) are outside our scope. *Without-replacement analyses* cover minimization (Mishchenko et al., 2020a; Safran & Shamir, 2020; Rajput et al., 2020; Cha et al., 2023) and finite-sum min–max, including adversarial ordering for GDA and proximal-point methods (Das et al., 2022; Cho & Yun, 2023; Chae et al., 2024) and SEG-RR itself (Emmanouilidis et al., 2024). SEG-FFA combines flip-flop sampling with anchoring and composes with herding and reuse (Section 4.5); FlipFlop (Rajput et al., 2021) reverses a random order on alternate epochs, and our antithetic schedule reverses a herded one. *Optimized orders* for minimization include GraB (Lu et al., 2022) and CD-GraB (Cooper et al., 2023), which balance previous-epoch gradients by herding (Welling, 2009; Harvey & Samadi, 2014) with vector-balancing primitives (Bárány & Grinberg, 1981; Alweiss et al., 2021). Here the epoch-bias expansion connects balancing, reversal, record reuse, and bias cancellation.

3 Setup

Standing assumptions on the components $F_i : \mathbb{R}^d \to \mathbb{R}^d$:

(A1) Each F_i is L-Lipschitz.

(A2) Each F_i is μ-strongly monotone, $\langle F_i(z) - F_i(z'), z - z' \rangle \geq \mu \|z - z'\|^2$; in particular the average operator F is μ-strongly monotone.

(A3) For the smooth results, each F_i has an M-Lipschitz Jacobian (Chae et al., 2024, Assumption 3.3).

Component strong monotonicity (A2) gives an elementary per-epoch contraction and is assumed throughout; relaxing it to F alone would go through the grouped-contraction argument of Chae et al. (2024), which we do not carry out. The heterogeneity scale is $R^2 = \frac{1}{n}\sum_{i=1}^n \|F_i(z^*)\|^2$; with $\rho_\infty = \max_i \|F_i(z^*)\|$, always $R \leq \rho_\infty \leq \sqrt{n}R$, and $\rho_\infty = \widetilde{O}(R)$ unless one component dominates.

Scope. The affine model $F_i(z) = M_i z + b_i$ ($\|M_i\| \le L$, $\sum_i b_i = 0$, so $z^* = 0$) gives the exact control mechanism and the strongest generic result. Smooth operators retain the worst-case herding mechanism under a curvature-dependent step cap (Theorem 8), and the generic smooth extension also assumes linearization genericity (Corollary 1).

For an order $\pi = (\pi_1, \dots, \pi_n)$, one SEG micro-step is $\widehat{z}_t = z_t - \alpha F_{\pi_t}(z_t)$, $z_{t+1} = z_t - \beta F_{\pi_t}(\widehat{z}_t)$. SEG-RR samples π uniformly each epoch; herded SEG chooses it to control prefix sums of centered component values. For u_i with $\sum_i u_i = 0$, the prefix discrepancy is $D(\pi; u) = \max_{1 \le t \le n} \left\| \sum_{s<t} u_{\pi_s} \right\|$. $\widetilde{O}(\cdot)$ hides factors polylogarithmic in n and polynomial in d (Appendix B makes the d-dependence explicit).

Ordering inputs. The analysis takes the ideal offsets $u_i = F_i(z^*)$. Our affine n-sweep orders by all components' current values (n extra evaluations per epoch, no z^*), the deployable order by values recorded in the previous epoch as in GraB (no extra evaluations), with indistinguishable plateaus (Appendix N); Section 4.3 states when records stand in for ideal offsets.

Ordering routines. Three routines appear. The greedy rule (Algorithm 3; $O(n^2 d)$) is an empirical heuristic. The self-balancing walk of Alweiss et al. (2021) ($O(nd)$) carries the unconditional $\widetilde{O}(\sqrt{d}\, R)$ discrepancy guarantee behind the worst-case theorems (Lemma 5(ii)). Recursive balanced bipartition with GD-D splits ($O(nd \log n)$) is the deployable default: a single split is proved (Theorem 2), whereas the generic results assume ($\mathrm{H}_{\mathrm{split}}$) at every node, which is checked empirically. Table 2's recursive pairing and the timing benchmark's sketch are separate empirical variants. Each routine inherits only the guarantee supported by its discrepancy bound. Empirically, all three gain with n; recursion is cheap only when an evaluation outweighs $O(d \log n)$ arithmetic (Appendix N).

3.1 The order-dependent bias: the object we act on

With a constant step size, reshuffled SEG converges to a neighbourhood of z^* whose scale is the per-epoch bias. Expanding one epoch around its starting point isolates the part the order controls; anchored at the solution, its order-dependent term is the functional X of (2).

Lemma 1 (Epoch expansion; anchor)**.** *Let $f_i = F_i(z_0)$, $R_0 = \max_i \|f_i\|$, $P_t = \sum_{s \le t} f_{\pi_s}$, and assume $\beta n L \le c_0$, $\alpha = \beta$. One epoch of* SEG *from z_0 displaces the iterate by*

$$z_n - z_0 \;=\; -\beta \sum_i f_i \;+\; \beta^2 \Big[\underbrace{\textstyle\sum_i M_i f_i}_{\textit{order-free}} + \underbrace{\textstyle\sum_t M_{\pi_t} P_{t-1}}_{X(\pi)\text{: order-dependent}} \Big] \;+\; E, \qquad \|E\| \le 20\beta^3 n^3 L^2 R_0, \tag{3}$$

with $M_i = DF_i$. In the affine case M_i is constant and (3) *holds as displayed. For smooth F_i* (A3) *the expansion holds with the* anchor *Jacobians $M_i = DF_i(z_0)$ and remainder $\|E\| = O(\beta^3 n^3 [L^2 R_0 + M R_0^2])$ (explicit form: Lemma 25). The anchor-change cost of moving $DF_i(z_0)$ to $DF_i(z^*)$, and the finite-step conditions under which it is absorbed, are recorded in Appendix H; no result below invokes the substitution.*

Because the plateau is (per-epoch bias)$/(1-q)$ with $1 - q = \Theta(\beta n \mu)$, controlling $X(\pi)$ controls the *leading* order-dependent plateau term $\mathcal{P}_{\mathrm{order}}(\pi) = \Theta\big(\beta \|X(\pi)\| / (n\mu)\big)$; the order-free term is the same for every order, and the remainder is of higher order. $X(\pi)$ is a *prefix-sum* functional, small when the order keeps prefix sums small—what herding does (Section 4.1) and why a discrepancy bound is the right currency. That bound is a *certificate* for $\|X(\pi)\|$ ($\|X(\pi)\| \le nL \max_t \|P_{t-1}\|$), not a per-order predictor of the realized bias direction, a distinction Section 5 tests on a deep model.

4 Controlling the Order-Dependent Bias

4.1 Shrink it: herded orders

The first action makes $X(\pi)$ small: a random order has $D(\pi) \asymp R\sqrt{n}$, a balancing routine keeps it at $\widetilde{O}(R)$, and the same discrepancy controls the cross-term in Lemma 1. Two scales govern the comparison: the heterogeneity R sizes both plateaus, and the mean-Jacobian nondegeneracy $\sigma_\star = \sigma_{\min}(\frac{1}{n}\sum_i M_i) \in [\mu, L]$ sizes the variance that re-drawing the order injects—$\Theta(L)$ on the rotation-dominated instances of our experiments, μ in the isotropic worst case.

Theorem 1 (Herded SEG vs. SEG-RR)**.** *Assume* (A1)–(A2) *and the affine model, with heterogeneity scale R and no single dominant component, $\rho_\infty = \widetilde{O}(R)$. Run plain* SEG *($\alpha = \beta$) for K epochs with a constant step $\beta \le \min\{1/(3nL),\, 2\mu/(9L^2)\}$. Then there is a contraction factor $q \in (0,1)$ with $1 - q \ge \beta n\mu/8$ such that*

$$\left\|z^K - z^*\right\| \le q^K \left\|z^0 - z^*\right\| + \mathcal{P}_{\mathrm{EG}} + \mathcal{P}_{\mathrm{order}}(\pi). \tag{4}$$

Here $\mathcal{P}_{\mathrm{EG}} = O(\alpha C_0/(n\mu))$, with $C_0 = \|\sum_i M_i b_i\|$, is an order-independent floor—the extragradient-correction bias, the target of anchoring—while $\mathcal{P}_{\mathrm{order}}$ is governed by the prefix sums of the centered values $F_i(z^*)$. A herded order from a balancing subroutine obeys an upper *bound, and* SEG-RR *a* lower *bound in the RMS sense, $\mathcal{P}^{\mathrm{RMS}}_{\mathrm{order}}(\mathrm{RR}) := (\mathbb{E}\,\|z^\infty - z^*\|^2)^{1/2}$ over the order redrawn each epoch), whose scale is set by the mean-Jacobian nondegeneracy $\sigma_\star := \sigma_{\min}\big(\frac{1}{n}\sum_i M_i\big) \in [\mu, L]$:*

$$\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{O}\left(\frac{\beta LR}{\mu}\right), \qquad \mathcal{P}^{\mathrm{RMS}}_{\mathrm{order}}(\mathrm{RR}) = \Omega\big(\beta^2 n^{3/2}\sigma_\star R/\sqrt{d}\big). \tag{5}$$

The RMS bound is by definition the second-moment statement $\mathbb{E}\,\|z^\infty - z^*\|^2 = \Omega(\beta^4 n^3 \sigma_\star^2 R^2/d)$ (no first-moment bound is claimed), valid for $\beta \lesssim \sigma_\star/(\sqrt{d}\,nL^2)$ (Appendix E). On rotation-dominated instances—$\sigma_\star = \Theta(L)$, including the families of Section 5 and Lemma 9—this admits $\beta = \Theta(1/(nL))$ up to a fixed-dimension constant, and the largest admissible step gives a herding gain of $\widetilde{\Omega}(\sqrt{n}\cdot\mu/L)$ in norm: a worst-case $\sqrt{n}$ for well-conditioned problems, $\sqrt{n}/\kappa$ in general. The SEG-RR *bound is a genuine lower bound—reshuffling variance injected by re-drawing the order, which a fixed herded order does not incur—and cannot vanish on any instance, since $\sigma_\star \ge \mu$ (Lemma 7, Appendices C and E). Without the no-dominant-component assumption the herded bound reads $\widetilde{O}(\beta L\rho_\infty/\mu)$, so the gain is $\sqrt{n}\,R/\rho_\infty$ up to κ.*

Proposition 1 (Generic instances)**.** *Suppose moreover the offsets are* centered generic*: latent i.i.d. samples $x_1, \dots, x_n$ from a density bounded above and below on a box of side $O(R_\nu)$ (the population scale; $R = \Theta_d(R_\nu)$ with high probability, and bounds are stated in R), drawn independently of the arbitrary fixed Jacobians $(M_i)_i$, with*

$$b_i = F_i(z^*) = x_i - \bar{x}, \qquad \bar{x} = \tfrac{1}{n}\textstyle\sum_j x_j,$$

so that $\sum_i b_i = 0$ holds identically, as the root condition forces (an i.i.d. model for the b_i themselves is inconsistent with it: the offsets are exchangeable, not independent). Assume $n = 2^q$; the box support is part of the hypothesis (Remark 1). Say that a recursive bipartition tree has per-node split accuracy *if*

$$\delta_v \;\le\; C_d\, R\, \ell(n)/\sqrt{m_v} \qquad \text{at every node } v \text{ of size } m_v, \tag{$\mathrm{H}_{\mathrm{split}}$}$$

with $\ell(n)$ polylogarithmic and δ_v the node's split imbalance. Then, at constant steps $\beta \le c\min\{\mu/(nL^2),\, 1/(n^{3/2}L)\}$ (under which the $O(\beta^3)$ epoch tail sits below the generic leading term; Appendix F carries the details and Corollary 1 the same caps), a herded order produced by any *label-blind recursive bipartition tree satisfying* ($\mathrm{H}_{\mathrm{split}}$) *obeys*

$$\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{O}\big(\beta LR/(\mu\sqrt{n})\big),$$

in expectation over the label matching and the child-order coins, with the unlabeled cloud, the priorities and the tree held fixed; the probability-space bookkeeping—three independent randomness sources, and why conditioning on ($\mathrm{H}_{\mathrm{split}}$) *disturbs neither the labels' uniformity nor the coins' fairness—is set out in Appendix F. The statement is* conditional structural*: it asserts nothing about the probability of* ($\mathrm{H}_{\mathrm{split}}$). *Two claims must be kept distinct. At the* coefficient *level, $\big\|X_{\mathrm{RR}} - \mathbb{E}X_{\mathrm{RR}}\big\|_{L_2} = \Omega(n^{3/2}\sigma_\star R/\sqrt{d})$ (Lemma 7) against a generic herded cross-term of $\widetilde{O}(LR\sqrt{n})$: an L_2 ratio of $\widetilde{\Omega}(n\sigma_\star/(L\sqrt{d}))$, a clean factor n when rotation-dominated at fixed d. The proven* finite-step *separation at the largest admissible step is $\widetilde{\Omega}\big(\min\{n/\kappa^2, \sqrt{n}/\kappa\}/\sqrt{d}\big)$, increasing in n throughout but nontrivial (> 1) only beyond $n \asymp \kappa^2$. Theorem 2 verifies* ($\mathrm{H}_{\mathrm{split}}$) *at the* root *node for every fixed $d \ge 2$, carried to the centered offsets by translation-equivariance (Lemma 21); a deeper node's input is a* value-dependent *selection of the sample, to which that theorem does not apply, so validity down the recursion is empirically supported (Appendix N, including a direct node-level audit) but not proved, and making the recursion self-certifying is open. The worst-case bound of Theorem 1 is otherwise tight: on the balanced signed rotational family of Lemma 9 the cross-term is $\Omega(nLR)$ for* every *order.*

Theorem 2 (Balanced splitting by grid differencing with deferral)**.** *Let $x_1, \dots, x_m$ be independent samples from densities bounded above and below on a box of side R, in fixed dimension $d \geq 2$.* GD-D *(Algorithm 2, Appendix F) runs $\lceil d/2 \rceil$ rounds of within-cell pairing by fresh random priorities on a grid anchored to the current cloud's own bounding box, defers odd cells' leftovers at their own scale, from the second round on recurses only the cells inside a certified central core (Lemma 17), and finishes with one coarse-to-fine greedy signing, uniformly random within each pile. The output is an exactly equal bipartition with imbalance $\widetilde{O}_d(R/\sqrt{m})$, with high probability, in near-linear time. The bound holds verbatim for the* centered *cloud $x_i - \bar{x}$ of Proposition 1, by translation-equivariance (Lemma 21). ($d = 1$ is not covered: the final chain inherits an entrance residual too large for the steering lemma at its own scale—Appendix F—and we make no scalar claim.)*

The GD-D **primitive.** Two rules carry Theorem 2: *pair only within cells* of a grid anchored to the cloud's bounding box, so the pairing is value-independent and a pair's difference has a two-sided density (which nearest-neighbour pairing does not admit); and *defer the leftovers*, since cross-cell repair injects long vectors that break the induction. Both failure modes are measured, and the primitive is Algorithm 2 (Appendix F).

4.2 CANCEL IT: ANTITHETIC HERDING

A fixed herded order makes $\|X(\pi)\|$ small but not zero: the residual repeats coherently each epoch. Unlike SEG-RR's self-averaging bias, this accumulation can erase the herding advantage under aggressive step decay; the measured gain falls below 1 (Appendix N). Re-randomizing within the balanced family leaves a nonzero mean bias. Reversal instead provides an exact cancellation identity.

Proposition 2 (Reversal negates the order-dependent bias)**.** *Let u_i be centered ($\sum_i u_i = 0$) and $\pi_t^R = \pi_{n+1-t}$. Then $P_t^{u,R} = -P_{n-t}^u$ for all t; consequently $D(\pi^R; u) = D(\pi; u)$—the reverse of a herded order is exactly as herded—and*

$$X_u(\pi^R) \;=\; -\,X_u(\pi) \;-\; \sum_i M_i u_i. \tag{6}$$

Hence $X(\pi) + X(\pi^R)$ is independent of π.

Antithetic herding runs π on even epochs and π^R on odd epochs. It reverses a chosen order across epochs, as FlipFlop (Rajput et al., 2021) does for random orders; SEG-FFA instead cycles a random order forward and back with iterate anchoring (Chae et al., 2024). The ordering routine, records, and call count are unchanged.

Theorem 3 (Antithetic cancellation)**.** *Under the affine assumptions of Theorem 1, alternating a herded order π with its reverse π^R cancels the* entire *leading β^2 order-dependent epoch bias: the antithetic iteration converges to a 2-periodic orbit whose surviving order-dependent plateau is one power of β higher,*

$$\mathcal{P}_{\mathrm{anti}}^{\mathrm{ord}} = \widetilde{O}\Big(\frac{\beta^2 n L^2 R}{\mu}\Big), \tag{7}$$

smaller by $O(\beta n L)$ than the worst-case herded scale $\widetilde{O}(\beta LR/\mu)$ of Theorem 1. Under a nonincreasing step schedule the only leading mismatch is proportional to $|\beta_k - \beta_{k+1}|$, so a fixed order's coherent accumulation is eliminated; the precise schedule bound, the general (non-herded) cubic residual, and the comparison's correct baseline are in Appendix K.

Beyond the repair, alternating an order with its reverse is a flip-flop-type cancellation *across* epochs at one pass per epoch, where SEG-FFA uses two, so it also improves the constant-step plateau ($4.4\times \to 8.1\times$ at $n = 64$; $7.7\times \to 8.4\times$ composed with SEG-FFA). Theorem 3 is affine; smooth and single-call antithetic variants are evaluated empirically only (Appendix K explains why reversal does not transfer to the augmented single-call state).

4.3 REUSE IT: SINGLE-CALL METHODS

Shrinking and cancelling leave two operator calls per component step. A herded order already keeps a table of recorded component values, and SC-HERD extrapolates from the component's record, stale since its previous visit, instead of from a fresh evaluation, so each step makes one call. The default recursive GD-D routine stays within 1.1–$1.6\times$ of the greedy heuristic at the full herding gain (Appendix N).

Theorem 4 (Single-call parity)**.** *Assume* (A1)–(A2) *and the affine model of Theorem 1 (smooth* SC-HERD *is an empirical extension; Appendix I). Let* SC-HERD (*Algorithm 1*) *maintain a per-component record table* $\{r_i\}$*, choose a herded order from it, and make the single call*

$$g_t = F_{\pi_t}(z_t - \alpha r_{\pi_t}), \qquad z_{t+1} = z_t - \beta g_t, \qquad r_{\pi_t} \leftarrow g_t. \tag{8}$$

For a fixed or eventually stabilized order π *that herds the* ideal offsets*—*$D(\pi; b) \le \bar{D}$ *with* $b_i = F_i(z^*)$*—*

$$\mathcal{P}_{\text{SC-HERD}} \le \mathcal{P}_{\textit{herded-}\text{SEG}} + O\left(\frac{\alpha^2 L^2 R}{\mu}\right), \tag{9}$$

where $\mathcal{P}_{\textit{herded-}\text{SEG}}$ *is the two-call herded-*SEG *plateau of Theorem 1 at discrepancy* $\bar{D}$*. Whenever* $\alpha L\sqrt{n} \lesssim 1$ *the additive term is of higher order: component-aligned reuse matches the two-call herded plateau while making one operator evaluation per component step. For an* arbitrary *switched sequence of per-epoch orders we prove boundedness of the augmented state* $(z; r_1, \ldots, r_n)$*, but not this sharp parity (Appendix I).*

The centered $u_i = r_i - \bar{r}$ are a temporary argument to the balancing routine; the lookahead $z - \alpha r_i$ reads the component's own *raw* record. Records yield the ideal-offset discrepancy through Lemma 30: at a fixed augmented point the records sum to zero, and a separately bounded record discrepancy transfers—$\widetilde{O}(R)$ for $\beta \lesssim \mu/(nL^2)$, $\widetilde{O}(\kappa R)$ under $\beta n L \le 1/3$. Sharp parity needs a settled order, whereas the default routine resamples each epoch. Replaying SC-HERD's recomputed orders with two-call SEG reproduces its plateau (0.998; Appendix N): an empirical match, not a proof of settling or transfer.

Algorithm 1: Single-call herded extragradient (SC-HERD)

Raw records $r_i \leftarrow 0$ for all i.
for epoch $k = 0, 1, 2, \ldots$ **do**
  $u_i \leftarrow r_i - \frac{1}{n}\sum_j r_j$ {temporary; r stays raw}
  $\pi \leftarrow \text{HERD}(u)$, or uniform if $k = 0$.
  for $t = 1, \ldots, n$ **do**
    $i \leftarrow \pi_t$; $g \leftarrow F_i(z - \alpha r_i)$ {raw r_i}
    $z \leftarrow z - \beta g$; $r_i \leftarrow g$.
  end for
end for

SPEG and SC-HERD differ in which value they reuse. SPEG reuses the previous step's value, usually from a *different* component, an error of order R; SC-HERD reuses the same component's entry, only extrapolation-stale, an $O(\alpha R)$ error. A 2×2 factorial over (reuse rule)$\times$(order) locates the effects (Appendix N): under a random order the reuse rule is immaterial (SEG-RR, SPEG-RR and single-call-RR agree to within 1%), under a herded order component-aligned reuse is worth a constant 1.3–$2.5\times$; what *grows* with n against SPEG-RR is the herding gain, transferred by a variance argument (Appendix E.1). Theorem 4 compares against two-call herding, not SPEG; Section 4.4 shows no order-generic SPEG lower bound can exist.

4.4 NULL IT: BIAS-CANCELLING ORDERS

The preceding actions control the prefix contribution, but a small $X(\pi)$ need not minimize the whole bias. Lemma 1 says more: for single-call past extragradient it holds with $X(\pi)$ replaced by $X(\pi) + A(\pi)$, $A(\pi) = \sum_t M_{\pi_t}(f_{\pi_{t-1}} - f_{\pi_t})$ (indices cyclic)—a sum over *adjacent* pairs whose sign the order can choose (for SC-HERD the extra term is $O(\beta)$ smaller). The order can therefore target SPEG's entire leading bias.

Theorem 5 (An order can null SPEG's leading bias)**.** *Let* $8 \mid n$*. There is a family of strongly monotone affine finite-sum VIs (rotational blocks* $M_i = \mu I + \delta s_i J$ *with balanced signs* s_i*, offsets* $b_i = \rho s_i v$*) and an* explicit *periodic order* $\pi^\star$ *for which the entire* β^2 *bias of* SPEG *vanishes, while for* every *order the* β^2 *bias of two-call* SEG *has norm* $\Omega(n\delta\rho)$*. Consequently, over the explicit finite-step range of Appendix L,*

$$\mathcal{P}(\text{SPEG}, \pi^\star) \;=\; O\big(\beta^2 n^2 L^2 \rho\big), \qquad \min_{\pi} \mathcal{P}(\text{SEG}, \pi) \;=\; \Omega\big(\beta\delta\rho\big)\,: \tag{10}$$

suitably ordered single-call SPEG *beats two-call* SEG *under* every *order, and no lower bound of the form "*SPEG *is worse for every order" can hold. The construction, the admissible step range, the constants, and the precise scope of the gain are in Appendix L.*

The order is the periodic pattern $(+, +, -, -, -, -, +, +)$, hand-checkable: both bias components vanish because the pattern has zero prefix area and exactly $n/4$ cyclic sign changes; at $n = 8$ it is the minimizer over all 40320 orders.

From impossibility to an algorithm. (i) Objective: by Lemma 1 the epoch bias coefficient is $\sum_i M_i f_i + T(\pi)$ with $T(\pi) = \sum_t M_{\pi_t}(P_{t-1} + f_{\pi_{t-1}} - f_{\pi_t})$, so the order should steer T to $-\sum_i M_i f_i$. (ii) Jacobians without additional operator evaluations: consecutive visits to component i give increments $\Delta F = M_i \Delta z$, so a running least-squares fit recovers M_i once they span $\mathbb{R}^d$ (then b_i and z^* too, which makes the certificate observable, Lemma 32); the fit and the search cost arithmetic, not operator calls. (iii) Optimizer: swap local search initialized at the herded order, so its value can only improve on herding's. The result is BC-SPEG.

Theorem 6 (A-posteriori plateau certificate)**.** *Assume the affine model of Theorem 1. Suppose the realized orders stabilize at π, take*

$$\beta \;\le\; \min\Big\{\frac{c_0}{nL},\; \frac{\mu}{64L^2}\Big\}, \tag{11}$$

and let $\varepsilon = \|\mathit{bias}_2(\mathrm{SPEG}, \pi; z^*)\|$. Then BC-SPEG *converges geometrically and*

$$\|F(z^\infty)\| \;\le\; \frac{2\beta\varepsilon}{n} \;+\; 160\,\beta^2 n^2 L^2 R_\infty. \tag{12}$$

Once every component has d linearly independent recorded increments, ε is observable from the trajectory by reconstructing the affine component models (Lemma 32); before then BC-SPEG *optimizes a surrogate. (Both caps, their roles, and the smooth model-error extension are discussed in Appendix L.)*

The guarantee is conditional on an *observable* quantity, not on the optimizer: under the step cap and a stabilized order, it certifies whatever bias the search achieved, with ε observable through the reconstruction $M_i \to b_i \to z^* = -\bar{M}^{-1}\bar{b}$ of Lemma 32 (reached at a median of d recorded increments per component in our runs). Empirically, at the recorded $\beta = 0.4/(nL)$ and equal operator calls (not equal time), BC-SPEG reaches 2.7–6.2× below single-call herding on generic instances, 11.7–13.4× on the rotation-aligned family, and 2.2–3.0× on a nonlinear VI; on the aligned family at small β it attains the enumerated optimum exactly (Appendix N). These cells test cancellation, not the certificate, whose cap $\mu/(64L^2)$ needs $n \ge 25.6\kappa$ at this step: three affine cells miss it, the fourth is unverified.

4.5 Compatibility with flip-flop anchoring

SEG-FFA (Chae et al., 2024) improves the epoch-count axis by a forward/backward sweep pair with anchoring; ordering improves the component-count axis, and the two compose without changing the controller (herded order forward, its reverse backward). Appendix G proves that flip-flop cancels the leading β^2 order-dependent term while the first surviving residual stays *prefix-controlled*, so the worst-case herding separation survives on top of SEG-FFA, lower bound included. Replacing both sweeps by the reuse step gives SC-FFA-Herd at $2n$ rather than $4n$ evaluations per outer epoch (Appendix M); Section 5 treats both as compatibility results.

5 Experiments: One Object, Four Actions

We organize the evidence by question: Figure 1 tests the scaling law and single-call parity; Table 1 cancellation, reuse on top of FFA, and nulling; Table 2 transfer across datasets and ordering routines; Table 3 wall-clock time; Figure 2 the object itself on a deep model. Instances span a strongly monotone heterogeneous affine VI with rotational blocks, smooth operators $F_i(z) = M_i z + b_i + \gamma_i \tanh(z)$, an adversarial $\{+v, -v\}$ construction, a five-player game, and five `sklearn` datasets (Pedregosa et al., 2011) (in the model class of the worst-case $\sqrt{n}\,\mu/L$ theory, per-run step caps unverified; the observed near-linear growth exceeds it). The metric is the tail mean of $\|F(z)\|$ under a constant step. Appendix N plots the convergence curves (Figure 7).

Scaling and parity. At the recorded step $\beta = 1/n$, which the theorems do not certify, the affine SEG-RR/herded ratio in Figure 1 climbs $1.09 \to 32.61$ over $n = 8 \to 256$, the SEG-FFA composition $1.62 \to 62.68$, and the nonlinear VI $1.50 \to 21.30$: the predicted direction and growth, ahead of the proven finite-step separation $\widetilde{\Omega}(\min\{n/\kappa^2, \sqrt{n}/\kappa\})$. Panel (d) is consistent with Theorem 4's fixed-order parity and extends it empirically to recomputed orders and the nonlinear family; panel (c)'s 17.97× separation from SPEG-RR at $n = 128$ rests on the variance-transfer argument of Appendix E.1, not on Theorem 4. Our affine code herds current-iterate values, and the smooth conditions are those of the Scope paragraph.

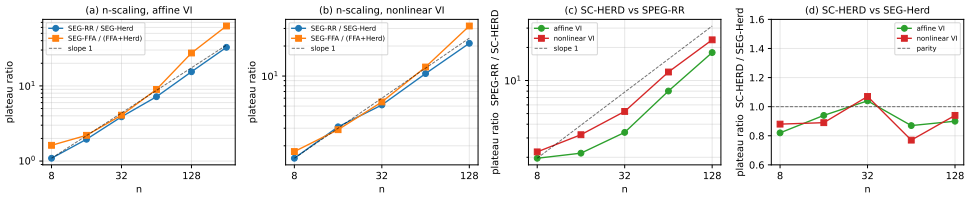

Figure 1: Empirical plateau ratios at the recorded $\beta = 1/n$: (a,b) SEG-RR/SEG-HERD and SEG-FFA/(FFA+HERD); (c) SPEG-RR/SC-HERD. In (a–c), larger is better; log–log slope one denotes a factor-n gain. (d) SC-HERD/SEG-HERD (linear y): below 1 favours SC-HERD; range 0.77–1.07, with no systematic growth in n.

Table 1: Plateau ratios on the affine VI, not time savings. *Cancel*: SEG-RR/ordered, at $n = 64$ (constant step) and $n = 16$ ($4\times$-halving). *Reuse*: row 1 checks parity at half the calls (< 1 favours single-call); row 2's 164.4 is a plateau gain. *Null*: SC-HERD/BC-SPEG at equal calls, an empirical comparison, not a certificate validation (Section 4.4). Larger is better except in reuse row 1.

cancel	const	sched	reuse (FFA)	n=64	n=128	null	n=16	n=32
herded	4.4	0.73	SC-FFA-HERD/FFA-HERD	0.96	0.77	generic	2.7	6.2
antithetic	8.1	1.28	SEG-RR/SC-FFA-HERD	45.0	164.4	aligned	11.7	13.4

Table 2: Neither the synthetic family nor one ordering rule is the source of the gain. **Left**: real-data saddles (rank-1 coupling from PCA features of five `sklearn` datasets), SEG-RR plateau over herded, median [IQR] over 24 draws. **Right**: three orderings targeting small prefix sums; each gain grows with n. The right block's recursive pairing is not the certified GD-D splitter.

dataset	n=8	n=64	n	greedy	recursive	walk
`digits`	0.75	3.58 [2.85, 3.88]	8	1.14	1.34	1.18
`wine`	1.41	7.39 [6.34, 8.31]	16	1.84	1.57	1.48
`breast_cancer`	1.29	6.65 [5.89, 7.74]	32	3.01	2.22	1.97
`diabetes`	0.81	4.02 [3.34, 4.48]	64	5.07	3.33	2.61
`iris`	1.75	10.91 [8.79, 15.30]	128	10.81	4.55	3.80

Worst-case boundary, transfer, and mechanism. The adversarial $\{+v, -v\}$ construction separates the two regimes: generic instances gain nearly linearly in n while the adversarial one tracks $\sqrt{n}$, the boundary the mechanism predicts (Figure 4; the proven tightness witness is Lemma 9). On generic instances the tail coefficient of variation is 0.36 under SEG-RR and numerically zero under herding—the reshuffling-variance mechanism—and cancelling the mean Jacobian's rotation drives $\sigma_\star$ from L to μ with the measured variance tracking the predicted $\sigma_\star^2$ across a $45\times$ range (Table 6). Replacing the Gaussian coupling by a rank-1 coupling built from real features reproduces the n-growth on all five datasets (3.6–$10.9\times$ at $n = 64$; Table 2), and the right block's three routines suggest the gain follows the discrepancy rather than the rule. Combining herding with SEG-FFA reaches 17–$48\times$ over vanilla SEG-RR at $n = 64$ (Figure 3), a measured composition.

Diagnostics. Appendix N audits both conditional bridges (Figure 5): at matched node size, internal splits have 0.80–1.20 times the imbalance of fresh root splits, and the replay of Section 4.3 compares one- and two-call plateaus under the same recomputed orders, without certifying that orders settle. More "extragradient-aware" orders—herding $G_i = (I - \alpha M_i)F_i$, fitting the cross-term, a rotation-aware objective—all fail to beat plain prefix control. And ordering acts on the finite sum, not on saddle structure: on a five-player monotone game with an odd interaction cycle (no two-player saddle-gradient representation under any bipartition) the same n-growing separation and single-call parity persist—at $n = 128$ greedy herding gains $7.5\times$, SC-HERD reaches $8.7\times$ over SPEG-RR and stays near parity with two-call herded SEG (Table 7, Appendix N).

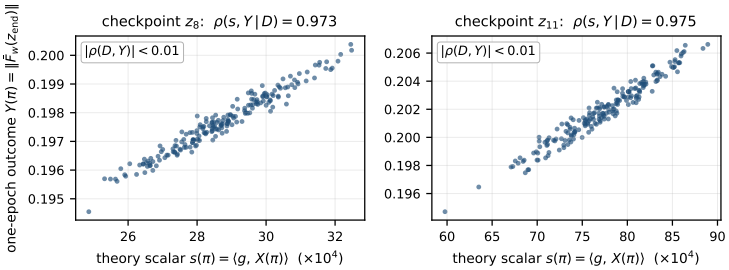

Figure 2: Deep validation of the transported bias. One SEG epoch from a frozen adversarial ResNet-18 (He et al., 2016) checkpoint on CIFAR-10 (Krizhevsky & Hinton, 2009) ($n = 128$, $\beta = 2 \cdot 10^{-4}$), 200 fresh orders per panel. The theory scalar $s(\pi) = \langle g, X(\pi) \rangle$, computed before any outcome is run, has partial Spearman correlation with the realized outcome, conditioned on D, of 0.973 at z_8 and 0.975 at z_{11}, while the prefix discrepancy $D(\pi)$ is rank-uncorrelated with the same outcomes ($|\rho(D, Y)| < 0.01$, insets).

Table 3: Covtype wall-clock ($p = 8192$, $n = 64$; epoch medians over five repeats, IQR $< 1.4\%$; Appendix N). *Speedup*: SEG-RR epoch time over the row's (baseline $1.00\times$). *Time-to-τ*: ordering and update seconds to the first epoch-end crossing of a shared residual threshold, excluding residual evaluation ("never": not by 400 epochs). *Final residual*: epoch-400 norm, not a tail mean.

Method	calls/epoch	epoch (ms)	speedup	time-to-τ (s)	final residual
SEG-RR (2-call)	$2n$	98.9	$1.00\times$	24.2	$2.56 \cdot 10^{-3}$
SPEG-RR (1-call)	n	49.8	$1.99\times$	never	$5.00 \cdot 10^{-3}$
Herded SEG (2-call)	$2n$	102.6	$0.96\times$	24.0	$2.40 \cdot 10^{-3}$
SC-HERD (1-call)	n	53.3	$1.86\times$	10.9	$2.51 \cdot 10^{-3}$

Robustness. Gains decrease with conditioning and component dominance, remain stable across the heterogeneity scale, and survive partial component non-monotonicity, matching the theorem's stated dependencies (Appendix N).

Wall-clock. On covtype ($p = 8192$), SC-HERD matches the two-call herded plateau with $1.86\times$ faster epochs and a $2.2\times$ faster threshold crossing than SEG-RR (Table 3). The timings include 4–15% ordering overhead. One-call reuse lowers epoch cost; threshold-crossing time also depends on residual progress, and herding alone gives little gain in this high-κ regime.

Deep validation of the transported bias. Beyond the strongly monotone setting, Figure 2 scores unseen random orders by $s(\pi)$ with $g = \nabla_z \frac{1}{2} \left\| \bar{F}_w \right\|^2$ at two frozen checkpoints, at a step where the expansion is empirically accurate (Appendix N). In the analysis, D certifies magnitude; $X(\pi)$ also retains the direction relevant to the observed order effect. This is a local one-epoch test, not an end-to-end training result.

6 LIMITATIONS AND CONCLUSION

The epoch expansion makes order a design choice: balancing reduces the Jacobian-weighted prefix term, and reversal cancels its order dependence. Component records also supply a one-call lookahead; Jacobian estimates guide cancellation of SPEG's full leading bias. Our theory concerns strongly monotone VIs. The generic factor-n reduction requires split accuracy ($\mathrm{H}_{\mathrm{split}}$) at every recursion node: proved for one split (Theorem 2), otherwise checked empirically. Worst-case $\sqrt{n}$ bounds do not require it. Fixed herding can lose to reshuffling under aggressive decay; antithetic ordering repairs this, without a claimed decreasing-step rate. Affine single-call parity requires fixed or stabilized orders herding the ideal offsets. BC-SPEG's certificate retains its step and stabilization conditions, with saturating cancellation on generic instances. The RR variance lower bound weakens as $\sigma_\star$ decreases (Lemma 7); Table 6's non-inversion is empirical on its tested family, not a general dominance guarantee.

AI USE STATEMENT

In line with the ICLR 2027 AI policy for authors, we describe how large language models (LLMs) were used. The research idea is ours: we noticed the rotational behaviour of extragradient dynamics and set out to optimize the order in which components are visited. The algorithms were designed by the authors. The theorem statements were formulated by the authors and refined in discussion with the LLMs. GPT-6 Astra (OpenAI), Claude Fable 5 and Claude Fable 5.1 (Anthropic) supplied proof steps throughout the appendices, working from the proof strategies and tools that we specified, and also audited the theoretical part; every proof was then checked by the authors and verified numerically where possible, for example by the exact enumerations and lemma-verification scripts of Appendix O. The experiment code, including the figure-generating scripts, was written with the assistance of Claude Opus 5 and Claude Opus 5.5 (Anthropic); the experiments were designed by the authors in discussion with the LLMs, and the code was reviewed by the authors and checked numerically. No generative AI was used to produce or preprocess data: the synthetic instances are sampled by our code from the distributions of Appendix N, and the public datasets are processed as described there. LLMs assisted the literature search alongside our own reading, and revised text that we drafted in English and Chinese; the structure of the paper was set by the authors. The authors take full responsibility for the content of the paper, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code will be released upon publication. The experiments can be rerun from the paper: Section 5 and Appendix N specify the problem instances, step sizes, metrics and protocols, and the theory is self-contained in the appendices. The theory checks and the synthetic and real-data experiments run on CPU with NumPy; Table 3 and Figure 2 come from separate GPU runs on a single NVIDIA A100. Appendix O gives the software and hardware of each evidence layer, and the scripts and seeds behind Figure 1 and Tables 1–2, with their appendix versions.

REFERENCES

Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4(4): 259–264, 1984.

Ryan Alweiss, Yang P. Liu, and Mehtaab Sawhney. Discrepancy minimization via a self-balancing walk. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pp. 14–20, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3450994. URL https://doi.org/10.1145/3406325.3450994.

Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998. doi: https://doi.org/10.1002/(SICI)1098-2418(199807)12:4⟨351::AID-RSA3⟩3.0.CO;2-S. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2418%28199807%2912%3A4%3C351%3A%3AAID-RSA3%3E3.0.CO%3B2-S.

Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. ISSN 0168-1699. doi: https://doi.org/10.1016/S0168-1699(99)00046-0. URL https://www.sciencedirect.com/science/article/pii/S0168169999000460.

I. Bárány and V.S. Grinberg. On some combinatorial questions in finite-dimensional spaces. *Linear Algebra and its Applications*, 41:1–9, 1981. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(81)90085-9. URL https://www.sciencedirect.com/science/article/pii/0024379581900859.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp.

3855–3912. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cha23a.html.

Jiseok Chae, Chulhee Yun, and Donghwan Kim. Stochastic extragradient with flip-flop shuffling & anchoring: Provable improvements. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 123675–123754. Curran Associates, Inc., 2024. doi: 10.52202/079017-3932. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/df658fe5097f65485ad80b06e6cb30dd-Paper-Conference.pdf.

Hanseul Cho and Chulhee Yun. Sgda with shuffling: faster convergence for nonconvex-pł minimax optimization, 2023. URL https://arxiv.org/abs/2210.05995.

Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64918–64956. Curran Associates, Inc., 2023. doi: 10.52202/075280-2833. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ccf02786d28730e8311676ffa842e216-Paper-Conference.pdf.

A. Feder Cooper, Wentao Guo, Duc Khiem Pham, Tiancheng Yuan, Charlie Ruan, Yucheng Lu, and Christopher De Sa. Coordinating distributed example orders for provably accelerated training. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 56198–56210. Curran Associates, Inc., 2023. doi: 10.52202/075280-2451. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/af9ac087ed9123957bb3a45dca56b9d4-Paper-Conference.pdf.

Aniket Das, Bernhard Schölkopf, and Michael Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6749–6762. Curran Associates, Inc., 2022. doi: 10.52202/068431-0489. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2ce4f0b8e24c45318352068603153590-Paper-Conference.pdf.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/937964195d6fb3a55cd7cc578165f058-Paper.pdf.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/diakonikolas21a.html.

Konstantinos Emmanouilidis, Rene Vidal, and Nicolas Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3682–3690. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/emmanouilidis24a.html.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks, 2020. URL https://arxiv.org/abs/1802.10551.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera

(eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7865–7901. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/gorbunov22b.html.

Nick Harvey and Samira Samadi. Near-optimal herding. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 1165–1182, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/harvey14.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4625d8e31dad7d1c4c83399a6eb62f0c-Paper.pdf.

Narenda Karmarker and Richard M. Karp. The differencing method of set partitioning. Technical Report UCB/CSD-83-113, 1983. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/1983/6353.html.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yucheng Lu, Wentao Guo, and Christopher De Sa. Grab: Finding provably better data permutations than random reshuffling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8969–8981. Curran Associates, Inc., 2022. doi: 10.52202/068431-0652. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3acb49252187efa352a1ae0e4b066ced-Paper-Conference.pdf.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik. Random reshuffling: Simple analysis with vast improvements. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8cc6e90ccbff44c9cee23611711cdc4-Paper.pdf.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtarik, and Yura Malitsky. Revisiting stochastic extragradient. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4573–4582. PMLR, 26–28 Aug 2020b. URL https://proceedings.mlr.press/v108/mishchenko20a.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7964–7973. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rajput20a.html.

Shashank Rajput, Kangwook Lee, and Dimitris Papailiopoulos. Permutation-based sgd: Is random optimal?, 2021. URL https://arxiv.org/abs/2102.09718.

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3250–3284. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/safran20a.html.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1121–1128, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553517. URL https://doi.org/10.1145/1553374.1553517.

A Affine Epoch Map and Bias Decomposition

Proof roadmap. Lemma 1 turns one epoch into an exact affine map plus an $O(\beta^3 n^3 L^2 R)$ remainder (Appendix A), so every plateau statement reduces to bounding a per-epoch bias and dividing by the contraction factor $1-q=\Theta(\beta n\mu)$. Theorem 1 needs a discrepancy bound for the herded order (Appendix B; the ℓ_2-versus-ℓ_∞ conversion costs one $\sqrt{d}$) and the bridge from discrepancy to $X(\pi)$, whose random-order value is $\Theta(n^{3/2})$ and herded value $\widetilde{O}(n)$ (Appendix C); the lower bound isolates a *reshuffling variance* that a fixed order removes (Appendix E). The single-call results repeat the two steps in an *augmented* state that includes the record table (Appendices I, L and M).

Throughout the appendix we use the affine model $F_i(z)=M_i z+b_i$, $\|M_i\|\le L$, $\sum_i b_i=0$, hence $z^*=0$ and $F_i(z^*)=b_i$; the general smooth case (Appendix H) is a perturbation. We build on the SEG epoch expansion of Chae et al. (2024) and the herding primitive of Lu et al. (2022). Unlabeled $O(\cdot)$ and $\Theta(\cdot)$ constants are absolute (independent of n, d, L, μ, R); step-size thresholds are stated explicitly.

For affine F_i the SEG micro-step is *exact*: $F_i(z-\alpha F_i(z)) = M_i(z-\alpha F_i(z))+b_i = (I-\alpha M_i)F_i(z)$, so the micro-step $z\mapsto z-\beta F_i(z-\alpha F_i(z))$ equals $z\mapsto z-\beta G_i(z)$ with the extragradient-corrected operator $G_i(z)=(I-\alpha M_i)F_i(z)=A_i z+g_i$, $A_i=(I-\alpha M_i)M_i$, $g_i=(I-\alpha M_i)b_i$; write $\|A\|_{\max}=\max_i\|A_i\|\le L+O(\alpha L^2)$.

Lemma 2 (Exact affine epoch map)**.** *One epoch with order π is the affine map $z\mapsto \Phi_\pi z-\psi_\pi$, where*

$$\Phi_\pi=\prod_{j=n}^{1}(I-\beta A_{\pi_j}),\qquad \psi_\pi=\beta\sum_{j=1}^{n}\Big[\prod_{\ell>j}(I-\beta A_{\pi_\ell})\Big]g_{\pi_j}.$$

Proof. The j-th micro-step is the affine map $R_j(z)=(I-\beta A_{\pi_j})z-\beta g_{\pi_j}$. Composing $R_n\circ\cdots\circ R_1$ and collecting the linear and constant parts gives Φ_π and ψ_π. □

Lemma 3 (Bias decomposition)**.** *The per-epoch bias $r_\pi=(\text{epoch map})(z^*)-z^*=-\psi_\pi$ satisfies*

$$r_\pi=-\beta\sum_i g_i+\beta^2\sum_{\ell=1}^{n}A_{\pi_\ell}P_\ell+E_3,\quad P_\ell=\sum_{j<\ell}g_{\pi_j},\quad \|E_3\|=O(\beta^3n^3L^2\rho_\infty)\text{ for }\beta nL\le\tfrac{1}{3}.$$

Moreover $-\beta\sum_i g_i=\alpha\beta\sum_i M_i b_i$, of norm $\alpha\beta C_0$ with $C_0=\|\sum_i M_i b_i\|$ — order-independent. The term $\beta^2\sum_\ell A_{\pi_\ell}P_\ell$ is the only order-dependent term at leading order; write $B(\pi)=\|\sum_\ell A_{\pi_\ell}P_\ell\|$.

Proof. Expand $\prod_{\ell>j}(I-\beta A_{\pi_\ell}) = I - \beta\sum_{\ell>j} A_{\pi_\ell} + O(\beta^2)$ in ψ_π. The β^1 part is $\beta\sum_j g_{\pi_j} = \beta\sum_i g_i$; the β^2 part is $-\beta^2\sum_j\big(\sum_{\ell>j} A_{\pi_\ell}\big)g_{\pi_j} = -\beta^2\sum_\ell A_{\pi_\ell}\sum_{j<\ell} g_{\pi_j}$. Negating gives r_π. For E_3: the β^k coefficient of ψ_π, $k\ge 3$, is a sum over at most $n^k/(k-1)!$ position tuples of products bounded by $\|A\|_{\max}^{k-1}\max_i\|g_i\|$ with $\|A\|_{\max}\le\frac{3}{2}L$, $\max_i\|g_i\|\le\frac{3}{2}\rho_\infty$; the tail is geometric for $\beta n\|A\|_{\max}\le\frac{1}{2}$, giving $\|E_3\| = O(\beta^3 n^3 L^2\rho_\infty)$ with an absolute constant. For the first term, $\sum_i g_i = \sum_i(I-\alpha M_i)b_i = \sum_i b_i - \alpha\sum_i M_i b_i = -\alpha\sum_i M_i b_i$ since $\sum_i b_i = 0$. □

Remark. Since $g_i = b_i + O(\alpha)$ and $A_i = M_i + O(\alpha)$, herding the raw centered values $F_i(z^*) = b_i$ controls P_ℓ to leading order; herding the corrected G_i instead changes nothing (Appendix N).

Lemma 4 (Plateau bound)**.** *Let $e^k = z_0^k - z^*$. If $\|\Phi_\pi\|\le q<1$ then $e^{k+1} = \Phi_\pi e^k + r_\pi$ gives $\|e^K\|\le q^K\|e^0\| + \|r_\pi\|/(1-q)$. Under* (A1)–(A2) *with $\alpha=\beta\le 2\mu/(9L^2)$ one has $q\le e^{-\beta n\mu/4}$, hence $1-q\ge\beta n\mu/8$ whenever $\beta n\mu\le 1$.*

Proof. Unrolling, $e^K = \Phi_\pi^K e^0 + \sum_{k=0}^{K-1}\Phi_\pi^k r_\pi$ and $\|\sum_k\Phi_\pi^k r_\pi\|\le\|r_\pi\|\sum_{k\ge0}q^k = \|r_\pi\|/(1-q)$. For the contraction, $A_i = (I-\beta M_i)M_i$ obeys $\langle A_i z, z\rangle\ge(\mu-\beta L^2)\|z\|^2\ge\frac{\mu}{2}\|z\|^2$ and $\|A_i\|\le(1+\beta L)L\le\frac{3}{2}L$ (both using $\beta\le 2\mu/(9L^2)\le 1/(2L)$). Hence $\|(I-\beta A_i)z\|^2\le(1-\beta\mu+\frac{9}{4}\beta^2L^2)\|z\|^2\le(1-\frac{1}{2}\beta\mu)\|z\|^2$, the last step by $\beta\le 2\mu/(9L^2)$; so $\|I-\beta A_i\|\le 1-\frac{1}{4}\beta\mu$ and $q\le(1-\frac{1}{4}\beta\mu)^n\le e^{-\beta n\mu/4}$. Finally $1-q\ge 1-e^{-\beta n\mu/4}\ge\beta n\mu/8$ for $\beta n\mu\le 1$. □

Combining Lemmas 3–4, with $\|r_\pi\|\le\alpha\beta C_0+\beta^2 B(\pi)+O(\beta^3)$ and $1-q\ge\beta n\mu/8$, the plateau obeys

$$\mathcal{P}(\pi) = \frac{\|r_\pi\|}{1-q} \le \underbrace{O\Big(\tfrac{\alpha C_0}{n\mu}\Big)}_{\mathcal{P}_{\mathrm{EG}}} + \underbrace{O\Big(\tfrac{\beta B(\pi)}{n\mu}\Big)}_{\mathcal{P}_{\mathrm{order}}(\pi)}.$$

Everything reduces to bounding $B(\pi)$.

B Herding Discrepancy

For centered $u_1,\dots,u_n$ ($\sum_i u_i = 0$) and $R^2 = \frac{1}{n}\sum_i\|u_i\|^2$, recall $D(\pi;u) = \max_t\|\sum_{s<t}u_{\pi_s}\|$.

Lemma 5 (Random vs. herded discrepancy)**.** (i) *For a uniformly random π, $\mathbb{E}\left\|\sum_{s\le t}u_{\pi_s}\right\|^2 = \frac{t(n-t)}{n-1}R^2$ exactly, and $\mathbb{E}D(\pi)^2 = \Theta(nR^2)$; in particular $\mathbb{E}D(\pi) = O(R\sqrt{n})$, and the typical prefix at $t=n/2$ is $\Theta(R\sqrt{n})$ in second moment.* (ii) *A balancing subroutine—the self-balancing walk of Alweiss et al. (2021) composed with the signing-to-permutation reduction of Harvey & Samadi (2014), or the latter's offline herding directly—produces an order whose ℓ_2 prefix discrepancy satisfies $D(\pi)\le c\sqrt{d}\,\rho_\infty\log(nd)$ for an absolute constant c, where $\rho_\infty = \max_i\|u_i\|$. When no single component dominates the heterogeneity ($\rho_\infty = \widetilde{O}(R)$), this reads $D(\pi) = \widetilde{O}(\sqrt{d}\,R)$.*

Proof. (i) The identity is the finite-population second moment. For the maximum, write $S_t = \sum_{s\le t}u_{\pi_s}$: sampling without replacement from a centered population makes $X_t = S_t/(n-t)$ a martingale ($\mathbb{E}[u_{\pi_{t+1}}\mid\mathcal{F}_t] = -S_t/(n-t)$), so Doob's L^2 inequality gives $\mathbb{E}\max_{t\le n/2}\|S_t\|^2\le n^2\cdot 4\,\mathbb{E}\|X_{n/2}\|^2\le 16\,\mathbb{E}\|S_{n/2}\|^2 = O(nR^2)$; the second half of the prefix path equals in law the first half under the reversed permutation, so $\mathbb{E}D(\pi)^2\le 32\,\mathbb{E}\|S_{\lceil n/2\rceil}\|^2 = O(nR^2)$. Conversely $D(\pi)\ge\|S_{\lfloor n/2\rfloor}\|$ pointwise, so $\mathbb{E}D(\pi)^2\ge\Theta(nR^2)$; Jensen gives $\mathbb{E}D(\pi) = O(R\sqrt{n})$.

(ii) The self-balancing walk of Alweiss et al. (2021) *signs* the centered vectors with every signed prefix sum bounded *in each coordinate* by $O(\rho_\infty\log(nd))$ with probability $1-1/\operatorname{poly}(nd)$; the signing-to-permutation reduction of Harvey & Samadi (2014) (their Theorem 10) converts any signing with bounded prefix discrepancy into a *permutation* with prefix discrepancy of the same order, and their offline herding result gives a deterministic counterpart. (The greedy rule, Algorithm 3, carries no worst-case guarantee and is not used here.) This is an ℓ_∞ bound. Since $\|x\|_2\le\sqrt{d}\,\|x\|_\infty$, the ℓ_2 prefix discrepancy obeys $D(\pi)\le\sqrt{d}\cdot O(\rho_\infty\log(nd))$. The factor $\sqrt{d}$ is the only place the dimension enters; it is n-independent, hence leaves every n-scaling in this paper unchanged.

(Running the balancing directly in ℓ_2 via a Banaszczyk-type bound (Banaszczyk, 1998) removes the $\sqrt{d}$; as the d-dependence does not affect the n-scaling, we do not pursue this.) □

In the idealized analysis $u_i = b_i$. A deployable method herds the previous epoch's recorded values; the staleness is $O(\beta)$ near the plateau (Appendix H) and empirically free (Appendix N).

C From Discrepancy to the Cross-Term

Recall $B(\pi) = \|\sum_\ell A_{\pi_\ell} P_\ell\|$. Writing $A_i = \bar{A} + \widetilde{A}_i$ with $\sum_i \widetilde{A}_i = 0$, the cross-term splits as $\bar{A}\sum_\ell P_\ell + \sum_\ell \widetilde{A}_{\pi_\ell} P_\ell$: the mean-Jacobian part is governed by the *outer sum* $\sum_\ell P_\ell$, while the centered part—*not* pointwise smaller, since $\left\|\widetilde{A}_i\right\|$ can be $\Theta(L)$—is controlled at the same scale by the random-labeling cancellation of Lemma 15 (Appendix F) for label-blind constructions on generic offsets. That lemma needs prefixes determined by the *values*, whereas P_ℓ collects the corrected offsets $g_i = (I - \alpha M_i) b_i$, which carry the labels; so split $P_\ell = P_\ell^b + (P_\ell - P_\ell^b)$ with $P_\ell^b = \sum_{j<\ell} b_{\pi_j}$. The lemma applies to P^b, and the label-carrying remainder obeys $\left\|\sum_\ell \widetilde{A}_{\pi_\ell}(P_\ell - P_\ell^b)\right\| \le 3\alpha n^2 L^2 \rho_\infty$, which with the standing convention $\alpha = \beta$ (Lemma 1, Theorem 1) and Proposition 1's cap $\beta \le c/(n^{3/2}L)$ is $\widetilde{O}(LR\sqrt{n})$—the same generic scale. (If α and β were decoupled, the cap would have to be imposed on α.) The mean part needs the *same* reduction, and for the same reason: $\sum_\ell P_\ell = \sum_j (n-j) g_{\pi_j}$ has $\sum_i g_i = -\alpha \sum_i M_i b_i \neq 0$, whereas Lemma 14's coin identity needs items summing to zero. Reduce it likewise to $\sum_\ell P_\ell^b = \sum_j (n-j) b_{\pi_j}$, at the cost of $\left\|\bar{A}\right\| \alpha \sum_j (n-j) \left\|M_{\pi_j} b_{\pi_j}\right\| \le \frac{3}{2}\alpha n^2 L^2 \rho_\infty$—the same $\widetilde{O}(LR\sqrt{n})$ scale under the same cap. Lemma 14 is therefore applied to $\sum_\ell P_\ell^b = \sum_{j=1}^n (n-j)\, b_{\pi_j}$.

Lemma 6 (Random cross-term). *For a uniformly random order, $\mathbb{E}\|\sum_\ell P_\ell\|^2 = \Theta(n^3 R^2)$, hence $\mathbb{E}B(\pi) = O(LRn^{3/2})$ and $\mathrm{Var}_\pi\big(\sum_\ell A_{\pi_\ell} P_\ell\big) = O(L^2 R^2 n^3)$; the matching instance-wise lower bound, at the scale set by the mean Jacobian, is Lemma 7 below.*

Proof. The prefixes P_ℓ collect the corrected offsets g_i, which are not centered: split $g_i = g_i' + \bar{g}$ with $\sum_i g_i' = 0$, $\|\bar{g}\| \le \alpha LR$, and RMS scale $R' \in [(1-2\alpha L)R, (1+2\alpha L)R]$ (as in Lemma 12). Then $\sum_\ell P_\ell = \sum_j w_j g'_{\pi_j} + \bar{g}\sum_j w_j$ with $w_j = n - j$; the second term is order-free of norm at most $\alpha LRn^2/2 \le nR/6$ under $\beta nL \le \frac{1}{3}$. The without-replacement identity applied to the centered g' gives $\mathbb{E}\left\|\sum_j w_j g'_{\pi_j}\right\|^2 = R'^2[\frac{n}{n-1}\sum_j w_j^2 - \frac{1}{n-1}(\sum_j w_j)^2]$; here $\sum_j w_j = \binom{n}{2}$ and $\sum_j w_j^2 = \Theta(n^3)$, so the bracket is $\Theta(n^3)$ and the order-free part is lower order. The mean $\|\mathbb{E}\sum_\ell P_\ell\| = O(Rn)$ is lower order than $\sqrt{\mathbb{E}\|\cdot\|^2} = \Theta(Rn^{3/2})$, so the variance is $\Theta(n^3R^2)$; the triangle inequality with $\|A_i\| \le \frac{3}{2}L$ then bounds the cross-term and its variance from above. The lower direction cannot be taken at the L-scale instance-wise—multiplying the fluctuating vector by $\bar{M}$ certifies only $\left\|\bar{M}v\right\| \ge \sigma_{\min}(\bar{M})\,\|v\|$—and is supplied at exactly that scale by the next lemma. □

Lemma 7 (Instance-wise reshuffling-variance lower bound). *Let $v_1, \dots, v_n \in \mathbb{R}^d$ with $\sum_a v_a = 0$, second-moment matrix $B = \sum_a v_a v_a^\top$ and RMS scale $\rho^2 = \frac{1}{n}\sum_a \|v_a\|^2$, and let $H_1, \dots, H_n \in \mathbb{R}^{d\times d}$ with mean $\bar{H} = \frac{1}{n}\sum_a H_a$. For a uniformly random order π set $X(\pi) = \sum_{s<t} H_{\pi_t} v_{\pi_s}$ and, for a unit vector u, the witness statistic $Y_u(\pi) = \sum_t \big(t - \frac{n+1}{2}\big)\langle u, v_{\pi_t}\rangle$. Then, exactly, for every unit w,*

$$\mathrm{Cov}\big(\langle w, X\rangle\,,\, Y_u\big) = -\frac{n+1}{12}\, w^\top \Big(\sum_a H_a\Big) B u, \qquad \mathrm{Var}(Y_u) = \frac{n(n+1)}{12}\, u^\top B u.$$

Consequently, by Cauchy–Schwarz,

$$\mathbb{E}\|X - \mathbb{E}X\|^2 \;\ge\; \frac{n(n+1)}{12}\,\sigma_{\min}(\bar{H})^2\,\lambda_{\max}(B) \;\ge\; \frac{n^3}{12\,d}\,\sigma_{\min}(\bar{H})^2\,\rho^2.$$

Proof. Write $y_a = \langle u, v_a \rangle$ and $c_t = t - \frac{n+1}{2}$, so $\sum_a y_a = 0$, $\sum_t c_t = 0$, $\sum_t c_t^2 = \frac{n(n^2-1)}{12}$. Exchangeability gives $\mathbb{E}[y_{\pi_t} y_{\pi_r}] = m_2\{\mathbb{1}[t = r] - \frac{1}{n-1}\mathbb{1}[t \neq r]\}$ with $m_2 = \frac{1}{n} u^\top B u$, whence $\mathrm{Var}(Y_u) = \frac{n}{n-1} m_2 \sum_t c_t^2 = \frac{n(n+1)}{12} u^\top B u$.

For the covariance put $K_1 = w^\top (\sum_a H_a) B u$ and $K_2 = \sum_a (w^\top H_a v_a) y_a$. Averaging over uniformly random distinct pairs and triples of components (inclusion–exclusion over coincident indices, each simplification using only $\sum_a v_a = 0$, $\sum_a y_a = 0$), for $s < t$,

$$\mathbb{E}\big[w^\top H_{\pi_t} v_{\pi_s} y_{\pi_r}\big] = \begin{cases} (K_1 - K_2)/\{n(n-1)\}, & r = s, \\ -K_2/\{n(n-1)\}, & r = t, \\ (2K_2 - K_1)/\{n(n-1)(n-2)\}, & r \notin \{s, t\}. \end{cases}$$

Summing $\sum_{s<t} \sum_r c_r(\cdot)$ with $\sum_r c_r = 0$ leaves only the $r = s$ and $r = t$ cases relative to the generic one; with $\sum_{s<t} c_s = -\Sigma$ and $\sum_{s<t} c_t = +\Sigma$, $\Sigma = \frac{n(n^2-1)}{12}$, every K_2 contribution cancels and $\mathrm{Cov}(\langle w, X \rangle, Y_u) = -\Sigma K_1/\{n(n-1)\} = -\frac{n+1}{12} K_1$. (Both identities were verified by exhaustive enumeration over S_n, $n \leq 7$.)

Finally take $u = u_1$, the top eigenvector of B, and $w = (\sum_a H_a) u_1 / \|(\sum_a H_a) u_1\|$; Cauchy–Schwarz gives

$$\begin{aligned} \mathbb{E}\,\|X - \mathbb{E}X\|^2 \;\geq\; \mathrm{Var}(\langle w, X \rangle) \;&\geq\; \frac{\mathrm{Cov}(\langle w, X \rangle, Y_{u_1})^2}{\mathrm{Var}(Y_{u_1})} = \frac{n+1}{12n} \lambda_{\max}(B) \Big\| \sum_a H_a u_1 \Big\|^2 \\ &\geq\; \frac{n(n+1)}{12} \sigma_{\min}(\bar{H})^2 \lambda_{\max}(B), \end{aligned}$$

and $\lambda_{\max}(B) \geq \mathrm{tr}(B)/d = n\rho^2/d$. □

Remark (scale and tightness). Under (A2), $\mathrm{sym}(\bar{M}) \succeq \mu I$ gives $\sigma_{\min}(\bar{M}) \geq \mu$: the reshuffling variance can never degenerate below the μ-scale $n^3\mu^2R^2/d$, on *any* instance. The L-scale of Lemma 6 is, however, not instance-wise: for $M_i \equiv \mu I$ the cross-term is exactly linear, $X = \mu \sum_j (n-j) b_{\pi_j}$, with variance $\Theta(n^3\mu^2R^2)$ for *every* draw of the offsets, however large the admissible L—so no condition on the offsets alone (in particular no genericity, cf. (H1) in Appendix H) can restore the L-scale. The correct instance constant is $\sigma_\star := \sigma_{\min}(\bar{M}) \in [\mu, L]$: rotation-dominated mean Jacobians—the rotational families of Section 5, and the balanced signed hard family of Lemma 9, where $\delta = \Theta(\mu)$ gives $\sigma_\star = \Theta(L)$—have $\sigma_\star = \Theta(L)$. The bound is attained up to constants (numerically, with equality for equal Jacobians and rank-one B).

Lemma 8 (Herded cross-term, worst case)**.** *For a herded order with $D(\pi) = \widetilde{O}(R)$, the triangle inequality gives $B(\pi) \leq \|A\|_{\max} \sum_\ell \|P_\ell\| \leq \|A\|_{\max}\, n\, D(\pi) = \widetilde{O}(LRn)$.*

Lemma 9 (Worst-case tightness)**.** *Let n be even and take the balanced signed rotational family of Theorem 5: $M_i = \mu I + \delta s_i J$, $b_i = \rho s_i v$ with $s_i \in \{\pm 1\}$, $\sum_i s_i = 0$, $J^\top = -J$, $J^2 = -I$, $\|v\| = 1$, and $\delta = \mu$ (so $L = \sqrt{2}\,\mu$, $R = \rho$, and $\sigma_\star = \sigma_{\min}(\bar{M}) = \mu = L/\sqrt{2}$). Then for* every *order π, under $\beta n L \leq \frac{1}{3}$,*

$$B(\pi) \;\geq\; \big|\langle Jv, \textstyle\sum_\ell A_{\pi_\ell} P_\ell \rangle\big| \;=\; \frac{\rho n}{2}\big[a\beta\delta(n-1) + c(1-\beta\mu)\big] \;\geq\; \frac{n\delta\rho}{4} \;=\; \Omega(nLR),$$

with $a = \mu - \beta(\mu^2 - \delta^2)$, $c = \delta(1 - 2\beta\mu)$. Since any low-discrepancy order also has $B(\pi) = O(nLR)$ (Lemma 8), the herded cross-term bound is tight, $B(\pi_{\mathrm{herd}}) = \Theta(nLR)$, and the herding gain on this family is $\Theta(\sqrt{n})$—while $\sigma_\star = \Theta(L)$ keeps the SEG-RR *floor of Lemma 7 at full scale, so the same family witnesses the worst-case $\sqrt{n}$ separation of Theorem 1.*

Proof. Write $\sigma_{t-1} = \sum_{r<t} s_{\pi_r}$. Since $s^2 = 1$ and $\sigma_n = 0$, summing $\sigma_t^2 - \sigma_{t-1}^2 = 2 s_{\pi_t} \sigma_{t-1} + 1$ gives the order-*invariant* identity $\sum_t s_{\pi_t} \sigma_{t-1} = -n/2$. Here $A_s = aI + scJ$ and $g_s = \rho[(1 - \beta\mu) sv - \beta\delta Jv]$, so $P_{t-1} = \rho[(1-\beta\mu)\sigma_{t-1} v - \beta\delta(t-1) Jv]$, and the Jv-component of $\sum_t A_{\pi_t} P_{t-1}$ collects two order-free sums, $\sum_t (t-1) = \binom{n}{2}$ and $\sum_t s_{\pi_t} \sigma_{t-1} = -n/2$:

$$\langle Jv, \textstyle\sum_t A_{\pi_t} P_{t-1} \rangle = -\frac{\rho n}{2}\big[a\beta\delta(n-1) + c(1-\beta\mu)\big].$$

Both bracket terms are positive under $\beta nL \le \frac{1}{3}$ and the bracket is at least $c(1-\beta\mu) \ge \delta/2$, giving the display; the identity was verified exactly over all 40320 orders at $n = 8$. (Identical Jacobians with $b_i \in \{\pm v\}$ do *not* give a witness: with $A_i \equiv A$ the order with sign pattern $(+,-,-,+)$ repeated has $\sum_\ell P_\ell = 0$, hence $B(\pi) = 0$, and a herded order need not alternate signs.) □

By Appendix A, $\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \Theta(\beta B(\pi_{\mathrm{herd}})/(n\mu)) = \widetilde{O}(\beta LR/\mu)$.

D Proof of Theorem 1

The herded upper bound is immediate from Appendices A and C: $\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \Theta(\beta B(\pi_{\mathrm{herd}})/(n\mu))$ with $B(\pi_{\mathrm{herd}}) = \widetilde{O}(LRn)$ (Lemma 8), giving $\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{O}(\beta LR/\mu)$. The matching SEG-RR lower bound is Appendix E. Together they give Theorem 1.

We record a second route to the herded upper bound, looser by a factor $1/\beta$ but the one that survives to general smooth F (Appendix H), where the exact expansion of Appendix A is unavailable. Within an epoch, $w_0 = z$, $w_{t+1} = w_t - \beta G_{\pi_t}(w_t)$; telescoping $\|w_{t+1} - z^*\|^2$ and writing $\delta_t = w_t - z = -\beta P_t^w$ with $P_t^w = \sum_{s<t} G_{\pi_s}(w_s)$, one splits $\langle w_t - z^*, G_{\pi_t}(w_t)\rangle$ into an order-free descent term plus drift terms; a Young split of the one term that would form a vector sum, with the resulting order-free $O(nL^2/\mu)\|z - z^*\|^2$ absorbed into the descent, yields the per-epoch inequality

$$\|z' - z^*\|^2 \le (1 - \beta n\mu)\|z - z^*\|^2 + C_1\beta^2 \sum_{t=1}^{n} \|P_t^w\|^2,$$

with $C_1 = O(L + L^2/\mu)$, valid for $\beta n \le 1/(3L)$.

Lemma 10 (Squared-prefix order term)**.** *For $\beta n \le 1/(3L)$, $\max_t \|P_t^w\| \le 2\max_t \left\|\sum_{s<t} g_{\pi_s}\right\|$. Hence a herded order gives $\sum_t \|P_t^w\|^2 = \widetilde{O}(nR^2)$, while* SEG-RR *gives $\mathbb{E}\sum_t \|P_t^w\|^2 = \Theta(n^2R^2)$.*

Proof. $P_t^w - \sum_{s<t} g_{\pi_s} = \sum_{s<t} A_{\pi_s}\delta_s$, of norm $\le \|A\|_{\max} n\beta \max_s \|P_s^w\| \le \frac{1}{2}\max_s \|P_s^w\|$ for $\beta n \le 1/(3L)$ (recall $\|A\|_{\max} \le \frac{3}{2}L$). Then $\left\|\sum_{s<t} g_{\pi_s}\right\| \le D(\pi) + O(\alpha nLR) = \widetilde{O}(R)$ for a herded order. For SEG-RR, the g_i are not centered: splitting $g_i = g_i' + \bar{g}$ as in Lemma 12, $\sum_{s<t} g_{\pi_s} = \sum_{s<t} g'_{\pi_s} + (t-1)\bar{g}$ with $(t-1)\|\bar{g}\| \le \alpha nLR \le R/3$; Lemma 5(i) on the centered g' gives $\mathbb{E}\left\|\sum_{s<t} g'_{\pi_s}\right\|^2 = \Theta(tR^2)$, and the drift part contributes $O(nR^2)$ over the epoch—lower order—so $\mathbb{E}\sum_t \|\cdot\|^2 = \Theta(n^2R^2)$. □

Unrolling the per-epoch inequality gives plateau $\le C_1\beta\,\mathbb{E}[\sum_t \|P_t^w\|^2]/(n\mu)$; with Lemma 10 this is $\widetilde{O}(\beta L^2R^2/\mu^2)$ for a herded order. This squared-prefix form has the same n-dependence as the exact route and extends verbatim to Appendices G–H.

E The Random-Reshuffling Lower Bound

SEG-RR redraws π each epoch, so in $e^{k+1} = \Phi^k e^k + r^k$ the biases r^k are i.i.d. across k. Write $r^k = \bar{\rho} + \xi^k$ with $\bar{\rho} = \mathbb{E}_\pi r^k$ and $\mathbb{E}\xi^k = 0$.

Lemma 11 (SEG full order-dependent tail)**.** *For affine* SEG *with $\beta nL \le \frac{1}{3}$, with $X_g(\pi) = \sum_\ell A_{\pi_\ell} P_\ell$ and $D_g(\pi) = \max_t \left\|\sum_{s<t} g_{\pi_s}\right\|$,*

$$\left\| r_\pi + \beta \textstyle\sum_i g_i - \beta^2 X_g(\pi)\right\| \;\le\; C\,\beta^3 n^2 L^2 D_g(\pi)$$

for an absolute constant C.

Proof. In $\psi_\pi = \beta\sum_j[\prod_{\ell>j}(I - \beta A_{\pi_\ell})]g_{\pi_j}$ the β^k coefficient, $k \ge 3$, is $(-1)^{k-1}\sum_{\ell_1>\cdots>\ell_{k-1}>j} A_{\pi_{\ell_1}}\cdots A_{\pi_{\ell_{k-1}}} g_{\pi_j}$: the innermost sum over j is the g-prefix $P_{\ell_{k-1}}$, bounded by $D_g(\pi)$, so the coefficient has norm at most $\binom{n}{k-1}(\frac{3}{2}L)^{k-1}D_g(\pi)$. The tail is geometric for $\frac{3}{2}\beta nL \le \frac{1}{2}$, giving the display. (Measured at the regime edge $\beta = 1/(3nL)$: the ratio to the stated bound is at most 0.09 over random instances and orders.) □

Lemma 12 (Injected variance)**.** *For affine* SEG *with* $\alpha = \beta \le \min\{1/(3nL),\, 2\mu/(9L^2)\}$, $\mathbb{E}\left\|\xi^k\right\|^2 = \beta^4 \operatorname{Var}_\pi\left(\sum_\ell A_{\pi_\ell} P_\ell\right) + O(\beta^5)$*, and with* $\sigma_\star = \sigma_{\min}(\bar{M}) \in [\mu, L]$*,*

$$c\,\frac{\beta^4 n^3 \sigma_\star^2 R^2}{d} \;\le\; \mathbb{E}\left\|\xi^k\right\|^2 \;\le\; C\,\beta^4 n^3 L^2 R^2,$$

the lower bound holding for $\beta \le c'\sigma_\star/(\sqrt{d}\,nL^2)$*—for which* $\beta \le c'\mu/(\sqrt{d}\,nL^2)$ *suffices; the* $\sigma_\star$ *condition enforces* $\beta nL \le c'\sigma_\star/(\sqrt{d}L) \le c'$*.*

Proof. By Lemma 3 the order-dependent part of r^k is $\beta^2 \sum_\ell A_{\pi_\ell} P_\ell$ up to $O(\beta^3)$; subtracting the mean and taking the second moment gives the variance, bounded above by Lemma 6. For the lower bound, recentre the offsets: $g_a = g'_a + \bar{g}$ with $\sum_a g'_a = 0$, $\|\bar{g}\| \le \alpha L R$, and RMS scale of $\{g'_a\}$ at least $R(1 - 2\alpha L) \ge R/2$. Then $\sum_\ell A_{\pi_\ell} P_\ell = \sum_{s<t} A_{\pi_t} g'_{\pi_s} + \sum_t (t-1) A_{\pi_t} \bar{g}$. Lemma 7 applied to the first term—with $\sigma_{\min}(\bar{A}) \ge \sigma_\star - \alpha L^2 \ge \frac{7}{9}\sigma_\star$ by Weyl, since $\alpha L^2 \le \frac{2}{9}\mu \le \frac{2}{9}\sigma_\star$—gives a standard deviation at least $n^{3/2}\sigma_\star R/(9\sqrt{d})$, while the second term deviates from its mean by at most $\frac{3}{2}\alpha n^2 L^2 R$ for every order ($\left\|A_a - \bar{A}\right\| \le 3L$). Minkowski's inequality absorbs it once $\alpha \le \sigma_\star/(27\sqrt{dn}\,L^2)$. The epoch bias also carries the genuine $O(\beta^3)$ tail beyond the β^2 coefficient: Lemma 11 bounds it pointwise by $C\beta^3 n^2 L^2 D_g(\pi)$, and $\mathbb{E}D_g(\pi)^2 = O(nR^2)$ (Lemma 5(i) on the centered g' plus the $O(R)$ drift part), so its L_2 size is $O(\beta^3 n^{5/2} L^2 R)$; Minkowski keeps half of the leading scale $\beta^2 n^{3/2}\sigma_\star R/(9\sqrt{d})$ once $\beta \le c\,\sigma_\star/(\sqrt{d}\,nL^2)$, for which $\beta \le c\,\mu/(\sqrt{d}\,nL^2)$ suffices. The $\sigma_\star$ tail cap also enforces $\beta nL \le c$; the common caps in the statement remain in force. □

Lemma 13 (SEG-RR stationary lower bound)**.** *For affine* SEG *with* $\alpha = \beta$ *and sufficiently small absolute constants* $c_1, c_2 > 0$*, suppose*

$$\beta \le \min\left\{\frac{c_1}{nL},\; \frac{2\mu}{9L^2},\; \frac{c_2\sigma_\star}{\sqrt{d}\,nL^2}\right\}.$$

Then the stationary SEG-RR *second moment satisfies* $\mathbb{E}\left\|z^\infty - z^*\right\|^2 = \Omega(\beta^4 n^3 \sigma_\star^2 R^2/d)$*. Replacing* $\sigma_\star$ *by* μ *in the third step cap is sufficient but not necessary. The lower bound is always* $\Omega(\beta^4 n^3 \mu^2 R^2/d)$*, and becomes* $\Omega(\beta^4 n^3 L^2 R^2)$ *up to the fixed dimension on rotation-dominated instances (*$\sigma_\star = \Theta(L)$*), still under all three displayed caps.*

Proof. The epoch recursion $e^{k+1} = \Phi^k e^k + \bar{\rho} + \xi^k$ has i.i.d. coefficient pairs (Φ^k, ξ^k)—each a function of the fresh π^k alone—with $\left\|\Phi^k\right\| \le q < 1$ (Lemma 4) and square-integrable noise. It is a contracting affine recursion, so the epoch-start error e^k converges in law to a unique stationary distribution (geometrically, since $\left\|\Phi^k\right\| \le q$), along which $\mathbb{E}\left\|e^k\right\|^2 \to V$, where $V := \mathbb{E}\left\|e^\infty\right\|^2$ is the stationary $\mathbb{E}\left\|z^\infty - z^*\right\|^2$ to be bounded. Set $\sigma^2 := \mathbb{E}\left\|\xi^k\right\|^2 \ge c\,\beta^4 n^3 \sigma_\star^2 R^2/d$ by Lemma 12.

At stationarity $e^\infty \overset{d}{=} \Phi e^\infty_- + \bar{\rho} + \xi$, with $e^\infty_- \overset{d}{=} e^\infty$ the previous state and the fresh epoch (Φ, ξ) independent of the past $\mathcal{F}_-$. Conditioning on $\mathcal{F}_-$ fixes e^∞_- but not (Φ, ξ). Writing $\Phi = I - \Delta$ and $\tilde{\Delta} := \Delta - \mathbb{E}\Delta$, the constant terms cancel in the centered increment, leaving the conditional fluctuation $e^\infty - \mathbb{E}[e^\infty \mid \mathcal{F}_-] = \xi - \tilde{\Delta}\, e^\infty_-$. Since $\mathbb{E}[\|\xi\|^2 \mid \mathcal{F}_-] = \sigma^2$, Cauchy–Schwarz gives

$$\operatorname{Var}(e^\infty \mid \mathcal{F}_-) = \mathbb{E}\big[\left\|\xi - \tilde{\Delta}\, e^\infty_-\right\|^2 \mid \mathcal{F}_-\big] \;\ge\; \sigma^2 - 2\sigma\sqrt{\mathbb{E}\left\|\tilde{\Delta}\right\|^2}\;\left\|e^\infty_-\right\|.$$

Lemma 4 gives $\sup_\pi \|\Delta_\pi\| = O(\beta nL)$, hence $\sup_\pi \left\|\widetilde{\Delta}_\pi\right\| \le 2\sup_\pi \|\Delta_\pi\|$ and $\sqrt{\mathbb{E}\left\|\tilde{\Delta}\right\|^2} \le c_0 \beta nL$ for an absolute constant c_0. Taking expectations, applying the law of total variance ($V \ge \operatorname{Var}(e^\infty) \ge \mathbb{E}\operatorname{Var}(e^\infty \mid \mathcal{F}_-)$) and $\mathbb{E}\left\|e^\infty_-\right\| \le \sqrt{V}$,

$$V \;\ge\; \sigma^2 - 2c_0\,\beta nL\,\sigma\sqrt{V}.$$

This is a closed inequality for V that assumes nothing about the size of e^k. If $V \ge \sigma^2$ the claim is immediate; otherwise $\sqrt{V} < \sigma$, and substituting on the right gives $V \ge \sigma^2(1 - 2c_0\beta nL)$. Either way $V \ge \sigma^2(1 - 2c_0\beta nL)$; for βnL below the absolute constant $1/(4c_0)$—a constant-factor subrange of Theorem 1's step regime, which is all a lower bound needs—this yields $V \ge \frac{1}{2}\sigma^2 = \Omega(\beta^4 n^3 \sigma_\star^2 R^2/d)$. No hypothesis on the magnitude of e^k enters: the bound is the stationary variance that re-drawing the order injects each epoch. □

A fixed herded order has $r^k \equiv r_{\mathrm{herd}}$ constant, so $\xi^k = 0$: no injected variance. Combining Lemma 13 with the herded plateau $\widetilde{O}(\beta^2 L^2 R^2/\mu^2)$ in squared norm (Appendix D),

$$\frac{\mathcal{P}_{\mathrm{order}}(\mathrm{RR})}{\mathcal{P}_{\mathrm{order}}(\textsc{Herd})} \geq \widetilde{\Omega}\Big(\beta^2 n^3 \mu^2 \cdot \frac{\sigma_\star^2}{dL^2}\Big),$$

which on rotation-dominated instances ($\sigma_\star = \Theta(L)$) lets Lemma 12's injection cap $\beta \leq c\sigma_\star/(\sqrt{d}\, nL^2)$ reach $\beta = \Theta(1/(nL))$ up to a fixed-dimension constant. The common caps $\alpha = \beta \leq \min\{c_1/(nL), 2\mu/(9L^2)\}$ must hold as well, and at that step they require $n \gtrsim \kappa$ up to absolute constants; the ratio is then $\widetilde{\Omega}(n\mu^2/(dL^2))$: the worst-case factor n in squared norm ($\sqrt{n}$ in norm), up to κ^2 and the fixed dimension. In the Jacobian worst case ($\sigma_\star = \mu$) the certified step shrinks by $\sigma_\star/L$ and further conditioning powers are lost in the ratio—the price of instance-wise validity (Lemma 7); the conditioning-limited gains measured in Section 5 are consistent with this dependence. This is the lower-bound half of Theorem 1.

E.1 Transfer to SPEG-RR

The same reshuffling-variance argument lower-bounds SPEG-RR. By Lemma 1 its epoch bias differs from SEG's by $\beta^2 A(\pi)$, $A(\pi) = \sum_t M_{\pi_t}(f_{\pi_{t-1}} - f_{\pi_t})$. Marginally the cyclic adjacent pairs (π_t, π_{t-1}) are uniform on ordered pairs, so with $\sum_i f_i = 0$,

$$\mathbb{E}_\pi\Big[\sum_t M_{\pi_t} f_{\pi_{t-1}}\Big] = \frac{n}{n(n-1)} \sum_{i \neq j} M_i f_j = -\frac{1}{n-1} \sum_i M_i f_i,$$

and a covariance count over position pairs (the $O(n)$ pairs sharing an index contribute $O(L^2 R_\infty^2)$ each; the $O(n^2)$ disjoint pairs contribute $O(L^2 R_\infty^2/n)$ each by the $O(1/n)$ total-variation gap between sampling with and without replacement on four indices) gives $\mathrm{Var}_\pi(A) = O(nL^2 R_\infty^2)$. Since $\mathrm{Var}_\pi(X) \geq c\, n^3 \sigma_\star^2 R^2/d$ (Lemma 7), Minkowski's inequality for the square root of the variance yields

$$\mathrm{Var}_\pi(\mathrm{bias}_{\mathrm{SPEG}})^{1/2} \;\geq\; \mathrm{Var}_\pi(\mathrm{bias}_{\mathrm{SEG}})^{1/2}\big(1 - \widetilde{O}(\sqrt{d}\, L/(n\sigma_\star))\big),$$

so SEG-RR's plateau floor transfers to SPEG-RR once $n \gg \sqrt{d}\, L/\sigma_\star$ (in particular $n \gg \sqrt{d}\,\kappa$ always suffices). As in Lemma 12, this leading-coefficient comparison becomes a finite-β statement under that lemma's caps, the tail condition $\beta \leq c\,\sigma_\star/(\sqrt{d}\, nL^2)$ together with $\alpha = \beta \leq \min\{1/(3nL), 2\mu/(9L^2)\}$: SPEG's epoch expansion shares SEG's prefix-controlled higher-order tail (Lemma 11 applies to the augmented one-sweep recursion with the same tuple structure), and the adjacency term $A(\pi)$ is a second-order perturbation only. Combined with Theorem 4, this theoretically explains an n-growing SPEG-RR/SC-Herd separation under the corresponding hypotheses; the observed near-linear scaling is stronger than the certified finite-step rate and remains empirical. (Measured at $n = 8$ over all orders: the variance ratio is 1.31, i.e. slightly larger, consistent with the finite-n size of the A-term.)

F Generic Instances: the Factor n, under Per-Node Split Accuracy

Under Proposition 1's per-node hypothesis ($\mathrm{H}_{\mathrm{split}}$), the leading coefficient on generic instances gains a full factor n and the finite-step plateau separation is $\widetilde{\Omega}(\min\{n/\kappa^2, \sqrt{n}/\kappa\})$. By the split of Appendix C the cross-term is $\bar{A}\sum_\ell P_\ell + \sum_\ell \widetilde{A}_{\pi_\ell} P_\ell$: the centered-Jacobian part is $\widetilde{O}(LR\sqrt{n})$ in L_2 by the random-labeling cancellation (Lemma 15 below), so the generic claim reduces to the *outer-sum* cancellation for the mean part.

Probability-space bookkeeping (Proposition 1). Three independent sources must be kept apart: ω_{split}, the splitter's internal randomness (priorities, pile order, signing); ω_{orient}, the fair child-order coins of Lemma 14; and ω_{label}, the uniform bijection carrying the fixed Jacobian labels onto the unlabeled values. The hypothesis depends on the values, the priorities and ω_{split}, but not on ω_{orient} or ω_{label}:

$$(\mathrm{H}_{\mathrm{split}}) \in \sigma\big(\text{values},\ \text{priorities},\ \omega_{\mathrm{split}}\big),$$

and $\omega_{\mathrm{label}}, \omega_{\mathrm{orient}}$ are independent of that σ-field and of each other, so $\mathcal{L}(\omega_{\mathrm{label}}, \omega_{\mathrm{orient}} \mid \text{cloud}, T, (\mathrm{H}_{\mathrm{split}})) = \mathcal{L}(\omega_{\mathrm{label}}) \otimes \mathcal{L}(\omega_{\mathrm{orient}})$: conditioning on the hypothesis changes neither the

Algorithm 2 Grid differencing with deferral (GD-D) for one balanced split

Input: $b_{1:m} \subset \mathbb{R}^d$ with even m; $K \leftarrow \lceil d/2 \rceil$; $\mathcal{A}_1 \leftarrow \{(b_i, e_i) : i \in [m]\}$; $\mathcal{P} \leftarrow \emptyset$
for $j = 1, \dots, K$ **do**
  $B_j \leftarrow \mathrm{bbox}(\mathcal{A}_j)$; $E_j \leftarrow$ items of $\mathcal{A}_j$ attaining a coordinate min or max ($|E_j| \le 2d$; for $j = 1$ add one further item if $|E_1|$ is odd) {every round: Lemma 16, extrema are not $f|_{B_j}$-distributed}
  $\mathcal{A}'_j \leftarrow \mathcal{A}_j \setminus E_j$; $\mathcal{P} \leftarrow \mathcal{P} \cup \begin{cases} \Delta(E_1), & j = 1 \\ E_j, & j > 1 \end{cases}$ {raw extrema priority-paired; later-round extrema are balanced blocks, deferred as is}
  $\{\mathcal{C}^j_q\}_q \leftarrow \mathrm{Grid}(B_j, \mathcal{A}'_j)$; assign fresh independent uniform priorities to $\mathcal{A}'_j$ {grid anchored to the *pre-removal* box B_j, never recomputed after removal}
  if $j \ge 2$: $\mathcal{P} \leftarrow \mathcal{P} \cup \{$blocks of cells $\not\subset Q_j\}$; keep only cells $\subset Q_j$ {Lemma 17: only the certified core carries two-sided bounds}
  $\mathcal{A}_{j+1} \leftarrow \bigcup_q \Delta(\mathcal{C}^j_q)$ {only core in-cell differences continue}
  $\mathcal{U}_j \leftarrow \bigcup_q \mathrm{Odd}(\mathcal{C}^j_q)$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \begin{cases} \Delta(\mathcal{U}_1), & j = 1, \\ \mathcal{U}_j, & j > 1. \end{cases}$ {defer; never repair the grid}
end for
$((w_\ell, a_\ell))_{\ell=1}^s \leftarrow \mathrm{Order}_{\mathrm{c}\to\mathrm{f}}(\mathcal{P} \cup \mathcal{A}_{K+1})$, piles coarse to fine and uniformly at random within each pile
$r_0 \leftarrow 0$, $c \leftarrow 0$
for $\ell = 1, \dots, s$ **do**
  $\sigma_\ell \leftarrow \arg\min_{\sigma \in \{-1,+1\}} \|r_{\ell-1} + \sigma w_\ell\|$
  $r_\ell \leftarrow r_{\ell-1} + \sigma_\ell w_\ell$, $c \leftarrow c + \sigma_\ell a_\ell$
end for
return $S_+ = \{i : c_i = +1\}$ and $S_- = \{i : c_i = -1\}$

uniformity of the labels nor the fairness of the coins, whereas conditioning on the *labeled* instance would make the label bijection degenerate and void Lemma 15. Both conclusions of Lemma 14—the outer-sum bound and the prefix bound $D(\pi) = \widetilde{O}(R)$—rest on ($\mathrm{H}_{\mathrm{split}}$); neither follows from $\rho_\infty = \widetilde{O}(R)$ alone, which bounds a node's imbalance only by $O(m_v \rho_\infty)$: a node holding $m_v/2$ copies of $+Rv$ and $m_v/2$ of $-Rv$ admits a cardinality-balanced split with $\delta_v = \frac{m_v}{2} R$.

The split primitive. Theorem 2 is constructive rather than existential—for a single split; the recursion's per-node accuracy is Proposition 1's hypothesis ($\mathrm{H}_{\mathrm{split}}$). Algorithm 2 is the split primitive used recursively to build the epoch order. Represent an active item as (v, a), where $a \in \{-1, 0, 1\}^m$ and $v = \sum_i a_i b_i$. For a priority-ordered cell $\mathcal{C} = ((v_r, a_r))_{r=1}^{|\mathcal{C}|}$ (the round's fresh priorities fix the order), define

$$\Delta(\mathcal{C}) := \{(v_{2\ell} - v_{2\ell-1},\ a_{2\ell} - a_{2\ell-1}) : 1 \le \ell \le \lfloor |\mathcal{C}|/2 \rfloor\}, \tag{13}$$

and let $\mathrm{Odd}(\mathcal{C})$ denote its unpaired final item, if one exists. The disjoint-support invariant gives $c \in \{-1, +1\}^m$ and $\mathbf{1}^\top c = 0$, while the returned imbalance is $\|\sum_{i \in S_+} b_i - \sum_{i \in S_-} b_i\| = \|r_s\|$.

Lemma 14 (Generic outer sum)**.** *For centered generic offsets $b_i = x_i - \bar{x}$ with latent i.i.d. x_i from a density bounded above and below on a box of side $O(R)$ (Proposition 1's model), $n = 2^q$, and a herded order π from a recursive balanced bipartition tree satisfying* ($\mathrm{H}_{\mathrm{split}}$)*, $\mathbb{E}_{\mathrm{coins}} \|\sum_{\ell=1}^n P_\ell\| = \widetilde{O}(R\sqrt{n})$—versus $\Theta(Rn^{3/2})$ for a random order—and $D(\pi) = \widetilde{O}(R)$. Both conclusions use* ($\mathrm{H}_{\mathrm{split}}$)*; the second does* not *follow from $\rho_\infty = \widetilde{O}(R)$.*

Construction ($n = 2^q$; the recursion needs an exact equal split at every node, which a non-dyadic n does not admit—already $6 \to 3 + 3$ stalls at the next level, and unequal children change the coin identity below to $\pm\frac{1}{2}(s_{c_2} T_{c_1} - s_{c_1} T_{c_2})$). Recursively split the n vectors into two equal halves; at each node flip a fair coin for child order—including at the size-two leaves—and recurse to depth $\log_2 n$. For an internal node v of size s_v with child sub-sums T_{c_1}, T_{c_2}, write $\delta_v = \|T_{c_1} - T_{c_2}\|/2$. Conditioning on the coins, the deterministic part of $Q = \sum_\ell P_\ell$ vanishes: each item's position is uniform on $[n]$ by the symmetry of the fair coins, so $\mathbb{E}_{\mathrm{coins}} Q = \frac{n-1}{2} \sum_i b_i = 0$—this step uses

the exact root condition $\sum_i b_i = 0$, which the centered model of Proposition 1 supplies identically, an i.i.d. offset model cannot, and the corrected offsets g_i do not satisfy (whence the reduction in Appendix C). (Both points were checked by exhaustive coin enumeration: with centered items $\mathbb{E}\,\|Q\|^2$ matches the identity to machine precision and $\mathbb{E}Q = 0$; with the g_i it exceeds it by exactly $\|\mathbb{E}Q\|^2$. The coefficient is exactly $\frac{n-1}{2}$; its value is immaterial when the items are centered but matters off-centre.) so the independent coins give the exact identity

$$\mathbb{E}\,\|Q\|^2 = \sum_v (s_v/2)^2 \delta_v^2 = \tfrac{n^2}{4} \sum_k 2^{-k} \delta_k^2,$$

δ_k the imbalance at level k (sub-problem size $m = n/2^k$). Thus $\mathbb{E}\,\|Q\| = \widetilde{O}(R\sqrt{n})$ *iff* the splits achieve $\delta_k = \widetilde{O}(R/\sqrt{m})$ — an accuracy that improves with sub-problem size. The same δ_v—bounded by ($\mathrm{H}_{\mathrm{split}}$), which gives $\delta_v \le C_d R \ell(n)$ at every node since $m_v \ge 1$—keep every subtree sum, hence every prefix, at $\widetilde{O}(R)$ by telescoping from $T_{\mathrm{root}} = \sum_i b_i = 0$, so one tree delivers both Lemma 5(ii) and Lemma 14.

Lemma 15 (Random labeling cancels the centered Jacobians)**.** *Let the offsets be centered generic, $b_i = x_i - \bar{x}$ with latent i.i.d. x_i independent of the (fixed) matrices $\{\widetilde{A}_i\}$, and let the ordering construction be* label-blind*: the value order—hence $D(\pi)$—is a function of the offset* values*, the coins, and independent pairing randomness only. Suppose moreover the prefixes $(P_\ell)_\ell$ are likewise value-determined, as the raw prefixes $P_\ell^b = \sum_{j<\ell} b_{\pi_j}$ are; the* corrected *prefixes $\sum_{j<\ell} g_{\pi_j}$ are not, because $g_i = (I - \alpha M_i) b_i$ carries the labels—Appendix C reduces those to P^b at the same scale. Then, conditional on the realized value order, the component labels occupy the positions as a uniform permutation σ independent of $(P_\ell)_\ell$, and for any matrices with $\sum_i \widetilde{A}_i = 0$, $\left\|\widetilde{A}_i\right\| \le \widetilde{L}$,*

$$\mathbb{E}\Big[\big\|\sum_\ell \widetilde{A}_{\sigma(\ell)} P_\ell\big\|^2 \,\Big|\, \textit{values}\Big] \;\le\; \frac{n}{n-1}\,\widetilde{L}^2 \sum_\ell \|P_\ell\|^2 \;\le\; 2\,\widetilde{L}^2\, n D(\pi)^2.$$

In particular, with $\widetilde{A}_i = A_i - \bar{A}$ (so $\widetilde{L} \le 3L$) and a herded value order ($D = \widetilde{O}(R)$), the centered-Jacobian cross-term is $\widetilde{O}(LR\sqrt{n})$ in L_2—the same scale as the mean-Jacobian part, even though $\left\|\widetilde{A}_i\right\|$ can be $\Theta(L)$ pointwise.

Proof. Probability space: three independent sources—(1) the latent i.i.d. draws x_i together with i.i.d. priorities r_i, from which the construction sees only the unlabeled *marked centered* cloud $\{(b_i, r_i)\} = \{(x_i - \bar{x}, r_i)\}$; (2) the label-to-marked-point bijection; (3) the child-order coins. The b_i are *not* independent—they sum to zero—but the centering map $(x_1, \dots, x_n) \mapsto (x_1 - \bar{x}, \dots, x_n - \bar{x})$ is permutation-equivariant, so the marked centered family $(b_i, r_i)_{i \le n}$ is exchangeable, and that is all the argument uses: conditional on the unlabeled marked cloud every label-to-marked-point bijection is equally likely, uniformity holding *because the Jacobians are fixed constants* (were M_i a function of b_i, the assignment would not be uniform). The grid and cell memberships are measurable in the unlabeled marked cloud alone; so are the priority-order pairing tree and leftover selections; the greedy signing is deterministic given these; the value order—hence every P_ℓ and $D(\pi)$—is measurable in the unlabeled marked cloud and the coins. The labels enter only through (2), which is independent of all the rest; hence, conditional on the value order, the label at position ℓ is $\sigma(\ell)$ for a uniform σ independent of the prefixes. For the moment bound: on the diagonal $\mathbb{E}[\widetilde{A}_{\sigma(\ell)}^\top \widetilde{A}_{\sigma(\ell)}] = \frac{1}{n}\sum_i \widetilde{A}_i^\top \widetilde{A}_i$, while for $\ell \ne \ell'$ centredness gives $\mathbb{E}\left\langle \widetilde{A}_{\sigma(\ell)} P_\ell, \widetilde{A}_{\sigma(\ell')} P_{\ell'} \right\rangle = -\frac{1}{n(n-1)} \sum_i \left\langle \widetilde{A}_i P_\ell, \widetilde{A}_i P_{\ell'} \right\rangle$; summing,

$$\mathbb{E}\,\|Z\|^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_i \sum_\ell \left\|\widetilde{A}_i P_\ell\right\|^2 - \frac{1}{n(n-1)} \sum_i \left\|\widetilde{A}_i \textstyle\sum_\ell P_\ell\right\|^2 \;\le\; \tfrac{n}{n-1}\,\widetilde{L}^2 \textstyle\sum_\ell \|P_\ell\|^2,$$

the subtracted term being nonnegative. $\|P_\ell\| \le D(\pi)$ finishes. (Verified on the realized construction: the centered term's median norm fits $n^{0.47}$ over $n = 32$–256, and the full heterogeneous cross-term fits $n^{0.59}$.) □

Finite-β remark. At the generic scale the worst-case tail bound of Lemma 11 is the binding constraint: $\widetilde{O}(\beta^3 n^2 L^2 R) \le \widetilde{O}(\beta^2 LR\sqrt{n})$ forces $\beta \lesssim 1/(n^{3/2}L)$, which together with $\beta \le c\mu/(nL^2)$

is the step regime of Proposition 1 and Corollary 1. Sharpening the tail at the generic scale—a maximal version of the coin identity for *partial* outer sums, plus a degree-two analogue of Lemma 15 for the doubly-centered term $\sum_{\ell>\ell'} \widetilde{A}_{\pi_\ell}\widetilde{A}_{\pi_{\ell'}}P_{\ell'}$—would extend the generic guarantee to $\beta = \Theta(1/(nL))$; we do not pursue this here.

Everything therefore rests on a splitting routine of accuracy $\widetilde{O}(R/\sqrt{m})$. We give one whose analysis needs no unproven hypothesis.

Why the natural construction is not enough. The obvious route iterates *nearest-neighbour* matching-and-differencing, a vector form of the differencing method of Karmarker & Karp (1983). It fails twice. First, the step "every nearest-neighbour distance is small, hence a nearest-neighbour *matching* has all edges small" is invalid: in a perfect matching, once close pairs are consumed the remaining points are far apart, and the Bárány–Grinberg signing is driven by the *largest* edge. Empirically the imbalance of both greedy nearest-neighbour variants does not decay with m at all — it sits at the Steinitz floor in every dimension (Appendix N). Naming the primitive a *minimum-weight* (or bottleneck) matching—the object of the Euclidean matching theorems (Ajtai et al., 1984)—repairs the max-edge scaling, but a second problem remains: rounds $2, \dots, \lceil d/2 \rceil$ act on *difference* vectors, and because a nearest-neighbour pairing is a function of the point values, conditioning on it distorts the law of those differences, so the density hypothesis the matching bound needs is not available. Our construction removes both obstacles by making the pairing independent of the point values.

A grid repair that fails, measurably. The obvious grid construction pairs within cells and repairs the odd cells' leftovers by pairing them *across* cells along a walk of the grid (GD-PR). This preserves exact balance, but its induction is *false*. A repair difference joins two cells, so it is centred at the inter-cell displacement and can vanish near the origin; the next round's cloud is then a mixture of an in-cell bulk with two-sided density and a spread-out repair minority with an upper bound only, and the per-cell count lower bound that any parity argument needs fails in the outer cells that only the minority reaches. We measured rather than argued (Table 23): at occupancy $t = 2$ the repair pairs are $\approx 25\%$ of all pairs and reach 8–22$\times$ the cell scale, the round-2 grid is 62–78% empty, and the longest run of non-odd cells is 88–203 against the $\approx \log_2(\#\text{cells}) \approx 10$ a parity lemma would need. Raising t restores the finite-m behaviour but not the asymptotics (the repair edges carry a $\log$ factor, so the round-2 support exceeds the bulk by $\Theta(\log m)$), and a hierarchical-promotion variant measures worse. The lesson is structural: *never pair across cells*.

The routine GD-D**.** One round on the current M vectors, in this order. *First* record the current cloud's bounding box B_j (origin at its coordinate minima, side $s = \Theta(\text{span} \cdot (t/M)^{1/d})$, $t = \Theta(1)$)—this makes the routine translation-equivariant, which Lemma 21 needs; *then* set aside the $\leq 2d$ items attaining the current cloud's coordinate extrema (Lemma 16; at every round, not only the first: each round's anchor manufactures its own extremal points, which are not distributed as the anchor-restricted law)—round-1 extrema are raw points, priority-paired into pile 1; later-round extrema are already balanced pair-difference blocks and defer to pile j as they are, $O_d(1)$ items per round over $O_d(1)$ rounds, so no rate changes—*then* lay the grid anchored to the *pre-removal* box B_j (never re-anchor after removal: a fresh bounding box would manufacture fresh extrema and void Lemma 16's conditioning), so that the recursion's input is conditionally *independent with common two-sided density bounds* (not i.i.d.: later rounds are not identically distributed). From round 2 on, only cells contained in the recursive core $Q_j = [-\rho_{j-1}/4, \rho_{j-1}/4]^d$—the strict inner half of the certified cube $[-\rho_{j-1}/2, \rho_{j-1}/2]^d$, both known deterministically from the previous round's cell side, not estimated from the current sample—are differenced again: the induction's lower density bound is *central*, and central minorization does not extend to the full bounding box (a uniform cell's difference has triangular density vanishing at its support boundary), so outer cells carry no lower bound and their blocks defer to pile j as they are (Lemma 17). Priorities: assign the active blocks fresh independent uniform *priorities* at the start of every round—fresh, so that the pairing rule is value-independent given the current cell structure with no appeal to inherited order; pair *within* cells by priority order, and take an odd cell's leftover to be its lowest-priority member—one rule, value-independent *and* label-blind, used by the analysis and the implementation alike; the odd cells' leftovers are *deferred* — round-1 leftovers (raw points) are likewise priority-paired among themselves into scale-R differences ("pile 1"), later-round leftovers are already differences and join pile j as they are — and only the in-cell differences are differenced again. After $k^* = \lceil d/2 \rceil$ rounds, sign coarse to fine—piles in decreasing scale, the final chain last—processing each pile in *uniformly*

random order with the greedy sign rule. Every signed vector is a pair difference, so any signs give an exactly equal bipartition; the cost is $O(md)$ per round.

Lemma 16 (Anchor conditioning)**.** *Let $Y_1, \dots, Y_M$ be* independent*—not necessarily identically distributed—with densities f_i obeying the* common *two-sided bounds $H_{\min} \le f_i \le H_{\max}$ on a box of side r and $f_i = 0$ outside. (Round 1 satisfies this by the model assumption, with $f_i \equiv f$. Later rounds do* not*: (IH_j) supplies a lower bound only near the origin, and that cannot be upgraded—a uniform cell's difference has triangular density vanishing at its support boundary—so rounds $j \ge 2$ are handled by Lemma 17 instead, whose independence part reuses* (a) *but whose geometry is restricted to a certified central core.) Let A be the $2d$ coordinatewise sample extrema, $E \subset [M]$ the (a.s. unique) indices attaining them ($|E| \le 2d$), and $\mathcal{B} = \mathcal{B}(A)$ the sample's bounding box. Then*

(a) *conditionally on the pair (A, E)—not on A alone—the $M - |E|$ non-extremal points remain* independent*, the i-th with density $f_i|_{\mathcal{B}} = f_i \mathbf{1}_{\mathcal{B}} / \Pr[Y_i \in \mathcal{B}]$, and these retain common two-sided bounds;*

(b) *on the good-anchor event $\mathcal{G} = \{$every coordinate span of $\mathcal{B}$ is $\ge r/2\}$ we have $\Pr[Y \in \mathcal{B}] \ge H_{\min}(r/2)^d$, so $f|_{\mathcal{B}}$ satisfies two-sided bounds with constants depending only on d and $H_{\max}/H_{\min}$—uniformly in the realized anchor;*

(c) $\Pr[\mathcal{G}^c] \le 4d\,(1 - H_{\min} r^d/4)^M = e^{-\Omega(M)}$.

The $|E|$ extremal points are not *$f|_{\mathcal{B}}$-distributed and are therefore excluded from the differencing recursion: they are priority-paired among themselves and deferred to pile 1 (one further lowest-priority point joins them if $|E|$ is odd). Since $|E| \le 2d+1 = O_d(1)$ while $N_1 = \Omega(M)$, this changes no pile's aggregate moment by more than $O_d(\rho/N_1)$ and leaves the exact cardinality balance intact.*

Proof. (a) With a density the extrema are a.s. attained uniquely. Given (A, E) the conditioning event is the product $\bigcap_{e \in E}\{Y_e = a_e\} \cap \bigcap_{j \notin E}\{Y_j \in \mathcal{B}\}$; the coordinates being independent across samples, the non-extremal factors renormalize individually to $f_i|_{\mathcal{B}}$—independence is preserved factor by factor, identical distribution is not needed anywhere below. Conditioning on A alone would leave the identity of the extremal indices random and the remaining points merely exchangeable, which is why E is included. (b) $\Pr[Y_i \in \mathcal{B}] \ge H_{\min}\,\mathrm{vol}(\mathcal{B}) \ge H_{\min}(r/2)^d$ on $\mathcal{G}$, and $\Pr[Y_i \in \mathcal{B}] \le 1$ always, so every $f_i|_{\mathcal{B}} \in [H_{\min}, H_{\max}/(H_{\min}(r/2)^d)]$ on $\mathcal{B}$—bounds independent of i and of the realized anchor. (c) If some point lies in the bottom quarter-slab of coordinate k and another in the top quarter-slab, that coordinate's span is $\ge r/2$; each slab has f-mass $\ge H_{\min} r^d/4$, so the failure probability per coordinate and per side is $(1 - H_{\min} r^d/4)^M$, and a union bound over $2d$ sides gives the claim. □

The induction hypothesis carried from round to round is *not* that the cloud is i.i.d.—it is not, for $j \ge 2$, since different source cells give different difference laws—but

> conditionally on $\mathcal{F}_{j-1}$ (all earlier anchors, cell structures, counts and priorities) the round-j vectors are *independent*, each with a density obeying the *common* two-sided bounds of scale ρ_{j-1}: bounded above, and bounded below on the *cube* $[-\rho_{j-1}/2, \rho_{j-1}/2]^d$ about the origin—the region the convolution bound actually certifies, via $\mathrm{vol}(c \cap (c - x)) \ge (\rho_{j-1}/2)^d$ for $\|x\|_\infty \le \rho_{j-1}/2$—with support in the $\sqrt{d}\,\rho_{j-1}$-ball. (IH_j)

The common lower bound is what survives non-identical distribution: a cell's difference law $f|_c \star \widetilde{f|_c}$ is centred at the origin and bounded below on the cube $[-s/2, s/2]^d$ (i.e. for $\|x\|_\infty \le s/2$) *whatever the cell*, and all cells of a round share the side s.

Lemma 17 (Core propagation for rounds $j \ge 2$)**.** *Assume (IH_j) at scale $\rho = \rho_{j-1}$: the M_j active blocks are conditionally independent given $\mathcal{F}_{j-1}$, each with a density $f_i \le C\rho^{-d}$, $f_i \ge c\rho^{-d}$ on the* certified cube *$K_j := [-\rho/2, \rho/2]^d$, and* $\mathrm{supp}\, f_i \subset B(0, \sqrt{d}\,\rho)$*. Let B_j be the sample bounding box and E_j the extrema, and let the* recursive core *be the strict inner half $Q_j := [-\rho/4, \rho/4]^d \subset K_j$—both $\mathcal{F}_{j-1}$-measurable, hence deterministic before round j's draws, with no estimated constants. Certified region and recursive core are* distinct *objects: the density bound lives on K_j, the recursion on Q_j. There is a good event of probability $1 - e^{-\Omega(M_j)}$ on which:*

(1) *every coordinate of B_j extends beyond $\pm 3\rho/8$, and every coordinate span lies in $[3\rho/4,\ 2\sqrt{d}\,\rho]$;*

(2) $B_j \supset [-3\rho/8, 3\rho/8]^d \supset Q_j$, *so* $\operatorname{dist}_\infty(Q_j, \partial B_j) \geq \rho/8$*;*

(3) *conditionally on* (B_j, E_j) *the non-extremal blocks are independent (Lemma 16*(a)*, whose proof uses only independence and measurability, not the density bounds) and* $\Omega(M_j)$ *of them land in* Q_j*;*

(4) *every grid cell contained in* Q_j *lies inside* K_j *and so carries the full two-sided bounds* $c'\rho^{-d} \leq f_i \leq C\rho^{-d}$ *on the cell, hence occupancy probabilities* $\Theta(t/M_j)$*, two-sided;*

(5) *restricting the recursion to cells contained in* Q_j*, the convolution step of Lemma 18*(i) *applies verbatim on each recursed cell and reproduces* (IH_j) *at scale* $\rho_j := s_j$*—the recursed cell side: a side-*s_j *cell's difference law has* $\mathrm{vol}(c \cap (c-x)) \geq (s_j/2)^d$ *for* $\|x\|_\infty \leq s_j/2$ *and support in* $B_2(0, \sqrt{d}\, s_j)$*, which is* (IH_j) *at scale* s_j *with* $K_{j+1} = [-s_j/2, s_j/2]^d$ *verbatim (setting* $\rho_j = \sqrt{d}\, s_j$ *instead would inflate the next certified cube beyond the difference support's coordinate range)—with population* $M_{j+1} = \Omega(M_j)$*;*

(6) *the blocks of non-core cells defer to pile* j *as they are; since* $|\langle V, u\rangle| \geq 0$ *they only dilute the pile's aggregate directional moment by a constant factor, while the* $\Omega(M_j)$ *core cells supply the odd-parity far-slab members of Lemma 19, so the pile bound survives with a smaller* $c(d,t)$*.*

Proof. (1) Upper: $\operatorname{supp} f_i \subset B(0, \sqrt{d}\,\rho)$. Lower: for each coordinate k and each sign, the strip $S_k^{\pm} := \{x \in K_j : \pm x_k \in [3\rho/8, \rho/2]\}$ lies inside the certified cube, so every f_i gives it mass $\geq c\rho^{-d} \cdot (\rho/8)\rho^{d-1} = c/8$; a Chernoff bound over the M_j independent blocks and a union over the $2d$ strips put a sample in every strip except with probability $2d\, e^{-\Omega(M_j)}$, and a sample in $S_k^{\pm}$ pushes the k-th coordinate of B_j beyond $\pm 3\rho/8$. (This replaces the quarter-slab argument of Lemma 16(c): at rounds $j \geq 2$ the support is not a box, so the strips must be strips *of the certified cube*.) (2) By (1), $B_j \supset [-3\rho/8, 3\rho/8]^d \supset Q_j$ with clearance $3\rho/8 - \rho/4 = \rho/8$ in every coordinate. Taking the recursive core to be the *full* certified cube would be false: certification of density on a region says nothing about samples reaching its boundary—for the uniform law on K_j, which satisfies (IH_j) verbatim, the sample bounding box is a proper subset of K_j almost surely, so $K_j \subset B_j$ has probability *zero*. Density certification on a region and containment of that region in the sample bounding box are different statements; the strict inner core buys the second from strips of the first. (3) Independence is Lemma 16(a); each block has $\Pr[V_i \in Q_j \mid B_j, E_j] \geq c\,\rho^{-d}\,\mathrm{vol}(Q_j) = c \cdot 4^{-d} > 0$ —dimension-explicit—and Bernstein gives $\Omega(M_j)$ hits. (4) A cell $c \subset Q_j \subset K_j$ lies where the unconditional bounds are two-sided; the anchor conditioning divides by $\Pr[V_i \in B_j] \in [c_3(d), 1]$, uniform by (1). (5) On a recursed cell both pair members have full two-sided conditional bounds, so $(g_1 \star \widetilde{g}_2)(x) \geq (c'\rho^{-d})^2\,\mathrm{vol}(c \cap (c-x))$ for $\|x\| \leq s_j/2$ and $\leq C\rho^{-d}$ everywhere—exactly Lemma 18(i), which was proved *for cells carrying two-sided bounds* and is here applied only to such cells. The population step needs care: $\Omega(M_j)$ core *hits* do not by themselves give $\Omega(M_j)$ *pairs*—if every hit sat in its own cell there would be none. Write $M_{j+1} = \sum_{C \subset Q_j} \lfloor N_C/2 \rfloor$. By (4) a core cell has $\lambda_C = \sum_i p_{i,C} = \Theta(t)$ with every $p_{i,C} = O(1/M_j)$, and there are $\Theta_d(M_j/t)$ core cells; for independent indicators with such λ_C, $\Pr[N_C \geq 2] \geq 1 - \prod_i (1 - p_{i,C}) - \sum_i p_{i,C} \prod_{i' \neq i} (1 - p_{i',C}) \geq c_4(t) > 0$ (elementary; no identical-distribution assumption), so $\mathbb{E}M_{j+1} \geq c_4 \cdot \Theta_d(M_j/t) = \Omega_d(M_j)$. Concentration is run at the anchor level, in the order of Lemma 19: conditionally on $(\mathcal{F}_{j-1}, B_j, E_j)$ the grid is deterministic and the blocks independent, and moving one block changes $\sum_C \lfloor N_C/2 \rfloor$ by at most 2 (one departure, one arrival), so McDiarmid gives $M_{j+1} \geq \frac{1}{2}\mathbb{E}M_{j+1} = \Omega_d(M_j)$ except with probability $e^{-\Omega(M_j)}$. (6) The deferred blocks enter the pile average with nonnegative terms and multiply N_j by at most a constant; the parity/concentration machinery of Lemma 19 runs over the core cells, where (4) supplies the $p_{i,C}$ bounds it needs, yielding $\Omega(M_j)$ odd interior far-slab core cells with centers $|\langle c_C, u\rangle| = \Omega(\rho)$ outside a slab of width $\rho/16$. □

Lemma 16 cannot be invoked at rounds $j \geq 2$: its full-support hypothesis is not supplied by (IH_j)—central minorization does not extend to the bounding box, and cannot (the triangular-density example above). The recursion is therefore restricted to the strict inner core Q_j of the certified cube K_j, which is what Algorithm 2 executes.

Lemma 18 (Clean propagation)**.** *Assume* (IH_j) *at round* j*. Work on the good-anchor event and condition on* $\mathcal{S}_j := \sigma\big((A,E),$ *the cell label* $C(i)$ *of* every *index* i*, the priority pairing*$\big)$*—the full*

membership pattern, not the counts alone: for non-identically-distributed points, conditioning only on $(N_C)_C$ leaves which index landed where random and does not *preserve independence; by Lemma 16(a) the round's points remain independent, and different cells' contents are independent given $\mathcal{S}_j$, each V_i following its own density restricted to its assigned cell; the* two-sided *bounds on a cell hold at round* 1 *on every cell (box model, Lemma 16(b)) and at rounds $j \geq 2$ on the* core *cells only (Lemma 17(4))—which is why only core cells are differenced again. All statements below are conditional on that σ-field. (i) In-cell differences of disjoint (priority-paired, value-independent) pairs are conditionally independent, with density $f|_c \star \widetilde{f|_c}$ — bounded above by $\|f|_c\|_\infty$ and below on the cube $[-s/2, s/2]^d$, support in the $\sqrt{d}\,s$-ball. (ii) The leftover of an odd cell, being its* lowest-priority *member (priorities independent of values), is a value-independent selection* within its cell*: it has the cell-conditioned law $f_{\mathrm{diff}}(\cdot \mid C)$. This is* not *two-sided at the pile scale—cell-conditioning destroys central density (a far cell's leftover is supported far from the origin, and symmetry of orientation does not restore mass near zero)—and what survives aggregation over the pile is a directional first moment, Lemma 19. (iii) The next round's input is exactly the in-cell differences of the recursed (core) cells — support $\leq \sqrt{d}\,s$, no mixture, population $\Omega(M_j)$ by Lemma 17(5). Hence the recursion's input model reproduces itself with $\rho_j := s_j = \widetilde{O}_d(\rho_{j-1}(t/M_j)^{1/d})$.*

Proof. (i) is the convolution bound, which needs only the common bounds and not identical laws: the difference of two *independent* draws with densities g_1, g_2 both in $[H_c^-, H_c^+]$ on c has density $(g_1 \star \widetilde{g_2})(x) \geq (H_c^-)^2 \operatorname{vol}(c \cap (c - x))$ for $\|x\| \leq s/2$ and $\leq H_c^+$ everywhere; this reproduces (IH_j) at scale ρ_j; disjoint pairs use disjoint points. (ii) Given the full membership pattern $\mathcal{S}_j$ (never the counts alone—see above), a cell's members are independent with common two-sided bounds on c (not necessarily identically distributed); the lowest-priority rule does not look at values, and the leftover is disjoint from the cell's pairs. (iii) is the definition of the routine. Measured: the round-2 support over s is 1.28–1.52 against $\sqrt{d} = 1.41$–1.73 — no log factor (Appendix N). □

Lemma 19 (Aggregate pile moment)**.** *Two conditioning levels are used, in a fixed order: the* anchor *σ-field $\sigma(\mathcal{F}_{j-1}, B_j, E_j)$ (prior rounds, the current bounding box, the extracted extrema), on which occupancy, parity and the concentration below live; and the finer full-membership σ-field $\mathcal{S}_j$ of Lemma 18, on which the cell-conditioned laws live. With probability $1 - 1/\operatorname{poly}(m)$ the membership realization lies in an $\mathcal{S}_j$-measurable good event $\mathcal{G}_j$, and on $\mathcal{G}_j$, jointly for every round j and every unit u,*

$$\frac{1}{N_j} \sum_{V \in \text{pile } j} \mathbb{E}\big[\, |\langle V, u\rangle| \,\big|\, \mathcal{S}_j \big] \;\geq\; c(d,t)\,\rho_{j-1};$$

the same bound holds for pile 1 *(priority-matched leftover pairs) at scale R, and for the final chain at scale ρ_{k^*} directly from Lemma 18(i). For $j \geq 2$ the pile also contains the deferred blocks of non-core cells (Lemma 17(6)); they enter the average with nonnegative terms and inflate N_j by at most a constant factor, so the bound holds with a smaller $c(d,t)$, the odd-parity far-slab members being supplied by the core cells.*

Proof. A pile-j member in cell C has $\mathbb{E}[\,|\langle V, u\rangle| \mid V \in C] \geq |\langle c_C, u\rangle| - \sqrt{d}\,s_j$. By Lemma 18(i) the round-$(j-1)$ difference law has two-sided density at scale ρ_{j-1}, so for every unit u at least a $c(d)$-fraction of its mass lies outside the slab $|\langle x, u\rangle| \leq \rho_{j-1}/(8\sqrt{d})$; hence $\Omega(M_j)$ of round j's input vectors land in far-slab cells. For the parities we avoid Poissonization altogether—the round-j occupancies are *not* multinomial for $j \geq 2$, since by (IH_j) the points are independent but not identically distributed, so no exchangeable (M_j, p) model is available and no de-Poissonization identity applies. Two elementary facts suffice instead, both valid for independent non-identically-distributed points. First, for a cell C the count $N_C = \sum_i \mathbf{1}\{V_i \in C\}$ is a sum of independent indicators with $p_{i,C} = \Pr[V_i \in C]$, so the generating function at -1 gives the *exact* parity probability

$$\Pr[N_C \text{ odd}] = \tfrac{1}{2}\Big(1 - \prod_i (1 - 2p_{i,C})\Big) \;\geq\; \tfrac{1}{2}\big(1 - e^{-2\lambda_C}\big) =: c_1(t) > 0, \qquad \lambda_C = \textstyle\sum_i p_{i,C},$$

where the far-slab cells are further restricted as follows. At round 1 the support *is* the box, and it suffices to take cells at distance $\geq s$ from the bounding box's boundary, so no cell is clipped by the support. At rounds $j \geq 2$ that inference is void—the support is *not* the bounding box, and distance from ∂B_j does not put a cell inside $\operatorname{supp} f_i$ (the same geometric error as the retired convex-support remark, one level down)—so we instead restrict to *core-interior* cells: $C \subset Q_j$ with

$\mathrm{dist}_\infty(C, \partial Q_j) \geq s_j$. Such a cell lies inside the certified cube K_j with margin, where the genuine two-sided bounds hold, giving $c(d,t) \leq \lambda_C \leq C(d,t)$, so $c_1(t) > 0$ is a genuine constant rather than a vanishing one. Discarding Q_j's width-s_j boundary shell costs an $O_d(s_j/\rho) = O_d((t/M_j)^{1/d})$ fraction of the core mass, so $\Omega(M_j)$ of the round's vectors still land in core-interior far-slab cells. The bound uses $0 \leq \prod_i (1 - 2p_{i,C}) \leq e^{-2\lambda_C}$, which requires every factor to be *nonnegative*: the bound is false otherwise—two cells with $p = 0.9$ give $\Pr[N \text{ odd}] = 0.18$ against the bound's 0.486, since negative factors multiply back up. The needed smallness is supplied by the grid itself: by (IH_j) each round-j density is bounded above by $C\rho_{j-1}^{-d}$ and a cell has volume $O_d(\rho_{j-1}^d\, t/M_j)$, so $p_{i,C} \leq C_{d,t}/M_j \leq \frac{1}{4}$ once $M_j \geq M_0(d,t) := 4C_{d,t}$. For the finitely many rounds with $M_j < M_0(d,t)$—a fixed-size base case, d and t being constants—the whole cloud is absorbed into the terminal pile and its contribution into the $\widetilde{O}(\cdot)$ constants. Hence $\mathbb{E}[\#\{\text{odd far-slab cells}\}] \geq c_1(t) \cdot \Omega(M_j)$. Second, the concentration is run at the *anchor* level, and the order of conditioning matters twice over. Condition on $\sigma(\mathcal{F}_{j-1}, B_j, E_j)$: by Lemma 16 the non-extremal vectors remain independent with the common bounds, their conditional supports lie in B_j, and the grid's origin and side are now *deterministic*. Only then is the bounded-difference property true: moving one non-extremal V_i within its conditional support relocates that point alone, flipping the parity of at most its old and new cells, so the odd-cell count changes by at most 2. For the unconditioned sample function this is *false*—the grid is anchored to the cloud's own bounding box, so moving a point that attains a coordinate extremum re-anchors the grid and can relabel every cell. McDiarmid, conditionally on the anchor, gives

$$\Pr\big[\#\{\text{odd far-slab cells}\} < \tfrac{1}{2}\mathbb{E}[\cdot] \;\big|\; \mathcal{F}_{j-1}, B_j, E_j\big] \;\leq\; \exp\big(-\Omega(M_j)\big),$$

with no multinomial, Poisson or identical-distribution assumption anywhere. Conditioning on the full membership $\mathcal{S}_j$ *first* would freeze the counts and leave nothing to concentrate; conditioning on it *afterwards* is legitimate, since each cell's odd/far-slab/interior status is a function of the membership pattern, so the event $\mathcal{G}_j = \{\#\{\text{odd interior far-slab cells}\} \geq \frac{1}{2}\mathbb{E}[\cdot]\}$ is $\mathcal{S}_j$-measurable and on it the displayed average of $\mathcal{S}_j$-conditional moments is bounded below deterministically. Averaging gives the display for a fixed u; a C^d-net of *constant* angular resolution $c(d,t)/8$ over the sphere—on which $u \mapsto \frac{1}{N_j}\sum \mathbb{E}|\langle V, u\rangle|$ is ρ_{j-1}-Lipschitz, so the net error is at most $\frac{1}{8}c(d,t)\rho_{j-1}$—plus a union bound over the $k^* = O_d(1)$ rounds upgrade it to the stated joint statement, with total failure probability $C^d k^* e^{-\Omega(\min_j M_j)} = 1/\operatorname{poly}(m)$. For pile 1: for every u at least a $c(d)$-fraction of the odd cells' centers lies on each side of the median slab of $\langle c_C, u\rangle$ at scale R, and the uniformly random priority matching creates $\Omega(N_1)$ pairs with $|\langle c_a - c_b, u\rangle| \geq cR$ with probability $1 - e^{-\Omega(N_1)}$. □

Lemma 20 (Steering, random pile order)**.** *Let $v_1, \ldots, v_N$ be conditionally independent given the grid anchor and cell structure $\mathcal{C}$ (not necessarily identically distributed) with $\|v_i\| \leq \rho$ and aggregate directional moment $\frac{1}{N}\sum_i \mathbb{E}|\langle v_i, u\rangle| \geq c_1\rho$ for every unit u. Process them in uniformly random order with the greedy signing $r_t = r_{t-1} + \sigma_t v_t$, $\sigma_t = -\operatorname{sign}\langle v_t, r_{t-1}\rangle$. Then, with probability $1-\delta$, from any r_0 with $N \geq C(d)(\|r_0\|/\rho + \log(N/\delta))$,*

$$\|r_N\| \;\leq\; C'\rho\log(N/\delta)$$

—the entire *pile processed, not half of it: the algorithm does not stop at the pile's midpoint, and a bound at $N/2$ says nothing about r_N, since the second half could deterministically add up to $N\rho/2$.*

Proof. (i) The greedy sign at step t points along the *adaptive* direction $\hat{r}_{t-1} = r_{t-1}/\|r_{t-1}\|$, which is $\mathcal{F}_{t-1}$-measurable and hence correlated with the history; a bound valid only for each *fixed* u may not be substituted at $u = \hat{r}_{t-1}$. We therefore establish the drift condition *uniformly over the sphere*, first, as an event depending on the processing permutation alone. The right objects to sample are not the realized $|\langle V, u\rangle|$ but the *conditional means*

$$m_i(u) := \mathbb{E}\big[\,|\langle v_i, u\rangle| \;\big|\; \mathcal{C}\big],$$

deterministic numbers given $\mathcal{C}$, with $0 \leq m_i(u) \leq \rho$ and $|m_i(u) - m_i(w)| \leq \rho\|u - w\|$. (Applying without-replacement Hoeffding to the realized $|\langle V, u\rangle|$ and comparing against the realized full average $\frac{1}{N}\sum_i |\langle v_i, u\rangle|$ would not suffice: the hypothesis bounds only the *expected* average and never lower-bounds the realized one.) The processing order is independent of the values given $\mathcal{C}$, so at step t the unprocessed *index set* $\mathcal{R}_t$ is a uniformly random $(N - t + 1)$-subset, and Hoeffding for sampling without replacement applies to the deterministic numbers

$\{m_i(u)\}_{i\le N}$: $\Pr[\frac{1}{|\mathcal{R}_t|}\sum_{i\in\mathcal{R}_t} m_i(u) < \frac{3}{4}\cdot\frac{1}{N}\sum_i m_i(u)] \le e^{-\Omega(c_1^2(N-t+1))}$, and the hypothesis gives $\frac{1}{N}\sum_i m_i(u) \ge c_1\rho$. The map $u \mapsto \frac{1}{|\mathcal{R}_t|}\sum_{\mathcal{R}_t} m_i(u)$ is ρ-Lipschitz, so a C^d-net of *constant* angular resolution $c_1/8$—the map is ρ-Lipschitz in u and the target is $c_1\rho$, so the resolution must be dimensionless, a fixed fraction of c_1, not a multiple of ρ. Define the *per-time* events

$$U_t := \Big\{\inf_{\|u\|=1} \tfrac{1}{|\mathcal{R}_t|}\textstyle\sum_{i\in\mathcal{R}_t} m_i(u) \;\ge\; \tfrac{1}{2}c_1\rho\Big\}, \qquad U_t \in \mathcal{F}_{t-1},$$

measurable at time $t-1$ because $\mathcal{R}_t$ is determined by the first $t-1$ draws of the permutation and the m_i are deterministic given $\mathcal{C}$—so no conditioning on the *future* of the permutation ever occurs (conditioning on a global good event over the whole permutation would destroy the uniformity of the step-t index). Set $L_0 := C(d)\log(N/\delta)$ and $T := N - L_0$: for every $t \le T$ the failure probability is $C^d e^{-\Omega(c_1^2|\mathcal{R}_t|)} \le C^d e^{-\Omega(c_1^2 L_0)}$, so a union bound over $t \le T$ gives $\Pr[\eta \le T] \le \delta/4$ for the *predictable failure time* $\eta := \inf\{t : U_t^c\}$ (note $\{\eta > t\} = \bigcap_{s\le t} U_s \in \mathcal{F}_{t-1}$), once $C(d)$ is large. On $\{\eta > t\}$ the infimum may be evaluated at the adaptive $\hat{r}_{t-1}$: given $\mathcal{F}_{t-1}$, the index processed at step t is uniform on $\mathcal{R}_t$ and the unprocessed vectors are conditionally independent of the history with their laws intact—the past reveals the permutation so far and the processed *values*, not the unprocessed ones—so on $\{\eta > t\}$

$$\mathbb{E}\big[\,|\langle v_t, \hat{r}_{t-1}\rangle| \,\big|\, \mathcal{F}_{t-1}\big] = \frac{1}{|\mathcal{R}_t|}\sum_{i\in\mathcal{R}_t} m_i(\hat{r}_{t-1}) \;\ge\; c_1\rho/2,$$

and $\mathbb{E}[\|r_t\|^2 \mid \mathcal{F}_{t-1}] \le \|r_{t-1}\|^2 - c_1\rho\|r_{t-1}\| + \rho^2$: negative drift above $\|r\| = 2\rho/c_1$. (ii) Stop *one step before* the failure: let $\eta_0 := \eta - 1 = \inf\{s \ge 0 : U_{s+1}^c\}$, so that every increment $r_{t-1} \to r_t$ of the stopped process corresponds to a time with U_t *known to hold*—stopping at η itself would include the step-η increment, whose drift U_η no longer guarantees. For the hitting time $\tau = \inf\{t : \|r_t\| \le 3\rho/c_1\}$, the drift-compensated *stopped* process $\|r_{t\wedge\tau\wedge\eta_0}\| + \frac{c_1\rho}{4}(t\wedge\tau\wedge\eta_0)$ is a genuine supermartingale with $C\rho$-bounded increments (the compensator adds $c_1\rho/4$ per step, so the increments are $O(\rho)$, not exactly ρ)—the drift bound is applied only on $\{t \le \tau\wedge\eta_0\} \in \mathcal{F}_{t-1}$, never on a global event—and Azuma gives $\Pr(\tau\wedge\eta_0 > T) \le \exp[-c(T - C\|r_0\|/\rho)] \le \delta/4$ under the stated N. (iii) Bounded increments and negative drift do *not* feed Doob's inequality directly—Doob wants a nonnegative supermartingale, which must be built. From the squared-norm drift and Jensen, for $R = \|r_{t-1}\| \ge h := 3\rho/c_1$,

$$\mathbb{E}[\|r_t\| \mid \mathcal{F}_{t-1}] \le \sqrt{R^2 - c_1\rho R + \rho^2} \;\le\; R - \tfrac{c_1\rho}{4},$$

so the norm increment Δ_t obeys $|\Delta_t| \le \rho$ and $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] \le -a\rho$ with $a = c_1/4$ above level h. Take $\lambda = a/(2\rho)$: Hoeffding's lemma gives $\mathbb{E}[e^{\lambda\Delta_t} \mid \mathcal{F}_{t-1}] \le e^{-\lambda a\rho + \lambda^2\rho^2/2} = e^{-3a^2/8} < 1$, so within each excursion above h (stopped at the excursion's return, at η_0, and at T) the process $\exp\{\lambda(\|r_{t\wedge\sigma}\| - h)\}$ *is* a nonnegative supermartingale, and Doob's maximal inequality gives $\Pr[\sup_{\text{excursion}}\|r_t\| \ge h + x] \le e^{\lambda\rho}e^{-\lambda x}$ (an excursion starts at most ρ above h). There are at most $T \le N$ excursions; a union bound and $x = C'\rho\log(N/\delta)$ bound $\max_{\tau\le t\le T\wedge\eta_0}\|r_t\|$ by $C'\rho\log(N/\delta)$ with probability $1 - \delta/4$. (iv) The last $N - T = L_0$ steps need no probability at all: $\|r_N - r_T\| \le L_0\rho = C(d)\rho\log(N/\delta)$ deterministically, since every increment has norm at most ρ. Collecting the three failure events ($\eta_0 < T$; $\tau\wedge\eta_0 > T$; the excursion maximum), $\|r_N\| \le C'\rho\log(N/\delta)$ with probability $1 - \delta$. □

Lemma 21 (Centering transfer)**.** *Run* GD-D *with the grid anchored to the current cloud's own bounding box—side $s = \Theta(\text{span}\cdot(t/M)^{1/d})$ and cell index $\lfloor(v - \min_i v_i)/s\rfloor$—and with pairing, leftover selection and signing driven by the priorities and coins alone. Then for every $c \in \mathbb{R}^d$—including a* data-dependent *shift such as $c = \bar{x}$, since the statement is pathwise, holding for each realization separately—the routine returns the* same *sign vector σ on $\{v_i - c\}$ as on $\{v_i\}$, and since the output is an exactly equal bipartition ($\sum_i \sigma_i = 0$),*

$$\Big\|\sum_i \sigma_i(v_i - c)\Big\| = \Big\|\sum_i \sigma_i v_i\Big\|.$$

Consequently Theorem 2, proved for i.i.d. x_i, applies verbatim to the centered cloud $b_i = x_i - \bar{x}$: the accuracy $\widetilde{O}_d(R/\sqrt{m})$ and the exact cardinality balance transfer, and no density model for the dependent centered vectors is required.

Proof. Round 1: $\min_i(v_i - c) = \min_i v_i - c$ and the coordinate span is unchanged, so s is unchanged and $\lfloor((v_i - c) - (\min_j v_j - c))/s\rfloor = \lfloor(v_i - \min_j v_j)/s\rfloor$—the cell memberships are identical. Priorities are attached to indices, so the within-cell pairings and the odd-cell leftover selections coincide; the resulting in-cell differences and the deferred pile-1 differences are identical because $(v_i - c) - (v_j - c) = v_i - v_j$. For even M the number of odd cells is even, so every deferred pile-1 item is such a difference and no raw point is signed alone. Every later round acts on differences, already independent of c, and the signing therefore sees the same vectors in the same order; hence σ is unchanged, and $\sum_i \sigma_i = 0$ gives the display. (Checked to machine precision—identical cell indices, pair lists, difference clouds, sign vectors and imbalances—at $d = 2, 3, 5$ and m up to 2048.) □

Proof of Theorem 2. Lemma 18, its induction fed through Lemma 17's certified-core recursion ($M_{j+1} = \Omega(M_j)$, so $M_j = \Omega(c^j m)$ and the per-round scale factor is still $\Theta((t/M_j)^{1/d})$), gives $\rho_{k^*} = \widetilde{O}(Rm^{-k^*/d}) \le \widetilde{O}(R/\sqrt{m})$ with no repair edges. Sign coarse to fine, each pile in uniformly random order: entering pile j the residual is $\le C'\rho_{j-2}\,\mathrm{polylog}$ by induction; pile j has $N_j \approx M_j/(2t)$ vectors of norm $\le \rho_{j-1}$, conditionally independent given the anchor and the full membership $\mathcal{S}_j$, carrying the aggregate directional moment $c(d,t)\rho_{j-1}$ (Lemma 19), and $N_j \gg (\rho_{j-2}/\rho_{j-1})\,\mathrm{polylog} = \widetilde{O}(M_j^{1/d})$, so Lemma 20's *full-pile* conclusion ends the pile at $\|r_{N_j}\| \le C'\rho_{j-1}\,\mathrm{polylog}$ (the scale is ρ_{j-1}, not ρ_j: a pile cannot floor below its own increment scale; and the bound is needed at the pile's *end*, which control at its midpoint would not give); the final chain—unconditioned in-cell differences, whose aggregate moment is direct from Lemma 18(i)—finishes at $C'\rho_{k^*}\,\mathrm{polylog}$. Every vector is a pair difference, so the bipartition is exactly equal. Measured rates, for the certified inner-core implementation the analysis describes: $\gamma = +0.44, +1.02, +0.49$ at $d = 2, 3, 5$ against the target $+0.5$ ($d = 2$ has $k^* = 1$, no core restriction; its dip from $+0.62$ is seed noise at the top m), and $+0.16$ at $d = 8$ over $m \le 65{,}536$—pre-asymptotic there twice over: the constants square each round, and at accessible m the cell side is comparable to the inner core, so most mass defers after two rounds; at $d = 3, 5$, where the repair variant's log-stretch bites, GD-D beats it ($+1.02, +0.49$ vs $+0.76, +0.39$; Appendix N). □

Remark 1 (Beyond box supports). Theorem 2 is proved for box-supported densities, and the box is used, not merely convenient: the anchor lemma's quarter-slab probabilities take a fixed fraction of the support's volume because the slabs are slabs *of the box*, and the interior-cell parity argument of Lemma 19 infers "unclipped by the support" from distance to the *bounding-box* boundary—an inference valid only when the support is that box. We expect the differencing-with-deferral construction to extend to sufficiently regular convex supports, but the corresponding boundary-geometry estimates (cap-volume lower bounds; boundary-shell mass) are not proved here and are left open.

Constants depend on d (the density condition number squares each round), so the statement is asymptotic in m at fixed d, as is the whole generic-instance discussion. Appendix N audits the recursion node by node (Figure 5(a)) and measures why cross-cell repair fails (Table 23). End to end, across four instance families, the uncertified grid variant GD-PR is the strongest of the three recursive orderings on the two uniform-density families — including the 8-dimensional one, where it needs twice the rounds — while on a 3:1 density mixture and on real PCA features consecutive pairing is stronger. All three keep the n-growth (fitted slopes $+0.45$ to $+0.61$). What distinguishes the grid family is GD-D's *provable* accuracy (Theorem 2); GD-PR itself is an empirical comparator whose induction closes only for $d \le 2$, and no routine is strongest on every family (Table 24, Appendix N).

G Analysis of Composition with SEG-FFA

Theorem 7 (Composition with SEG-FFA)**.** *Assume* (A1)–(A2) *and the affine model. In* SEG-FFA *the forward and backward sweeps cancel the leading β^2 order-dependent bias of Theorem 1. The first surviving order-dependent term of the per-epoch bias is*

$$-\frac{\beta^3}{2}\Big[\, n\,\bar{M}\,C(\pi) \;+\; T(\pi) + T(\pi^R)\,\Big], \qquad C(\pi) = \sum_j M_{\pi_j} P_j, \quad T(\pi) = \sum_{\ell > \ell'} M_{\pi_\ell} M_{\pi_{\ell'}} P_{\ell'}, \tag{14}$$

with $P_j = \sum_{s<j} F_{\pi_s}(z^*)$ the same prefix sums as in Theorem 1, $\bar{M} = \frac{1}{n}\sum_i M_i$, and the factor $\frac{1}{2}$ from the anchoring. Every term of (14) *is a prefix functional bounded by $O(n^2 L^2 D(\pi))$, so, under*

the same no-dominant-component condition as Theorem 1 ($\rho_\infty = \widetilde{O}(R)$; without it, replace R by ρ_∞), a herded order reduces the surviving SEG-FFA *order residual by the worst-case discrepancy factor of Theorem 1—$\sqrt{n}$. The separation is instance-wise, not only an upper-bound comparison: the triple term is reversal-symmetric and therefore cannot cancel the witness covariance of the $\bar{M}$-part, so the reshuffling variance of* (14) *is at least $\beta^6 n^5 \sigma_\star^4 R^2/(48\,d)$ on* every *instance (Lemma 23), a floor that* SEG-FFA-RR *inherits at stationarity for $\beta \le c\,\sigma_\star^2/(\sqrt{d}\,nL^3)$—on rotation-dominated instances a constant-factor subrange of $\beta = \Theta(1/(nL))$—and a fixed herded order does not incur. Dividing by the contraction gap, the proven separation is $\mathcal{P}_{\mathrm{order}}(\mathrm{RR})/\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{\Omega}\big(\beta n^{3/2}\mu\sigma_\star^2/(\sqrt{d}\,L^2)\big)$, which at the largest admissible step is $\widetilde{\Omega}\big(\sqrt{n}\,\mu\sigma_\star^4/(d\,L^5)\big)$: the worst-case $\sqrt{n}$ of Theorem 1, up to conditioning powers and the fixed dimension. The epoch-count improvement of* SEG-FFA *is preserved throughout. The generic factor-n transfer applies to the $\bar{M}$-part of* (14)*, which is Theorem 1's functional; we do not establish a separate generic cancellation for the triple term T, and the full generic-scale composition gain is confirmed empirically (Tables 4 and 12).*

SEG-FFA (Chae et al., 2024) runs each epoch as a forward sweep with order π, a backward sweep with the reversed order, and anchoring $z^{k+1} = \frac{1}{2}(z_2 + z)$, with $\alpha = \beta/2$.

Flip-flop cancels the β^2 term. Let $C_{\mathrm{fwd}} = \sum_{j>i} DF_{\pi_j}F_{\pi_i}$ and C_{bwd} be the forward and backward cross-terms. A direct index computation gives $C_{\mathrm{fwd}} + C_{\mathrm{bwd}} = J\,S_F - S_{DFF}$ with $J = \sum_a DF_a$, $S_F = \sum_a F_a$, $S_{DFF} = \sum_a DF_aF_a$, all order-independent; hence every β^2 term of the two-sweep map is order-independent.

Lemma 22 (FFA residual)**.** *The first order-dependent term of the anchored two-sweep map is at order β^3. In the per-epoch bias $r_{\mathrm{FFA}} = (\text{epoch map})(z^*) - z^*$ it equals*

$$-\frac{\beta^3}{2}\Big[\,n\bar{M}\,C(\pi)\;+\;T(\pi)+T(\pi^R)\,\Big], \qquad C(\pi)=\sum_j M_{\pi_j}P_j, \quad T(\pi)=\sum_{\ell>\ell'} M_{\pi_\ell}M_{\pi_{\ell'}}P_{\ell'},$$

with $P_j = \sum_{i<j} b_{\pi_i}$.

Proof. Write both sweeps in the corrected-operator form of Appendix A with $\alpha = \beta/2$: $A_i = M_i - \frac{\beta}{2}M_i^2$, $g_i = b_i - \frac{\beta}{2}M_ib_i$, forward map $z \mapsto \Phi^F z - \psi^F$, backward map $z \mapsto \Phi^B z - \psi^B$ over the reversed order, so that $r_{\mathrm{FFA}} = -\frac{1}{2}(\Phi^B\psi^F + \psi^B)$ at $z^* = 0$. Expand $\psi = \beta\sum_j g_j - \beta^2\sum_{\ell>j} A_\ell g_j + \beta^3\sum_{\ell>\ell'>j} A_\ell A_{\ell'} g_j + O(\beta^4)$ and $\Phi^B = I - \beta\sum_i M_i + O(\beta^2)$, and sort by order dependence. At order β^2, $\psi^F + \psi^B$ carries $C(\pi) + C(\pi^R)$, order-free by the index identity above. At order β^3, the $O(\alpha)$ corrections $\sum_{\ell>j}(M_\ell^2 b_j + M_\ell M_j b_j)$ likewise cancel in the π/π^R pair, and two order-dependent sources survive: the composition term $(\Phi^B - I)\psi^F \ni \beta^3(\sum_i M_i)C(\pi) = \beta^3 n\bar{M}C(\pi)$, and the third-order sweep terms $\beta^3[T(\pi) + T(\pi^R)]$—the reversal maps the triple pattern $\ell > \ell' > j$ to $\ell < \ell' < j$, so the pair sum excludes the mixed patterns and remains genuinely order-dependent. Halving by the anchor gives the display. The expression, with both sources and the anchoring factor, was verified against the exact affine epoch polynomial, whose residual difference between orders matches it to relative error $O(\beta)$. □

Lemma 23 (FFA residual nondegeneracy)**.** *Let $\Xi(\pi) = n\bar{M}\,C(\pi) + T(\pi) + T(\pi^R)$ be the order-dependent coefficient of* (14)*, π uniform. The symmetrized triple term is reversal-*symmetric*, while the witness $Y_u(\pi) = \sum_t (t - \frac{n+1}{2})\,\langle u, b_{\pi_t}\rangle$ of Lemma 7 is reversal-*antisymmetric *($c_{n+1-t} = -c_t$); hence $\mathrm{Cov}(\langle w, T + T^R\rangle, Y_u) = 0$ for all w, u—no covariance cancellation between the two parts of Ξ is possible. Consequently, exactly,*

$$\mathrm{Cov}(\langle w, \Xi\rangle, Y_u) = -\frac{n+1}{12}\,w^\top S^2 B u, \qquad S = \sum_i M_i, \quad B = \sum_i b_i b_i^\top,$$

and by Cauchy–Schwarz, with u_1 the top eigenvector of B,

$$\mathbb{E}\,\|\Xi - \mathbb{E}\Xi\|^2 \;\ge\; \frac{n+1}{12n}\,\lambda_{\max}(B)\,\big\|S^2u_1\big\|^2 \;\ge\; \frac{n^5\,\sigma_\star^4 R^2}{12\,d}, \qquad \sigma_\star = \sigma_{\min}(\bar{M}) \ge \mu.$$

Proof. Reversal fixes $T + T^R$ and negates Y_u; since $\pi \mapsto \pi^R$ preserves the uniform law, $\mathbb{E}[\langle w, T + T^R\rangle\, Y_u] = -\mathbb{E}[\langle w, T + T^R\rangle\, Y_u] = 0$. The C-part covariance is Lemma 7's identity with $H_a = M_a$, premultiplied by $n\bar{M} = S$. The bound follows with $w = S^2u_1/\big\|S^2u_1\big\|$,

$\|S^2 u_1\| \geq \sigma_{\min}(S)^2 = n^2\sigma_\star^2$, and $\lambda_{\max}(B) \geq nR^2/d$. (The triple term alone also satisfies an exact covariance identity, $\mathrm{Cov}(\langle w, T\rangle, Y_u) = \frac{n+1}{24}\, w^\top[-S^2B + Q_2B + SN - \Lambda]u$ with $Q_2 = \sum_a M_a^2$, $N = \sum_a M_a b_a b_a^\top$, $\Lambda = \sum_a M_a(\sum_c M_c b_c) b_a^\top$, verified by exhaustive enumeration; it is not needed above.) □

Lemma 24 (Quartic FFA remainder decomposition)**.** *Write the anchored two-sweep bias as $r_{\mathrm{FFA}}(\pi) = \beta^2 A_2 + \beta^3 A_3(\pi) + \beta^4 A_4(\pi) + O(\beta^5)$, so that $A_3(\pi) = -\frac{1}{2}\Xi(\pi)$ up to order-free terms. For an order o let $P_\ell = \sum_{j<\ell} b_{o_j}$ and $P_\ell^{Mb} = \sum_{j<\ell} M_{o_j} b_{o_j}$, and set*

$$C^{(2)} = \sum_\ell M_{o_\ell}^2 P_\ell, \qquad C^{mb} = \sum_\ell M_{o_\ell} P_\ell^{Mb},$$

$$T^{(2,1)} = \sum_{\ell>\ell'} M_{o_\ell}^2 M_{o_{\ell'}} P_{\ell'}, \qquad T^{(1,2)} = \sum_{\ell>\ell'} M_{o_\ell} M_{o_{\ell'}}^2 P_{\ell'},$$

$$T^{mb} = \sum_{\ell>\ell'} M_{o_\ell} M_{o_{\ell'}} P_{\ell'}^{Mb}, \qquad Q = \sum_{\ell>\ell'>\ell''} M_{o_\ell} M_{o_{\ell'}} M_{o_{\ell''}} P_{\ell''}, \qquad \Pi_2(o) = \sum_{\ell>\ell'} M_{o_\ell} M_{o_{\ell'}}.$$

Then, exactly and up to order-free terms,

$$A_4(\pi) = -\frac{1}{2}\Big[-\tfrac{1}{2} \sum_{o\in\{\pi,\pi^R\}} \big(T^{(2,1)} + T^{(1,2)} + T^{mb}\big)(o) - \big(Q(\pi) + Q(\pi^R)\big) \\ - S\big(\tfrac{1}{2}C^{(2)} + \tfrac{1}{2}C^{mb} + T\big)(\pi) - \big(\tfrac{1}{2}S_2 + \Pi_2(\pi^R)\big)\big(\tfrac{1}{2}W + C(\pi)\big)\Big],$$

with $S = \sum_i M_i$, $S_2 = \sum_i M_i^2$, $W = \sum_i M_i b_i$; the remaining pair class $\sum_{\ell>j} M_{o_\ell}^2 M_{o_j} b_{o_j}$ is order-free in the π/π^R pair and is dropped. Using $\|P_\ell\| \leq D(\pi) = D(\pi^R)$ and $\|P_\ell^{Mb}\| \leq nL\rho_\infty$, the classes obey

$$\|Q\| \leq \tfrac{n^3}{6}L^3 D, \quad \big\|T^{(2,1)}\big\|, \big\|T^{(1,2)}\big\| \leq \tfrac{n^2}{2}L^3 D, \quad \|T^{mb}\| \leq \tfrac{n^3}{2}L^3\rho_\infty, \quad \big\|S\,C^{(2)}\big\| \leq n^2 L^3 D,$$

$$\big\|S\,C^{mb}\big\| \leq n^3 L^3\rho_\infty, \qquad \|S\,T\| \leq \tfrac{n^3}{2}L^3 D, \qquad \|S_2\,C\| \leq n^2 L^3 D,$$

$$\|\Pi_2 W\| \leq \tfrac{n^3}{2}L^3\rho_\infty, \qquad \|\Pi_2 C\| \leq \tfrac{n^3}{2}L^3 D,$$

whence the order-dependent part satisfies $\big\|A_4^{\mathrm{ord}}(\pi)\big\| \leq 5\, n^3 L^3\big(D(\pi) + \rho_\infty\big)$. Moreover, for every $k \geq 3$ the β^k coefficient obeys, up to order-free terms, $\|A_k(\pi)\| \leq k^3\, n^{k-1} L^{k-1}(D(\pi) + \rho_\infty)$, so for $\beta nL \leq c_0$ the entire *order-dependent remainder beyond the cubic term resums geometrically:*

$$\Big\|\sum_{k\geq 4} \beta^k A_k^{\mathrm{ord}}(\pi)\Big\| \;\leq\; C\,\beta^4 n^3 L^3\big(D(\pi) + \rho_\infty\big).$$

Proof. Collect the β^4 terms of $r_{\mathrm{FFA}} = -\frac{1}{2}(\Phi^B\psi^F + \psi^B)$ with $A_i = M_i - \frac{\beta}{2}M_i^2$, $g_i = b_i - \frac{\beta}{2}M_i b_i$ and $\sum_i b_i = 0$: the sweeps' own fourth order contributes the first two brackets (the β-corrections of the triple sums, and the quadruple sums); the composition contributes $(-\beta S)\cdot\psi^F_{\beta^3}$ and $\Phi^B_{\beta^2}\,\psi^F_{\beta^2} = \big(\frac{1}{2}S_2 + \Pi_2(\pi^R)\big)\cdot\big(-\frac{1}{2}W - C(\pi)\big)$; ψ^F has no β^1 term because the offsets are centered. Completeness of the enumeration was verified against the Richardson-extracted fourth-order coefficient of the exact affine epoch polynomial: the relative deviation is $O(\beta^2)$ ($7\cdot 10^{-5} \to 2\cdot 10^{-5}$ under halving β), i.e. the display *is* the fourth-order coefficient. The class bounds are the triangle inequality applied tuple by tuple; measured class ratios reach at most 0.04 of the stated constants.

For the full tail, note the structural fact that generates all the classes above: every β^k term of $\psi(o)$ is a sum over decreasing position tuples $\ell_1 > \cdots > \ell_m > j$ of products $W_{\ell_1}\cdots W_{\ell_m} v_j$ with $W_\ell \in \{M_{o_\ell}, \frac{1}{2}M_{o_\ell}^2\}$ and $v_j \in \{b_{o_j}, \frac{1}{2}M_{o_j}b_{o_j}\}$ (degrees summing to k), so the innermost sum over j is *always* a b- or Mb-prefix, bounded by $D(\pi)$ or $nL\rho_\infty$; the triangle inequality over the at most $n^m/m!$ tuples and the at most $2k$ degree patterns gives $\|\psi_{\beta^k}\| \leq 2k\, n^{k-1}L^{k-1}(D + \rho_\infty)$— the only prefix-free class, $m = 0$, is the order-free $-\frac{\beta^2}{2}W$. Likewise $\|\Phi_{\beta^a}\| \leq 2a\, n^a L^a$, and composition preserves the form: $\|(\Phi^B\psi^F)_{\beta^k}\| \leq \sum_{a+c=k} 4ac\, n^a L^a \cdot n^{c-1}L^{c-1}(D + \rho_\infty) \leq$

$k^3 n^{k-1} L^{k-1}(D + \rho_\infty)$. Summing $\sum_{k\geq 5} k^3(\beta nL)^{k-4} \leq C$ for $\beta nL \leq c_0$ gives the display. Numerically, the residual after subtracting the exact cubic and quartic coefficients scales as $\beta^{4.99}$ with order-fluctuation exponent $n^{4.26}$—consistent with the prefix-controlled $n^{k-1}\sqrt{n}$ at $k = 5$ and excluding the prefix-free n^k scale. □

Proof of Theorem 7. By Lemma 22, flip-flop demotes the order dependence by one power of β while leaving prefix-sum functionals as the controlling objects: $\|C(\pi)\| \leq L\, nD(\pi)$ and $\|T(\pi)\| \leq nL \cdot L\, nD(\pi)$ (the inner sum of T is a truncated C), so the residual (14) is $O(\beta^3 n^2 L^2 D(\pi))$: $\widetilde{O}(\beta^3 n^2 L^2 R)$ for a herded order (Lemma 5) against $\Theta(\beta^3 n^2 L^2 \cdot R\sqrt{n})$ at the random-order discrepancy scale. For the matching lower bound, SEG-FFA redraws the order each epoch and its anchored epoch is an affine contraction for *every* order (the weighted-norm argument of Appendix M), so the injection argument of Appendix E applies verbatim with the per-epoch fluctuation $\frac{\beta^3}{2}(\Xi(\pi) - \mathbb{E}\Xi)$. Only the *order-dependent* part of the β^4 remainder enters the fluctuation, and Lemma 24 bounds the *entire* order-dependent remainder beyond the cubic term by $C\beta^4 n^3 L^3(D(\pi) + \rho_\infty)$ for $\beta nL \leq c_0$ (geometric resummation of prefix-controlled coefficients); at the random-order discrepancy scale $D_\pi = \widetilde{O}(R\sqrt{n})$—with the deterministic bound $D_\pi \leq 2n\rho_\infty$ on the $(nd)^{-k}$ tail, as in Corollary 1—this is $\widetilde{O}(\beta^4 n^{7/2} L^3 R)$ (measured exponent $n^{3.50}$), so Minkowski's inequality retains half of the cubic fluctuation scale $\frac{\beta^3}{2} n^{5/2}\sigma_\star^2 R/\sqrt{12d}$ once $\beta \leq c\,\sigma_\star^2/(\sqrt{d}\, nL^3)$—on rotation-dominated instances a constant-factor subrange of $\beta = \Theta(1/(nL))$, which is all a lower bound needs. Lemma 23 then gives $\mathbb{E}\,\|z^\infty - z^*\|^2 = \Omega(\beta^6 n^5 \sigma_\star^4 R^2/d)$ under SEG-FFA-RR, while a fixed herded order injects no reshuffling variance. Combining the two sides: the herded order plateau is $\widetilde{O}(\beta^3 n^2 L^2 R)/(\beta n\mu) = \widetilde{O}(\beta^2 nL^2 R/\mu)$ and the RR floor in norm is $\Omega(\beta^3 n^{5/2}\sigma_\star^2 R/\sqrt{d})$, so the separation is $\widetilde{\Omega}(\beta n^{3/2}\mu\sigma_\star^2/(\sqrt{d}\, L^2))$; at $\beta = \Theta(\sigma_\star^2/(\sqrt{d}\, nL^3))$ this is $\widetilde{\Omega}(\sqrt{n}\,\mu\sigma_\star^4/(dL^5))$—the claimed instance-wise worst-case $\sqrt{n}$, up to conditioning and the fixed dimension. For the M-part, Lemmas 6–9 and Appendix F apply verbatim to $C(\pi)$, giving the generic factor-n transfer stated for that part; the triple term T is controlled by the same discrepancy at the same order, but its generic outer-sum cancellation is not separately established here. The two-sweep expansion remainder is $O(\beta^4)$, and the order-dependent part of the within-epoch remainder of Chae et al. (2024) (their Appendix E.2) is higher order and herding-irrelevant. □

H General Smooth Operators

Anchor-change cost (deferred from Lemma 1). Replacing $DF_i(z_0)$ by $DF_i(z^*)$ costs a further $\beta^2 \sum_i \|DF_i(z_0) - DF_i(z^*)\|\,\|f_i\| = O(\beta^2 nMR_0\,\|z_0 - z^*\|)$, which is $O(\beta^3)$ only at anchors with $\|z_0 - z^*\| = O(\beta)$, not at an arbitrary one. No result below invokes the substitution: the smooth statements anchor the epoch expansion at z^* itself, where the cost vanishes identically, and absorb the error-proportional corrections into the contraction instead (Appendix H). We record what invoking it at a *plateau* anchor would cost, since "$\|z - z^*\| = O(\beta)$ on the plateau" is not by itself a licence: with $\|z_0 - z^*\| \leq \mathcal{P} = \widetilde{O}(\beta LQ_{\mathrm{H}}/\mu)$ and $\rho \leq Q_{\mathrm{H}}$ the per-epoch cost is $O(\beta^2 nMQ_{\mathrm{H}}\mathcal{P})$, i.e. $\widetilde{O}(\beta^2 MLQ_{\mathrm{H}}^2/\mu^2)$ after dividing by the gap $\Theta(\beta n\mu)$; relative to the order term $\widetilde{O}(\beta LQ_{\mathrm{H}}/\mu)$ this is a factor $\widetilde{O}(\beta MQ_{\mathrm{H}}/\mu)$, which Theorem 8's step cap $\beta \leq c\mu/(MQ_{\mathrm{H}})$ does absorb. That is precisely why the cap carries μ and not L: the same comparison governs the epoch's own contraction, and an L-scaled cap would leave the effective gap $c'\beta n\mu - C\beta^2 nMQ_{\mathrm{H}}$ negative for large κ (proof of Theorem 8).

Fix an epoch anchor z and write $f_i = F_i(z)$, $\bar{f} = F(z)$, $u_i = f_i - \bar{f}$, and $\rho(z) = \max_i \|f_i\|$. For an order π, let

$$D_\pi(z) = \max_t \left\| \sum_{s<t} u_{\pi_s} \right\|, \qquad Q_\pi(z) = \rho(z) + D_\pi(z) + n\,\|F(z)\|\,.$$

The expansion of Chae et al. (2024, Propositions D.1–D.2), specialized to $\alpha = \beta$, is

$$\begin{aligned} z^+ - z = &-\beta \sum_j F_{\pi_j}(z) + \beta^2 \sum_j DF_{\pi_j}(z) F_{\pi_j}(z) \\ &+ \beta^2 \sum_{i<j} DF_{\pi_j}(z) F_{\pi_i}(z) + \varepsilon_\pi(z). \end{aligned}$$

Their result gives the fixed-n asymptotic order of ε_π. The following bound exposes the dependence on the epoch length needed here.

Lemma 25 (Explicit smooth epoch remainder)**.** *Under* (A1) *and* (A3)*, if $\beta n L \le c_0$ for a sufficiently small absolute constant, then*

$$\|\varepsilon_\pi(z)\| \le C\beta^3 \left[n^2 L^2 Q_\pi(z) + nM Q_\pi(z)^2\right] \tag{15}$$

for an absolute constant C. Since $D_\pi(z) \le 2n\rho(z)$ and $\|F(z)\| \le \rho(z)$, the order-uniform consequence

$$\|\varepsilon_\pi(z)\| \le C'\beta^3 n^3 \left[L^2 \rho(z) + M\rho(z)^2\right] \tag{16}$$

also holds.

Proof. Let z_j and $\widehat{z}_j = z_j - \beta F_{\pi_j}(z_j)$ be the two within-epoch iterates and put $W = \max_j\{\|z_j - z\|, \|\widehat{z}_j - z\|\}$. Adding and subtracting the frozen values f_{π_j} in the epoch recursion gives

$$\max_j \|z_j - z\| \le \beta\{D_\pi(z) + n\,\|F(z)\|\} + \beta n L W,$$

while $\|\widehat{z}_j - z\| \le (1 + \beta L)\,\|z_j - z\| + \beta\rho(z)$. Thus $\beta n L \le c_0$ yields $W \le C_0 \beta Q_\pi(z)$ by absorption.

Equation (30) of Chae et al. (2024) splits the remainder into three terms. Lemma C.6 of that paper bounds the Taylor term by $\beta \sum_j (M/2)\,\|\widehat{z}_j - z\|^2 \le C\beta^3 n M Q_\pi(z)^2$. Lipschitzness bounds the other two terms by $C\beta^2 n L^2 W$ and $C\beta^2 n^2 L^2 W$, respectively. Substituting the bound on W proves (15); the order-uniform estimate follows from the displayed elementary bounds on $D_\pi(z)$ and $F(z)$. □

Lemma 26 (Nonlinear reduction)**.** *The only order-dependent term at order β^2 is*

$$\beta^2 \sum_{i<j} DF_{\pi_j}(z) F_{\pi_i}(z) = \beta^2 \sum_j DF_{\pi_j}(z) P_j, \qquad P_j = \sum_{i<j} F_{\pi_i}(z),$$

which is linear in the frozen prefix sums, exactly as in Lemma 3. All remaining composition and curvature terms are third order and obey Lemma 25; in particular their dependence on n, L, and M is polynomial, not polylogarithmic.

The explicit bound separates two statements that fixed-n notation conflates. The total remainder is cubic in β, but it need not be negligible uniformly in n without a step-size comparison. At the solution, $F(z^*) = 0$ and a herded order with $D_\pi(z^*) \le D_{\mathrm{H}}$ has $Q_\pi(z^*) \le Q_{\mathrm{H}} := \rho_\infty + D_{\mathrm{H}}$, removing one power of n from the order-uniform estimate.

Discrepancy stability, not order stability. The deployable order is computed from records rather than from $\{F_i(z^*)\}$. We do not claim that the resulting permutation equals the ideal one; only its discrepancy must transfer.

Lemma 27 (Discrepancy stability)**.** *Let an ordering routine H satisfy $D(H(v); v) \le D_\star$ for every centered input v. Then for any centered records $\widetilde{u}$ and centered targets u, the order $\pi = H(\widetilde{u})$ obeys*

$$D(\pi; u) \le D_\star + \sum_{i=1}^n \|\widetilde{u}_i - u_i\|\,.$$

Proof. $D(\pi; u) \le D(\pi; \widetilde{u}) + \max_t \left\|\sum_{s<t} (u - \widetilde{u})_{\pi_s}\right\| \le D_\star + \sum_i \|u_i - \widetilde{u}_i\|$. □

For stale records $\widetilde{u}_i = F_i(w_i)$, assumption (A1) gives $\|\widetilde{u}_i - F_i(z^*)\| \le L\,\|w_i - z^*\|$. Hence the ideal discrepancy D_{H} below may be replaced by the realized bound supplied by Lemma 27. Its part proportional to the epoch-start error enters the recurrence multiplicatively and is absorbed by $\beta \lesssim \mu/(nL^2)$; the record-induced $O(\beta)$ part changes the displayed floor only by an absolute constant. This transfer concerns the worst-case discrepancy guarantee. We do not use it to claim the sharper generic factor-n rate for nonlinear operators.

Theorem 8 (Worst-case smooth extension of herded SEG)**.** *Assume* (A1)–(A3) *and let the herded orders satisfy $D_\pi(z^*) \le D_{\mathrm{H}}$. Set $Q_{\mathrm{H}} = \rho_\infty + D_{\mathrm{H}}$ and use the convention $\mu/(MQ_{\mathrm{H}}) = +\infty$ when $M = 0$. If*

$$\beta \le c \min\left\{ \frac{\mu}{nL^2}, \frac{\mu}{MQ_{\mathrm{H}}} \right\} \qquad \Big(\textit{equivalently } \beta \lesssim \tfrac{\mu}{nL^2 + MQ_{\mathrm{H}}}\Big), \tag{17}$$

then herded SEG *with $\alpha = \beta$ contracts geometrically to a plateau satisfying*

$$\mathcal{P}(\textsc{Herd}) \le \mathcal{P}_{\mathrm{EG}} + \widetilde{O}\left(\frac{\beta L Q_{\mathrm{H}}}{\mu}\right) + O\left(\frac{\beta^2}{\mu}\left[nL^2 Q_{\mathrm{H}} + MQ_{\mathrm{H}}^2\right]\right). \tag{18}$$

The last term is absorbed by the second term under (17)*. Consequently, if no component dominates and the balancing routine gives $D_{\mathrm{H}} = \widetilde{O}(\rho_\infty) = \widetilde{O}(R)$, then*

$$\mathcal{P}(\textsc{Herd}) \le \mathcal{P}_{\mathrm{EG}} + \widetilde{O}\left(\frac{\beta L R}{\mu}\right).$$

The corresponding random-reshuffling lower bound is worst-case, not instance-wise: for every n the smooth class contains instances for which $\mathbb{E}\,\|z^\infty - z^*\|^2 = \Omega(\beta^4 n^3 L^2 R^2)$. Indeed the balanced signed affine family of Lemma 9 (in dimension two) already belongs to (A1)–(A3) *with $M = 0$; arbitrarily small genuinely nonlinear perturbations preserve the lower bound for each fixed n by continuity; instance-wise, Lemma 7 moreover puts a μ-scale floor under* every *smooth instance via its linearization at the running anchor (Corollary 1); a $\sigma_\star$-scale smooth floor survives only along the continuity argument just given, for perturbations of a fixed affine instance. Two quantifiers must not be mixed here. The* class-level *worst-case separation is witnessed already by the affine hard family of Lemma 9, for which $M = 0$: on that witness the curvature cap $\mu/(MQ_{\mathrm{H}})$ is void, $\beta = \Theta(\mu/(nL^2))$ is admissible, and the $\widetilde{\Omega}(\sqrt{n}/\kappa^2)$ scaling stands. For a* specified *genuinely nonlinear instance, however, the admissible step may instead be limited by $\mu/(MQ_{\mathrm{H}})$—when that cap binds, the same β-dependent comparison shrinks by $nL^2/(MQ_{\mathrm{H}})$—and we do* not *use the affine L-scale witness to claim an instance-wise L-scale separation in that regime: the uniform instance-wise floor is the μ-scale one of Corollary 1, giving $\widetilde{\Omega}\big[\frac{1}{\sqrt{d}} \min\{\sqrt{n}/\kappa^3,\ n^{3/2}\mu^3/(LMQ_{\mathrm{H}})\}\big]$ at that instance's largest admissible step. No* unconditional *general-smooth generic factor-n claim is made; Corollary 1 below recovers a conditioning-dependent generic separation under two hypotheses—genericity of the linearization (H1) and the per-node split accuracy* ($\mathrm{H}_{\mathrm{split}}$) *inherited from Proposition 1—the required variance nondegeneracy being proved instance-wise (Lemma 7).*

Proof. At z^*, Lemma 26 gives the same order-free second-order term and the same leading cross-term as the affine linearization. The latter is at most $O(\beta^2 nLD_{\mathrm{H}})$; Lemma 25 contributes at most $C\beta^3[n^2L^2Q_{\mathrm{H}} + nMQ_{\mathrm{H}}^2]$. Under $\beta \le c\mu/L^2$, each component extragradient map is a contraction, so an epoch contracts by q with $1 - q \ge c'\beta n\mu$. Dividing the epoch forcing by this gap gives (18). The two comparisons $\beta nL \le c\mu/L \le c$ and $\beta MQ_{\mathrm{H}} \le c\mu$ absorb its last term. The second carries μ, not L, because it is also what keeps the epoch contracting: the micro-step $z \mapsto z - \beta G_i(z)$ has $\operatorname{sym} DG_i \succeq (\mu - \beta L^2 - \beta M\rho)I$ with $\rho \le Q_{\mathrm{H}}$, so the error-proportional coefficient is $O(\beta^2 n[nL^2 + MQ_{\mathrm{H}}])$ against a gap $c'\beta n\mu$, and an L-scaled cap $\beta \le cL/(MQ_{\mathrm{H}})$ would leave the effective gap $c'\beta n\mu - C\beta^2 nMQ_{\mathrm{H}}$ *negative* once κ is large—the plateau identification would fail outright, not merely lose a constant. Lemma 27 and the preceding record calculation give the stated stale-record transfer. Finally, the affine hard family proves the existential RR statement because affine maps are a subclass of smooth maps; continuity of the finite-dimensional epoch map and its stationary second moment gives the optional genuinely nonlinear perturbation statement. □

Closing the generic gap. Theorem 8 is unconditional but worst-case. Of the two statements it declines to make, the *instance-wise* RR lower bound turns out to need no hypothesis at all: Lemma 7 proves that the reshuffling variance of the linearization can never degenerate below the scale set by $\sigma_\star = \sigma_{\min}(DF(z^*)) \in [\mu, L]$. One might hope that genericity of the offsets restores the sharper L-scale with high probability—it cannot: for $DF_i(z^*) \equiv \mu I$ the fluctuation is exactly linear in the offsets and of scale $\beta^2 n^{3/2}\mu R$ for *every* draw, however large L (remark after Lemma 7). The L-scale is a property of the Jacobians, not of the offsets, and $\sigma_\star$ is the correct instance constant. What still requires assumptions is only the *generic herded upper bound*, and there are two of them:

(H1) *Centered generic linearized offsets.* There are latent i.i.d. samples x_i from a density bounded above and below on a box of side $O(R)$, drawn independently of the fixed Jacobians $DF_i(z^*)$, with $F_i(z^*) = x_i - \bar{x}$—the model of Proposition 1, whose centering is what makes $\sum_i F_i(z^*) = 0$ compatible with genericity; constants depend on the fixed dimension d, and $n = 2^q$ as there.

(H2) *Per-node split accuracy.* The ordering recursion satisfies ($\mathrm{H}_{\mathrm{split}}$) at every node—proved for a single split by Theorem 2, and empirically validated down the recursion (Appendix N) rather than proved there.

Corollary 1 (Generic smooth gain)**.** *Assume* (A1)–(A3)*, put* $\sigma_\star = \sigma_{\min}(DF(z^*))$*, and let the herded order be a label-blind recursive balanced bipartition tree satisfying Proposition 1's per-node accuracy* ($\mathrm{H}_{\mathrm{split}}$)*, built from the ideal linearization values* $\{F_i(z^*)\}$ *(so this corollary, like Proposition 1, is conditional on* ($\mathrm{H}_{\mathrm{split}}$)*, which Theorem 2 supplies at the root node only, for* $d \geq 2$*). If*

$$\beta \leq c_d \min\left\{ \frac{\mu}{nL^2},\ \frac{1}{n^{3/2}L},\ \frac{\mu}{\sqrt{n}\,MR} \right\} \tag{19}$$

with c_d *depending only on the fixed dimension* d *(convention* $\mu/(\sqrt{n}\,MR) = +\infty$ *for* $M = 0$*; since* $Q_{\mathrm{H}} = \widetilde{O}(R)$ *here, the third cap implies Theorem 8's nonlinear-contraction cap* $\beta \leq c\mu/(MQ_{\mathrm{H}})$ *up to logarithms, so no separate condition is needed), then, with no further hypothesis,*

$$\mathbb{E}\,\|z^\infty - z^*\|^2 \ \geq\ c\,\frac{\beta^4 n^3 \mu^2 R^2}{d} \quad \textit{under SEG-RR,}$$

and under (H1) *additionally* $\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{O}\big(\beta LR/(\mu\sqrt{n})\big)$*. The RMS norm gain is therefore* $\widetilde{\Omega}\big(\beta n^2\mu^2/(L\sqrt{d})\big)$*; at the largest admissible step* $\beta = c_d \min\{\cdot\}$ *of* (19)*, all three caps enter and the gain is*

$$\widetilde{\Omega}\Big[\frac{1}{\sqrt{d}}\,\min\Big\{\frac{n}{\kappa^3},\ \frac{\sqrt{n}}{\kappa^2},\ \frac{n^{3/2}\mu^3}{LMR}\Big\}\Big],$$

the third entry coming from the curvature cap $\mu/(\sqrt{n}\,MR)$*. When that cap is inactive—*$MR \lesssim \sqrt{n}\,L^2$ *in the first regime,* $MR \lesssim n\mu L$ *in the second—this reduces to* $n/(\kappa^3\sqrt{d})$ *for* $\sqrt{n} \leq \kappa$ *and* $\sqrt{n}/(\kappa^2\sqrt{d})$ *beyond; for the affine limit* $M = 0$ *the cap is void and the two-regime form is unconditional. The floor is* μ*-scale, not* $\sigma_\star$*-scale: the fluctuation is a stationary object, linearized at the* running *anchor, where* (A2) *certifies only* $\sigma_{\min}(DF(z)) \geq \mu$*—unlike the affine case (Proposition 1), whose Jacobians are fixed and whose* $\sigma_\star$*-scale floor stands. Recovering* $\sigma_\star$ *here would need a local-retention argument—*$\sigma_{\min}(DF(z)) \geq \sigma_\star - M\,\|z - z^*\|$ *together with an* RR *stationary-radius bound making* $M\,\|z - z^*\| \leq \sigma_\star/2$*—which we do not prove; it is left open.*

Proof. Herded upper bound. The linearization at z^* is an affine instance with $M_i = DF_i(z^*)$, $b_i = F_i(z^*)$. Under (H1), Proposition 1 and Appendix C give the ideal generic cross-term $\widetilde{O}(\beta^2 LR\sqrt{n})$, hence a floor $\widetilde{O}(\beta LR/(\mu\sqrt{n}))$ after dividing by the gap $1 - q \geq c'\beta n\mu$. The herded order has $Q_{\mathrm{H}} = \widetilde{O}(R)$, so Lemma 25 adds per-epoch forcing $C\beta^3[n^2L^2R + nMR^2]$ up to logarithms, i.e. a floor $C\beta^2[nL^2R + MR^2]/\mu$; the second and third comparisons in (19) bound this by a constant times $\beta LR/(\mu\sqrt{n})$. Error-proportional corrections are absorbed into the contraction under $\beta \leq c\mu/(nL^2)$ exactly as in Theorem 8.

RR *lower bound.* The fluctuation is a *stationary* object, evaluated at the running iterate rather than at z^*, so we apply Lemma 7 to the affine instance linearized at that anchor z—not at z^*—which avoids the Jacobian substitution of Lemma 1 entirely and costs nothing: (A2) gives $\sigma_{\min}(DF(z)) \geq \mu$ at *every* z, and at stationarity $\|z - z^*\| = O(\mathcal{P})$ gives $\frac{1}{n}\sum_i \|F_i(z)\|^2 \geq R^2/4$ once $\mathcal{P} \leq R/(4L)$, which (19) enforces. Hence, through Lemma 12—whose step condition $\beta \leq c'\mu/(\sqrt{d}\,nL^2)$ is implied by (19) for c_d small—the anchored affine fluctuation obeys $\mathbb{E}\,\|\xi_{\mathrm{aff}}\|^2 \geq c\,\beta^4 n^3 \sigma_\star^2 R^2/d$ with $\sigma_\star = \sigma_{\min}(DF(z^*))$ replaced by its uniform lower bound μ where the anchor moves. Write the stationary epoch fluctuation as $\xi = \xi_{\mathrm{aff}} + \xi_{\mathrm{rem}}$. For a uniformly random order, finite-population concentration gives $D_\pi(z) = \widetilde{O}(R\sqrt{n})$ except with probability $(nd)^{-k}$ (any fixed k); on this event Lemma 25 with $Q_\pi = \widetilde{O}(R\sqrt{n})$ gives $\|\xi_{\mathrm{rem}}\| \leq C\beta^3[n^{5/2}L^2R + n^2MR^2] \leq \frac{1}{2}\sqrt{c/d}\;\beta^2 n^{3/2}\mu R$—the μ-scale being the largest the affine part supports at the moving anchor, so absorption must be against it under (19). On the complement the deterministic bound $D_\pi \leq 2n\rho_\infty = \widetilde{O}(nR)$ applies,

and the event contributes $\widetilde{O}((nd)^{-k}\beta^6 n^6[L^2R + MR^2]^2) = o(\beta^4 n^3 \mu^2 R^2/d)$ for k a sufficiently large absolute constant, again under (19). Since $\|\xi\| \geq \|\xi_{\mathrm{aff}}\| - \|\xi_{\mathrm{rem}}\|$ and $(x-b)_+^2 \geq x^2/2 - b^2$,

$$\mathbb{E}\,\|\xi\|^2 \;\geq\; \tfrac{1}{2}\mathbb{E}\,\|\xi_{\mathrm{aff}}\|^2 - \mathbb{E}\,\|\xi_{\mathrm{rem}}\|^2 \;\geq\; \tfrac{c}{4}\,\frac{\beta^4 n^3 \mu^2 R^2}{d},$$

since at the running anchor Lemma 7 yields $\mathbb{E}\,\|\xi_{\mathrm{aff}}\|^2 \geq c\,\beta^4 n^3 \sigma_{\min}(DF(z))^2 R^2/d \geq c\,\beta^4 n^3 \mu^2 R^2/d$ and no more: once $\sigma_{\min}(DF(z))$ has been replaced by μ, the display cannot be restated with $\sigma_\star$, which $\mu \leq \sigma_\star$ does not permit. The stationary second moment inherits the floor by the injection argument of Appendix E. Finally, the ratio of the RR RMS floor $c''\beta^2 n^{3/2}\mu R/\sqrt{d}$ to the herded floor $\widetilde{O}(\beta LR/(\mu\sqrt{n}))$ is $\widetilde{\Omega}(\beta n^2\mu^2/(L\sqrt{d}))$; among the first two terms of (19) the first binds when $\sqrt{n} \leq \kappa$ and the second beyond; the third (curvature) cap can bind instead, contributing the $n^{3/2}\mu^3/(LMR)$ entry of the displayed three-way minimum. □

What the corollary does not cover. It is stated for the ideal linearization order; the stale-record transfer of Lemma 27 preserves the *worst-case* guarantee of Theorem 8, but at the generic scale the record perturbation would additionally have to stay below the cell scale of the bipartition recursion, which we do not prove. And the scale of the RR floor is genuinely instance-dependent: Lemma 7 certifies the $\sigma_\star$-scale unconditionally for *fixed* Jacobians (the affine case), while at the smooth corollary's moving anchor only the uniform μ-scale survives, and isotropic-Jacobian instances show that no condition on the offsets can improve on either; the unconditional worst-case statement remains Theorem 8.

I SINGLE-CALL SEG: PROOF OF THEOREM 4

For affine F_i, one SC-HERD micro-step on component i at iterate w with record r_i computes $g = F_i(w - \alpha r_i) = F_i(w) - \alpha M_i r_i$ and sets $w^+ = w - \beta g$, $r_i^+ = g$. Since the two-call direction is $G_i(w) = (I - \alpha M_i)F_i(w)$,

$$\begin{aligned} w^+ &= w - \beta G_i(w) + \alpha\beta M_i e_i, \quad e_i = r_i - F_i(w), \\ r_i^+ &= F_i(w) - \alpha M_i r_i. \end{aligned}$$

Deferred from Theorem 4. The switched-stability clause—boundedness under an *arbitrary* per-epoch order sequence, $\limsup_k \|s^k\|_\gamma \leq \sup_k \|\varphi_{\pi^k}\|_\gamma / \Theta(\beta n\mu)$—is derived below, together with why the sharp parity needs the order to settle (the fixed-point identity of Lemma 31 requires a genuine augmented fixed point s^∞, which a switched sequence need not possess). Proving (9) for a genuinely switched sequence would need a non-fixed-point staleness bound, e.g. $\|e_i^k\| \lesssim \alpha LR + L \cdot$ (motion of z between component i's last two visits), and is open.

Every micro-step is affine in the *augmented state* $s = (z; r_1, \ldots, r_n)$, so one epoch with order π is an affine map $s \mapsto \Psi_\pi s - \varphi_\pi$. Lemma 29 bounds the linear part by $\|\Psi_\pi\|_\gamma \leq 1 - \Theta(\beta n\mu)$ in one fixed norm $\|\cdot\|_\gamma$ and *for every* π; its proof uses no property of the order. A common contraction norm makes even the switched system—the order recomputed each epoch—converge: $\|s^{k+1}\|_\gamma \leq (1 - \Theta(\beta n\mu))\,\|s^k\|_\gamma + \max_\pi \|\varphi_\pi\|_\gamma$, so the iterate enters and stays in a ball of radius $\max_\pi \|\varphi_\pi\|_\gamma / \Theta(\beta n\mu)$. This is Theorem 4's switched-stability clause, and it is all the switched recursion gives: with $\|\Psi_\pi\|_\gamma \leq q$ for every π, $\limsup_k \|s^k\|_\gamma \leq \sup_k \|\varphi_{\pi^k}\|_\gamma/(1-q)$, and $\|\varphi_{\pi^k}\|_\gamma$ is controlled by $D(\pi^k; b)$ through Appendix C. Two cautions. First, for a *fixed* π the map $v \mapsto D(\pi; v)$ is Lipschitz with constant $O(n)$ in ℓ_∞—that n is the content of Lemma 30 and must not be dropped—whereas the composite $r \mapsto \pi(r) \mapsto D(\pi(r); b)$ is not Lipschitz at all, since $\pi(\cdot)$ is piecewise constant. Second, identifying the plateau with the two-call value uses the fixed-point identity of Lemma 31, which needs a genuine s^∞; that is why the sharp clause (b) is stated for a fixed or eventually stabilized order rather than for an arbitrary switched sequence. Stabilization is recorded in Remark 2 for the deterministic greedy rule and measured in Appendix N: clause (a)'s boundedness does not use it, but clause (b)'s sharp parity does.

Lemma 28 (Records stay bounded)**.** $\|r_i^k\| = O(R + L\,\|z_0^k\|)$ *at every epoch k, where z_0^k is the epoch-start iterate.*

Proof. $r_i^k = F_i(\hat{w})$ for the extrapolation point $\hat{w} = w - \alpha r_i^{k-1}$ at i's visit, so $\|r_i^k\| \le L\|\hat{w}\| + R_\infty \le L\|w\| + \frac{1}{2}\|r_i^{k-1}\| + R_\infty$ (using $\alpha L \le \frac{1}{2}$). A within-epoch Grönwall bound gives $\|w\| = O(\|z_0^k\| + \beta n R)$ for $\beta n L = O(1)$; unrolling the recursion in k gives the claim. □

Lemma 29 (Augmented contraction). *In the norm $\|s\|_\gamma = \|z\| + \gamma \max_i \|r_i\|$ with $\gamma = c\,\beta n\mu/L$ for a small absolute constant c, the linear part of Ψ_π satisfies $\|\Psi_\pi\|_\gamma \le 1 - \Theta(\beta n\mu)$. Hence, for any fixed order π,* SC-HERD *converges to a unique fixed point $s^\infty = (z_{\mathrm{SC}}^\infty; r^\infty)$.*

Proof. Over an epoch the z-component obeys $z^+ = \Phi_\pi z + \eta + \text{const}$, $\|\eta\| \le O(\alpha\beta n L)\max_i \|r_i\|$ (the accumulated $\alpha\beta M_i r_i$ terms), and each record obeys $r_i^+ = M_i w_{t_i} - \alpha M_i r_i + b_i$ with $\|w_{t_i}\| \le O(\|z\|) + O(\alpha\beta n L)\max_i \|r_i\|$. Taking deviations from s^∞ and writing $\|\Delta r\| = \max_i \|\Delta r_i\|$,

$$\|\Delta z^+\| \le q\,\|\Delta z\| + O(\alpha\beta n L)\,\|\Delta r\|,$$
$$\|\Delta r^+\| \le O(L)\,\|\Delta z\| + O(\alpha L)\,\|\Delta r\|,$$

with $q = 1 - \Theta(\beta n\mu)$ (Lemma 4). With $\gamma = c\,\beta n\mu/L$ the $\|\Delta z\|$-coefficient of $\|\Psi_\pi \Delta s\|_\gamma$ is $q + \gamma O(L) = 1 - \Theta(\beta n\mu)$ for c small, and the $\|\Delta r\|$-coefficient is $O(\alpha\beta n L) + \gamma O(\alpha L) \le (1 - \Theta(\beta n\mu))\gamma$ once $\alpha \le c'\mu/L^2$—which holds under Theorem 1's step size $\beta = \alpha \le 2\mu/(9L^2)$. □

Remark 2 (Conditional freezing of the greedy rule; not used by any bound). If the eventual record region lies in one greedy-order cell, the order freezes. Otherwise the switched-stability clause gives boundedness, not freezing or sharp parity. A fixed order permits Theorem 4's matching bound only with that theorem's ideal-offset discrepancy hypothesis; neither boundedness nor the name of the greedy routine certifies that hypothesis.

Why it holds. The greedy rule (Algorithm 3) is deterministic and piecewise constant on record space—finitely many cells, one permutation each, the boundaries being the quadrics $\|p + u_i\| = \|p + u_j\|$. By Lemma 29 the switched system enters a ball $\mathcal{B}$ of radius $O(\max_\pi \|\varphi_\pi\|_\gamma / (\beta n\mu))$ within $O\big((\beta n\mu)^{-1}\log(\|s^0\|_\gamma / \operatorname{rad}\mathcal{B})\big)$ epochs. If $\mathcal{B}$ lies in a single cell, the order is a fixed π at every later epoch; we do not prove that this happens: measure-zero boundaries make it unlikely for a *fixed* point to sit on one, but for a genuinely switched trajectory we have not established that the records possess a limiting point, so "the records' limiting distance to the nearest boundary" is not available as a quantity. We therefore report freezing as an empirical observation (Appendix N measures its frequency) rather than a generic fact. Otherwise the order may visit multiple cells meeting $\mathcal{B}$. This observation alone gives neither a small ideal-offset discrepancy nor the fixed-order parity bound. Lemma 30 applies at a genuine augmented fixed point and requires a separate bound on its actual record discrepancy.

Lemma 30 (Record-to-offset discrepancy transfer). *Let $F_i(z) = M_i(z - z^*) + b_i$, $\sum_i b_i = 0$, $\|M_i\| \le L$, and $\frac{1}{2}(\bar{M} + \bar{M}^\top) \succeq \mu I$, with $\kappa = L/\mu$. For any records r_i and the same permutation π, writing $\widetilde{r}_i = r_i - \bar{r}$ gives the elementary bridge*

$$D(\pi; b) \le D(\pi; \widetilde{r}) + n \max_i \|b_i - \widetilde{r}_i\| \le D(\pi; \widetilde{r}) + 2n \max_i \|r_i - b_i\|.$$

Now suppose $(z; r_1, \dots, r_n)$ is a genuine fixed point of the augmented SC-HERD *epoch map with fixed order π, $\beta > 0$ and $\alpha \ge 0$. Then $\sum_i r_i = 0$, so $\widetilde{r}_i = r_i$. Set $D_r = D(\pi; r)$ and $D_b = D(\pi; b)$. Without assuming a herded plateau bound,*

$$\|z - z^*\| \le \frac{(\beta + 2\alpha)L}{\mu}\, D_r, \qquad D_b \le \big[1 + (\beta + 2\alpha)nL(1 + \kappa)\big] D_r.$$

In particular, for $\alpha = \beta$ and $\beta n L \le 1/3$, $D_b \le (2 + \kappa)D_r$; if also $\beta n L^2 \le c\mu$, then $D_b \le (1 + 6c)D_r$. Thus an independently established $D_r = \widetilde{O}(R)$ transfers to $D_b = \widetilde{O}(R)$ in the latter sub-range and to $D_b = \widetilde{O}(\kappa R)$ under the former cap. These are fixed-point implications, not guarantees that a recomputed order freezes or that a particular balancing heuristic achieves $D_r = \widetilde{O}(R)$.

Proof. The elementary bridge follows by summing prefix differences and using $\|\bar{r}\| \le \max_i \|r_i - b_i\|$. At an augmented fixed point each component's updated record equals its stored record; closure of the epoch therefore gives $0 = z_{n+1} - z_1 = -\beta \sum_i r_i$. Define $Q_t = \sum_{s<t} r_{\pi_s}$ for

Algorithm 3 Greedy herded ordering

1: Input centered records $u_1, \dots, u_n$ with $\sum_i u_i = 0$.
2: $S \leftarrow \{1, \dots, n\}$, $p \leftarrow 0$, $\pi \leftarrow ()$.
3: **for** $t = 1, \dots, n$ **do**
4: Choose $i_t \in \arg\min_{i \in S} \|p + u_i\|$.
5: Append i_t to π; set $p \leftarrow p + u_{i_t}$ and $S \leftarrow S \setminus \{i_t\}$.
6: **end for**
7: Return π.

$t = 1, \dots, n+1$. Then $Q_1 = Q_{n+1} = 0$, $\|Q_t\| \le D_r$, and $\|r_{\pi_t}\| = \|Q_{t+1} - Q_t\| \le 2D_r$. Writing $e = z - z^*$, the iterate at visit t is $z - \beta Q_t$, so the exact record equation is

$$r_{\pi_t} = M_{\pi_t}(e - \beta Q_t - \alpha r_{\pi_t}) + b_{\pi_t}.$$

Summing its rearrangement yields

$$n\bar{M}e = \beta \sum_t M_{\pi_t} Q_t + \alpha \sum_t M_{\pi_t} r_{\pi_t}.$$

Since $\left\|\bar{M}^{-1}\right\| \le 1/\mu$, the first displayed bound follows. For any prefix, sum $b_{\pi_t} = r_{\pi_t} - M_{\pi_t} e + \beta M_{\pi_t} Q_t + \alpha M_{\pi_t} r_{\pi_t}$ and bound its norm by $D_r + nL\|e\| + (\beta + 2\alpha) nLD_r$. This proves the transfer bound. For $\alpha = \beta$, the multiplier is $1 + 3\beta nL(1+\kappa)$. It is at most $2 + \kappa$ when $\beta nL \le 1/3$, and at most $1 + 6c$ when $\beta nL^2/\mu \le c$, since $\kappa \ge 1$. □

Lemma 31 (Exact staleness at the fixed point)**.** *At s^∞, the staleness is $e_i^\infty = -\alpha M_i r_i^\infty$ exactly, with $\|r_i^\infty\| = O(R)$.*

Proof. At s^∞ the record update leaves r_i fixed, so $r_i^\infty = F_i(w - \alpha r_i^\infty)$ for the fixed iterate w at i's visit. Hence $e_i^\infty = r_i^\infty - F_i(w) = F_i(w - \alpha r_i^\infty) - F_i(w) = -\alpha M_i r_i^\infty$. The norm bound is Lemma 28 at the plateau, where $\|z_{\mathrm{SC}}^\infty\| = O(\beta LR/\mu)$. □

Proof of Theorem 4 (sharp parity). The switched-stability clause is the recursion above. For (b), the order being fixed (or eventually so), Lemma 29 makes the epoch map a contraction with a genuine fixed point, so SC-HERD reaches s^∞—this is exactly where a switched sequence would fail, no common s^∞ being available. By the micro-step identity its epoch map is the two-call herded-SEG epoch map plus $r_{\mathrm{extra}} = \sum_\ell \alpha\beta M_{\pi_\ell} e_{\pi_\ell}^\infty$ (up to $O(\beta)$ higher-order factors); by Lemma 31, $r_{\mathrm{extra}} = -\alpha^2\beta \sum_\ell M_{\pi_\ell}^2 r_{\pi_\ell}^\infty$, so $\|r_{\mathrm{extra}}\| \le \alpha^2 \beta nL^2 \max_i \|r_i^\infty\| = O(\alpha^2 \beta nL^2 R)$. Writing z_{2c}^∞ for the two-call herded-SEG fixed point under the same order, both fixed points satisfy $z^\infty = \Phi_\pi z^\infty + r_\pi + (\cdot)$, the extra term being r_{extra} for SC-HERD and 0 for two-call SEG; subtracting,

$$\begin{aligned} z_{\mathrm{SC}}^\infty - z_{\mathrm{2c}}^\infty &= (I - \Phi_\pi)^{-1} r_{\mathrm{extra}}, \\ \|z_{\mathrm{SC}}^\infty - z_{\mathrm{2c}}^\infty\| &\le \|r_{\mathrm{extra}}\| / (1 - q) \\ &\le O(\alpha^2 \beta nL^2 R)/(\beta n\mu/8) = O(\alpha^2 L^2 R/\mu). \end{aligned}$$

Hence $\mathcal{P}_{\text{SC-HERD}} = \|z_{\mathrm{SC}}^\infty - z^*\| \le \|z_{\mathrm{2c}}^\infty - z^*\| + O(\alpha^2 L^2 R/\mu) = \mathcal{P}_{\text{herded-SEG}} + O(\alpha^2 L^2 R/\mu)$. The herded-SEG floor is of order $\alpha\beta$ or β^2 (Appendix A), so for $\alpha L\sqrt{n} \lesssim 1$ the added $O(\alpha^2)$ term is higher order. □

The contrast with SPEG is structural: SPEG reuses the previous *step's* value, from a *different* component, so its staleness is $O(R)$—the full heterogeneity, not the $O(\alpha R)$ of Lemma 31. For smooth F_i the staleness itself is still small—$e_i^\infty = F_i(w - \alpha r_i^\infty) - F_i(w)$ is $O(L\alpha R)$ there too, and the $O(\beta^3)$ remainder of Appendix H is unchanged—but that does not deliver a smooth version of Theorem 4: the augmented-state contraction of Lemma 29 is proved for the affine map $s \mapsto \Psi_\pi s - \varphi_\pi$, and a nonlinear augmented contraction argument would be needed. Nonlinear SC-HERD is therefore an empirical extension (Appendix N).

J Extended Proof Overview

We outline the affine argument. In the affine model one epoch is exactly an affine map; its bias has an explicit expansion up to a third-order remainder, and the smooth case is a perturbation of this picture. Full proofs are in the appendices. Let $F_i(z) = M_i z + b_i$ and define the extragradient-corrected affine operator

$$\begin{aligned} G_i(z) &= (I - \alpha M_i)F_i(z) = A_i z + g_i, \\ A_i &= (I - \alpha M_i)M_i, \qquad g_i = (I - \alpha M_i)b_i. \end{aligned} \tag{20}$$

One epoch with order π is the affine map

$$z \mapsto \Phi_\pi z - \psi_\pi, \quad \Phi_\pi = \prod_{j=n}^{1}(I - \beta A_{\pi_j}), \tag{21}$$

where

$$\psi_\pi = \beta \sum_{j=1}^{n} \left[\prod_{\ell>j} (I - \beta A_{\pi_\ell}) \right] g_{\pi_j}. \tag{22}$$

At $z^* = 0$, the per-epoch bias $r_\pi = -\psi_\pi$ admits the expansion

$$\begin{aligned} r_\pi &= -\beta \sum_i g_i + \beta^2 \sum_{j=1}^{n} A_{\pi_j} P_j + O(\beta^3), \\ P_j &= \sum_{s<j} g_{\pi_s}. \end{aligned} \tag{23}$$

The first term in Eq. (23) is order independent. It is the extragradient-correction bias, and herding cannot change it. The second term is the only leading term that depends on the order. It is the SEG analogue of the prefix-sum term controlled by GraB, but here it appears inside the per-epoch bias of an extragradient method.

Strong monotonicity gives an epoch contraction $\|\Phi_\pi\| \le q < 1$ for small enough α, β. Thus

$$\left\|z^K - z^*\right\| \le q^K \left\|z^0 - z^*\right\| + \frac{\|r_\pi\|}{1-q}. \tag{24}$$

Since $1 - q$ is of order $\beta n \mu$ for a full epoch, the bias-to-plateau conversion preserves the relative size of the order-dependent cross-term. Writing $C_0 = \|\sum_i M_i b_i\|$ and $B(\pi) = \left\|\sum_j A_{\pi_j} P_j\right\|$, the leading-order plateau has the schematic form

$$\mathcal{P}(\pi) \asymp \frac{1}{n\mu}\big\{\alpha C_0 + \beta B(\pi)\big\}. \tag{25}$$

The C_0 term is common to every order and is the part targeted by anchoring. The herding question is exactly whether $B(\pi)$ can be made smaller than its random-reshuffling value.

Herding achieves this. For a random order, the prefix process has size $\Theta(R\sqrt{n})$ and the cross-term has size $\Theta(LRn^{3/2})$ on rotation-dominated heterogeneous instances; instance-wise it is at least $\sigma_\star R n^{3/2}/\sqrt{d}$ with $\sigma_\star = \sigma_{\min}(\bar{M}) \ge \mu$ (Appendix C). A herded order has $D(\pi) = \widetilde{O}(\sqrt{d}\,R)$ in the worst case, giving an $\widetilde{O}(LRn)$ cross-term by a direct triangle bound. The resulting plateau gain—an RR *lower* bound divided by a herded *upper* bound—is $\widetilde{\Omega}(\sqrt{n}\,\mu/L)$ in norm on rotation-dominated instances at the largest admissible step, under the step range and non-dominance conditions of Theorem 1. On generic instances, under Proposition 1's centered model, $n = 2^q$, its step caps and the per-node hypothesis ($\mathrm{H}_{\mathrm{split}}$), recursive balanced bipartition gives an additional outer-sum cancellation. The herded cross-term is then only $\widetilde{O}(LR\sqrt{n})$, an n-fold coefficient reduction, and the finite-step plateau separation it yields is $\widetilde{\Omega}(\min\{n/\kappa^2, \sqrt{n}/\kappa\})$ (Proposition 1; Appendices B–C and F). The near-linear growth in the scaling experiments is empirical evidence for this coefficient mechanism, not a verification of the certified finite-step rate.

The random-reshuffling lower bound comes from re-randomization itself. In SEG-RR, the order-dependent part of r_π is redrawn each epoch. This creates an i.i.d. mean-zero bias fluctuation with

variance proportional to the random-order cross-term. In stationarity, the AR(1)-type recursion retains that variance as a plateau. A fixed herded order has no such per-epoch fluctuation; its trajectory converges to a point. This is visible in the experiments through the tail coefficient of variation, and is made precise in Appendix E.

For SEG-FFA, the forward and backward sweeps cancel the leading β^2 order-dependent term in Eq. (23). Expanding the two-sweep map shows that the leading residual is the β^3 term of Eq. (14), plus order-independent and higher-order terms. The same prefix sums P_j remain, so under the same no-dominant-component condition a herded order reduces the surviving order residual by the same worst-case discrepancy factor (Theorem 7). As in Section 4.5, the two act on compatible axes: flip-flop and anchoring on the epoch axis, herding on the component axis.

For SC-HERD, in the affine case the two-call SEG direction is $(I-\alpha M_i)F_i(z)$, whereas the single-call direction is

$$\begin{aligned} F_i(z-\alpha r_i) &= (I-\alpha M_i)F_i(z) - \alpha M_i e_i, \\ e_i &= r_i - F_i(z). \end{aligned} \tag{26}$$

Thus the micro-step differs from two-call SEG by $\alpha\beta M_i e_i$. For a fixed or eventually stabilized order, SC-HERD reaches a genuine augmented fixed point, where r_i is the same component's operator value from its previous visit and the staleness is $O(L\alpha R)$ (Lemma 31). Summing over an epoch and dividing by the contraction gap gives the additive $O(\alpha^2 L^2 R/\mu)$ plateau term of Theorem 4, whose parity statement also needs the order to herd the ideal offsets. For an order sequence that keeps switching, only boundedness is proved (Appendix I).

An equivalent Lyapunov view explains why *prefix sums*—not the final permutation sum—are the controlling object. Summing the one-step descent inequality over an epoch, the within-epoch drift is a scaled prefix sum and the order-dependent error collects as $\beta^2 L^2 \sum_{t=1}^{n} \|P_t\|^2$ with $P_t = \sum_{s<t} F_{\pi_s}(z^*)$. For a random order $\mathbb{E}\,\|P_t\|^2$ follows the finite-population Brownian-bridge law $R^2 t(1-t/n)$, summing to $\Theta(R^2 n^2)$; herding to $D(\pi) = \widetilde{O}(R)$ makes the sum $\widetilde{O}(R^2 n)$. This squared-prefix form is the SEG analogue of GraB's descent term, it yields the worst-case theorem, and it is the form that extends cleanly to general smooth F (Appendices D and H).

The smooth nonlinear case is a perturbation of the affine one. Linearizing each F_i at z^* reproduces (23) with $M_i = DF_i(z^*)$ and $b_i = F_i(z^*)$, and Jacobian-Lipschitzness bounds the remainder at order β^3 (Lemma 25). Being third order does not by itself make the remainder negligible uniformly in n: it is persistent epoch forcing and accumulates as its size divided by the contraction gap. Appendix H therefore states the comparison $\beta \lesssim \min\{\mu/(nL^2),\, \mu/(MQ_{\mathrm{H}})\}$—equivalently $\beta \lesssim \mu/(nL^2 + MQ_{\mathrm{H}})$—under which it falls below the herding-controlled term and the epoch still contracts. The mechanism is thus not tied to quadratic games. For general smooth operators, however, the statement without genericity hypotheses is the worst-case bound. The generic factor n needs both the linearization genericity hypothesis (H1) of Appendix H and the per-node split accuracy ($\mathrm{H}_{\mathrm{split}}$); its variance-nondegeneracy half is proved instance-wise (Lemma 7).

Two remarks place SC-HERD. First, it is not SAGA-style variance reduction: the record table is not an unbiased estimate of F, but a device that simultaneously selects a balanced order and supplies a *component-aligned* lookahead. The extragradient lookahead is component-specific—in a two-call step the direction is $F_i(z-\alpha F_i(z))$—so the table's one-value-per-component structure is exactly what SPEG lacks: SPEG reuses the previous step's value, from a different component. Second, reuse removes SEG's second operator call. An incremental gradient step already makes a single call, so there is no second call to remove; the second call is native to extragradient. Once herding stores component values, the lookahead reuses them without an additional operator call, at the price of the staleness term of Theorem 4; the table's memory and the ordering arithmetic are still paid. SC-HERD thus occupies its own point in the design space—the operator-call count of SPEG, the order quality of herded SEG, and the component-aligned lookahead of extragradient. The conservative perturbation proof is given in Appendix I.

K ANTITHETIC HERDING: PROOFS

Proof of Proposition 2. $P_t^{u,R} = \sum_{s\le t} u_{\pi_{n+1-s}} = \sum_{k\ge n+1-t} u_{\pi_k} = P_n^u - P_{n-t}^u = -P_{n-t}^u$ since $P_n^u = 0$; the discrepancy statement is immediate. For the bias, $X_u(\pi^R) = \sum_t M_{\pi_{n+1-t}} P_{t-1}^{u,R} = -\sum_t M_{\pi_{n+1-t}} P_{n+1-t}^u = -\sum_m M_{\pi_m} P_m^u = -\sum_m M_{\pi_m}(P_{m-1}^u + u_{\pi_m}) = -X_u(\pi) - \sum_i M_i u_i$.

Reindexing $t \mapsto n+1-t$ also shows $\sum_t (t-1)M_{\pi_t^R} = (n-1)\sum_i M_i - \sum_t (t-1)M_{\pi_t}$, so the position-weighted mean term of X is likewise order-free after summing π and π^R. □

Proof of Theorem 3. Let T_π, T_{π^R} be the epoch maps and $T_2 = T_{\pi^R} \circ T_\pi$. Applying Lemma 1 at z_0 and then at $z_{\mathrm{mid}} = T_\pi(z_0)$, and using Proposition 2 to sum the two β^2 brackets, gives

$$T_2(z_0) - z_0 = -2\beta n F(z_0) + 2\beta^2 \Phi_{\mathrm{free}}(z_0) + E_2, \qquad \|E_2\| \le 50\beta^3 n^3 L^2 R_0,$$

where Φ_{free} does not depend on π; the two extra sources beyond the lemma's remainders are the evaluation of the second bracket at z_{mid} (Lipschitz in z, and $\|z_{\mathrm{mid}} - z_0\| \le 2\beta n R_0$) and the composition cross-term $(\Psi_{\pi^R} - I)$ applied to the first displacement, both $O(\beta^3 n^3 L^2 R_0)$. This is affine-only: for smooth F_i the two brackets are taken at different anchors, and the Jacobian drift across the pair adds $\beta^2 nM \|z_{\mathrm{mid}} - z_0\| \cdot nR_0 \le 2\beta^3 n^3 M R_0^2$, so E_2 would read $O(\beta^3 n^3 [L^2 R_0 + M R_0^2])$—Lemma 1's smooth remainder shape. We do not state a smooth antithetic theorem; the nonlinear antithetic runs in Appendix N are therefore empirical. Each epoch map contracts by $1 - \beta n\mu/2$, so T_2 contracts by at least $1 - \beta n\mu$; dividing the bias of T_2 at z^* by $1 - q_2 = \Theta(\beta n\mu)$ gives the general bound $\mathcal{P}_{\mathrm{anti}}^{\mathrm{ord}} \lesssim \beta^2 n^2 L^2 R/\mu$ (the herded refinement below is the displayed (7)). For the herded refinement both sweeps of the pair carry the same discrepancy ($D(\pi^R) = D(\pi)$, Proposition 2), so the surviving cubic forcing is prefix-controlled: Lemma 11 bounds each sweep's order-dependent tail by $C\beta^3 n^2 L^2 D(\pi)$ and the composition cross-term inherits that scale, giving $\widetilde{O}(\beta^3 n^2 L^2 R)$ and hence $\widetilde{O}(\beta^2 n L^2 R/\mu)$ after division.

The residual is not proportional to $X(\pi)$. It is tempting to state the surviving term as a multiple of the fixed order's own contribution $\beta \|X(\pi)\|/(n\mu)$; that is false. Take $n = 3$ in dimension one with $M = (\frac{3}{2}, 1, 1)$, $b = (1, -2, 1)$ (so $\sum_i b_i = 0$, $\mu = 1$, $L = \frac{3}{2}$) and $\pi = (1, 2, 3)$: then $X(\pi) = M_2 b_1 + M_3(b_1 + b_2) = 1 - 1 = 0$, while the exact two-epoch bias retains an order-*dependent* cubic coefficient—against $\pi' = (1, 3, 2)$ the difference is $6\beta^3 + O(\beta^4)$, computed in exact rational arithmetic. A bound multiplicative in $\|X(\pi)\|$ would force the antithetic forcing to vanish on this instance; it does not. Leading-coefficient cancellation does not make the higher-order residual proportional to the cancelled coefficient—the same point as for FFA in Appendix G. With a nonincreasing schedule the pair's β^2 terms combine as $\beta_{2k+1}^2(\mathrm{bias}_2(\pi) + \mathrm{bias}_2(\pi^R)) + (\beta_{2k}^2 - \beta_{2k+1}^2)\mathrm{bias}_2(\pi)$: the first bracket is order-free by Proposition 2 and the second is the schedule residual. Unrolling the recursion with the smaller step's contraction factor gives the two-mechanism bound deferred from the theorem's statement:

$$\mathcal{P}_{\mathrm{anti},k}^{\mathrm{ord}} \lesssim \underbrace{\frac{|\beta_k - \beta_{k+1}|\ \|X\|}{n\mu}}_{\text{unequal steps: incomplete cancellation of } X} + \underbrace{\widetilde{O}\Big(\frac{\beta_k^2 n L^2 R}{\mu}\Big)}_{\text{finite } \beta\text{: cubic residual}},$$

versus $\Theta(\beta_k \|X\|/(n\mu))$ for a fixed order. □

No single-call antithetic theorem. For the single-call method of Section 4.3 we claim no antithetic theorem: reversal is an involution on the component order, but not on the augmented dynamics $(z; r_1, \ldots, r_n)$—the reverse epoch inherits a record table already updated by the forward one, so its staleness errors $\alpha\beta M_i e_i$ are not the reversal of the forward epoch's, and Proposition 2's identity does not transfer. Antithetic SC-HERD is evaluated empirically (Appendix N); a theorem would need a two-epoch expansion of $\Psi_{\pi^R}\Psi_\pi$ in the augmented state.

The variant that chooses π or π^R by a fair coin each epoch makes the order bias exactly mean-zero across epochs and performs the same within noise; deterministic alternation telescopes rather than averaging, and is the default.

L Bias-Cancelling Orders: Proofs

Scope of Theorem 5 (deferred from the statement). Why $8 \mid n$: the periodic block has length 8, and $\mathcal{A}(\pi^\star) = 0$ together with $m_-(\pi^\star) = n/4$ hold by exact repetition of it; truncating the pattern would break $\sum_i s_i = 0$, $\mathcal{A} = 0$ and $m_- = n/4$ simultaneously, so no other n is claimed—an existential counterexample along $n = 8, 16, 24, \ldots$ suffices. The admissible step range is $\beta \le c\min\{\mu/(nL^2),\ \delta/(n^2L^2)\}$: the first cap absorbs the move of the bias coefficient from z^*

to z^∞, the second keeps the cubic tail $O(\beta^3 n^3 L^2 \rho)$ below the leading $\Theta(\beta^2 n \delta \rho)$. With constants, $\mathcal{P}(\mathrm{SPEG}, \pi^\star) \le 160\,\beta^2 n^2 L^2 \rho$ and $\min_\pi \mathcal{P}(\mathrm{SEG}, \pi) \ge \frac{1}{8}\beta\delta\rho$, so the gain is

$$\frac{\min_\pi \mathcal{P}(\mathrm{SEG}, \pi)}{\mathcal{P}(\mathrm{SPEG}, \pi^\star)} \;\ge\; \frac{\delta}{1280\,\beta n^2 L^2},$$

which is $\Theta(1/\beta)$ at fixed n, L, δ as $\beta \to 0$—not a gain uniform in n. The theorem makes no comparison against the single-call *herded* method: Theorem 4 bounds SC-HERD from above, which cannot yield the every-order lower bound such a comparison would need; the empirical comparison is in Appendix N.

L.1 THE EXPLICIT CANCELLING ORDER (THEOREM 5)

On the family $M_i = \mu I + \delta s_i J$, $b_i = \rho s_i v$ with $\|v\| = 1$, J the rotation generator and $\sum_i s_i = 0$, write $\sigma_t = \sum_{s \le t} s_{\pi_s}$, $\mathcal{A}(\pi) = \sum_t \sigma_{t-1}$ and $m_-(\pi)$ for the number of cyclic adjacent sign changes. At $z^* = 0$,

$$X(\pi) = \mu\rho\,\mathcal{A}(\pi)\,v - \tfrac{n}{2}\delta\rho\,Jv, \qquad A(\pi) = -2m_-(\pi)\,\delta\rho\,Jv,$$

because $P_{t-1} = \rho\sigma_{t-1}v$ and, using $s^2 = 1$, $\sum_t s_{\pi_t}\sigma_{t-1} = (\sigma_n^2 - n)/2 = -n/2$ for *every* order. Hence the β^2 bias of two-call SEG has Jv-component $\frac{n}{2}\delta\rho$ regardless of the order, while for SPEG it is $(\frac{n}{2} - 2m_-)\delta\rho$. The periodic pattern $(+,+,-,-,-,-,+,+)$ has running sums $1, 2, 1, 0, -1, -2, -1, 0$ per period, so $\mathcal{A} = 0$, and exactly two sign changes per period, so $m_- = n/4$: both components vanish. The plateau bounds follow from the fixed-point displacement identity $\beta n F(z^\infty) = \beta^2 \mathrm{bias}_2(\pi; z^\infty) + E$ together with $\|\mathrm{bias}_2(z^\infty) - \mathrm{bias}_2(z^*)\| = O(n^2 L^2 \|z^\infty - z^*\|)$ and $\|z^\infty - z^*\| \le \|F(z^\infty)\|/\mu$. At $n = 8$ this order coincides with the minimizer over all 40320 orders at every step size tested.

L.2 AUGMENTED CONTRACTION AND THE CERTIFICATE (THEOREM 6)

SPEG's state is (z, p) with p the previous returned value; each micro-step is affine, $\Delta z^+ = (I - \beta M_i)\Delta z + \beta^2 M_i \Delta p$, $\Delta p^+ = M_i \Delta z - \beta M_i \Delta p$. In the norm $V = \|\Delta z\| + \gamma \max \|\Delta p\|$ with $\gamma = \beta n\mu/(32L)$ and β obeying (11), one epoch contracts by $1 - \beta n\mu/8$ *for every order*—the cap $\beta \le c_0/(nL)$ is needed both to make $\beta n\mu \le 1$, without which the displayed factor is not a contraction factor at all, and to justify the geometric epoch expansion used below — the proof uses no property of the order, only $\|I - \beta M_i\| \le 1 - \beta\mu/2$ and $\beta L \le 1/4$. This gives existence of and geometric convergence to a fixed point; the plateau then follows from the same displacement identity as above, which converts the achieved $\varepsilon = \|\mathrm{bias}_2\|$ into $\|F(z^\infty)\| \le 2\beta\varepsilon/n + O(\beta^2 n^2 L^2 R_\infty)$. It remains to justify calling ε *observable*: it is defined at z^*, which the algorithm does not know a priori. In the affine model it can reconstruct it.

Lemma 32 (Certificate observability)**.** *Suppose each component i has been visited at iterates whose increments $\Delta z_i^{(1)}, \ldots, \Delta z_i^{(d)}$ are* linearly independent *(a rank condition on the recorded trajectory, not a stochastic independence assumption). Then the recorded pairs determine the instance exactly, with no extra operator evaluation:*

$$M_i \text{ from } \Delta F_i^{(k)} = M_i \Delta z_i^{(k)}, \qquad b_i = F_i(x_i) - M_i x_i, \qquad z^* = -\bar{M}^{-1}\bar{b},$$

$\bar{M} = \frac{1}{n}\sum_i M_i$ being invertible because (A2) *makes it strongly monotone. Hence $\varepsilon = \|\mathit{bias}_2(\mathrm{SPEG}, \pi; z^*)\|$ is computable from the records alone and Theorem 6 is a genuine a-posteriori certificate. (Checked end to end at machine precision on random affine instances with $z^* \ne 0$: $\|\widehat{z}^* - z^*\| \approx 7 \cdot 10^{-16}$ and $|\widehat{\varepsilon} - \varepsilon| \approx 2 \cdot 10^{-13}$ over 50 orders.)*

Before the rank condition is met the algorithm's $\widehat{\varepsilon}$ is a model-based *surrogate objective*, not a certificate: a small least-squares residual does not by itself bound $\left\|\widehat{M}_i - M_i\right\|$ in operator norm without further design-conditioning assumptions. Given a certified $\left\|\widehat{M}_i - M_i\right\| \le \eta_i$ the bound gains a corresponding model-error term. Thus BC-SPEG optimizes a surrogate until every component has accumulated d linearly independent increments, after which the reported bias is rigorous. If the optimizer changes the order across epochs, one extra $O(\beta^2 L R_\infty)$ term appears only when the *last* component changes; pinning π_n removes it.

M Analysis of Single-Call SEG-FFA

Theorem 9 (Single-call SEG-FFA)**.** *Replace every two-call micro-step of* SEG-FFA *by the single-call step of Algorithm 1, keeping the flip-flop sweep pair and anchoring. Under the assumptions of Theorem 4—affine, fixed or stabilized order, and the ideal-offset discrepancy hypothesis—the resulting method* SC-FFA-Herd *satisfies*

$$\mathcal{P}(\text{SC-FFA-HERD}) \;\le\; \mathcal{P}(\text{SEG-FFA-HERD}) + O\Big(\frac{\alpha\beta n L^2 R}{\mu}\Big).$$

The single-call correction is thus of the same order as the surviving herded FFA *residual, which for a prefix-controlled order is $O(\beta^3 n^2 L^2 R)$ before division and $O(\beta^2 n L^2 R/\mu)$ after it. Consequently,* under the additional step-size conditions of Theorem 7*—in particular $\beta \le c\,\sigma_\star^2/(\sqrt{d}\,nL^3)$, which that theorem needs to absorb the full $k \ge 4$ tail—the proved worst-case component-axis separation is preserved up to constants, at $2n$ instead of $4n$ operator evaluations per outer epoch.*

The augmented state is $(z, r_1, \dots, r_n)$ and each of the $2n$ micro-steps of an outer epoch is affine in it, as in Appendix I; the anchoring is the affine map $z \mapsto \frac{1}{2}(z_{\text{start}} + z)$. Three lemmas transfer. Records stay bounded (Appendix I's argument with $2n$ steps, so constants double). The two-sweep map contracts for *every* order in the weighted norm of Appendix L, and anchoring composes $\frac{1}{2}(1+q)$, so the anchored epoch contracts by $1 - \beta n\mu/8$. At the fixed point each record equals the operator value at its refresh point, so the lookahead error obeys $\|e_i\| = O(\beta n L R_0)$ on both sweeps — on the backward sweep the refresh happened in the *same* epoch, which is the freshness advantage. Writing the single-call micro-step as the two-call one plus $\alpha\beta M_i e_i$ and summing over the $2n$ micro-steps, the extra per-epoch *forcing* is $O(\alpha\beta^2 n^2 L^2 R_0)$; dividing by $1 - q = \Theta(\beta n\mu)$ gives the stated plateau correction $O(\alpha\beta n L^2 R/\mu)$—the $1/\mu$ appears only after this division. It is of the same order as the β^3 residual surviving flip-flop cancellation for a prefix-controlled order, which divides to $O(\beta^2 n L^2 R/\mu)$ under $\alpha = \beta/2$.

N Experiment Details and Tables

This appendix gives the settings and tables behind the main-text evidence. Appendix O states which evidence layer the results belong to and their dependencies, and lists the scripts and seeds behind Figure 1 and Tables 1–2, with their appendix versions. The default instance is a strongly monotone heterogeneous affine VI with rotational blocks $M_i = \left[\begin{smallmatrix} d_i I & r_i I \\ -r_i I & d_i I \end{smallmatrix}\right]$ and mean-zero offsets b_i, so $z^* = 0$. The plateau metric is the tail mean of $\|F(z)\|$ over a constant-step run (not the minimum, which would flatter the oscillating random-order methods). Reported numbers are medians over 8–24 seeds.

Table 4 gives the affine n-scaling at the recorded $\beta = 1/n$, an empirical setting: our code herds current-iterate values of all components, at n extra evaluations per epoch. Table 5 gives the single-call scaling. A nonlinear VI $F_i(z) = M_i z + b_i + \gamma_i \tanh(z)$ shows the same linear-in-n growth (Figure 1(b)). This is empirical evidence that the mechanism persists beyond the affine model; the smooth theory of Appendix H covers the worst case without genericity hypotheses, and the generic factor n only under (H1) together with the per-node split accuracy ($\mathrm{H}_{\mathrm{split}}$). An adversarial construction with $b_i \in \{\pm v\}$ and identical Jacobians tracks the worst-case $\sqrt{n}$ boundary empirically (fitted slope $+0.52$ in Figure 4(a)); the proven tightness witness is Lemma 9's signed rotational family. Because its components repeat exactly, greedy herding meets exact ties. Figure 4(a) breaks them by perturbing each offset by a relative 10^{-12} (8 draws per instance; 10^{-9} gives the same curve); greedy then settles on a fixed order whose prefix sums accumulate almost coherently, $\|\sum_\ell P_\ell\|$ being 83–100% of its maximum $\frac{n}{2}\|v\|$ over orders of prefix discrepancy $\|v\|$. Left unbroken, the ties are resolved by rounding noise that varies with the iterate: the herded order then changes between epochs (15% of plateau epochs at $n = 64$), the switching partly cancels the order bias, and the gain is higher (7.65 against 4.36 at $n = 64$) while still growing like $\sqrt{n}$. On the generic instances the tail coefficient of variation is ≈ 0.36 for SEG-RR versus ≈ 0 for herded SEG (the runs of Figure 4(a)), a direct exhibit of the reshuffling variance of Appendix E. Diagnostic runs indicate that controlling raw prefix sums is the operative lever: herding the corrected operator G_i instead of F_i, directly fitting the cross-term, and a rotation-aware greedy objective all fail to improve on plain prefix control.

Table 4: Affine n-scaling, $\beta = 1/n$. Ratios random/herded: prefix discrepancy D, plain-SEG plateau, and SEG-FFA plateau.

n	D ratio	SEG ratio	SEG-FFA ratio
8	1.68	1.09	1.62
16	1.94	1.95	2.20
32	2.86	3.88	4.08
64	3.65	7.18	8.97
128	4.74	15.42	27.40
256	6.50	32.61	62.68

Table 5: Single-call scaling, $\beta = 1/n$, affine VIs. Entries are plateau ratios: SC-HERD holds parity with two-call herded SEG, while the single-call baseline SPEG-RR falls behind by a factor that grows with n.

n	SC-HERD / herded SEG	SPEG-RR / SC-HERD
8	0.82	1.96
16	0.94	2.18
32	1.04	3.37
64	0.87	8.02
128	0.90	17.97

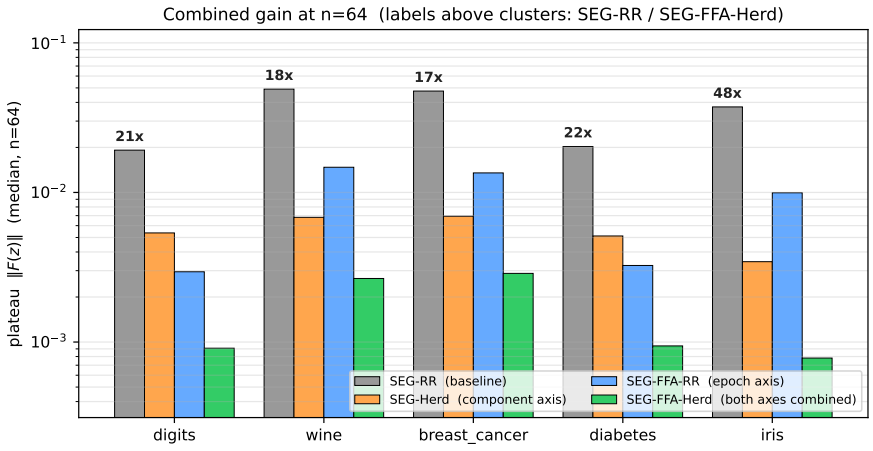

Figure 3: Combined epoch-axis (SEG-FFA) and component-axis (herding) gain at $n = 64$. Vanilla SEG-RR (gray) is the baseline; herding alone (orange) gives the component-axis improvement; SEG-FFA alone (blue) gives the epoch-axis improvement; combining both (green) reaches 17–48$\times$ plateau reduction over SEG-RR (labels above each cluster). The two measured factors compose here; no theorem asserts that they multiply in general.

Ordering routines: cost and comparison. Three ordering routines appear below, summarized in the table. *Guarantees.* The streaming self-balancing walk of Alweiss et al. (2021) carries the $\widetilde{O}(R)$ discrepancy guarantee of Lemma 5 used by the worst-case theorems. The generic factor n uses recursive balanced bipartition (Appendix F) and assumes the per-node split accuracy ($\mathrm{H}_{\mathrm{split}}$) at every node; Theorem 2 proves that accuracy for a single split only. Algorithm 3 (greedy) is an $O(n^2 d)$ heuristic with no standard worst-case discrepancy bound; it is the simplest to state and empirically attains the balancing bound. *Evidence.* This appendix compares all three on the n-scaling: the herding gain grows with n for each, an empirical pattern consistent with low prefix discrepancy rather than with the greedy rule in particular. *Cost.* The deployable order is computed from recorded values, so ordering adds no operator evaluation; its cost is arithmetic. Each SC-HERD epoch uses n operator evaluations, and the near-linear $O(nd \log n)$ ordering cost is small against them only once one evaluation is heavier than a low-dimensional matrix–vector product, as minibatch operators

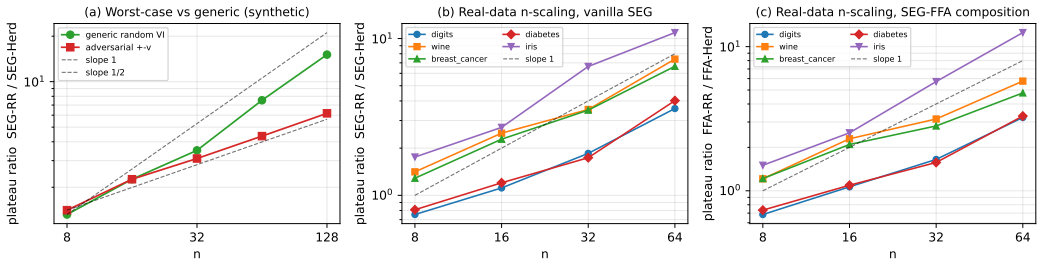

Figure 4: Scaling boundary and real-data reproduction. (a) Worst-case vs generic herding gain on synthetic VIs: generic instances gain approximately linearly in n (empirically; Proposition 1's n-fold coefficient reduction and its finite-step separation $\min\{n/\kappa^2, \sqrt{n}/\kappa\}$ both hold only under its per-node hypothesis ($\mathrm{H}_{\mathrm{split}}$) and step caps); the adversarial $\{+v, -v\}$ construction, its exact ties broken by a relative 10^{-12} perturbation of the offsets (see text), tracks the worst-case $\sqrt{n}$ boundary. (b,c) Real-data n-scaling on five sklearn datasets: vanilla SEG-RR/SEG-HERD (b) and the SEG-FFA composition ratio SEG-FFA-RR/FFA-Herd (c) both grow approximately linearly with n on every dataset. Theorems 1 and 7 certify only worst-case $\sqrt{n}$-scale separations on these instances, so this is evidence that the ordering *mechanism* carries to real feature distributions, not a confirmation of the certified rates.

typically are and the affine test operators here are not. In wall-clock terms, the timing benchmark measures 4–28% ordering overhead, with break-even near 60–70 μs per call. The $O(n^2 d)$ greedy rule is an experimental convenience, not the deployable routine.

routine	guarantee	cost/epoch	exp.
greedy (Alg. 3)	empirical	$O(n^2 d)$	yes
self-balancing walk	$\widetilde{O}(\sqrt{d}\,R)$	$O(nd)$	yes
recursive bipart.	per-split only†	$O(nd \log n)$	yes

†Theorem 2 bounds the imbalance of a *single* split; an order-level discrepancy guarantee would need ($\mathrm{H}_{\mathrm{split}}$) at every node of the recursion, which is verified empirically rather than proved.

Supplementary experiments. The studies below draw a fresh problem instance per seed, so the median and inter-quartile range $[\cdot]$ span the instance distribution; 24 instances each.

Ordering cost. Per epoch, greedy herding (Algorithm 3, $O(n^2 d)$) costs 6.8× the n operator evaluations at $n = 16$ and 193× at $n = 512$; the recursive balanced order ($O(nd \log n)$) costs 3.3× rising only to 7.9×. SC-HERD makes n operator calls per epoch against $2n$ for two-call SEG—a 2× cut in operator evaluations; it becomes a 2× wall-clock saving when an evaluation dominates the per-component ordering work ($\approx 22\,\mu$s in this pure-NumPy timing), as a finite-sum operator over a real minibatch typically does.

Ordering routines. Table 14 reruns the affine n-scaling with three orders that all target small prefix sums: greedy herding (Algorithm 3); recursive balanced bipartition; and the same recursion with each split signed by the self-balancing walk of Alweiss et al. (2021). The herding gain $\mathcal{P}(\mathrm{RR})/\mathcal{P}(\mathrm{HERD})$ grows with n for all three—reaching 10.8× (greedy), 4.6× (recursive) and 3.8× (walk) at $n = 128$—so it is a property of low prefix discrepancy, not of the greedy rule in particular. Greedy grows the steepest, attaining an empirically $O(1)$ discrepancy; of the other two, only the self-balancing walk carries the provable worst-case $\widetilde{O}(\sqrt{d}\,R)$, the GD-D recursion having a per-split guarantee only. Both recursive routines here pair indices consecutively and so realize the worst-case $\sqrt{n}$ class; the generic factor-n of Proposition 1 additionally requires a short-edge (nearest-neighbor-style) pairing, isolated next. The recursive routines must *concatenate* their ordered halves (interleaving gives discrepancy $O(n)$) and flip a fair per-node coin for child order—including at the size-two leaves; omitting the leaf coins re-correlates the outer sum and destroys the cancellation—as in Appendix F.

Generic recursive construction. Proposition 1's factor n needs its per-node accuracy ($\mathrm{H}_{\mathrm{split}}$), for any label-blind tree; short edges are a useful empirical *proxy* for that condition, not a proof of it:

in fixed dimension d a local matching has characteristic edge scale $\widetilde{O}(Rm^{-1/d})$, which shrinks with the sub-problem size while consecutive pairs stay $O(R)$, and only at $d = 2$—the diagnostic below—does it read $\widetilde{O}(R/\sqrt{m})$; GD-D reaches its $m^{-1/2}$ *split imbalance* through multiple differencing-and-steering rounds, not from one pairing's edge length, and edge length alone does not establish ($\mathrm{H}_{\mathrm{split}}$) in higher dimensions. A direct node-level audit of ($\mathrm{H}_{\mathrm{split}}$) itself is in the diagnostics paragraph below (Figure 5(a)). Table 15 isolates this on bounded-domain generic instances (uniform offsets in $\mathbb{R}^2$, centered—the $d = 2$ case of Proposition 1's model), against the same recursion with consecutive pairing. The short-edge matching—a kd-tree median bisection, every pair inside a leaf cell—makes the herding gain grow markedly faster: $13.9\times$ versus $6.4\times$ at $n = 128$, an empirical exponent $\approx n^{0.8}$ against consecutive pairing's $\sqrt{n}$. This trends toward the coefficient mechanism's factor n (Proposition 1 proves the n-fold coefficient reduction; its finite-step plateau guarantee is conditioning-dependent), the residual gap reflecting finite n and the polylog/κ factors the $\widetilde{O}$ hides—support for Proposition 1's construction, measured at $\beta = 1/(nL)$ rather than inside its certified step regime.

Conditioning and heterogeneity. Table 8 sweeps the condition number $\kappa = L/\mu$: the herding gain $\mathcal{P}_{\mathrm{order}}(\mathrm{RR})/\mathcal{P}_{\mathrm{order}}(\mathrm{HERD})$ falls from $\approx 5\times$ to $\approx 1.3\times$ as κ grows $4.5 \to 78$, the degradation the certified bounds anticipate ($\sqrt{n}/\kappa$ worst case; generic $\min\{n/\kappa^2, \sqrt{n}/\kappa\}$ under ($\mathrm{H}_{\mathrm{split}}$)). Sweeping the heterogeneity instead, across b-scale $\in \{0.25, 1, 4, 16\}$ (a $64\times$ range of R), the gain is $3.0\,[2.4, 3.6]$—unchanged, since R cancels in the ratio. With a quarter of the components anti-monotone but the average strongly monotone, the gain is still $2.4\,[1.8, 2.8]$: herding helps outside the component-wise assumption of Theorem 1.

Mean-Jacobian nondegeneracy ($\sigma_\star$-sweep). Lemma 7 gives an instance-wise lower bound proportional to $\sigma_\star^2 = \sigma_{\min}(M)^2$, not an equality for the variance. Table 6 varies this scale at fixed L, μ, n and R: components $M_i = \mu I \pm cJ$ share $\|M_i\| = L$, while the sign mix cancels the mean rotation. The left block measures the corrected cross-term $X_g = \sum_{s<t} A_{\pi_t} g_{\pi_s}$. Its reference bound applies Lemma 7 to centered offsets $g_i' = g_i - \bar{g}$, whereas the measured statistic uses g_i. Thus the reported reference/measured ratios 0.44–0.52 are numerical diagnostics, not a direct same-statistic verification: X_g also contains the order-dependent term $\sum_t (t-1) A_{\pi_t} \bar{g}$. The measured variance broadly follows the $\sigma_\star^2$ trend on this family. In the right block, the RR/herded second-moment ratio falls from 1304 to 3.0 without reversing. These dynamics at $n = 64$ and $\beta = 1/(3nL)$ are empirical; their step is not certified by the sufficient conditions used in the proof of Lemma 12. Neither a general non-inversion guarantee nor an attribution of the finite-step discrepancy to a particular remainder term follows.

Table 6: $\sigma_\star$-sweep: $M_i = \mu I \pm cJ$ with a $p : (1-p)$ sign mix cancelling the mean rotation at fixed $L = 2.02$, $\mu = 0.3$, $R = 1$, $d = 2$. Left block: permutation variance of the cross-term X at $n = 256$ (1500 sampled orders), its ratio to the $p = 1$ row, the predicted $\sigma_\star^2$ ratio, and the centered-offset reference bound divided by the measured variance of X_g (see text). Right block: stationary SEG at $n = 64$, $\beta = 1/(3nL)$ (median over seeds 100–104)—tail-fluctuation ratio and the SEG-RR/herded second-moment ratio.

		$\mathrm{Var}_\pi(X)$, $n = 256$			stationary, $n = 64$	
p	$\sigma_\star$	ratio	$\sigma_\star^2$ ratio	bound/measured	fluct. ratio	RR / herded
1.000	2.02	1.000	1.000	0.51	1.000	1304
0.875	1.53	0.577	0.572	0.51	0.537	470
0.750	1.04	0.262	0.267	0.52	0.270	175
0.625	0.58	0.084	0.083	0.51	0.099	11.6
0.500	0.30	0.026	0.022	0.44	0.035	3.0

Heterogeneity outliers. Theorem 1's herded plateau is, in general, $\mathcal{P}_{\mathrm{order}}(\mathrm{HERD}) = \widetilde{O}(\beta L \rho_\infty/\mu)$, the displayed $\sqrt{n}$ gain being the no-dominant-component case $\rho_\infty = \widetilde{O}(R)$. Table 16 builds the failure mode directly: scaling one component's offset by a factor c drives ρ_∞/R from 1.9 to 6.9, and the herding gain falls monotonically, $5.6\times$ to $2.9\times$—a single dominant component erodes the gain, as the general form predicts. The fall is slower than the leading R/ρ_∞, since the order-independent extragradient floor also grows with the outlier.

A bilinear matrix game. Table 9 runs the finite-sum saddle point $\min_x \max_y \frac{1}{n}\sum_i (x^\top A_i y + c_i^\top x - e_i^\top y) + \frac{\lambda_i}{2}(\|x\|^2 - \|y\|^2)$ with dense random coupling A_i and heterogeneous regularization λ_i. The herding gain grows with n and SC-HERD beats SPEG by up to $3.2\times$, while SC-HERD holds within 4% of two-call herded SEG—the n-scaling and single-call results survive on a genuine min-max task, if more modestly than on the rotational instance.

Real-data finite-sum saddles, multiple datasets. Table 10 reruns the same saddle structure with the random Gaussian coupling replaced by a *real-data* rank-1 coupling $A_i = \phi_i \phi_i^\top$, where ϕ_i is the i-th sample's feature vector, PCA-reduced to $d = \min(8, \text{raw_dim})$ components and unit-normalized. We instantiate this on five sklearn datasets—`digits`, `wine`, `breast_cancer`, `diabetes`, `iris`—spanning raw dimensions 4–64 and sample counts 150–1797. All other choices match Table 9 ($\lambda_i \sim \text{Uniform}[0.15, 0.5]$, centered offsets, $\beta = 1/(nL)$, 24 fresh draws of n records per cell). The herding n-axis gain reproduces on every dataset: SEG-RR/herded climbs monotonically with n to $3.6\times$–$10.9\times$ at $n = 64$. The two datasets whose $n = 8$ ratio is below 1 (digits at 0.75, diabetes at 0.81) cross 1 already at $n = 16$—the small-n shortfall is consistent with the worst-case $\sqrt{n}/\kappa$ bound exceeding one only once $n \gtrsim \kappa^2$—the square of the condition number, not the condition number itself. SPEG-RR/SC-HERD shows the same growth pattern and SC-HERD again stays within a few percent of two-call herded SEG across all five datasets (Table 11). The herding mechanism is thus not an artifact of any one feature distribution—it reproduces across digit pixels, wine chemistry, breast-cancer measurements, diabetes readings, and iris geometry.

Composition with SEG-FFA*: Theorem 7 empirically.* Theorem 7 predicts that herding's discrepancy improvement carries through SEG-FFA: flip-flop cancellation kills the β^2 order-dependent bias of Theorem 1, and the surviving β^3 residual carries the same prefix sums P_j that herding controls. Table 12 tests this on the same five real-data saddles, with SEG-FFA implemented exactly as in Chae et al. (2024)'s released strongly-monotone code: forward sweep with stepsize pair $(\beta/2, \beta)$, backward sweep over the *reversed* permutation (same π, not a fresh shuffle), and equal-weight anchoring $z \leftarrow \frac{1}{2}(z_{\text{start}} + z_{\text{end}})$. The herded variant replaces the per-epoch random shuffle with our balancing-subroutine order, recording F_i during the forward sweep's extrapolation step. SEG-FFA-RR/SEG-FFA-HERD grows monotonically with n on every dataset, reaching $3.2\times$–$12.5\times$ at $n = 64$—essentially the same range as the SEG-RR/herded gain on top of vanilla SEG (Table 10: $3.6\times$–$10.9\times$), consistent with the predicted transfer of the discrepancy factor on top of FFA—proved in the worst case (Theorem 7), and here confirmed at the generic scale, where the theorem defers to experiment. Table 13 compares absolute plateau scales: vanilla SEG-RR sits at 2–5 $\times$ 10^{-2}, SEG-FFA shaves 3–7$\times$ off (the epoch-axis improvement), and SEG-FFA with herding shaves a further 3–13$\times$ (the component-axis improvement). The combined gain SEG-RR/SEG-FFA-HERD reaches 17–48$\times$ at $n = 64$—an empirical demonstration that the epoch-axis and component-axis improvements compose in practice; the compatibility statement of Section 4.5 certifies only that herding retains its worst-case discrepancy gain on top of SEG-FFA, not that the two factors multiply.

Order stabilization. Theorem 4's sharp parity assumes the herded order stabilizes; Remark 2 explains why we expect it but does not prove it, and Table 17 measures the order's *discrepancy* stability (its "D-stable" column is not exact permutation freezing). Measuring the permutations themselves (diagnostics below) shows the two are different: under the greedy rule the order freezes exactly in $20/12/11$ of 24 runs at $n = 16/64/128$ on the affine VI ($19/9/10$ on the bilinear game), so at $n \geq 64$ the median run is still changing its order at the last epoch, cycling among 2–3 herded orders with a median 9% wobble in ideal discrepancy; under the deployable recursive routine the order *never* freezes, because its fair per-node coin re-randomises the child concatenation every epoch. The stabilized-order regime of Theorem 4 is therefore not the typical one for the default routine, and the sharp parity is not covered by the theorem there. It holds nonetheless: two-call herded SEG replaying SC-HERD's own recomputed orders reproduces SC-HERD's plateau to 0.998 $[0.992, 1.000]$ across all 216 runs and both routines, independently of order turnover (Figure 5(b)); the wide scatter of the raw ratio (0.30–2.90) is order *selection*—the two runs herd different records and settle on different orders—not single-call error. The bounded-plateau clause covers the switched regime; the parity itself is here an empirical property of the switched dynamics.

Beyond two-player saddles: a multi-player game (Table 7). Ordering acts on the finite sum, not on saddle structure, and a multi-player game makes the distinction concrete. Five players with two-dimensional actions interact on an odd directed cycle: component i has pseudo-gradient $F_i(x) = (\lambda_i I + \gamma_i K)x + \eta_i \tanh(x) + b_i$, $\sum_i b_i = 0$, with K the skew-symmetric five-cycle coupling, so

every component is strongly monotone (checked: minimum symmetric-Jacobian eigenvalue 0.21) while the odd cycle leaves at least one internal skew edge under *every* bipartition of the players—no partition writes the game as a two-player saddle-gradient problem. Table 7 reports the deployable stale-record orders at $\beta = 0.5/n$, the common step at which all 600 runs are finite. The picture of Figure 1 repeats: the herding gain grows with n, reaching $7.5\times$ at $n = 128$ for greedy herding (log–log slope $+0.78$) and $3.2\times$ for the GD-D recursion ($+0.39$); single-call herding separates from SPEG-RR at the same rate ($8.7\times$, slope $+0.67$); and SC-HERD tracks two-call herded SEG at a flat ratio (slope $+0.02$)—reuse costs nothing here either. At the aggressive step $\beta = 1/n$ the separations widen ($11.2\times$ and $11.7\times$ at $n = 128$) but SC-HERD approaches divergence on 2 of 24 instances at $n = 8$ (where $\beta L \approx 0.7$), so we report the conservative step.

Table 7: Five-player monotone game on an odd interaction cycle (no two-player saddle-gradient representation under any player bipartition): plateau ratios at $\beta = 0.5/n$ with deployable stale-record orders, median [IQR] over 24 instances per n. "GD-D" is recursive bipartition with GD-D splits.

n	SEG-RR / greedy	SEG-RR / GD-D	SPEG-RR / SC-HERD	SC-HERD / greedy SEG
8	0.91 [0.73, 1.18]	1.10 [0.95, 1.23]	1.38 [1.17, 1.50]	0.96 [0.86, 0.99]
16	1.33 [1.11, 1.74]	1.30 [1.23, 1.40]	1.47 [1.30, 2.26]	0.96 [0.90, 1.00]
32	2.08 [1.76, 2.43]	1.62 [1.54, 1.69]	2.25 [1.96, 2.65]	0.94 [0.80, 1.09]
64	4.36 [3.75, 5.12]	2.28 [2.09, 2.33]	3.84 [3.38, 4.77]	1.07 [0.97, 1.18]
128	7.48 [6.58, 9.53]	3.21 [3.06, 3.35]	8.68 [7.21, 10.35]	0.96 [0.83, 1.05]

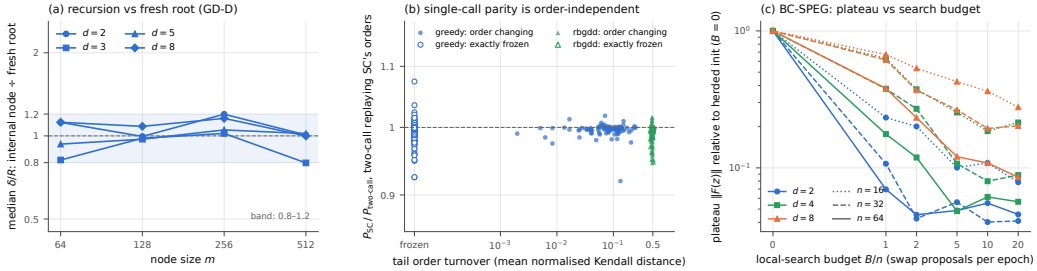

Figure 5: Direct diagnostics. (a) GD-D recursion vs fresh root: at matched node size m, the median split imbalance δ/R of an internal recursion node divided by that of a fresh root split of the same size (canonical primitive; 24 seeds; $d = 2, 3, 5, 8$); band 0.8–1.2. (b) SC-HERD single-call parity against tail order turnover, with two-call SEG *replaying* the single-call run's own per-epoch orders (affine VI, bilinear game, five-player game; $n \in \{16, 64, 128\}$; 24 seeds; greedy and recursive routines; the legend's "rbgdd" is the recursive GD-D routine); hollow markers are runs whose order froze exactly. (c) BC-SPEG plateau against local-search budget B (swap proposals per epoch) relative to the herded initialization $B = 0$ (generic families, 12 seeds).

Direct tests of the conditional hypotheses. *Recursive split accuracy.* We run the deployed GD-D primitive (Algorithm 2, copied with a sign output and checked to reproduce Theorem 2's implementation exactly) inside the full recursive bipartition on the centered generic model, $d \in \{2, 3, 5, 8\}$, $n \in \{64, \ldots, 1024\}$, 24 seeds, and log every node's size m_v and imbalance δ_v (284,256 nodes; imbalance zero at every node). Two facts. First, the recursion does not degrade the primitive: at matched node size the internal-node/fresh-root ratio of median δ/R lies in $[0.80, 1.20]$ for every d and $m \in \{64, \ldots, 512\}$ (Figure 5(a)), and at fixed size $\delta_v\sqrt{m_v}/R$ is insensitive to depth and to n. Second, the $m^{-1/2}$ law is dimension-dependent within the accessible range: the all-node slope of $\log(\delta_v/R)$ against $\log m_v$ is -0.63 $[-0.64, -0.62]$ at $d = 2$ (-0.49 over $m \geq 64$; the excess is the $O(R)$ floor of tiny nodes), -0.37 at $d = 3$, -0.14 at $d = 5$ and $+0.03$ at $d = 8$ (bootstrap over seeds). A root-only scan to $m = 65{,}536$ gives slopes -0.52, -1.07, -0.41, -0.17 for $d = 2, 3, 5, 8$: the regime is directly visible at $d = 2$, emerges at larger sizes at $d = 3$, approaches the target at $d = 5$, and is not visible at $d = 8$ over the tested range—consistent with the pre-asymptotic caveat of Theorem 2. The consecutive-pairing recursion (negative control) has slopes -0.02 to $+0.12$ and is worse than GD-D at the root by $22\times$ / $7.2\times$ / $4.3\times$ / $2.1\times$. Because ($\mathrm{H}_{\mathrm{split}}$) carries an unknown

$C_d\,\ell(n)$, none of this is a pass/fail test; what it rules out is the interpretation that the value-dependent recursion destroys the root guarantee.

Single-call order dynamics. For every epoch we record the permutation, its normalised Kendall distance to the previous one, the record discrepancy $D(\pi_k;\widetilde{r})$ and the ideal-offset discrepancy $D(\pi_k;b)$ (with z^* solved to 10^{-15}), on the affine VI, the bilinear game and the five-player game, $n\in\{16,64,128\}$, 24 seeds, greedy and recursive routines (864 runs). Table 17 reproduces exactly. Exact freezing is rarer than discrepancy stability (pooled $128/216$ vs $150/216$ under greedy) and never occurs under the recursive routine; $D(\pi;\widetilde{r})$ tracks $D(\pi;b)$ within 1% on the plateau. The order-matched parity is 0.998 $[0.992,1.000]$ for both routines, with no run outside $[0.92,1.08]$ and all $n\geq 64$ runs inside $[0.97,1.02]$, identical for frozen and changing runs; the raw ratio is uncorrelated with the ratio of the two runs' ideal discrepancies ($\rho=0.01$): orders of equal maximum-prefix discrepancy give plateaus differing by up to $3\times$.

BC-SPEG *maturity.* On the generic family generalised to $d\in\{2,4,8,16\}$, $n\in\{16,32,64\}$, 12 seeds, the per-component increment design matrix reaches numerical rank d (thresholds 10^{-6} and 10^{-8} on $\sigma_{\min}/\sigma_{\max}$ agree) at a median of d recorded increments—the earliest possible—in all 288 runs; the unregularised reconstruction of Lemma 32 then gives $\|\widehat{z}^*-z^*\|$ between $9\cdot10^{-15}$ ($d=2$) and $3\cdot10^{-9}$ ($d=16$) and $\widehat{\varepsilon}$ within $3\cdot10^{-11}$ relative of ε, and the certificate's leading term $2\beta\varepsilon/n$ is within $2\times$ of the plateau for $d\geq 4$. Two caveats: the running *ridge*-regularised Jacobian estimate is 70–99.9% wrong at onset for $d\geq 4$ (the ridge dominates the tiny early Gram), so the certificate must be computed from the unregularised reconstruction; and the theorem's stabilized-order hypothesis fails in many runs (median 30–446 order changes after warm-up). For the search budget (Figure 5(c)), $B\approx 5n$ proposals per epoch attain $\geq 95\%$ of the $20n$ improvement for $d\leq 4$, $n\geq 32$ (search 2–11 ms per epoch); at $d=8$ the plateau is still falling at $20n$ (-14% to -22% from $10n$ to $20n$), so BC-SPEG at $d\geq 8$ is proof-of-principle. Drawing fresh random proposals every epoch lowers the plateau a further 1.3–$3.3\times$ at equal budget—an across-epoch order-ensemble effect that Theorem 6 does not certify.

Wall-clock benchmark (Table 3). *Protocol.* Data preloaded on one A100 80GB; warm-up epochs before timing; `cuda.synchronize` around every timed region; identical precision, partition, initialization, step sizes and seeds across methods; per epoch we record operator time, ordering time (projection plus recursion), other time, total, and peak allocated memory, over five timed repeats (epoch-time IQR below 1.4%). Time-to-threshold is the cumulative training wall-clock to a common threshold $\tau=1.05\times$ the larger of the two herded methods' mean residuals over epochs $k>\lfloor 2E/3\rfloor$, where E is the run length; these tail means differ from the epoch-E residuals in Table 3. Evaluation is excluded from all timings. The herded orders are computed by the deployable recursive routine on a $k=16$-dimensional Gaussian random-projection sketch of the centered records, a practical device outside the theorems, timed separately.

covtype saddle (inside the model class, not the step range). covtype (Blackard & Dean, 1999) (581k×54, 7 classes) lifted by a fixed random-feature map (Rahimi & Recht, 2007) to $p\in\{2048,8192\}$; $n\in\{32,64,128\}$ fixed minibatches; multiclass logistic loss plus a random Gaussian coupling A_i ($q=64$) and $\lambda_w=\lambda_y=10^{-2}$ ($\kappa\approx 150$); closed-form operator; $\beta=\alpha=2/(nL)$, the largest common stable step. SC-HERD's epoch-time speedup over SEG-RR is $1.93\times/1.86\times/1.71\times$ at $p=8192$ and $1.74\times/1.55\times/1.22\times$ at $p=2048$ for $n=32/64/128$ (per-call times 1.5 ms down to 0.15 ms), with time-to-threshold speedups 1.7–$2.2\times$ and plateaus within 1–5% of two-call herded SEG; ordering costs $1.9/4.1/8.9$ ms per epoch independently of p (interpreter-bound), i.e. 4–28% of an epoch, and the measured speedups fit $2/(1+t_{\mathrm{ord}}/(n\,t_{\mathrm{call}}))$ with a break-even near 60–70 μs per call; the record table adds under 0.3% memory. SPEG-RR is 1.9–$2.0\times$ faster per epoch but its plateau is 1.4–$7.6\times$ worse and it never reaches τ, diverging for $\beta L\geq 0.125$. Herded SEG equals SEG-RR within 5% here: the sketch order's prefix discrepancy is indistinguishable from random (0.44–0.57 vs 0.51–0.54), because independent random A_i make the offsets near-orthogonal in $D\sim 10^4$–$6\cdot10^4$ dimensions and $\kappa\approx 150$ lies outside Theorem 1's range; SC-HERD's numbers are therefore conservative (it pays for ordering without benefit).

Antithetic herding under three step schedules. Table 18 gives the full schedule study behind Theorem 3: constant $\beta_0=1/(nL)$; *drops*, halving β four times over the run; *harmonic*, $\beta_k=\beta_0\cdot 50/(50+k)$. Metric: tail mean of $\|z\|$ over $K=3000$ passes, median over 8 instances; all columns from one run, so the herded and antithetic rows are directly comparable.

Table 8: Condition-number sweep: herding gain vs. $\kappa = L/\mu$ ($n = 32$, $\beta = 1/(nL)$; median [IQR] over 24 instances).

κ	SEG-RR / herded
4.5	4.99 [4.33, 6.53]
9.8	2.57 [2.07, 3.01]
29	1.41 [1.16, 2.09]
78	1.28 [0.95, 2.16]

Table 9: Finite-sum bilinear matrix game: herding n-scaling ($\beta = 1/(nL)$; median [IQR] over 24 instances).

n	SEG-RR / herded	SPEG-RR / SC-HERD
8	1.01 [0.87, 1.37]	1.56 [1.24, 1.83]
16	1.14 [0.95, 1.35]	1.33 [1.10, 1.57]
32	1.57 [1.41, 1.91]	1.81 [1.59, 2.18]
64	2.81 [2.36, 3.39]	3.17 [2.39, 3.63]

Table 10: Real-data bilinear saddle across five sklearn datasets: rank-1 coupling $A_i = \phi_i \phi_i^\top$ from PCA features ($d = \min(8, \text{raw_dim})$, unit-normalized), $\lambda_i \sim \text{Uniform}[0.15, 0.5]$, $\beta = 1/(nL)$. SEG-RR / herded plateau ratio, median [IQR] over 24 fresh samples of n records per cell.

dataset	d	$n = 8$	$n = 16$	$n = 32$	$n = 64$
`digits`	8	0.75 [0.67, 0.87]	1.11 [1.07, 1.22]	1.85 [1.64, 2.12]	3.58 [2.85, 3.88]
`wine`	8	1.41 [1.14, 1.83]	2.49 [2.07, 2.65]	3.53 [3.29, 4.00]	7.39 [6.34, 8.31]
`breast_cancer`	8	1.29 [1.12, 1.74]	2.28 [2.07, 2.62]	3.50 [3.11, 3.96]	6.65 [5.89, 7.74]
`diabetes`	8	0.81 [0.69, 0.92]	1.20 [1.06, 1.36]	1.74 [1.49, 2.00]	4.02 [3.34, 4.48]
`iris`	4	1.75 [1.38, 2.27]	2.71 [2.23, 3.12]	6.64 [4.52, 7.70]	10.91 [8.79, 15.30]

What separates the single-call methods: a 2×2 factorial. Table 19 varies the reuse rule (fresh second call / previous step's value / own record) against the order (random / herded), on three settings. Under a random order all three reuse rules coincide; the growth with n lives entirely in the order.

At $n = 128$ the three random-order plateaus agree to within 1% in every setting (e.g. 1.25e-2, 1.24e-2, 1.25e-2 for two-call SEG-RR, SPEG-RR and single-call-RR in the first row), while at a herded order component-aligned reuse is worth 1.3–$2.5\times$.

Bias-cancelling orders. (Certificate-onset and search-budget diagnostics: the diagnostics paragraph above and Figure 5(c).) Table 20 reports BC-SPEG against the herded methods at an equal number of operator calls (not equal wall-clock), with Jacobians learned from the trajectory ($\beta = 0.4/(nL)$, medians over 8 instances). At this step Theorem 6's cap $\beta \le \mu/(64L^2)$ holds iff $n \ge 25.6\kappa$. Since $\kappa \ge 1$ both $n = 16$ affine cells fail it; the aligned family ($M_i = \frac{1}{2} I + s_i J$, $J^\top J = I$) has $\kappa = \sqrt{5}$, so its $n = 32$ cell runs at $1.79\times$ the cap; the generic $n = 32$ cell's κ is not recorded. The certificate also needs a stabilized order and a completed reconstruction. The table is an empirical bias-cancellation comparison, not a validation of the certificate. The table reports the learned-Jacobian variant. On the generic family it agrees with the exact-Jacobian variant to four significant digits; on the aligned family its plateau is lower (1.04e-3 against 1.15e-3 at $n = 16$, 4.57e-4 against 7.78e-4 at $n = 32$).

The deployable default, end to end. Theorem 1's worst-case herding bound assumes an order carrying a discrepancy guarantee; Theorem 4's switched-stability clause covers arbitrary order sequences, while its sharp two-call comparison assumes a fixed or eventually stabilized order with bounded realized ideal-offset discrepancy. The orders differ by experiment: our affine n-sweep applies the greedy rule to current-iterate values, Table 2's right block compares three routines, Table 22 uses the GD-D recursion, and the timing benchmark a 16-dimensional sketch of the records. Table 22 closes the loop: two-call herded SEG and single-call SC-HERD run with the recursive-bipartition/GD-D order (the stated default of Algorithm 1) against the greedy heuristic, on the affine

Table 11: Single-call ratios on the same five datasets at $n = 64$: median [IQR] over 24 instances. SC-HERD stays within a few percent of two-call herded SEG on every dataset, while SPEG-RR/SC-HERD grows from 3.3× to 11.7×.

dataset	SPEG-RR / SC-HERD	SC-HERD / herded
digits	3.34 [3.03, 4.01]	0.98 [0.93, 1.06]
wine	6.87 [6.14, 8.69]	1.00 [1.00, 1.02]
breast_cancer	6.53 [6.00, 7.54]	1.00 [1.00, 1.08]
diabetes	3.57 [3.44, 4.02]	1.01 [0.90, 1.25]
iris	11.71 [9.35, 14.17]	1.00 [0.83, 1.20]

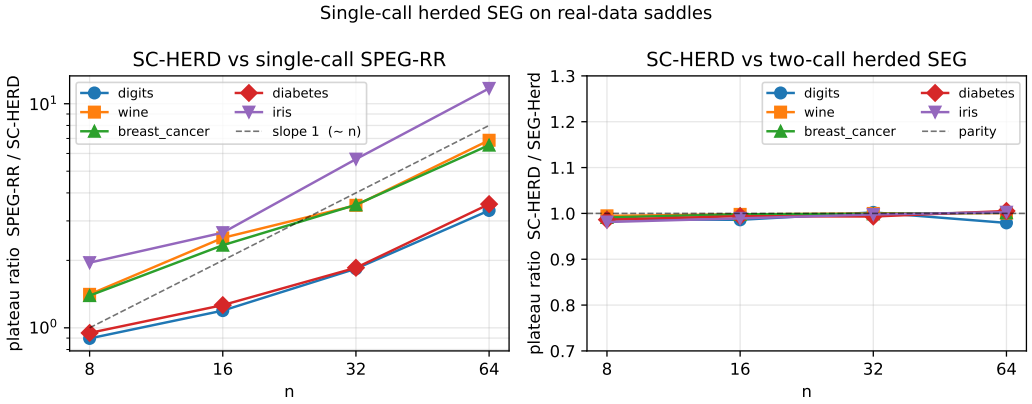

Figure 6: Single-call herded SEG on real-data saddles. Left: SPEG-RR / SC-HERD grows roughly linearly with n on every dataset, matching Figure 1(c)'s synthetic result. Right: SC-HERD / two-call herded SEG sits near 1 across all five datasets and all n. Theorem 4 covers this only for a fixed or eventually stabilized order and under its ideal-offset discrepancy hypothesis; the runs recompute the order each epoch, so the agreement is evidence for the parity rather than an instance of the theorem.

Table 12: SEG-FFA composition (Theorem 7 empirically): herding gain on top of flip-flop+anchoring. Ratio = SEG-FFA-RR plateau / SEG-FFA-HERD plateau, same five datasets as Table 10; FFA implemented per Chae et al. (2024)'s released code (forward sweep with $(\beta/2, \beta)$, reversed-permutation backward sweep, equal-weight anchoring); $\beta = 1/(nL)$, 400 outer epochs ($\equiv$ 800 passes); median [IQR] over 24 instances.

dataset	d	$n = 8$	$n = 16$	$n = 32$	$n = 64$
digits	8	0.68 [0.60, 0.80]	1.07 [0.91, 1.18]	1.65 [1.52, 1.82]	3.22 [2.78, 3.65]
wine	8	1.21 [0.87, 1.76]	2.30 [1.45, 2.74]	3.15 [2.40, 3.65]	5.77 [4.00, 7.16]
breast_cancer	8	1.22 [0.94, 1.67]	2.10 [1.54, 2.45]	2.82 [2.15, 4.15]	4.77 [3.83, 6.62]
diabetes	8	0.73 [0.61, 0.87]	1.09 [0.94, 1.45]	1.57 [1.32, 1.85]	3.30 [2.92, 4.02]
iris	4	1.50 [1.06, 2.39]	2.52 [1.53, 4.23]	5.72 [3.43, 9.84]	12.51 [9.97, 17.48]

family at $\beta = 1/(nL)$. Empirically, this deployed recursion retains the full n-growing gain (SEG-RR/GD-D-order $= 1.5 \to 8.2$ over $n = 16 \to 128$) and sits within 1.1–1.6× of greedy. Analysis and implementation thus agree at the *split-primitive* level: the code runs exactly the single-split GD-D primitive that Theorem 2 analyzes. An end-to-end *order-level* guarantee is a different matter—it would require (H_{split}) at every recursive node, which we verify empirically rather than prove—and greedy is an optional heuristic upgrade.

The failed repair variant, measured; and the fix. Appendix F's design rule — never pair across cells — came from testing the obvious alternative (GD-PR, parity repair along a walk) and watching its induction fail. Table 23 records the failure: the two columns a parity-mixing argument would need are the per-cell occupancy of the next round's grid and the longest run of non-odd cells, and both are catastrophically off at $t = 2$ (runs of 88–203 against the ≈ 10 a logarithmic law would give),

Table 13: Absolute plateau scales at $n = 64$ (median over 24 instances): vanilla SEG-RR, SEG-FFA, and SEG-FFA+herded. The rightmost column is the total combined gain over vanilla SEG-RR (epoch-axis factor from FFA times component-axis factor from herding).

dataset	SEG-RR	SEG-FFA-RR	SEG-FFA-Herd	SEG-RR / SEG-FFA-Herd
`digits`	1.92×10^{-2}	2.95×10^{-3}	9.10×10^{-4}	21.1×
`wine`	4.91×10^{-2}	1.47×10^{-2}	2.66×10^{-3}	18.5×
`breast_cancer`	4.76×10^{-2}	1.35×10^{-2}	2.88×10^{-3}	16.6×
`diabetes`	2.03×10^{-2}	3.25×10^{-3}	9.42×10^{-4}	21.5×
`iris`	3.73×10^{-2}	9.95×10^{-3}	7.82×10^{-4}	47.7×

Table 14: Ordering routines, full version of Table 2's right block, with inter-quartile ranges ($\beta = 1/(nL)$, 24 instances).

n	greedy	recursive	walk
8	1.14 [0.96, 1.33]	1.34 [1.15, 1.49]	1.18 [1.11, 1.31]
16	1.84 [1.42, 2.79]	1.57 [1.47, 1.74]	1.48 [1.37, 1.52]
32	3.01 [2.44, 3.55]	2.22 [2.15, 2.49]	1.97 [1.84, 2.07]
64	5.07 [4.09, 6.41]	3.33 [2.85, 3.77]	2.61 [2.49, 2.76]
128	10.81 [9.71, 15.69]	4.55 [4.25, 4.77]	3.80 [3.66, 3.95]

Table 15: Recursive bipartition, consecutive vs. geometric short-edge pairing, on bounded-domain generic instances (uniform offsets in $\mathbb{R}^2$, centered—the $d = 2$ case of Proposition 1's model). Entries: SEG-RR plateau over the herded plateau ($\beta = 1/(nL)$; median [IQR] over 24 instances).

n	consecutive	geometric
8	1.39 [1.14, 1.58]	1.58 [1.22, 1.67]
16	1.94 [1.66, 2.18]	2.22 [1.92, 2.98]
32	2.85 [2.57, 3.38]	4.14 [3.44, 5.14]
64	4.25 [3.74, 4.70]	7.28 [6.12, 8.55]
128	6.37 [6.00, 7.57]	13.92 [11.46, 15.51]

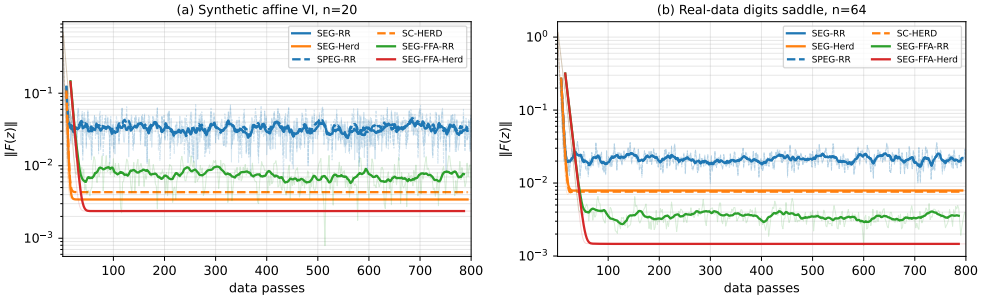

Figure 7: Convergence behavior in both regimes; same six algorithms in each panel, x-axis is the number of data passes (one SEG-FFA outer iteration is two passes). (a) Synthetic heterogeneous affine VI with $n = 20$. (b) Real-data digits saddle ($n = 64$, PCA $d = 8$). In both regimes herded SEG reaches a much lower plateau than SEG-RR, SC-HERD closely tracks two-call herded SEG, and the SEG-FFA composition opens two further plateau levels, with SEG-FFA+herded reaching the lowest—a separation larger than the worst-case $\sqrt{n}$-scale one Theorem 7 certifies.

for the reason the table's left half shows — a quarter of all pairs are cross-cell repairs reaching 8–22× the cell scale. Larger occupancy restores the finite-m numbers but not the asymptotics. The deferral routine GD-D removes the mechanism: its round-2 support over the cell scale is 1.28–1.52 against $\sqrt{d} = 1.41$–1.73 (no log factor), the signing residual tracks the current scale (median ratio 1.0–1.3, max 2.4–5.6 at $d = 3, 5$), and its measured splitting exponents are $\gamma = +0.44, +1.02, +0.49, +0.16$

Table 16: Heterogeneity-outlier sweep: one component's offset scaled by c, so the largest component norm ρ_∞ separates from the RMS scale R ($n = 64$ affine VI, $\beta = 1/(nL)$; median [IQR] over 24 instances). The herding gain degrades as a single component dominates—the failure mode of Theorem 1's general form $\mathcal{P}_{\mathrm{order}}(\textsc{Herd}) = \widetilde{O}(\beta L \rho_\infty/\mu)$.

c	ρ_∞/R	SEG-RR / herded
1	1.87	5.62 [4.25, 6.30]
2	2.02	5.39 [4.22, 7.42]
4	3.24	4.58 [3.74, 6.01]
8	5.29	3.66 [3.06, 5.86]
16	6.93	2.93 [2.45, 3.26]

Table 17: SC-HERD order stabilization ($\beta = 1/(nL)$; 24 instances). *D settled*: median [IQR] last epoch the order's prefix discrepancy is $> 5\%$ off its tail median. *D-stable*: runs whose last-20% discrepancy spread is below 2%—a statement about the order's *discrepancy*, not about the permutation itself freezing. *tail spread*: median of that spread.

setting	D settled	D-stable	tail spread
affine n=16	6 [4, 11]	23/24	0.0%
affine n=64	23 [20, 32]	16/24	0.0%
game n=16	19 [16, 30]	19/24	0.0%
game n=64	63 [45, 999]	10/24	6.4%

Table 18: Antithetic vs. fixed herded orders under step schedules. Entries are plateau ratios over SEG-RR (larger is better; < 1 means the ordered method *loses* to random reshuffling). The FFA columns apply the same alternation to SEG-FFA's forward order.

		plain SEG		SEG-FFA	
n	schedule	herded	antithetic	herded	antithetic
16	constant	1.61	**2.65**	1.98	**3.16**
16	drops	0.73	**1.28**	0.75	**1.51**
16	harmonic	0.56	**0.92**	0.42	**0.83**
64	constant	4.39	**8.05**	7.73	**8.41**
64	drops	1.94	**3.09**	2.48	**3.45**
64	harmonic	0.84	**1.31**	1.00	**1.11**

at $d = 2, 3, 5, 8$ against the target $+0.5$ (the $d = 8$ range $m \le 65{,}536$ is pre-asymptotic: at accessible m the cell side is comparable to the inner core, so most mass defers after two rounds and the certified variant trails both comparators there) — against GD-PR's $+0.90, +0.76, +0.39, +0.37$: at $d = 2$ the repaired grid is competitive, and at $d = 3, 5$, where its round-2 log-stretch bites, GD-D wins, at lower cost (no walk). These numbers are for the certified inner-core implementation of Algorithm 2 with the two-layer K_j/Q_j structure the proof defines—the variant the analysis covers; recursing *all* cells (no core restriction) is an uncertified heuristic upgrade in the same sense that greedy ordering is, and measures $+0.62, +0.69, +0.59, +0.39$ on the same harness (with a single initial shuffle; seed-level differences only).

Two further measurements. At occupancy $t = 2$ the splitting rates of GD-PR are $\gamma = +1.07, +0.71, +0.46$ at $d = 2, 3, 5$ (target $+0.5$), so the imbalance rate that Theorem 2 proves for GD-D is also observed for this uncertified variant; and the promotion variant GD-PR/H, which forces every difference to be an in-cell difference at its own scale and would make the induction go through level by level, measures *worse* ($\gamma = +1.01, +0.65, +0.32$), so it is not a free repair.

Does the analysed routine also win in practice? Four families. Table 24 runs the three deployable recursive orderings end-to-end on four families chosen to stress Theorem 2's hypotheses in turn: offsets uniform on a box in $\mathbb{R}^4$ (the reference case, density condition number $\kappa_f = 1$); a 3:1 mixture of uniforms over the two halves of the box ($\kappa_f = 3$, still within the theorem); uniform on a

Table 19: Reuse rule × order. Left: plateaus at $n = 128$ (median, 16 instances). Right: fitted slope p in (ratio) $\sim n^p$ over $n \in \{8, \ldots, 128\}$ — $p \approx 1$ means the ratio grows like n, $p \approx 0$ that it is constant.

	plateau at $n = 128$			slope p		
setting	SPEG-RR	SPEG-HERD	SC-HERD	$\frac{\text{SPEG-RR}}{\text{SC-HERD}}$	$\frac{\text{SPEG-HERD}}{\text{SC-HERD}}$	$\frac{\text{SC-RR}}{\text{SC-HERD}}$
affine, $\beta = 1/n$	1.24e-2	1.40e-3	7.11e-4	$+0.86$	$+0.08$	$+0.95$
affine, $\beta = 1/(nL)$	2.90e-3	5.64e-4	2.91e-4	$+0.82$	$+0.13$	$+0.89$
nonlinear, $\beta = 1/n$	1.23e-2	1.69e-3	8.26e-4	$+0.79$	-0.01	$+0.86$

Table 20: BC-SPEG plateaus. "gain" is SC-HERD over BC-SPEG. The aligned family is the rotational two-group construction of Section 4.4; the nonlinear family adds $\gamma_i \tanh(z)$, where the learner tracks the local Jacobian.

family	n	SEG-HERD	SC-HERD	BC-SPEG	gain
generic	16	3.83e-3	3.79e-3	1.40e-3	2.7×
generic	32	1.08e-3	1.13e-3	1.82e-4	6.2×
aligned	16	1.25e-2	1.21e-2	1.04e-3	11.7×
aligned	32	6.24e-3	6.13e-3	4.57e-4	13.4×
nonlinear	16	3.35e-3	3.09e-3	1.39e-3	2.2×
nonlinear	32	8.35e-4	8.75e-4	2.96e-4	3.0×

Table 21: The central prediction, on the affine VI at $\beta = 1/n$: the ordering gain grows with the component count. Columns: prefix-discrepancy ratio random/herded; plain-SEG plateau ratio; the same on top of SEG-FFA (the setting of Theorem 7, which certifies a worst-case $\sqrt{n}$-scale separation there); and the SPEG-RR/SC-HERD comparison, whose n-growth comes from the variance transfer of Appendix E.1 rather than from Theorem 4. Medians over 8 instances; full tables and the corresponding figures in this appendix.

n	D ratio	SEG-RR/HERD	SEG-FFA-RR/FFA-HERD	SPEG-RR/SC-HERD
8	1.68	1.09	1.62	1.96
16	1.94	1.95	2.20	2.18
32	2.86	3.88	4.08	3.37
64	3.65	7.18	8.97	8.02
128	4.74	15.42	27.40	17.97
256	6.50	32.61	62.68	—

box in $\mathbb{R}^8$, which doubles the round count $\lceil d/2 \rceil$; and offsets taken from the principal components of `sklearn`'s `digits`, a real distribution satisfying no density hypothesis at all.

Fitted slopes in n are $+0.54/ + 0.51/ + 0.58$ (uniform $\mathbb{R}^4$), $+0.54/ + 0.46/ + 0.45$ (mixture), $+0.49/ + 0.49/ + 0.57$ (uniform $\mathbb{R}^8$) and $+0.61/ + 0.47/ + 0.53$ (`digits`) for consecutive/kd/GD-PR respectively. Two readings. First, the n-gain is a property of low prefix discrepancy and not of any one pairing rule, consistent with Table 14. Second, the grid family's theoretical advantage—GD-D's per-split *guarantee* (Theorem 2); GD-PR is the empirical comparator—is not empirical dominance: where the offset density is far from uniform the grid's occupancy varies, its parity repair travels farther, and a cruder pairing can do better in practice. We report this rather than select the family that flatters the routine.

Outside strong monotonicity: a weak-Minty probe. Theorem 1 assumes strong monotonicity. Table 25 probes what survives without it, on an affine family whose *average* operator has a negative symmetric part and strong rotation (weak-Minty, solvable by extragradient+ (Diakonikolas et al., 2021) with extrapolation $4\times$ the update step); `spread` is the dispersion of the components' monotonicity, so larger values mean more anti-monotone components.

Deep validation of the transported bias (Figure 2). *Workload.* Adversarial training (Madry et al., 2019) of a ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky & Hinton, 2009) written as

Table 22: Ordering-routine closure (medians over 8 instances). Plateaus for two-call herded SEG and single-call SC-HERD under the GD-D-split recursive-bipartition order versus the greedy heuristic.

n	call	SEG-RR	greedy	GD-D-order	SEG-RR/GD-D
16	two-call	1.30e-2	9.51e-3	8.62e-3	1.51
16	single-call	1.28e-2	7.89e-3	8.42e-3	1.52
64	two-call	5.73e-3	1.30e-3	1.44e-3	3.99
64	single-call	5.73e-3	1.23e-3	1.53e-3	3.74
128	two-call	2.88e-3	2.24e-4	3.50e-4	8.20
128	single-call	2.87e-3	2.69e-4	3.54e-4	8.12

Table 23: Why cross-cell repair is forbidden (uniform i.i.d. points, medians over 4–6 draws). Left: the repair minority produced by one GD-PR round. Right: round-2 parity health under GD-PR, longest non-odd run over $\log_2(\#\text{cells})$ — a parity-mixing law would make this $O(1)$. GD-D (Theorem 2) avoids the mechanism entirely by deferring leftovers.

d	m	repair share	reach / s	d	t	empty cells	run / $\log_2$
2	4096	25.0%	10.1	2	2	56%	5.3
2	16384	24.6%	14.6	2	8	46%	2.3
2	65536	24.5%	29.7	2	32	44%	1.5
3	65536	24.5%	12.9	3	2	38%	1.7
5	65536	24.6%	7.9	3	32	29%	1.0

Table 24: SEG-RR plateau over the ordered method's plateau (larger is better), median over 8 instances, $\beta = 1/(nL)$. **Bold** marks the strongest routine per family at $n = 256$. GD-PR wins on both uniform families but not on the non-uniform mixture or on real features; every routine retains the n-growth.

family	routine	n=32	n=64	n=128	n=256
uniform, $\mathbb{R}^4$	consecutive	2.03	2.94	4.57	6.06
	kd-median	1.79	2.76	4.08	5.15
	GD-PR	2.22	3.38	4.93	**7.44**
mixture, $\mathbb{R}^4$	consecutive	2.02	3.17	4.11	**6.45**
	kd-median	1.66	2.08	3.04	4.22
	GD-PR	1.53	2.11	2.77	3.95
uniform, $\mathbb{R}^8$	consecutive	1.76	2.52	3.47	4.85
	kd-median	1.75	2.55	4.02	4.69
	GD-PR	1.82	2.71	4.34	**5.83**
`digits` PCA	consecutive	2.09	2.90	4.50	**7.37**
	kd-median	1.95	2.80	3.83	5.23
	GD-PR	2.16	2.97	4.77	6.32

a finite-sum saddle: $z = (w, r_1, \ldots, r_n)$ with the network weights w shared and one adversarial-perturbation block r_i per component, $F_i = (\nabla_w \mathcal{L}_i + \lambda w,\ -\nabla_{r_i} \mathcal{L}_i + \gamma r_i)$ on $n = 128$ i.i.d. blocks of 390 images ($\lambda = 5 \cdot 10^{-4}$, $\gamma = 1$, $\|r_i\|_\infty \le 8/255$), plain two-call SEG with $\alpha = \beta$. Two checkpoints z_8 and z_{11} (epochs 8 and 11 of a burn-in run) are frozen; every arm restores the identical state bit for bit, and a repeated order reproduces its outcome to 0.000. The outcome is the shared-block residual $Y(\pi) = \left\|\bar{F}_w(z^\pi_{\text{end}})\right\|$ after one epoch; the prefix discrepancy $D(\pi) = \max_t \left\|\sum_{s \le t} u_{\pi_s}\right\|$ uses the centred records $u_i = F_i(z_0) - \bar{F}$ at the checkpoint. *The scalar.* By Lemma 1, with uncentred prefixes $\tilde{P}_{t-1} = \sum_{s<t} F_{\pi_s}(z_0)$, the order-dependent displacement is $\beta^2 X(\pi)$, so to leading order $\delta Y^2 = 2\beta^2 \langle g, X(\pi)\rangle$ with $g = (\partial \bar{F}_w/\partial z)^\top \bar{F}_w$. Writing $v_i = M_i^\top g$ (one vector–Jacobian product per component; exact automatic differentiation in `float32`, with `float64` cross-checks agreeing to 10^{-6}) and $A_{ij} = \langle v_i, F_j(z_0)\rangle$, the scalar is $s(\pi) = \sum_{r<t} A_{\pi_t \pi_r}$; only the antisymmetric part $A_{ji} - A_{ij}$ depends on the order. It is frozen to disk before any outcome

Table 25: Weak-Minty probe ($n = 32$, 12 seeds, 1500 epochs). Median plateau $\|F(z)\|$ with the number of diverged seeds in parentheses; anchoring is the epoch-wise average against the epoch's starting point. Herding does not change the divergence boundary, but lowers the anchored plateau by a stable $\approx 4\times$.

spread	RR	herded	RR+anchor	herded+anchor	gain (anchored)
0.00	3.61e-1 (0)	4.08e-2 (0)	1.33e-1 (0)	3.18e-2 (0)	4.18×
0.25	4.18e-1 (0)	5.15e-2 (0)	1.36e-1 (0)	3.40e-2 (0)	3.98×
0.50	4.84e-1 (5)	7.22e-2 (4)	1.44e-1 (0)	3.75e-2 (0)	3.84×
1.00	2.03e-1 (11)	2.12e-2 (11)	1.30e-1 (3)	3.20e-2 (3)	4.05×

exists, with no re-anchoring of g and no alternative scalar. *(A) Expansion validity.* For two orders, the predicted difference $\beta^2[X^w(\pi_0) - X^w(\pi_1)]$ (Jacobian–vector products) is compared with the actual $z_{\mathrm{end}}(\pi_0) - z_{\mathrm{end}}(\pi_1)$ from two `float64` epochs. At z_8, with the diagnostic directional scale $L_g := \mathrm{median}_i \left\|M_i^\top g\right\| / \|g\|$ (a median along the outcome direction, not the Lipschitz constant L of the theorems), $\beta n L_g = 0.37,\ 0.74,\ 1.11$ give cosine $0.989,\ 0.990,\ 0.985$ and norm ratio $1.17,\ 1.33,\ 1.52$—the ratio grows approximately linearly in $\beta n L_g$, consistent with the relative scaling expected from the cubic tail in Lemma 1; at z_{11}, $\beta = 2 \cdot 10^{-4}$ ($\beta n L_g = 1.04$) gives cosine 0.984 and ratio 1.42. At $\beta = 10^{-2}$ ($\beta n L_g \approx 37$) the same check gives cosine 0.25 and ratio 42: there the second-order epoch truncation no longer describes the realized dynamics. The validation therefore runs at $\beta = 2 \cdot 10^{-4}$, an empirically validated perturbative regime. *(B) Detectability.* At $\beta = 2 \cdot 10^{-4}$ the cross-order spread of Y over eight `float64` orders is $7.9 \cdot 10^{-4}$ (on $Y \approx 0.198$), 10^8 above the numerical floor; `float32` epochs reproduce a $6 \cdot 10^{-5}$ cross-order difference to 0.2% when paired against `float64`, so the 200-order runs use `float32`. The leading-order predicted spread $\beta^2 \,\mathrm{sd}_\pi(s)/Y$ overstates the measured one by $3.4\times$: the outcome linearisation is taken at the anchor while Y itself moves by 29% within the epoch. A common scale mismatch would not affect a rank statistic; the predictive ranking itself is tested directly in (C). *(C) Predictive replication.* 200 fresh permutations per checkpoint, disjoint from every order used anywhere else; predictors frozen before the first epoch. Spearman $\rho(s, Y \mid D) = 0.973\ [0.960, 0.981]$ at z_8 and $0.975\ [0.963, 0.981]$ at z_{11} (percentile bootstrap, 10^4 draws); a permutation null (shuffling s across orders, 10^4 draws, $q_{95} = 0.12$) gives $p \le 1/10001$ at both. The discrepancy carries no rank information about the same outcomes, $\rho(D, Y) = -0.000\ [-0.14, 0.14]$ and $0.004\ [-0.15, 0.16]$, and is uncorrelated with s. Everything at z_{11} is recomputed at z_{11}; nothing is transported between checkpoints. *Mechanism (replicated; not a headline claim).* At both checkpoints A is numerically near rank one ($\sigma_1/\sigma_2 = 65$ and 72; the top singular pair carries 0.9996 of $\|A\|_F^2$), with a leading left factor nearly constant across components (coefficient of variation 0.02): the pull-backs $M_i^\top g$ are nearly one common vector $\bar{v} \approx \bar{M}^\top g$, so $s(\pi) \approx \langle \bar{v}, \sum_t \tilde{P}_{t-1} \rangle = \sum_r (n - 1 - r)\, b_{\pi_r}$ with $b_j = \langle \bar{v}, F_j \rangle$—the mean-Jacobian, outer-sum part of the cross-term decomposition of Appendix C. By the rearrangement inequality, sorting the scores b_j in the order opposite to the positional weights minimises the leading-order outcome change. D is blind to this term twice over: it is isotropic where the outcome reads one direction, and it takes a maximum where the outcome takes a position-weighted sum. *Scope.* One epoch from a frozen state, plain SEG; no end-to-end training claim, and nothing about adaptive optimizers. At $\beta = 10^{-2}$ the runs sit far outside the perturbative regime, and no prefix functional predicts the one-epoch outcome; the results above are obtained in the empirically validated perturbative regime described in (A). The directional diagnostic $\beta n L_g$ is a measured scale along the outcome direction; it does not certify Lemma 1's Lipschitz-based step condition, whose constant L is not the same object.

O Reproducibility Details

The evidence comes from three layers with different requirements. *(i) Theory and synthetic/real-data checks (CPU).* The solvers, orderings and problem generators are implemented in NumPy alone, single-threaded, no GPU; scikit-learn appears only to load the five public datasets and to take their principal components, matplotlib only in the figure builders, and networkx only for the reference minimum-weight matching in the appendix comparison of splitting primitives. *(ii) Wall-clock measurement (GPU).* Table 3 and its appendix cells come from a PyTorch benchmark on a single NVIDIA A100 80GB (torch 2.10.0+cu128, CUDA 12.8, `float32`), on the covtype saddle

with the random-feature operator held in device memory, 400 epochs after a 20-epoch warm-up and five timing repeats at seed 0; medians and inter-quartile ranges are over those repeats. *(iii) Deep-network mechanism (GPU).* Figure 2 is produced by a separate three-stage pipeline (prepare orders and freeze predictors on CPU; run 200 one-epoch arms per checkpoint; analyse) on the same A100 under a strict deterministic protocol, from frozen checkpoints that are identified by hash and are not re-trained. The theory statements are self-contained in Appendices A–I and K–M, and the numerical checks of the expansions and constants (including the exact enumeration over all 40320 orders at $n = 8$ used in Section 4.4, and the exhaustive-permutation check of the covariance identities behind Lemma 7) are reproduced by the exact-machinery and lemma-verification scripts. The saved outputs of the real-data runs, `realdata_full.npz`, reproduce the twenty printed real-data entries in Table 2 by its median-of-paired-ratios reduction and will be part of the code release.

Scripts behind Figure 1 and Tables 1–2. The values below were regenerated by running the listed scripts of the code release (single-threaded, Python 3.12, NumPy 2.5) at the seeds each script fixes. Figure 1 and Tables 4, 5 and 21: `pg_nscale.py` (affine; seeds 0–7), `pg_nlvi.py` (nonlinear; 0–23), `pg_single2.py` and `pg_single_nl.py` (single-call; 0–7); the figure is drawn from these outputs by `make_compact_figures.py`. Table 1: `pg_antithetic.py` (with Table 18), `pg_scffa.py`, and `pg_bcspeg.py` with `pg_bcnl.py` (with Table 20), all at seeds 0–7; the first two print medians to three digits, and the cancel and reuse ratios of Table 1 and the ratios of Table 18 are formed from the full-precision medians by `regen/antithetic_ratios.py` and `regen/scffa_ratios.py`. The right block of Table 2 and Table 14: `pg_order.py`, seeds 0–23. Also regenerated: Table 19 (`pg_specomp.py`; 0–15) and Figure 4(a) (`pg_adv.py` for the generic curve and coefficients of variation, `pg_adv_ties.py` for the adversarial one; 0–23).