THE COMPOSITION GAP IN DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation compresses a training set into a small synthetic set, usually evaluated one at a time. In federated and data-governance settings, several parties distill their own data and a user trains on their union. We ask whether the union of separately distilled sets reproduces training on the union of the real data (*composability*) and show that it can fail even when every source is distilled exactly and the total budget admits an exact joint distillate. Compressing a training trajectory into fewer steps transforms the source statistics nonlinearly, so averaging compressed sources differs from compressing their average. For quadratic objectives we derive the exact composition error for two-to-one step compression in terms of the source-Hessian variance and the linear terms of the losses, and on a smooth network at small step sizes this prediction captures the local endpoint discrepancy in magnitude and direction. For learned synthetic sets, however, the composed error decomposes exactly into this local discrepancy and an aggregate source residual. Under endpoint matching the residual exceeds the structural term by more than an order of magnitude, and under distribution matching the two terms partly cancel. Joint distillation also retains an accuracy advantage when both sets are distilled from the same dataset, where the local discrepancy is exactly zero. Training fidelity and downstream accuracy are therefore distinct requirements, neither established by evaluating each set on its own. Code will be released.

1 INTRODUCTION

Dataset distillation (DD), i.e., the compression of a large training set into a small synthetic set that supports effective model training (Wang et al., 2018; Zhao et al., 2021; Cazenavette et al., 2022), is usually evaluated one distillate at a time. In a distributed application, several data holders instead release independently distilled sets, and a downstream user combines them for training, as in a consortium of hospitals that share only distilled versions of their images (Li et al., 2022; Holland et al., 2024). The union of the synthetic sets is meant to stand in for the union of the real data. This raises a natural question: *Can independently distilled datasets be combined without loss of training fidelity?* Evaluating each set separately does not answer it (Figure 1).

We answer this question through the training maps that gradient descent induces. A source distillate is faithful when a prescribed number of synthetic updates reproduces the endpoint of training on its real source, and composability asks whether the same fidelity holds after the sources and their distillates are combined with matching weights. Our central result is that exact source-wise fidelity does not guarantee composability, even when a joint solution exists within the same total budget.

The discrepancy arises from the order of compression and aggregation. Specifically, for quadratic regression the sufficient statistics defining the loss combine linearly across sources (Izzo & Zou, 2023; Gupta et al., 2026). Matching two gradient steps with one synthetic step instead preserves a nonlinear transformation of the source curvature. Distilling sources separately applies this transformation before averaging, and distilling their union applies it after averaging (Figure 1), so the two orders can disagree even when the sources share a minimizer.

For two-to-one compression of quadratic objectives, we derive an exact endpoint discrepancy that scales with the squared step size (Theorem 4.2). It depends on the source-Hessian variance, which captures differences in both eigenvalues and eigenvectors, and on an affine term coupling the Hessians with the linear loss coefficients.

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.

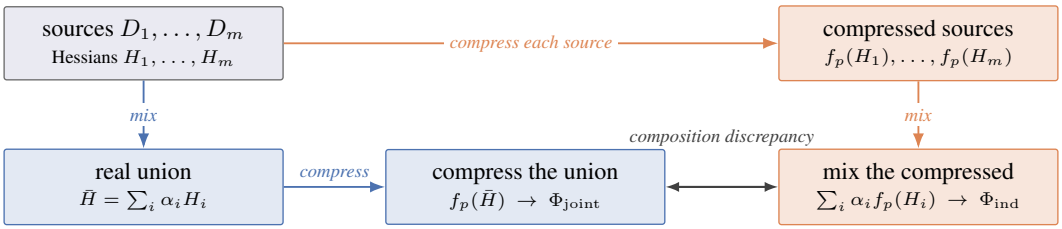

Figure 1: Compression and mixing do not commute. Distilling each source and then mixing the distillates (orange) applies the finite-step compression map f_p to each source Hessian before averaging. Distilling the union (blue) averages first and compresses once. The two paths can yield different training maps, and their difference is the composition discrepancy. The diagram is a schematic of exact quadratic compression at a general ratio p for sources with a shared optimum; the local endpoint discrepancy $\Delta\Phi$ measured in Section 6 is its two-step instance on real data.

With distribution matching on CIFAR-10, SVHN and CIFAR-100 (Zhao & Bilen, 2023), the excess endpoint error of the union is near zero under IID partitioning and positive under label skew, whereas the accuracy gain of joint distillation persists at every level of skew and even when both distillates are learned from the same dataset, so the curvature mechanism characterizes the training map and not the accuracy gain. Our contributions are summarized as follows:

- We show that exact source-wise fidelity does not guarantee composability, even when the total budget admits an exact joint distillate. For quadratic objectives the composition error has an exact form in the source-Hessian variance and an affine term that allows noncommuting source Hessians (Theorem 4.2), with a zero-error condition, an extension to other compression ratios, a certificate under approximate distillation (Corollary 4.3) and a local statistic P_{comp} computable on a network from source gradients and Hessian-vector products.
- We decompose the observed error of a union of distillates exactly into the local endpoint discrepancy of the real sources and the aggregate source residual, and measure both on the network. The quadratic prediction is accurate at small step sizes, but for learned synthetic sets the source residual is the larger term: under endpoint matching it exceeds the structural term by more than an order of magnitude, and under distribution matching it is anti-aligned with the local discrepancy under label skew and partly cancels it.
- We show that training fidelity and downstream accuracy are distinct requirements: the accuracy gain of joint distillation persists for two distillates of the same dataset, where the structural term is exactly zero, and reweighting the constituents after the fact narrows the excess endpoint error, removes at most about half of the squared composed error and gives no detectable accuracy change.

2 Related Work

Additional related work can be found in Appendix A.

Dataset distillation. DD constructs synthetic sets by unrolled training (Wang et al., 2018), gradient matching (Zhao et al., 2021; Zhao & Bilen, 2021), distribution matching (Zhao & Bilen, 2023), kernel-ridge solutions (Nguyen et al., 2021a) or trajectory matching (Cazenavette et al., 2022), among many refinements (Lei & Tao, 2024). Theory characterizes what a distillate preserves: a solution at convergence (Izzo & Zou, 2023; Nguyen et al., 2021a), a loss over random regressors (Gupta et al., 2026), or a spectrum of the feature correlation matrix (Bo et al., 2026).

Distilled data across sources. Federated DD trains on synthetic data transmitted by clients (Zhou et al., 2020; Goetz & Tewari, 2020; Hu et al., 2022; Xiong et al., 2023; Liu et al., 2023a; Wang et al., 2024; Holland et al., 2024; Arazzi et al., 2025; Shi et al., 2026). We characterize the post-hoc composition of such sets under finite-step training. The closest analysis is the drift of local updates in federated optimization (Li et al., 2020), whose endpoint-averaging identity our one-step case

recovers, and poor composition of separately distilled sets has been observed empirically (Holland et al., 2024). We add that the failure survives exact constituents and an equal-budget joint solution, and we characterize its size for noncommuting source Hessians and general compression ratios.

3 SETUP

Training maps and distillation. Let $\Phi_D^{K,\eta} : \Theta \to \Theta$ denote the map induced by K full-batch gradient-descent steps of size η on the loss of a dataset D. A DD algorithm returns $S = \mathcal{A}(D)$ such that $\Phi_S^{K,\eta}$ approximates the reference $\Phi_D^{T,\eta}$ over initializations $\theta_0 \sim P_0$, typically with $K < T$.

Definition 3.1 (Endpoint error)**.** For two training maps Φ_1, Φ_2, define $\mathcal{C}_{P_0}(\Phi_1, \Phi_2) := \left(\mathbb{E}_{P_0}\|\Phi_1(\theta_0) - \Phi_2(\theta_0)\|_2^2\right)^{1/2}$ and $\mathcal{E}_{P_0}(\Phi_1, \Phi_2) := \mathcal{C}_{P_0}(\Phi_1, \Phi_2)^2$.

For sources D_1, D_2 with distillates $S_i = \mathcal{A}(D_i)$, the constituent error is $\epsilon_i = \mathcal{E}_{P_0}(\Phi_{S_i}^{K,\eta}, \Phi_{D_i}^{T,\eta})$. Write $D_1 \oplus_\alpha D_2$ for the weighted union with loss $\alpha L_{D_1} + (1-\alpha)L_{D_2}$, where α is the weight in the union loss rather than the share of stored samples.

Definition 3.2 (Composition gap)**.** For synthetic weight β and total budget $M \geq |S_1| + |S_2|$, with the infimum over weighted synthetic sets of at most M samples (Appendix B), define the independent composition error, the oracle joint error and the composition gap by

$$\mathcal{E}_{\text{ind}}(\alpha, \beta) := \mathcal{E}_{P_0}\left(\Phi^{K,\eta}_{S_1 \oplus_\beta S_2}, \Phi^{T,\eta}_{D_1 \oplus_\alpha D_2}\right),$$
$$\mathcal{E}^\star_{\text{joint}}(M) := \inf_{|S| \leq M} \mathcal{E}_{P_0}\left(\Phi^{K,\eta}_{S}, \Phi^{T,\eta}_{D_1 \oplus_\alpha D_2}\right),$$
$$\Gamma^\star_{\text{joint}} := \mathcal{E}_{\text{ind}}(\alpha, \alpha) - \mathcal{E}^\star_{\text{joint}}(M) \geq 0,$$

and abbreviate $\mathcal{E}_{\text{ind}} := \mathcal{E}_{\text{ind}}(\alpha, \alpha)$.

The gap is nonnegative since the composed set is an admissible joint candidate, and every exact joint candidate used below is realizable within the same budget (Appendix B). The empirical excess $\widehat{\Gamma}_{\text{RMS}} = \mathcal{C}_{\text{ind}} - \mathcal{C}_{\text{joint}}$ of Section 6 replaces the infimum by one joint distillation run and the difference of squared errors by the difference of RMS errors, and can therefore be negative.

Assumptions. Each theorem states which of the following assumptions it requires.

(A1) *Quadratic, symmetric, positive branch:* full-batch fixed-step gradient descent on $L_i(\theta) = \frac{1}{2}\theta^\top H_i \theta - b_i^\top \theta + c_i$, with symmetric H_i, $\|H_i\|_2 \leq L$ and $\eta L < 1$.

(A2) *Realizability:* $H_i \succeq 0$ and $b_i \in \text{range}(H_i)$, so that the transformed quadratics are realizable as finite squared-loss sample sets (Appendix B).

(A3) *Common optimum, full coverage:* all sources share θ^*, and the second moment $M_0 = \mathbb{E}[(\theta_0 - \theta^*)(\theta_0 - \theta^*)^\top]$ is positive definite.

(A4) *Slowly varying Hessian:* for a non-quadratic F, $\nabla^2 F$ is ν-Lipschitz on a ball containing the K-step trajectory.

The two-to-one identity requires only symmetry. The positive-branch condition is required for the arbitrary-ratio and lower-bound results, (A2) ensures sample realizability, (A3) is invoked where stated and (A4) supports the local nonlinear analysis. Appendix D.4 extends the two-to-one identity to heavy-ball momentum, where the compressed curvature becomes $(2+\beta)H - \eta H^2$ and the discrepancy is unchanged, and to the expectation over independent minibatches.

Universal composability requires affinity. A compression rule that composes for every source collection and every weighting must preserve convex averages. Under the positive-branch condition, compressing T real steps into $K < T$ synthetic steps transforms a quadratic Hessian through $H \mapsto f_p(H) = (I - (I - \eta H)^p)/\eta$ with $p = T/K$, a matrix function (Higham, 2008) that satisfies $(I - \eta f_p(H))^K = (I - \eta H)^T$ and is not affine for $p > 1$.

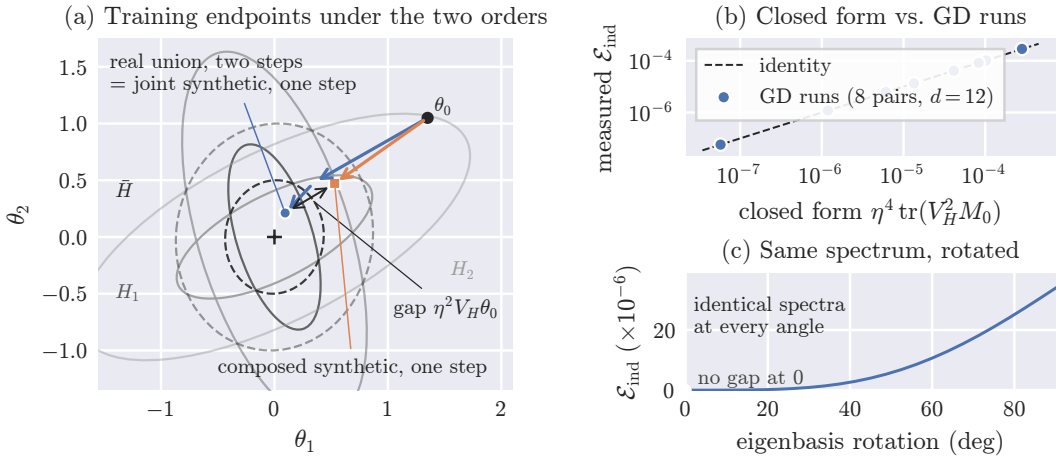

Figure 2: Exact constituents compose with an error set by curvature disagreement (Theorem 4.2, quadratic models with a shared optimum at zero). (a) Two sources with different curvature ellipses. The real two-step and joint one-step endpoints coincide (blue), and the composed endpoint is displaced by $\eta^2 V_H \theta_0$ (orange). (b) Endpoint discrepancies of gradient-descent runs on random source pairs against the closed form. (c) Identical spectra with rotated eigenbases still yield a nonzero discrepancy, since V_H includes eigenvector alignment.

Proposition 3.3 (Composability–affinity principle)**.** Let $\mathcal{X}$ be convex and let $C : \mathcal{X} \to \mathcal{Y}$ be a continuous source-wise compression map into a normed space of synthetic statistics averaged under union. The following are equivalent: (i) every equal-mass pair composes, $C((x+y)/2) = (C(x) + C(y))/2$; (ii) every finite weighted collection composes; (iii) C is affine. The same holds for training endpoints when the endpoint map is injective on the relevant convex hull of compressed statistics.

This is the midpoint form of Jensen's functional equation (Aczél, 1966). The proof can be found in Appendix B.

4 The Composition Law

4.1 One dimension

Let $L_a(\theta) = \frac{a}{2}(\theta - u)^2$ with $0 < \eta a < 1$. After T steps the centered parameter is multiplied by $(1 - \eta a)^T$. For $p = T/K > 1$, training for K steps with the synthetic curvature $g_p(a) := (1 - (1 - \eta a)^p)/\eta$ reproduces this factor exactly, so each source admits a zero-error distillate.

Example 4.1 (Exact constituents need not compose)**.** Take two sources with distinct curvatures $a_1 \neq a_2$ and common optimum u, distill each exactly, and mix the distillates with the correct mass $\alpha \in (0, 1)$. Since $g_p''(a) = -p(p-1)\eta(1 - \eta a)^{p-2} < 0$, strict Jensen's inequality gives

$$\alpha g_p(a_1) + (1 - \alpha) g_p(a_2) < g_p(\alpha a_1 + (1 - \alpha) a_2)\,. \tag{1}$$

The composed synthetic curvature is smaller than the exact joint curvature $g_p(\bar{a})$, which a single example realizes (Appendix Figure 5). Hence $\epsilon_1 = \epsilon_2 = 0$ and $\Gamma^\star_{\text{joint}} > 0$ whenever $\mathbb{E}_{P_0}[(\theta_0 - u)^2] > 0$ (Theorem C.1, Appendix C).

For $a_1 = 1$, $a_2 = 3$, $\eta = 0.1$ and $p = 2$, the exact synthetic curvatures 1.9 and 5.1 average to 3.5, short of the joint value $g_2(2) = 3.6$. For $p = 2$ the shortfall is exactly $\eta\alpha(1 - \alpha)(a_1 - a_2)^2$, the scalar form of the curvature variance introduced below (Figure 2a shows the two-dimensional case). Both distillates are exact and a joint distillate of the same budget exists, so the failure cannot be attributed to poorly optimized source distillates or to an insufficient synthetic budget.

4.2 EXACT LAW FOR TWO-TO-ONE COMPRESSION

The scalar result extends to an exact matrix identity without a shared eigenbasis. Let source i have objective $L_i(\theta) = \frac{1}{2}\theta^\top H_i\theta - b_i^\top\theta + c_i$ with symmetric H_i. For $T = 2$ and $K = 1$, the individually exact synthetic statistics are

$$\widetilde{H}_i = 2H_i - \eta H_i^2, \qquad \widetilde{b}_i = (2I - \eta H_i)b_i. \tag{2}$$

For source masses $\alpha_i > 0$ with $\sum_i \alpha_i = 1$, let $\bar{H} := \sum_i \alpha_i H_i$, $\bar{b} := \sum_i \alpha_i b_i$ and

$$V_H := \sum_i \alpha_i H_i^2 - \bar{H}^2, \qquad w := \sum_i \alpha_i H_i b_i - \bar{H}\bar{b}. \tag{3}$$

V_H is the variance of the source Hessians under the mass weights and w the corresponding covariance between the Hessians and the linear coefficients; both vanish when the sources coincide.

Theorem 4.2 (Structural composition law)**.** Let Φ_{ind} be the one-step map of the union of the exact distillates, with statistics $(\sum_i \alpha_i \widetilde{H}_i, \sum_i \alpha_i \widetilde{b}_i)$, and Φ_R the two-step map of the real union, with statistics $(\bar{H}, \bar{b})$. For symmetric source Hessians, without any commutativity assumption,

$$V_H = \sum_i \alpha_i (H_i - \bar{H})^2 = \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j (H_i - H_j)^2 \succeq 0. \tag{4}$$

The independently composed one-step map and the real-union two-step map satisfy

$$\Phi_{\text{ind}}(\theta_0) - \Phi_R(\theta_0) = \eta^2(V_H\theta_0 - w). \tag{5}$$

For an initialization distribution with mean μ_0 and covariance Σ_0, the composition error is

$$\mathcal{E}_{\text{ind}} = \eta^4\Big[\underbrace{\operatorname{tr}(V_H^2\Sigma_0)}_{\text{curvature variance}} + \underbrace{\|V_H\mu_0 - w\|_2^2}_{\text{mean affine discrepancy}}\Big]. \tag{6}$$

The full proof can be found in Appendix D. In brief, expanding equation 2 shows that the composed statistics $\sum_i \alpha_i(\widetilde{H}_i, \widetilde{b}_i)$ and the exact joint statistics $(2\bar{H} - \eta\bar{H}^2, (2I - \eta\bar{H})\bar{b})$ differ by exactly $(-\eta V_H, -\eta w)$. Substituting this difference into the one-step update yields equation 5, and the expected squared norm gives equation 6. The joint quadratic is realizable under (A2) and the positive branch (Lemma B.1).

Under a common optimum θ^*, $w = V_H\theta^*$ and the error reduces to

$$\mathcal{E}_{\text{ind}} = \eta^4 \operatorname{tr}(V_H^2 M_0), \qquad M_0 := \mathbb{E}[(\theta_0 - \theta^*)(\theta_0 - \theta^*)^\top]. \tag{7}$$

Agreement on the optimum thus does not ensure agreement on the endpoint: the discrepancy $\eta^2 V_H(\theta_0 - \theta^*)$ vanishes at the optimum and grows with the curvature differences along the directions the initialization occupies. V_H is the mass-weighted variance of the Hessians as matrices, so it records differences in eigenvalues and eigenvectors alike (Figure 2c), and M_0 records the directions the initialization covers. The matrix law is not a restatement of the scalar Jensen gap: the Hessians need not commute, the affine term couples them with the linear coefficients, and the lower bound for other ratios (Theorem E.2) holds without commutativity. The squared error grows as η^4, the log–log learning-rate slope of four that Section 6 tests.

Zero-error condition. The composition error vanishes if and only if $V_H\Sigma_0^{1/2} = 0$ and $V_H\mu_0 = w$. Under a common optimum and $M_0 \succ 0$, it vanishes if and only if all sources have the same Hessian.

With $M_0 \succ 0$ no direction of curvature disagreement is avoided by the initialization, which makes the second statement an equivalence. With $\Sigma_0 \succ 0$ the first condition forces $V_H = 0$ and hence $w = 0$, so identical source Hessians are necessary and sufficient whatever the optima, and the affine condition is independent only for a rank-deficient initialization (Appendix D).

4.3 Other compression ratios

Under a common optimum and the positive-branch condition, for any $p = T/K > 1$ and integer $K \geq 1$, the joint and the composed transformed Hessians differ by

$$J_p = \binom{p}{2}\eta V_H + R_p, \qquad \|R_p\|_F \leq \eta^2 L c_p(\eta L)\sqrt{r_H}\,\|V_H\|_F, \tag{8}$$

where c_p is the remainder series of Appendix E and $r_H = \operatorname{tr}(V_H)^2/\|V_H\|_F^2$ is the effective rank of $V_H \neq 0$. The K-step endpoint discrepancy admits a lower bound without commutativity (Theorem E.2), positive for $\eta L c_p(\eta L)\sqrt{r_H} < \binom{p}{2}$ and unconditional for $p = 2$, so curvature variance governs the leading operator term for every ratio and step count.

4.4 Approximate distillates

Perturbing the composed statistics by $(\Delta H, \Delta b)$ changes the RMS error by at most a computable Δ_{aff} (Theorem F.1, Appendix F). Writing $\mathcal{C} = \sqrt{\mathcal{E}}$, $\mathcal{C}_{\text{struct}}$ for the composition error with exact constituents and $\widehat{\mathcal{C}}_{\text{comp}}$ for the error with the constituents actually produced, the bound yields a certificate.

Corollary 4.3 (Sufficient detectability condition)**.** Assume $T = 2$, $K = 1$, a common optimum θ^*, and that both the ideal and the actual source statistics fix θ^*. Let $\mathcal{U}_{\text{src}} = \sum_i \alpha_i\sqrt{\epsilon_i}$ and $\mathcal{C}_{\text{struct}} = \eta^2\|V_H M_0^{1/2}\|_F$. Then

$$\left|\widehat{\mathcal{C}}_{\text{comp}} - \mathcal{C}_{\text{struct}}\right| \leq \mathcal{U}_{\text{src}}, \qquad \mathcal{C}_{\text{struct}} > \mathcal{U}_{\text{src}} \Longrightarrow \widehat{\mathcal{C}}_{\text{comp}} \geq \mathcal{C}_{\text{struct}} - \mathcal{U}_{\text{src}} > 0. \tag{9}$$

The corollary compares two quantities that can be estimated separately: $\mathcal{C}_{\text{struct}}$ from the source statistics before any distillation, and $\mathcal{U}_{\text{src}}$ from the distillates actually produced. When the first exceeds the second, the composed union misses the real-union endpoint by at least their difference. On a network, where the common-optimum assumption is not established, the inequality is a diagnostic rather than a certificate. Constituent errors alone cannot upper-bound the composition error, since exact constituents with separating optima have unbounded composition error (Proposition F.3).

5 From Quadratic Models to Networks

Example 5.1 (Networks as local quadratic proxies)**.** For a neural loss F at a reference point θ°, the frozen quadratic Q has symmetric Hessian $H = \nabla^2 F(\theta^\circ)$, which need not be positive semidefinite, so Q is a proxy and a realizable synthetic set only when (A2) holds.

For the objective F and the frozen quadratic Q of Example 5.1, if $\nabla^2 F$ is ν-Lipschitz on a ball of radius r centered at θ° that contains the K-step trajectory, and $\|I - \eta H\|_2 \leq \rho$, then

$$\|\Phi_F^{K,\eta}(\theta_0) - \Phi_Q^{K,\eta}(\theta_0)\|_2 \leq \frac{\eta\nu r^2}{2}\sum_{t=0}^{K-1}\rho^t. \tag{10}$$

The nonlinear RMS composition error thus differs from its quadratic counterpart by at most $\delta_S + \delta_R$, the sum of the synthetic and real training remainders (Theorem G.1, Appendix G).

The structural term of the local model, that is, of the frozen quadratic at the initialization, can be evaluated at a shared parameter without a common optimum and without forming Hessians. Writing $g_i = H_i\theta_0 - b_i$ and $\bar{g} = \sum_i \alpha_i g_i$, define

$$r_{\text{comp}}(\theta_0) := \sum_i \alpha_i H_i(g_i - \bar{g}) = V_H\theta_0 - w, \qquad P_{\text{comp}} := \eta^4\mathbb{E}_{P_0}\|r_{\text{comp}}(\theta_0)\|_2^2, \tag{11}$$

so that $P_{\text{comp}} = \mathcal{E}_{\text{ind}}$ for ideal two-to-one quadratic compression and $P_{\text{comp}} = \eta^4\operatorname{tr}(V_H^2 M_0)$ under a common optimum. Given the source gradients, each evaluation of r_{comp} costs one Hessian-vector product per source (Pearlmutter, 1994) (Appendix J.2). The observed RMS error satisfies

$$\left|\mathcal{C}_{\text{obs}} - \sqrt{P_{\text{comp}}}\right| \leq \Delta_{\text{aff}} + \delta_S + \delta_R \tag{12}$$

(Corollary H.3), where Δ_{aff} is the error of the distilled statistics and δ_S, δ_R are the nonlinear remainders of synthetic and real training. Thus $\sqrt{P_{\text{comp}}}$ approximates the observed RMS error when all three are small relative to it, and Section 6 measures the source matching errors and their contribution.

6 Experimental Results

The experiments have three tasks (settings and additional results in Appendix J): verifying the exact law on quadratic sources with learned synthetic sets, testing the local approximation on a network by comparing the local endpoint discrepancy $\Delta\Phi(\theta_0)$ of the real sources, which involves no synthetic set, with the prediction $\eta^2 r_{\text{comp}}(\theta_0)$, and decomposing the composition error of learned synthetic sets, by endpoint matching on the same network and by distribution matching on CIFAR-10, SVHN and CIFAR-100.

Learned quadratic distillates reproduce the law once each source is matched. For twenty random pairs of quadratic sources in $d = 12$ dimensions, n_{syn} squared-loss samples per source are optimized on the two-step versus one-step endpoint discrepancy over random initializations, without access to the target statistics (Appendix C.1). Once the budget reaches the dimension, the learned Gram matrices recover $2H - \eta H^2$ to machine precision, and the union of the learned sets then incurs exactly the composition error of Theorem 4.2, with $V_H = \frac{1}{4}(H_1 - H_2)^2$, in every pair (Appendix Table 2). Below the dimension the constituents are inexact, and the composed error exceeds the closed form by a factor of 4.8 to 11.7 that grows as the budget shrinks.

The quadratic prediction captures local network dynamics at small step sizes. Network experiments use two-source CIFAR-10 (Krizhevsky, 2009) partitions, IID or class-wise Dirichlet with concentration 1.0, 0.5 or 0.1 (Dir-1.0 to Dir-0.1), and a three-block ConvNet (Appendix J). The main endpoint measurements use two real steps ($T = 2$; Appendix J.3 also examines $T = 4$, $K = 2$). The local endpoint discrepancy $\Delta\Phi(\theta_0)$ applies the construction of Theorem 4.2 to the network, and its RMS is the local reference $\mathcal{C}_{\text{loc}}$, equal to $\mathcal{C}_{\text{struct}} = \sqrt{P_{\text{comp}}}$ for quadratics:

$$\mathcal{C}_{\text{loc}} = \left(\mathbb{E}_{\theta_0}\|\Delta\Phi(\theta_0)\|^2\right)^{1/2}, \qquad \Delta\Phi(\theta_0) = \sum_i \alpha_i \Phi^{2,\eta}_{D_i}(\theta_0) - \Phi^{2,\eta}_{D_1 \oplus_\alpha D_2}(\theta_0). \tag{13}$$

For a union of synthetic sets the discrepancy against the same target decomposes exactly as

$$d_{\text{ind}}(\theta_0) = \Delta\Phi(\theta_0) + r_{\text{src}}(\theta_0), \qquad r_{\text{src}}(\theta_0) = \sum_i \alpha_i \left(\Phi^{1,\eta}_{S_i}(\theta_0) - \Phi^{2,\eta}_{D_i}(\theta_0)\right), \tag{14}$$

so three quantities appear below: $\eta^2 r_{\text{comp}}(\theta_0)$ is the prediction from local gradients and curvature, $\Delta\Phi(\theta_0)$ is the real-data discrepancy between separate training followed by averaging and training on the union, and $d_{\text{ind}}(\theta_0)$ is the error of the union of the distillates against the real-union target. The weighted source error $\mathcal{U}_{\text{src}} = \sum_i \alpha_i \|\Phi^{1,\eta}_{S_i} - \Phi^{2,\eta}_{D_i}\|_{L^2(P_0)}$ bounds $\|r_{\text{src}}\|_{L^2(P_0)}$ from above, and the residual falls below the bound when the source errors partly cancel.

We compare $\Delta\Phi(\theta_0)$ with its prediction $\eta^2 r_{\text{comp}}(\theta_0)$ at $\eta_{\text{ref}} = 7.8 \times 10^{-4}$ and above. Within $\eta L < 1$, SiLU and tanh networks give squared-error exponents of 3.71 ± 0.08 and 3.86 ± 0.06 at Dir-0.1, against the quadratic value of four, and the calibration ratio is within about three percent of one at η_{ref} (Appendix Figure 9). On the SiLU network at η_{ref} the prediction also matches the direction of $\Delta\Phi(\theta_0)$, with cosine 1.000 and a relative residual of 2 to 4 percent, and it degrades as the step size grows, overshooting $\Delta\Phi$ by a factor of 1.9 at $16\eta_{\text{ref}}$ (Appendix J.3, Figure 12b).

Source residuals exceed the local discrepancy in the endpoint-matching experiments. Synthetic sets learned on the same network by matching one synthetic step to the two real steps at one, five or ten images per class per source retain a relative source error of 0.5 at one image per class and 0.25 to 0.39 at five and ten (Appendix J.3, Figure 12). At Dir-0.1 the aggregate source residual $\|r_{\text{src}}\|$ is 24, 13 and 13 times $\mathcal{C}_{\text{struct}}$ at one, five and ten images per class, the inequality of Corollary 4.3 is satisfied for none of 96 initializations at η_{ref}, and the composed discrepancy is nearly orthogonal to the prediction. For these sets the aggregate source residual r_{src}, not the structural term, sets the composed error.

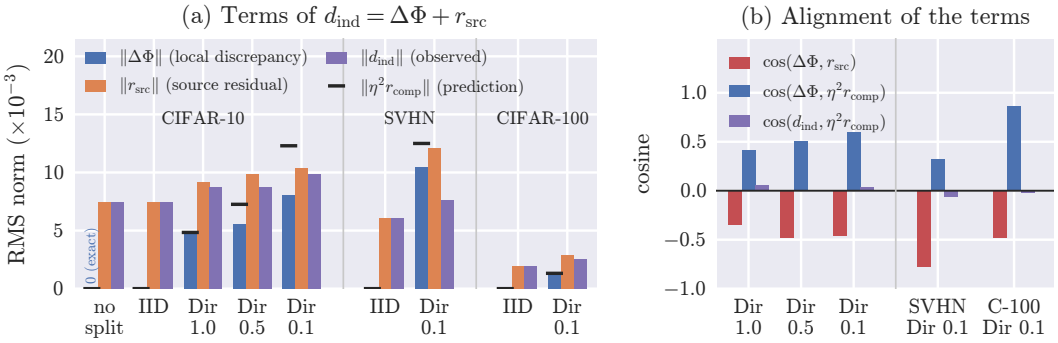

Figure 3: The observed error of a union of distillates is the net of the local discrepancy and the source residual. Exact decomposition $d_{\text{ind}} = \Delta\Phi + r_{\text{src}}$ on the distribution-matching sets of Section 6, with the step calibration described there (CIFAR-10, SVHN and CIFAR-100; level means over partitions, Appendix Table 8). (a) RMS norms of the local discrepancy $\Delta\Phi$, the aggregate source residual r_{src} and the observed error d_{ind}, with the quadratic prediction $\eta^2 r_{\text{comp}}$ marked; the local discrepancy is exactly zero in the no-split control and 0.013–0.018×10^{-3} at IID, below the resolution of the axis. The residual is the larger term at every level, and $\|d_{\text{ind}}\| < \|r_{\text{src}}\|$ under skew. (b) Cosines at the skewed levels: the two terms are anti-aligned, the prediction is partly aligned with $\Delta\Phi$ at this step size ($\eta L \approx 1$) and nearly orthogonal to d_{ind}.

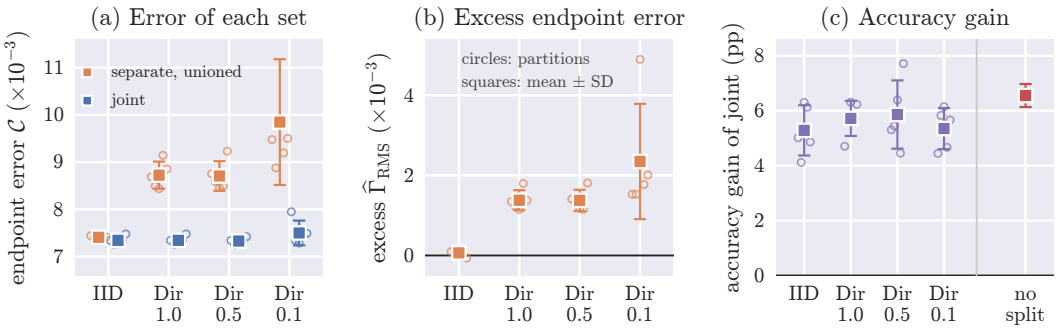

Figure 4: The excess endpoint error grows with label skew, whereas the accuracy gain of joint distillation persists in the no-split control. Independent versus joint distillation on CIFAR-10 with distribution matching at ten images per class per source. Each skew level has five partitions (circles), and squares with bars give their mean and SD. Within a partition, endpoint errors (after the per-set step calibration of Section 6) are RMS over three initializations and accuracies average three evaluation seeds. (a) RMS endpoint error of the separately distilled union and of the joint set against the real-union two-step map. (b) Their difference, the excess endpoint error. (c) Accuracy gain of joint over separate-then-union distillation. The no-split control (red) has three replicates of the full dataset, and its bar is the SD over replicates.

The excess endpoint error of the union grows with label skew. Distribution matching (DM) produces synthetic sets as in practice, here on a ReLU network at $\eta = 6.25 \times 10^{-3}$ at its official schedule with ten images per represented class per source. The joint set is distilled from the union at the same total budget (Appendix J.4). DM sets are not trained for any step size, so each set is applied with the scalar step multiplier that best matches its own real two-step target at each evaluation initialization, the joint set being calibrated against the union target (Appendix H). Without it all errors rise and the excess shrinks but stays positive at every skewed level (Appendix J.5). Below, $\Phi_S^{1,\eta}$ denotes this calibrated map and $\mathcal{C}_{\text{ind}}$, $\mathcal{C}_{\text{joint}}$ and $\widehat{\Gamma}_{\text{RMS}} = \mathcal{C}_{\text{ind}} - \mathcal{C}_{\text{joint}}$ the resulting endpoint errors (Definition 3.1).

Joint distillation uses the same full training set at every level and its endpoint error stays within 7.3–7.5×10^{-3} (Figure 4a). The error of the separately distilled union matches it under IID and rises to 9.85×10^{-3} at Dir-0.1. The excess is near zero under IID partitioning and positive and an order of

magnitude larger on every label-skewed partition (Figure 4b). SVHN and CIFAR-100 replicate this pattern, the CIFAR-100 excess rising from 0.09 at IID to 0.56×10^{-3} at Dir-0.1 and positive in nine of nine skewed cells (Appendix J.10).

Both errors are comparable to $\mathcal{C}_{\text{struct}}$ and $\mathcal{U}_{\text{src}}$ is about three times $\mathcal{C}_{\text{struct}}$, so the inequality of Corollary 4.3 is not satisfied.

The observed discrepancy is the net of the local discrepancy and the source residual. By equation 14 the excess is $\widehat{\Gamma}_{\text{RMS}} = \|\Delta\Phi + r_{\text{src}}\|_{L^2(P_0)} - \|d_{\text{joint}}\|_{L^2(P_0)}$ (Appendix J.5), and $\|d_{\text{ind}}\| < \|r_{\text{src}}\|$ requires $\langle \Delta\Phi, r_{\text{src}} \rangle < -\frac{1}{2}\|\Delta\Phi\|^2$, a net cancellation between the two terms. The DM sets show this cancellation (Figure 3): the measured $\|d_{\text{ind}}\|$ is smaller than $\|r_{\text{src}}\|$ at every skewed level, and the cosine between $\Delta\Phi$ and r_{src} averages -0.35 to -0.48 and is negative for all 45 skewed partition–seed pairs. At this step size ($\eta L \approx 1$) the quadratic model is a proxy rather than a calibrated predictor: $\eta^2 r_{\text{comp}}$ matches $\Delta\Phi$ in magnitude but only partly in direction and is nearly orthogonal to d_{ind} (Figure 3b). Reweighting the constituents narrows the excess without removing it: class weights fitted on held-out initializations remove 15 to 25 percent of the union's squared error at the skewed levels and cut the mean excess over the joint set to 0.1–0.7×10^{-3}, still positive at every skewed level, and accuracy does not change detectably (Appendix J.6).

The accuracy advantage persists without source heterogeneity. On CIFAR-10, joint distillation improves accuracy by 5.3 to 5.9 points at every heterogeneity level (Figure 4c, Appendix Table 7). With both independent distillates learned from the full training set, the local discrepancy is exactly zero, yet joint distillation still improves accuracy by 6.6 ± 0.4 points (Appendix J.7). Between IID and Dir-0.1 the local reference $\mathcal{C}_{\text{loc}}$ grows five-hundredfold while the mean gain stays between 5 and 6 points: heterogeneity is not necessary for the gain. Matching the per-class counts of the joint set to those of the composed set at Dir-0.1 (Appendix J.8) leaves the excess positive while the accuracy gain on the two affected partitions falls by 3.5 and 1.9 points, so much of that gain reflects the class allocation. Continuing distillation from the union of two distillates for 2,000 iterations recovers 0.4 points of the no-split gap, against 1.8 for a set initialized from real images over the same iterations (Appendix J.7). On the ConvNet the CIFAR-10 and CIFAR-100 gains do not increase with skew, whereas on SVHN (Netzer et al., 2011) the gain increases with skew, and on ResNet-18, where the same sets gain 1.8 to 3.4 points, the Dir-0.1 gain exceeds the IID gain on every partition (Appendices J.10 and J.11).

7 Conclusion and Limitations

We showed that independently distilled datasets can each reproduce training on their own source exactly and still fail when combined, even when the total budget admits an exact joint distillate. For quadratic objectives the failure has an exact form, the difference between averaging compressed source statistics and compressing their average, and learned quadratic distillates reproduce it. On a smooth network at small step sizes the quadratic prediction captures the local endpoint discrepancy in magnitude and direction. For learned synthetic sets the source residual, not the structural term, sets the composed error. Under endpoint matching it exceeds the structural term by more than an order of magnitude. Under distribution matching it partly offsets the local discrepancy. The accuracy gain of joint distillation persists even for two distillates of the same dataset. Single-source fidelity is therefore insufficient to judge composite fidelity, and training fidelity and downstream accuracy are distinct requirements, neither of which is established by evaluating each distillate on its own.

The exact results concern fixed-step gradient descent on quadratic objectives (A1) and apply to a network only locally (A4); adaptive optimizers and long horizons lie outside them (Appendix K). One distillation method was used at the official schedule, and the network experiments have not entered the regime in which the structural term dominates; whether a distillation method can reach it is open, as is whether multi-step trajectory matching couples short-horizon fidelity to accuracy. For a downstream user the practical consequence is that a union of independently produced distillates should be validated as a whole against a reference on the union, since the fidelity of each constituent certifies neither the endpoint nor the accuracy of their combination.

Ethics Statement

All datasets used in this work are publicly available benchmarks (CIFAR-10, CIFAR-100 and SVHN) obtained from their official distribution channels and used under the providers' terms. No personally identifiable or sensitive data is involved, and no new data involving human subjects were collected. The paper combines theoretical analysis with benchmark experiments. Its practical implication, that independently produced distillates should be validated after union, is a caution for federated and data-governance settings rather than a new capability.

Reproducibility Statement

All assumptions, theorem statements and complete proofs are given in Appendices B–I. Appendix J specifies the partition constructions, distillation schedules, hyperparameters, evaluation protocols, estimator checks and partition seeds of every experiment (E1–E15), including the exact decomposition of the observed discrepancy (E9), the endpoint-matching distillation on the network (E10) and the class-allocation-matched joint set (E11), together with the additional results referenced in the main text. The code, the scripts of every experiment and their row-level results are included in the supplementary material. All datasets are publicly available.

AI Use Statement

Large language models were used to aid and polish the writing and to check the proofs and derivations for errors and gaps. They were not used to develop the research question or the theoretical framework, to formulate the theorems, to implement the code, to run experiments, to search the literature, or to generate data. The authors verified every AI-assisted passage and derivation and take responsibility for the final content of this work.

References

János Aczél. *Lectures on Functional Equations and Their Applications*. Academic Press, New York, 1966.

Pankaj K. Agarwal, Graham Cormode, Zengfeng Huang, Jeff M. Phillips, Zhewei Wei, and Ke Yi. Mergeable summaries. *ACM Transactions on Database Systems*, 38(4), 2013.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Marco Arazzi, Mert Cihangiroglu, Serena Nicolazzo, and Antonino Nocera. Secure federated dataset distillation. *Engineering Applications of Artificial Intelligence*, 2025.

Deyu Bo, Songhua Liu, and Xinchao Wang. Understanding dataset distillation via spectral filtering. In *International Conference on Learning Representations*, 2026.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yanda Chen, Gongwei Chen, Miao Zhang, Weili Guan, and Liqiang Nie. Curriculum coarse-to-fine selection for high-IPC dataset distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20437–20446, 2025.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. DC-BENCH: Dataset condensation benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34391–34404, 2022.

Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3758, 2023.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM Journal on Computing*, 49(3), 2020.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *International Conference on Machine Learning*, 2019.

Jack Goetz and Ambuj Tewari. Federated learning via synthetic data. *arXiv preprint arXiv:2008.04489*, 2020.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *International Conference on Learning Representations*, 2024.

Aaryan Gupta, Rishi Saket, and Aravindan Raghuveer. Algorithmic guarantees for distilling supervised and offline RL datasets. In *International Conference on Learning Representations*, 2026.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, 2008.

William Holland, Chandra Thapa, Wei Shao, Seyit Camtepe, and Sarah Ali Siddiqui. One-shot collaborative data distillation. In *European Conference on Artificial Intelligence*, 2024.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Shengyuan Hu, Jack Goetz, Kshitiz Malik, Hongyuan Zhan, Zhe Liu, and Yue Liu. FedSynth: Gradient compression via synthetic data in federated learning. *arXiv preprint arXiv:2204.01273*, 2022.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 18(3):1059–1076, 1989.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S. Mirrokni. Composable coresets for diversity and coverage maximization. In *ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 100–108, 2014.

Zachary Izzo and James Zou. A theoretical study of dataset distillation. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2023.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11102–11118, 2022.

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

Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):17–32, 2024.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Compressed gastric image generation based on soft-label dataset distillation for medical data sharing. *Computer Methods and Programs in Biomedicine*, 227:107189, 2022.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation for medical dataset sharing. In *AAAI Conference on Artificial Intelligence (AAAI) Workshop*, pp. 1–6, 2023.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Importance-aware adaptive dataset distillation. *Neural Networks*, 172, 2024a.

Longzhen Li, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Generative dataset distillation: Balancing global structure and local details. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, pp. 7664–7671, 2024b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020.

Ping Liu, Xin Yu, and Joey Tianyi Zhou. Meta knowledge condensation for federated learning. In *International Conference on Learning Representations*, 2023a.

Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. DREAM: Efficient dataset distillation by representative matching. In *IEEE/CVF International Conference on Computer Vision*, pp. 17314–17324, 2023b.

Michael Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations*, 2021a.

Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

Bincheng Peng, Guang Li, Ping Liu, Takahiro Ogawa, and Miki Haseyama. Closed-form linear-probe dataset distillation for pre-trained vision models. *arXiv preprint arXiv:2605.07194*, 2026.

Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Distilled replay: Overcoming forgetting through synthetic samples. In *Continual Semi-Supervised Learning*, pp. 104–117. Springer, 2022.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z. Liu, Yuri A. Lawryshyn, and Konstantinos N. Plataniotis. DataDAM: Efficient dataset distillation with attention matching. In *IEEE/CVF International Conference on Computer Vision*, pp. 17097–17107, 2023.

Xiufang Shi, Wei Zhang, Yuheng Li, Mincheng Wu, Zhenyu Wen, Shibo He, Tejal Shah, and Rajiv Ranjan. Dataset distillation-based hybrid federated learning on non-IID data. *IEEE Transactions on Network Science and Engineering*, 2026. to appear; arXiv:2409.17517v3.

Rui Song, Dai Liu, Dave Zhenyu Chen, Andreas Festag, Carsten Trinitis, Martin Schulz, and Alois Knoll. Federated learning via decentralized dataset distillation in resource-constrained edge environments. In *International Joint Conference on Neural Networks*, 2023.

Duo Su, Junjie Hou, Guang Li, Ren Togo, Rui Song, Takahiro Ogawa, and Miki Haseyama. Generative dataset distillation based on diffusion model. In *European Conference on Computer Vision (ECCV) Workshop*, 2024.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. CAFE: Learning to condense dataset by aligning features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Peiran Wang and Haohan Wang. DistDD: Distributed data distillation aggregation through gradient matching. *arXiv preprint arXiv:2410.08665*, 2024.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. An aggregation-free federated learning for tackling data heterogeneity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26233–26242, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022.

Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. FedDM: Iterative distribution matching for communication-efficient federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16323–16332, 2023.

Enneng Yang, Li Shen, Zhenyi Wang, Tongliang Liu, and Guibing Guo. An efficient dataset condensation plugin and its application to continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

William Yang, Ye Zhu, Zhiwei Deng, and Olga Russakovsky. What is dataset distillation learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56812–56834, 2024.

Zeyuan Yin, Eric P. Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at ImageNet scale from a new perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2024.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12674–12685, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.

Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020.

Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Appendix

Contents

A Extended Related Work 14

B Formal Framework and Finite-Sample Realizability 15

C One-Dimensional Exact Separation 17

D Exact Matrix Composition Laws 19

E General Finite-Step Jensen Gap 22

F Approximate Distillation and Identifiability 27

G Local Nonlinear Remainder 29

H Calibration and Components That Scalar Retiming Cannot Repair 30

I HVP Estimator and Concentration 32

J Additional Experimental Results 34

K Limitations and Future Work 50

A Extended Related Work

Dataset distillation objectives. Beyond the objectives cited in the main text, feature-alignment and contrastive objectives (Wang et al., 2022; Sajedi et al., 2023), kernel and random-feature regression (Nguyen et al., 2021b; Zhou et al., 2022), difficulty- and representativeness-aware matching (Guo et al., 2024; Liu et al., 2023b), and large-scale, generative and sample-weighting formulations (Kim et al., 2022; Du et al., 2023; Yin et al., 2023; Li et al., 2024a;b; Su et al., 2024) have broadened the setting (benchmark: Cui et al., 2022; surveys: Yu et al., 2024; Lei & Tao, 2024), and closed-form linear-probe distillation for pre-trained backbones (Peng et al., 2026) shares our aim of making part of the problem exactly solvable. In all of these settings the evaluation unit is a single synthetic set.

Sharing distilled data across institutions. Distilled sets have been proposed as an anonymized, compact substitute for sharing clinical images across institutions (Li et al., 2022; 2023), where the real union is never available and the downstream user receives independently distilled sets. Federated variants (building on federated averaging; McMahan et al., 2017) apply a similar approach to client data, either once (Zhou et al., 2020; Song et al., 2023) or iteratively (Goetz & Tewari, 2020; Hu et al., 2022; Xiong et al., 2023; Wang & Wang, 2024), with the server training on the union. Liu et al. (2023a) observe that locally condensed sets remain heterogeneous and reweight them, Wang et al. (2024) condense collaboratively, and Holland et al. (2024) report that naive merging underperforms. Two recent designs are closest to our setting. Arazzi et al. (2025) distill a single synthetic set collaboratively, coordinating clients during condensation and applying a differential-privacy mechanism to the shared updates; because the set is jointly optimized, no post-hoc union occurs. Shi et al. (2026) partition clients into clusters, transmit independently distilled sets to a cluster head, and train on their union at the head, with clusters arranged so that the union is approximately label-balanced. This is a post-hoc union of independent distillates, but their convergence analysis concerns the federated model rather than the training map induced by the union. In continual learning, condensed examples are accumulated as replay memories (Rosasco et al., 2022; Deng & Russakovsky, 2022; Yang et al., 2023), and Yang et al. (2024) and Chen et al. (2025) report that mixing distilled and real samples can degrade performance.

Merging and mergeability. Mergeable summaries and composable coresets provide task-specific guarantees that survive union by construction (Agarwal et al., 2013; Indyk et al., 2014; Feldman et al., 2020). Standard source-wise distillation objectives do not provide such guarantees on their

own, because they compress a training procedure rather than a statistic. Model-merging studies address mode connectivity and weight averaging (Wortsman et al., 2022), task arithmetic (Ilharco et al., 2023), permutation symmetry (Ainsworth et al., 2023), and curvature-aware weighting (Matena & Raffel, 2022). The last is the closest analogue: Fisher-weighted averaging corrects, to first order, a curvature mismatch between separately fine-tuned parameters. In our setting the mismatch lies in the data, and Theorem 4.2 expresses it in closed form as the variance of the source Hessians.

Theory of distillation. Existing analyses ask what a distillate can represent at a converged solution, for example exact recovery of the empirical-risk minimizer in linear and kernel settings (Izzo & Zou, 2023; Nguyen et al., 2021a), or in closed form for a frozen backbone (Peng et al., 2026). We instead fix an optimizer and a finite horizon and ask whether compression commutes with source mixing.

B Formal Framework and Finite-Sample Realizability

Table 1: Notation.

Symbol	Meaning
Maps	
$\Phi_D^{K,\eta}$	K-step full-batch GD map at step size η on dataset D
$\Phi_{\text{ind}},\ \Phi_R$	composed synthetic one-step map; real-union two-step map
$f_p(H)$	time-compression transform $(I-(I-\eta H)^p)/\eta$, $p=T/K$
Errors	
$\mathcal{E},\ \mathcal{C}=\sqrt{\mathcal{E}}$	squared endpoint error over $\theta_0\sim P_0$; its RMS form
ϵ_i	squared endpoint error of constituent i
$\mathcal{E}_{\text{ind}},\ \mathcal{C}_{\text{ind}}$	error of the independently composed union
$\Gamma^\star_{\text{joint}}$	oracle composition gap, $\mathcal{E}_{\text{ind}}-\mathcal{E}^\star_{\text{joint}}(M)$
$\widehat{\Gamma}_{\text{RMS}}$	empirical excess endpoint error, $\mathcal{C}_{\text{ind}}-\mathcal{C}_{\text{joint}}$
$\widehat{\Gamma}_{\text{mean}}$	mean over seeds of the per-seed endpoint-distance difference (Appendix J)
Structural quantities	
$V_H,\ w$	curvature variance and affine conflict, Eq. (3)
$\mathcal{C}_{\text{struct}},\ \widehat{\mathcal{C}}_{\text{comp}}$	quadratic-model RMS gap with exact / actual constituents
$\mathcal{C}_{\text{loc}}$	measured local reference (network, exact constituents)
P_{comp}	HVP-computable estimate of $\mathcal{C}^2_{\text{struct}}$ for $T=2, K=1$
r_H	effective rank of V_H
M_0	second-moment matrix of the initialization about θ^*
Bounds	
$\mathcal{U}_{\text{src}}$	weighted constituent RMS, $\sum_i \alpha_i\sqrt{\epsilon_i}$
$\Delta_{\text{aff}},\ \delta_S,\ \delta_R$	affine-stability and nonlinear remainders
ζ	their sum, the constituent-plus-nonlinear uncertainty

B.1 Weighted Unions and Training Maps

A weighted dataset is a finite collection of examples with nonnegative weights that sum to one. The loss of $D_1 \oplus_\alpha D_2$ is $\alpha L_{D_1} + (1-\alpha)L_{D_2}$. This definition separates the intended source mass from the cardinality of a stored synthetic set. Naive concatenation is the special case in which empirical sample counts determine the weights.

Throughout the theory, $\Phi_D^{K,\eta}$ denotes full-batch gradient descent with fixed learning rate η, initialized from P_0. All endpoint comparisons use the same parameterization and initialization distribution. The RMS quantity $\mathcal{C}_{P_0}$ obeys the triangle inequality as an $L_2(P_0)$ norm. Its square $\mathcal{E}_{P_0}$ is used for exact error formulas.

The oracle joint error is no larger than the independent error when $M \geq \sum_i |S_i|$, since the independently composed set is itself feasible in the oracle optimization. For the exact quadratic constructions, the next realizability lemma (Lemma B.1) supplies a joint candidate within this budget rather than relying only on the trivial feasibility argument.

B.2 Proof of the composability–affinity principle

Proof of Proposition 3.3. Condition (ii) immediately implies (i), while (iii) implies (ii). It remains to show that (i) implies (iii). Repeated midpoint composition gives, for every dyadic rational $q \in [0,1]$,

$$C(qx + (1-q)y) = qC(x) + (1-q)C(y). \tag{15}$$

Continuity of C extends the identity from dyadic rationals to every $q \in [0,1]$, so C is affine on $\mathcal{X}$. If composition is specified only through a downstream endpoint map E, injectivity of E on the relevant convex hull turns

$$E(C((x+y)/2)) = E((C(x) + C(y))/2) \tag{16}$$

back into the statistic-level midpoint identity, and the same argument applies. Injectivity is a property of the induced operator map $\theta \mapsto \Phi^{K,\eta}(\theta)$ on Θ, or of its restriction to an initialization distribution with $M_0 \succ 0$, not of an endpoint at a single θ_0; for quadratics it holds under (A3) together with the positive-branch condition on the compressed statistics, whereas for overparameterized networks the map from surrogate statistics to trajectories is generally not injective, so the equivalence is an algebraic property of the quadratic model rather than an impossibility result for networks. □

B.3 Finite-sample realizability

Lemma B.1 (Quadratic objectives are finite synthetic datasets)**.** Let

$$L(\theta) = \frac{1}{2}\theta^\top H\theta - b^\top\theta + c, \qquad H \succeq 0, \quad b \in \mathrm{range}(H). \tag{17}$$

Up to an additive constant, L is the sum of $r = \mathrm{rank}(H)$ scalar squared-loss examples. For $p = T/K > 1$ and $0 \preceq \eta H \prec I$, the transformed statistics

$$\widetilde{H} = f_p(H), \qquad \widetilde{b} = f_p(H)H^\dagger b \tag{18}$$

have the same range and rank as H, reproduce T real steps in K synthetic steps, and are realizable with the same budget. For positive-mass sources (H_i, b_i), the exact transformed real-union objective needs at most $\sum_i \mathrm{rank}(H_i)$ examples.

Proof. Write the positive eigendecomposition

$$H = \sum_{j=1}^{r} \lambda_j u_j u_j^\top, \qquad b = \sum_{j=1}^{r} b_j u_j. \tag{19}$$

For $x_j = \sqrt{\lambda_j}u_j$ and $y_j = b_j/\sqrt{\lambda_j}$,

$$\frac{1}{2}\sum_{j=1}^{r}(x_j^\top\theta - y_j)^2 = \frac{1}{2}\theta^\top H\theta - b^\top\theta + \frac{1}{2}\sum_{j=1}^{r} y_j^2. \tag{20}$$

The scalar map $\lambda \mapsto [1-(1-\eta\lambda)^p]/\eta$ is positive exactly when $\lambda > 0$ on the stable interval, so $\widetilde{H}$ has the same range and rank. Since $b \in \mathrm{range}(H)$, $\widetilde{b} = f_p(H)H^\dagger b$ belongs to that range. If $u = H^\dagger b$, both real and synthetic objectives have optimum u on $\mathrm{range}(H)$, while

$$(I - \eta\widetilde{H})^K = (I - \eta H)^T. \tag{21}$$

They therefore induce the same affine endpoint map.

For PSD matrices,

$$\ker\left(\sum_i \alpha_i H_i\right) = \bigcap_i \ker(H_i). \tag{22}$$

Because $b_i \in \mathrm{range}(H_i) = \ker(H_i)^\perp$, the union linear statistic $\bar{b} = \sum_i \alpha_i b_i$ belongs to $\mathrm{range}(\bar{H})$. Finally,

$$\mathrm{rank}(\bar{H}) \le \sum_i \mathrm{rank}(H_i), \tag{23}$$

which proves the budget statement. □

C One-Dimensional Exact Separation

For $L_a(\theta) = \frac{a}{2}(\theta - u)^2$, gradient descent gives

$$\Phi_a^{T,\eta}(\theta_0) - u = (1 - \eta a)^T(\theta_0 - u). \tag{24}$$

The time-compression transform $g_p(a) := (1 - (1 - \eta a)^p)/\eta$ of Section 3 satisfies

$$(1 - \eta g_p(a))^K = (1 - \eta a)^T, \tag{25}$$

so a single synthetic squared-loss example with $x = \sqrt{g_p(a)}$ and $y = xu$ is individually exact.

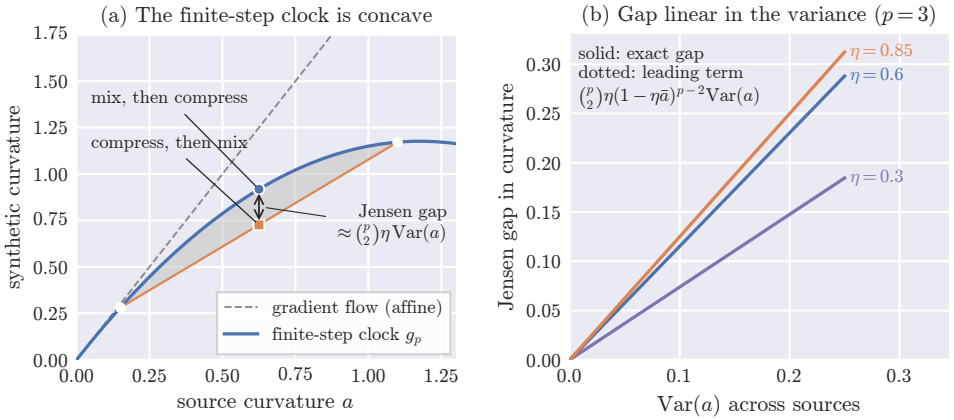

Figure 5: Concavity of the finite-step clock. g_p is concave in the source curvature, so the average of two compressed sources (compress, then mix) lies below the compression of their average (mix, then compress); under gradient flow (dashed) the two coincide.

Theorem C.1 (Exact constituents need not compose)**.** Let $a_1 \neq a_2$, $\alpha \in (0, 1)$, and distill each source exactly using g_p. Under the correct union mass α,

$$\alpha g_p(a_1) + (1 - \alpha)g_p(a_2) < g_p(\alpha a_1 + (1 - \alpha)a_2)\,. \tag{26}$$

Thus $\epsilon_1 = \epsilon_2 = 0$ but $\Gamma^\star_{\text{joint}} = \mathcal{E}_{\text{ind}} > 0$ for every non-degenerate initialization distribution around u.

Proof of Theorem C.1. Direct differentiation yields

$$g_p''(a) = -p(p - 1)\eta(1 - \eta a)^{p-2} < 0. \tag{27}$$

Strict concavity gives

$$\bar{s} := \alpha g_p(a_1) + (1 - \alpha)g_p(a_2) < g_p(\bar{a}) =: s_{\text{joint}}, \quad \bar{a} := \alpha a_1 + (1 - \alpha)a_2. \tag{28}$$

The independent endpoint error is therefore

$$\mathcal{E}_{\text{ind}} = \mathbb{E}[(\theta_0 - u)^2]\left[(1 - \eta\bar{s})^K - (1 - \eta\bar{a})^T\right]^2 > 0. \tag{29}$$

The exact joint curvature is s_{joint}, realizable by one squared-loss example, so $\mathcal{E}^\star_{\text{joint}} = 0$. Each constituent is exact by construction, hence $\epsilon_1 = \epsilon_2 = 0$. □

Figure 6 illustrates the single-mode repair and the two-mode residual floor.

Proposition C.2 (Mode-wise reweighting compatibility)**.** For two scalar modes with $a_{1j} < a_{2j}$, define

$$\beta_j^* = \frac{g_p(a_{2j}) - g_p(\alpha a_{1j} + (1 - \alpha)a_{2j})}{g_p(a_{2j}) - g_p(a_{1j})}. \tag{30}$$

Then $\beta_j^* \in (0, 1)$ is the unique source-1 synthetic weight that repairs mode j. A diagonal multi-mode problem is exactly repairable by one global reweighting if and only if all nontrivial β_j^*

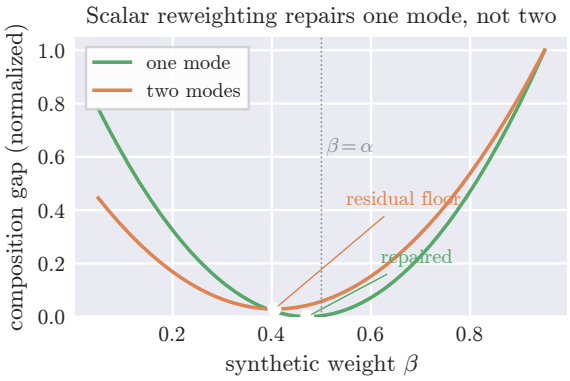

Figure 6: Scalar reweighting repairs one curvature mode, not two. Composition gap (normalized) as a function of the synthetic weight β: for a single mode there is a $\beta^* \neq \alpha$ at which the gap is zero; for two modes with different β_j^* the minimum over β is a residual floor.

coincide. For $p = 2$, this weight has the closed form

$$\beta_j^* = \alpha \frac{2 - \eta(a_{2j} + \bar{a}_j)}{2 - \eta(a_{1j} + a_{2j})}, \qquad \bar{a}_j = \alpha a_{1j} + (1 - \alpha) a_{2j}. \tag{31}$$

Consequently, in a nondegenerate mode $\beta_j^* = \alpha$ if and only if $\eta = 0$, and equality of the repair weights of two distinct modes is a measure-zero, codimension-one condition on their four curvatures.

Proof. Strict monotonicity of g_p places $g_p(\alpha a_{1j} + (1-\alpha)a_{2j})$ strictly between the two endpoint values, so solving the affine interpolation equation gives the unique weight equation 30. Simultaneous repair requires the same scalar weight to solve every modal equation, which is equivalent to equality of the β_j^*. For $p = 2$, substituting $g_2(a) = a(2 - \eta a)$ and factoring differences gives equation 31. Its denominator is positive on the stable branch. Since $\bar{a}_j - a_{1j} = (1-\alpha)(a_{2j} - a_{1j}) > 0$, the equality $\beta_j^* = \alpha$ holds exactly when $\eta = 0$. Equating the two rational expressions for modes j and j', then clearing their positive denominators, gives one nonzero polynomial equation in $(a_{1j}, a_{2j}, a_{1j'}, a_{2j'})$. Its regular zero set has codimension one, and the full zero set has Lebesgue measure zero. For general real $p > 1$, g_p and β_j^* are real analytic on the stable branch. The equality of two repair weights is again a nontrivial analytic equation: if the weight were constant over all endpoint pairs, the coincident-endpoint limit would force that constant to be α, contradicting strict concavity of g_p. Its zero set therefore also has Lebesgue measure zero. □

For any fixed real $p > 1$, the generalized binomial expansion gives

$$g_p(a) = pa - \binom{p}{2} \eta a^2 + O(\eta^2 a^3). \tag{32}$$

The leading Jensen gap is therefore

$$g_p(\bar{a}) - \sum_i \alpha_i g_p(a_i) = \binom{p}{2} \eta \operatorname{Var}_\alpha(a_i) + O(\eta^2), \tag{33}$$

which vanishes in the gradient-flow limit while predicting the curvature variance that appears in the matrix law.

C.1 Quadratic endpoint matching (E5)

The identities above describe what an exact distillate must contain. E5 asks whether a distillation algorithm finds it. For twenty random pairs of quadratic sources in $d = 12$ dimensions, a synthetic set of n_{syn} squared-loss samples per source is optimized by endpoint matching (gradient descent on

the two-step versus one-step endpoint discrepancy over random θ_0, 6,000 steps, random initialization, no access to the target statistics). Table 2 reports, per budget, the relative Frobenius error of the learned Gram matrix against the transformed statistic $2H - \eta H^2$, and the ratio of the measured composed defect to the closed form $\frac{\eta^2}{4}(H_1 - H_2)^2$ of Theorem 4.2 (Appendix D).

Table 2: E5: endpoint matching recovers the transformed statistic once the budget reaches the dimension.

n_{syn}	rank of Gram	rel. Frobenius error	endpoint loss	measured/closed-form defect	log-log corr.
4	4	48.0%	4.8×10^{-1}	11.7 ± 1.4	0.12
6	6	31.7%	2.0×10^{-1}	7.6 ± 0.9	0.16
8	8	19.9%	8.9×10^{-2}	4.8 ± 0.6	0.16
12	12	$< 10^{-6}$	1.3×10^{-31}	1.000 ± 0.000	1.00
24	12	$< 10^{-6}$	1.6×10^{-31}	1.000 ± 0.000	1.00

Twenty source pairs, $d = 12$, mean $\pm$ SD over pairs. Log-log correlation is between the measured defect and $\|(H_1 - H_2)^2\|_F$ across pairs.

At $n_{\text{syn}} \geq d$ the algorithm recovers $2H - \eta H^2$ to machine precision from a random start and the composed defect equals the closed form exactly. Below the dimension the constituents are inexact and the composed error is dominated by per-source approximation error, the budget-limited situation of the CIFAR-10 runs (Appendix J.9). In this experiment every budget-limited run gave a composed error above the closed form, but that is not a general property of approximate quadratic constituents: per-source approximation errors can cancel the structural term. With $H_1 = 1$, $H_2 = 3$, $\eta = 0.1$, equal masses, and a common optimum, the exact constituents have transformed curvatures 1.9 and 5.1, whose average 3.5 differs from the joint value 3.6. Approximate constituents with curvatures 2.0 and 5.2 each carry an error, yet their average is exactly 3.6, so $\mathcal{C}_{\text{ind}} = 0 < \mathcal{C}_{\text{struct}}$, which is consistent with Corollary 4.3 (whose positivity condition is sufficient, not necessary). Two quantities should be kept apart here: $\mathcal{C}_{\text{ind}} = \|s + e_{\text{ind}}\|$ can fall below $\|s\|$ only through cancellation between the composed set's error and the structural term, whereas $\widehat{\Gamma}_{\text{RMS}} = \|s + e_{\text{ind}}\| - \|e_{\text{joint}}\|$ also depends on the joint set's error. The ratios near one in the network budget sweep (Table 13) concern the first quantity.

D Exact Matrix Composition Laws

D.1 Two-source identity

For equal source masses and common optimum θ^*, let $\bar{H} = (H_1 + H_2)/2$. The independent one-step contraction is

$$A_{\text{ind}} = I - \eta \left(2\bar{H} - \eta \frac{H_1^2 + H_2^2}{2} \right), \tag{34}$$

whereas the real-union two-step contraction is

$$A_R = (I - \eta \bar{H})^2 = I - 2\eta \bar{H} + \eta^2 \bar{H}^2. \tag{35}$$

Thus

$$A_{\text{ind}} - A_R = \eta^2 \left(\frac{H_1^2 + H_2^2}{2} - \bar{H}^2 \right) = \frac{\eta^2}{4}(H_1 - H_2)^2. \tag{36}$$

No commutativity is used in the last identity. For $K = 1$ the composed one-step map is linear in the synthetic losses, so with exact constituents $\Phi_{\text{ind}}(\theta_0) = \sum_i \alpha_i \Phi_{D_i}^{T,\eta}(\theta_0)$ and the defect coincides with the drift between averaging local T-step updates and taking T steps on the pooled data, the parameter-averaging drift of federated optimization (Li et al., 2020). The departure specific to distillation appears through the transformed statistics and, for $K > 1$, through the compounded contractions of Theorem E.2.

D.2 Multi-source identity

For arbitrary positive masses,

$$\sum_i \alpha_i (H_i - \bar{H})^2 = \sum_i \alpha_i H_i^2 - \bar{H}^2, \tag{37}$$

$$\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j (H_i - H_j)^2 = \sum_i \alpha_i H_i^2 - \bar{H}^2. \tag{38}$$

Every summand in the first representation is of the form X^2 for a symmetric X, proving $V_H \succeq 0$. Moreover, for a vector v,

$$v^\top V_H v = \sum_i \alpha_i \|(H_i - \bar{H})v\|_2^2. \tag{39}$$

It follows that $V_H = 0$ exactly when all positive-mass Hessians are equal.

Scope of positive semidefiniteness. The identities above require only symmetric H_i: each $(H_i - \bar{H})^2$ remains positive semidefinite even if H_i is indefinite. Positive semidefiniteness is instead used for finite-dataset realizability, the positive contraction branch, and the rank-budget interpretation. Consequently, an exact-loss Hessian can still define the local algebraic proxy, while a generalized Gauss–Newton matrix supplies a PSD surrogate aligned with the signed and realizability results.

D.3 Curvature and affine statistics

Proof of Theorem 4.2. The independent synthetic union has statistics

$$H_{\text{ind}} = 2\bar{H} - \eta \sum_i \alpha_i H_i^2, \qquad b_{\text{ind}} = 2\bar{b} - \eta \sum_i \alpha_i H_i b_i. \tag{40}$$

The exact joint one-step statistics are

$$H_J = 2\bar{H} - \eta \bar{H}^2, \qquad b_J = (2I - \eta \bar{H})\bar{b}. \tag{41}$$

Hence $H_{\text{ind}} = H_J - \eta V_H$ and $b_{\text{ind}} = b_J - \eta w$. A one-step affine GD map is $(I - \eta H)\theta + \eta b$, which gives equation 5. Write $\theta_0 = \mu_0 + u_0$, with $\mathbb{E}[u_0] = 0$ and $\mathbb{E}[u_0 u_0^\top] = \Sigma_0$. Squaring and taking expectation yields equation 6. □

If $b_i = H_i \theta^*$ for all i, then

$$w = \left(\sum_i \alpha_i H_i^2 - \bar{H}^2 \right) \theta^* = V_H \theta^*, \tag{42}$$

which gives equation 7.

Corollary D.1 (Distribution-aware zero-gap characterization)**.** In the setting of Theorem 4.2, the composition error is zero if and only if

$$V_H \Sigma_0^{1/2} = 0 \quad \text{and} \quad V_H \mu_0 = w. \tag{43}$$

Under a common optimum this is equivalent to $V_H M_0^{1/2} = 0$. If $M_0 \succ 0$, exact composition holds iff all positive-mass source Hessians are equal.

Proof of Corollary D.1. The identity equation 6 is a sum of two nonnegative terms, with

$$\operatorname{tr}(V_H^2 \Sigma_0) = \|V_H \Sigma_0^{1/2}\|_F^2. \tag{44}$$

It is therefore zero exactly under the two conditions in equation 43. With a common optimum, $w = V_H \theta^*$, and the same sum is $\|V_H M_0^{1/2}\|_F^2$. If $M_0 \succ 0$, this vanishes exactly when $V_H = 0$, which by the sum-of-squares identity in Appendix D holds exactly when every positive-mass H_i equals $\bar{H}$. The same conclusion holds without a common optimum whenever $\Sigma_0 \succ 0$: the first condition in equation 43 then forces $V_H = 0$, hence $H_i = \bar{H}$ for all positive-mass sources, and substituting into w gives $w = \bar{H}\bar{b} - \bar{H}\bar{b} = 0$, so the second condition holds for every μ_0. Affine compatibility is an independent requirement only when Σ_0 is rank-deficient, for instance at a fixed initialization $\Sigma_0 = 0$, where $V_H \mu_0 = w$ can hold with $w \neq 0$. □

Why union does not average minimizers. Dataset union acts linearly on (H, b), not on $u = H^{-1}b$. If two Hessians are uniformly positive definite with minimum eigenvalue at least $m > 0$, then

$$\|\widehat{u} - u^{\circ}\|_2 \leq \frac{1}{m} \left(\|\Delta b\|_2 + \|\Delta H\|_2 \|u^{\circ}\|_2 \right). \tag{45}$$

Without this margin, small affine-statistic error does not imply minimizer closeness.

Relation to Jensen's inequality. Strict concavity of the scalar clock g_p supplies the one-dimensional case (Theorem C.1) and nothing more. The matrix law (Theorem 4.2) is an exact identity for noncommuting Hessians rather than an inequality: it characterizes the endpoint defect, yields a positive-semidefinite curvature discrepancy, separates a curvature-variance term from an affine-conflict term, and identifies which components admit scalar repair (Appendix H). The arbitrary-ratio decomposition (Theorem E.2) shows that curvature variance remains the leading term under a common optimum, and the strong-monotonicity lower bound (Lemma E.1) rules out endpoint cancellation across K steps without any Loewner ordering. When the Hessians commute, spectral Jensen recovers a signed refinement (Theorem E.5). In general no operator-concavity statement is assumed or needed, and P_{comp} follows from the identity, not from the inequality.

D.4 Momentum and minibatch stochasticity

The two-to-one identity of Theorem 4.2 is stated for full-batch gradient descent. The following lemma records that the same discrepancy arises under heavy-ball momentum and, in expectation, under independent minibatches. Adaptive preconditioners such as Adam, whose effective step depends on the gradient history, are not covered.

Lemma D.2 (Two-to-one identity under momentum and minibatches)**.** Let the sources satisfy (A1) with masses α_i.

(i) Let the real training use heavy-ball momentum with a fixed coefficient $\beta \geq 0$ and zero initial momentum, $\theta_{t+1} = \theta_t - \eta(H\theta_t - b) + \beta(\theta_t - \theta_{t-1})$ with $\theta_{-1} = \theta_0$, and let one synthetic step be a plain gradient step. The synthetic statistics that reproduce two real steps on source i exactly are $\widetilde{H}_i = (2+\beta)H_i - \eta H_i^2$ and $\widetilde{b}_i = ((2+\beta)I - \eta H_i)b_i$, and the union of these exact distillates satisfies $\Phi_{\mathrm{ind}}(\theta_0) - \Phi_R(\theta_0) = \eta^2(V_H\theta_0 - w)$, the same discrepancy as in equation 5.

(ii) Let each real step use an independent unbiased minibatch, with statistics $(H^{(t)}, b^{(t)})$, $\mathbb{E}[H^{(t)}] = H$, $\mathbb{E}[b^{(t)}] = b$, independent of θ_0 and of each other. Then the expected two-step endpoint equals the full-batch endpoint, $\mathbb{E}[\theta_2 \mid \theta_0] = (I - \eta H)^2\theta_0 + \eta(2I - \eta H)b$, so the statistics of Theorem 4.2 reproduce it exactly and the discrepancy of the composed one-step map from the expected real-union endpoint is again $\eta^2(V_H\theta_0 - w)$. Against the stochastic reference, the mean squared endpoint error of any deterministic one-step candidate is its error against the expected endpoint plus the reference variance $\mathbb{E}\|\theta_2 - \mathbb{E}[\theta_2 \mid \theta_0]\|^2$, which does not depend on the candidate and cancels in the composition gap $\Gamma^{\star}_{\mathrm{joint}}$.

Proof. (i) With zero initial momentum the first step is $\theta_1 = (I - \eta H)\theta_0 + \eta b$ and $\theta_1 - \theta_0 = -\eta(H\theta_0 - b)$. The second step gives $\theta_2 = (I - \eta H)\theta_1 + \eta b + \beta(\theta_1 - \theta_0) = [(I - \eta H)^2 - \beta\eta H]\theta_0 + \eta[(2+\beta)I - \eta H]b$. A plain step with statistics $(\widetilde{H}, \widetilde{b})$ maps θ_0 to $(I - \eta\widetilde{H})\theta_0 + \eta\widetilde{b}$, and equating the two maps for all θ_0 gives $\widetilde{H} = (2+\beta)H - \eta H^2$ and $\widetilde{b} = ((2+\beta)I - \eta H)b$. Applying the same rule to the union, whose statistics are $(\bar{H}, \bar{b})$, the composed and joint statistics differ by $\sum_i \alpha_i \widetilde{H}_i - \widetilde{H}_{\cup} = -\eta(\sum_i \alpha_i H_i^2 - \bar{H}^2) = -\eta V_H$ and $\sum_i \alpha_i \widetilde{b}_i - \widetilde{b}_{\cup} = -\eta(\sum_i \alpha_i H_i b_i - \bar{H}\bar{b}) = -\eta w$; the terms carrying β are linear in (H, b) and cancel. One synthetic step with these differences yields $\eta^2(V_H\theta_0 - w)$, as in the proof of Theorem 4.2. (ii) Two minibatch steps give $\theta_2 = (I - \eta H^{(2)})(I - \eta H^{(1)})\theta_0 + \eta(I - \eta H^{(2)})b^{(1)} + \eta b^{(2)}$. Taking the conditional expectation and using the independence of the two batches gives the full-batch expression. For a deterministic candidate endpoint $\phi(\theta_0)$, $\mathbb{E}\|\phi - \theta_2\|^2 = \|\phi - \mathbb{E}[\theta_2 \mid \theta_0]\|^2 + \mathbb{E}\|\theta_2 - \mathbb{E}[\theta_2 \mid \theta_0]\|^2$ because the cross term has zero mean; the second term is the same for the composed candidate and for every joint candidate, so it cancels in $\mathcal{E}_{\mathrm{ind}} - \mathcal{E}^{\star}_{\mathrm{joint}}$. □

E General Finite-Step Jensen Gap

E.1 General noncommutative arbitrary-ratio decomposition

Lemma E.1 (Strong monotonicity of positive matrix powers)**.** If symmetric matrices $X, Y \succeq qI$, with $q > 0$, then for every integer $K \geq 1$,

$$\|X^K - Y^K\|_F \geq Kq^{K-1}\|X - Y\|_F. \tag{46}$$

No commutativity between X and Y is required.

Proof. Let $E = X - Y$ and $Z_t = Y + tE \succeq qI$. The Fréchet derivative is

$$D(Z_t^K)[E] = \sum_{\ell=0}^{K-1} Z_t^\ell E Z_t^{K-1-\ell}. \tag{47}$$

For each summand,

$$\left\langle E, Z_t^\ell E Z_t^{K-1-\ell} \right\rangle_F = \left\| Z_t^{\ell/2} E Z_t^{(K-1-\ell)/2} \right\|_F^2 \tag{48}$$

$$\geq q^{K-1}\|E\|_F^2. \tag{49}$$

Integrating the derivative along Z_t and applying Cauchy–Schwarz gives

$$\|X^K - Y^K\|_F \|E\|_F \geq \langle X^K - Y^K, E\rangle_F \geq Kq^{K-1}\|E\|_F^2, \tag{50}$$

which proves equation 46. □

Centered matrix-power bound. Let symmetric H_i satisfy $\|H_i\|_2 \leq L$, $\bar{H} = \sum_i \alpha_i H_i$, and $\Delta_i = H_i - \bar{H}$. For every integer $k \geq 2$,

$$\left\| \bar{H}^k - \sum_i \alpha_i H_i^k \right\|_F \leq \binom{k}{2} L^{k-2} \sum_i \alpha_i \|\Delta_i\|_2 \|\Delta_i\|_F$$

$$\leq \binom{k}{2} L^{k-2} \operatorname{tr}(V_H). \tag{51}$$

Proof. For $F_k(X) = X^k$, the second Fréchet derivative is

$$D^2 F_k(X)[E, E] = 2 \sum_{0 \leq a < b \leq k-1} X^a E X^{b-a-1} E X^{k-1-b}. \tag{52}$$

Along $X_t = \bar{H} + t\Delta_i = (1-t)\bar{H} + tH_i$, one has $\|X_t\|_2 \leq L$. Each of the $k(k-1)$ ordered insertion terms is therefore bounded in Frobenius norm by $L^{k-2}\|\Delta_i\|_2\|\Delta_i\|_F$. Taylor's formula with integral remainder gives

$$H_i^k = \bar{H}^k + DF_k(\bar{H})[\Delta_i] + \int_0^1 (1-t) D^2 F_k(X_t)[\Delta_i, \Delta_i] dt. \tag{53}$$

Averaging cancels the first derivative since $\sum_i \alpha_i \Delta_i = 0$. The integral contributes the factor $\int_0^1 (1-t)dt = 1/2$, leaving $\binom{k}{2}$. Finally,

$$\sum_i \alpha_i \|\Delta_i\|_2 \|\Delta_i\|_F \leq \sum_i \alpha_i \|\Delta_i\|_F^2 = \operatorname{tr}(V_H), \tag{54}$$

which proves equation 51. □

For $p = T/K > 1$, symmetric H with $\|H\|_2 \leq L$, and $\eta L < 1$, the spectral transform $f_p(H) := (I - (I - \eta H)^p)/\eta$ satisfies $(I - \eta f_p(H))^K = (I - \eta H)^T$. The following theorem separates the universal leading term from higher-order, possibly noncommutative, corrections for every real compression ratio.

Theorem E.2 (Arbitrary-ratio defect decomposition)**.** Let real $p > 1$, symmetric H_i with $\|H_i\|_2 \le L$, $\eta L < 1$, and suppose the sources share an optimum. Define

$$H_{\text{ind}} = \sum_i \alpha_i f_p(H_i), \quad H_J = f_p(\bar{H}), \quad J_p = H_J - H_{\text{ind}}. \tag{55}$$

Let $r_H := \operatorname{tr}(V_H)^2/\|V_H\|_F^2$ be the effective rank (participation ratio) of $V_H \neq 0$, so that $\operatorname{tr}(V_H) = \sqrt{r_H}\,\|V_H\|_F$ for $V_H \succeq 0$ and $1 \le r_H \le \operatorname{rank}(V_H)$, and, for $0 \le x < 1$, define

$$c_p(x) := \sum_{k=3}^{\infty} \left|\binom{p}{k}\right| \binom{k}{2} x^{k-3}. \tag{56}$$

Without any commutativity assumption,

$$J_p = \binom{p}{2}\eta V_H + R_p, \tag{57}$$

$$R_p = \sum_{k=3}^{\infty}(-1)^{k+1}\binom{p}{k}\eta^{k-1}\left(\bar{H}^k - \sum_i \alpha_i H_i^k\right),$$

$$\begin{aligned}\|R_p\|_F &\le \eta^2 L c_p(\eta L)\operatorname{tr}(V_H)\\ &= \eta^2 L c_p(\eta L)\sqrt{r_H}\|V_H\|_F.\end{aligned} \tag{58}$$

Let $B_{\text{ind}} = I - \eta H_{\text{ind}}$ and $B_J = I - \eta H_J$, and write $A_{\text{ind}} := B_{\text{ind}}^K$ and $A_R := B_J^K$ for the K-step contractions. The endpoint defect is exactly

$$A_{\text{ind}} - A_R = \eta \sum_{\ell=0}^{K-1} B_{\text{ind}}^{K-1-\ell} J_p B_J^{\ell}. \tag{59}$$

For $K = 1$,

$$\mathcal{E}_{\text{ind}} = \eta^2 \|J_p M_0^{1/2}\|_F^2, \tag{60}$$

For arbitrary K, if $M_0 \succeq \lambda_0 I$ and $q := (1-\eta L)^p$, then

$$\sqrt{\mathcal{E}_{\text{ind}}} \ge K\eta q^{K-1}\sqrt{\lambda_0}\|J_p\|_F \ge K\eta^2 q^{K-1}\sqrt{\lambda_0}\|V_H\|_F\left[\binom{p}{2} - \eta L c_p(\eta L)\sqrt{r_H}\right]_+. \tag{61}$$

Proof of Theorem E.2. Because $\|\eta H\|_2 < 1$, the generalized binomial series converges absolutely in operator norm and gives

$$f_p(H) = \sum_{k=1}^{\infty}(-1)^{k+1}\binom{p}{k}\eta^{k-1}H^k. \tag{62}$$

For integer p, the series terminates at $k = p$. No product between distinct source Hessians occurs, so no commutativity assumption is needed. Subtracting the weighted source expansion from the expansion at $\bar{H}$, the $k = 1$ terms cancel. Since

$$\bar{H}^2 - \sum_i \alpha_i H_i^2 = -V_H, \tag{63}$$

the $k = 2$ term is $\binom{p}{2}\eta V_H$, and the remaining terms are exactly R_p in equation 57. Applying equation 51 termwise gives

$$\begin{aligned}\|R_p\|_F &\le \eta^2 L \operatorname{tr}(V_H)\sum_{k=3}^{\infty}\left|\binom{p}{k}\right|\binom{k}{2}(\eta L)^{k-3}\\ &= \eta^2 L c_p(\eta L)\operatorname{tr}(V_H).\end{aligned} \tag{64}$$

The series defining c_p converges for $\eta L < 1$. Since $V_H \succeq 0$, $\operatorname{tr}(V_H) = \sqrt{r_H}\,\|V_H\|_F$ by the definition of the effective rank, proving equation 58. (With the algebraic rank $\rho = \operatorname{rank} V_H$ one has instead $\operatorname{tr}(V_H) \le \sqrt{\rho}\|V_H\|_F$. Since $r_H \le \rho$, the effective-rank form is the sharper of the two.)

The exact joint contraction is $B_J = I - \eta f_p(\bar{H}) = (I - \eta\bar{H})^p$, hence $B_J^K = (I - \eta\bar{H})^T = A_R$. Also $B_{\text{ind}} - B_J = \eta J_p$. The difference-of-powers identity gives

$$B_{\text{ind}}^K - B_J^K = \eta \sum_{\ell=0}^{K-1} B_{\text{ind}}^{K-1-\ell} J_p B_J^\ell, \tag{65}$$

establishing equation 59. When $K = 1$, the centered endpoint difference is $\eta J_p(\theta_0 - \theta^*)$, which yields equation 60.

For general K, set $D = B_{\text{ind}}^K - B_J^K$. Since D is symmetric and $M_0 \succeq \lambda_0 I$,

$$\sqrt{\mathcal{E}_{\text{ind}}} = \|DM_0^{1/2}\|_F \geq \sqrt{\lambda_0}\|D\|_F. \tag{66}$$

Moreover,

$$B_{\text{ind}} = \sum_i \alpha_i (I - \eta H_i)^p \succeq qI, \qquad B_J = (I - \eta\bar{H})^p \succeq qI, \quad q = (1 - \eta L)^p. \tag{67}$$

Lemma E.1, together with $B_{\text{ind}} - B_J = \eta J_p$, implies

$$\|D\|_F \geq K\eta q^{K-1}\|J_p\|_F. \tag{68}$$

Finally, the reverse triangle inequality combined with equation 58 gives

$$\|J_p\|_F \geq \eta\|V_H\|_F \left[\binom{p}{2} - \eta L c_p(\eta L)\sqrt{r_H}\right]_+, \tag{69}$$

proving equation 61. □

Figure 7 visualizes the sufficient-separation bracket $s_{p,r}(x) = \binom{p}{2} - xc_p(x)\sqrt{r}$. Positivity certifies a nonzero lower bound, whereas nonpositivity only means that this sufficient condition is uninformative. The exact $p = 2$ case is omitted because $c_2 \equiv 0$ and the bracket is one.

E.2 Proof of the flow-limit boundary

Proposition E.3 (Flow-limit boundary)**.** Fix $p > 1$, bounded symmetric source Hessians, and a common optimum. For exact quadratic constituents, the RMS composition gap is $O(\eta^2)$ for fixed K, and $O(\eta)$ under $K\eta \to \tau/p$. Hence the finite-step Jensen mechanism vanishes as $\eta \to 0$.

Proof of Proposition E.3. For bounded symmetric H and fixed real $p > 1$, the generalized binomial series is uniform for sufficiently small η and gives

$$f_p(H) = pH - \binom{p}{2}\eta H^2 + O(\eta^2). \tag{70}$$

Consequently,

$$J_p = f_p(\bar{H}) - \sum_i \alpha_i f_p(H_i) = \binom{p}{2}\eta V_H + O(\eta^2), \tag{71}$$

so $\|J_p\|_2 = O(\eta)$. Under a common optimum, the endpoint identity in Theorem E.2 and bounded contraction factors imply

$$\mathcal{C}_{\text{struct}} \leq \eta K\|J_p\|_2\sqrt{\text{tr}(M_0)}. \tag{72}$$

This is $O(\eta^2)$ for fixed K. If $K\eta \to \tau/p$, the prefactor ηK remains bounded and the same expression is $O(\eta)$. In both regimes the finite-step composition mechanism vanishes. □

E.3 Different optima via affine augmentation

For source i, introduce the homogeneous-coordinate one-step map and its generator

$$\mathsf{A}_i = \begin{bmatrix} I - \eta H_i & \eta b_i \\ 0 & 1 \end{bmatrix} = I_{d+1} + \eta\mathsf{G}_i, \qquad \mathsf{G}_i = \begin{bmatrix} -H_i & b_i \\ 0 & 0 \end{bmatrix}. \tag{73}$$

Write $\bar{\mathsf{A}} = \sum_i \alpha_i \mathsf{A}_i$ and $\bar{\mathsf{G}} = \sum_i \alpha_i \mathsf{G}_i$.

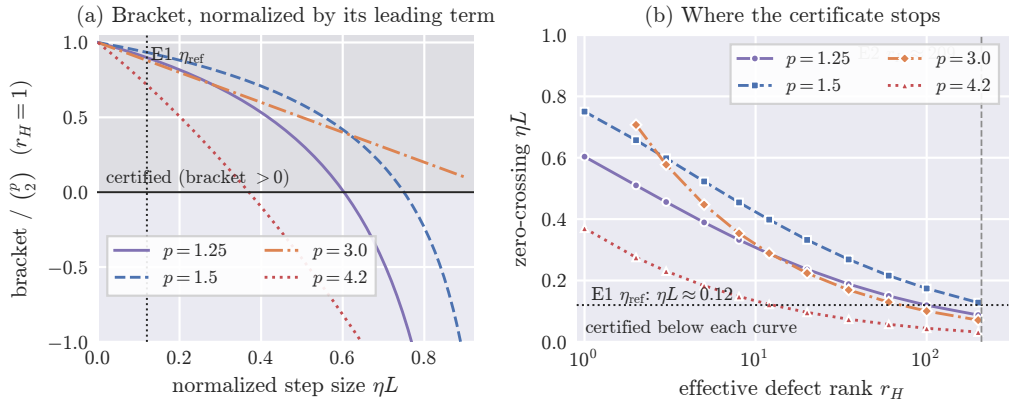

Figure 7: Where the arbitrary-ratio lower bound is certified. The zero-crossing of Theorem E.2's sufficient-separation bracket as a function of effective defect rank, for several compression ratios; the bound is certified below each curve, and larger p, step size, or rank can make it uninformative even though the exact composition gap remains nonnegative. The dotted line marks $\eta L \approx 0.12$, the value of the ReLU 2,000-example subset diagnostic at η_{ref} (the calibration cells of Figure 9 lie at $\eta L = 0.04$–0.11); the independent-versus-joint comparison operates at $\eta L \approx 1$, outside the range, where only the exact $p = 2$ identity is invoked.

Proposition E.4 (Exact affine arbitrary-ratio law)**.** Let $H_i \succeq 0$, $b_i \in \operatorname{range}(H_i)$, and $\eta\|H_i\|_2 < 1$. For any real $p > 1$, exact source-wise time compression replaces A_i by its principal power A_i^p. Let

$$\mathsf{A}_{\text{ind}} = \sum_i \alpha_i \mathsf{A}_i^p, \qquad \mathsf{A}_J = \left(\sum_i \alpha_i \mathsf{A}_i\right)^p, \qquad \mathsf{D}_p = \mathsf{A}_{\text{ind}} - \mathsf{A}_J. \tag{74}$$

Then, without common optima or commutativity,

$$\mathsf{D}_p = \sum_{k=2}^{\infty} \binom{p}{k} \eta^k \left(\sum_i \alpha_i \mathsf{G}_i^k - \bar{\mathsf{G}}^k\right), \tag{75}$$

$$\mathsf{A}_{\text{ind}}^K - \mathsf{A}_J^K = \sum_{\ell=0}^{K-1} \mathsf{A}_{\text{ind}}^{K-1-\ell} \mathsf{D}_p \mathsf{A}_J^{\ell}. \tag{76}$$

For $p = 2$, the augmented defect matrix is exactly

$$\mathsf{D}_2 = \eta^2 \begin{bmatrix} V_H & -w \\ 0 & 0 \end{bmatrix}, \tag{77}$$

recovering Theorem 4.2.

Proof. Because $b_i \in \operatorname{range}(H_i)$, take $u_i = H_i^{\dagger} b_i$. Translation by u_i makes A_i similar to $\operatorname{diag}(I - \eta H_i, 1)$, so its principal real power exists and is precisely the affine quadratic map in Lemma B.1. The same argument applies to the weighted union since $\bar{b} \in \operatorname{range}(\bar{H})$. Weighted union averages the affine statistics, hence it averages the one-step homogeneous-coordinate matrices. Compressing time before mixing gives $\sum_i \alpha_i \mathsf{A}_i^p$, whereas mixing before applying the same principal power gives $\bar{\mathsf{A}}^p$. Expanding $(I + \eta \mathsf{G}_i)^p$ by the generalized binomial series makes the constant and linear terms cancel. Convergence is blockwise: by induction on k,

$$\mathsf{G}_i^k = \begin{bmatrix} (-H_i)^k & (-H_i)^{k-1} b_i \\ 0 & 0 \end{bmatrix}, \qquad k \geq 1,$$

so $\|\eta^k \mathsf{G}_i^k\| \leq (\eta L)^k + \eta \|b_i\| (\eta L)^{k-1}$, and with $\eta L < 1$ the generalized binomial series $\sum_k \binom{p}{k} \eta^k \mathsf{G}_i^k$ converges absolutely in every block. This proves equation 75. The identity equation 76 is the

difference-of-powers identity. For $p = 2$, direct block multiplication gives

$$\sum_i \alpha_i \mathsf{G}_i^2 - \bar{\mathsf{G}}^2 = \begin{bmatrix} \sum_i \alpha_i H_i^2 - \bar{H}^2 & -\sum_i \alpha_i H_i b_i + \bar{H}\bar{b} \\ 0 & 0 \end{bmatrix}, \tag{78}$$

whose top blocks are V_H and $-w$. □

The matrices A_i are generally non-symmetric. Consequently, Proposition E.4 supplies an exact identity for different source optima, but Lemma E.1 does not furnish an arbitrary-K lower bound in this augmented space.

E.4 Commuting arbitrary-ratio lower bound

Theorem E.5 (Multi-source finite-step Jensen gap)**.** Let $H_1, \dots, H_m$ be pairwise commuting positive-semidefinite Hessians that share an optimum. Define $H_{\text{ind}} = \sum_i \alpha_i f_p(H_i)$ and $H_{\text{joint}} = f_p(\sum_i \alpha_i H_i)$. Then $H_{\text{ind}} \preceq H_{\text{joint}}$, strictly on every heterogeneous common eigendirection. Let $\bar{h}_j = \sum_i \alpha_i h_{ij}$, $v_j = \sum_i \alpha_i (h_{ij} - \bar{h}_j)^2$, and $m_f = \min_{x \in [0,L]} p(p-1)\eta(1-\eta x)^{p-2}$. If U is the common eigenbasis and $\sigma_j^2 = (U^\top M_0 U)_{jj}$, then

$$\mathcal{E}_{\text{ind}} \geq \sum_j \sigma_j^2 \left[\frac{K\eta m_f}{2} (1 - \eta \bar{h}_j)^{p(K-1)} v_j \right]^2. \tag{79}$$

Since all source Hessians commute, they admit a common eigenbasis. On a common mode, define the scalar function

$$f_p(x) = \frac{1 - (1 - \eta x)^p}{\eta}. \tag{80}$$

Its second derivative is

$$f_p''(x) = -p(p-1)\eta(1-\eta x)^{p-2} \leq -m_f < 0. \tag{81}$$

Lemma E.6 (Strong-concavity Jensen remainder)**.** For masses $\alpha_i \geq 0$ with $\sum_i \alpha_i = 1$, values $x_i \in [0, L]$, mean $\bar{x} = \sum_i \alpha_i x_i$, and variance $v = \sum_i \alpha_i (x_i - \bar{x})^2$,

$$f_p(\bar{x}) - \sum_i \alpha_i f_p(x_i) \geq \frac{m_f}{2} v. \tag{82}$$

Proof. The function $q(x) = f_p(x) + \frac{m_f}{2} x^2$ is concave. Jensen's inequality for q, followed by expansion of the quadratic term, proves the result. □

Proof of Theorem E.5. Applying scalar Jensen mode by mode gives $H_{\text{ind}} \preceq H_{\text{joint}}$ and the curvature difference

$$\delta_j := f_p(\bar{h}_j) - \sum_i \alpha_i f_p(h_{ij}) \geq \frac{m_f}{2} v_j. \tag{83}$$

Let $q_j = 1 - \eta f_p(\bar{h}_j) = (1 - \eta \bar{h}_j)^p$. The independent one-step contraction on mode j is $q_j + \eta\delta_j$, whereas the joint contraction is q_j. Therefore

$$(q_j + \eta\delta_j)^K - q_j^K \geq K q_j^{K-1} \eta \delta_j \tag{84}$$

$$\geq \frac{K\eta m_f}{2} (1 - \eta \bar{h}_j)^{p(K-1)} v_j. \tag{85}$$

Squaring this difference, weighting by the initialization mode variance σ_j^2, and summing over modes gives equation 79. □

The arbitrary-ratio signed theorem uses simultaneous diagonalization. The arbitrary-ratio decomposition in Theorem E.2 does not, but its higher-order remainder need not have a fixed Loewner sign. Extending the global ordering to every real $p > 1$ and noncommuting sources would require an operator-concavity statement that is not assumed here. This does not weaken the Frobenius lower bound in equation 61: strong monotonicity of positive matrix powers prevents endpoint cancellation without asserting a Loewner ordering for J_p.

F Approximate Distillation and Identifiability

F.1 Difference-of-powers bound

For square matrices X, Y,

$$X^K - Y^K = \sum_{\ell=0}^{K-1} X^{K-1-\ell}(X-Y)Y^\ell. \tag{86}$$

If $\|X\|_2, \|Y\|_2 \le \rho$, then

$$\|X^K - Y^K\|_2 \le K\rho^{K-1}\|X-Y\|_2. \tag{87}$$

This identity does not require commutativity.

F.2 Proof of affine-statistic stability

Define

$$L_K := K\eta\|\Delta H\|_2, \tag{88}$$

$$C_K := K\eta\|\Delta b\|_2 + \frac{\eta^2 K(K-1)}{2} B_b\|\Delta H\|_2, \tag{89}$$

$$\Delta_{\text{aff}} := \left[L_K^2 \operatorname{tr}(\Sigma_0) + (L_K\|m_0\|_2 + C_K)^2\right]^{1/2}. \tag{90}$$

Theorem F.1 (Affine-statistic stability)**.** Let $\mathcal{C}_{\text{struct}}$ be the RMS gap between the ideal composed map and the real-union map, and let $\widehat{\mathcal{C}}_{\text{comp}}$ be the corresponding RMS error of the actual composed map. If the ideal and actual one-step affine contractions have operator norm at most one and $\|b^\circ\|_2 \le B_b$, then

$$\left|\widehat{\mathcal{C}}_{\text{comp}} - \mathcal{C}_{\text{struct}}\right| \le \Delta_{\text{aff}}.$$

For contractions bounded by ρ, the same conclusion holds with $L_K(\rho)$ and $C_K(\rho)$ defined below.

Proof of Theorem F.1. The K-step affine map associated with (H, b) is

$$\Phi_{H,b}^{K,\eta}(\theta_0) = (I-\eta H)^K\theta_0 + \eta\sum_{t=0}^{K-1}(I-\eta H)^t b. \tag{91}$$

Equation 87 gives the linear-map difference

$$\|\widehat{B}^K - (B^\circ)^K\|_2 \le K\eta\|\Delta H\|_2 = L_K. \tag{92}$$

For the affine intercept, telescope each power to obtain

$$\left\|\eta\sum_{t=0}^{K-1}\widehat{B}^t\widehat{b} - \eta\sum_{t=0}^{K-1}(B^\circ)^t b^\circ\right\|_2 \tag{93}$$

$$\le K\eta\|\Delta b\|_2 + \frac{\eta^2 K(K-1)}{2} B_b\|\Delta H\|_2 = C_K. \tag{94}$$

Writing $\theta_0 = m_0 + u_0$ makes the centered cross term vanish, so

$$\mathbb{E}\|\widehat{\Phi}(\theta_0) - \Phi^\circ(\theta_0)\|_2^2 \le L_K^2\operatorname{tr}(\Sigma_0) + (L_K\|m_0\|_2 + C_K)^2. \tag{95}$$

The reverse triangle inequality in $L_2(P_0)$ proves the theorem's bound. □

For contractions bounded by a general ρ, replace the simplified constants by

$$L_K(\rho) = K\eta\rho^{K-1}\|\Delta H\|_2, \tag{96}$$

$$C_K(\rho) = \eta S_K(\rho)\|\Delta b\|_2 + \eta^2 T_K(\rho) B_b\|\Delta H\|_2, \tag{97}$$

$$S_K(\rho) = \sum_{t=0}^{K-1}\rho^t, \qquad T_K(\rho) = \sum_{t=1}^{K-1} t\rho^{t-1}. \tag{98}$$

F.3 From endpoint error to statistic error

Let the ideal and actual one-step contraction matrices for source i be $Y_i = I - \eta H_i^\circ$ and $X_i = I - \eta \widehat{H}_i$, and define

$$R_i = X_i^K - Y_i^K, \qquad \epsilon_i = \mathrm{tr}(R_i^2 M_0). \tag{99}$$

Assume $M_0 \succeq \lambda_0 I$, and that X_i, Y_i are symmetric positive definite with spectra in $[q_{\min}, \rho]$, where $q_{\min} > 0$. Lemma E.1, applied directly to X_i, Y_i, yields

$$\|\widehat{H}_i - H_i^\circ\|_F \le \frac{\|R_i\|_F}{K\eta q_{\min}^{K-1}} \le \frac{\sqrt{\epsilon_i}}{K\eta q_{\min}^{K-1}\sqrt{\lambda_0}}. \tag{100}$$

Thus endpoint fidelity controls curvature statistics only under full-rank initialization coverage and a fixed positive dynamics branch.

Corollary F.2 (Stability from identifiable endpoint fidelity)**.** Suppose $M_0 \succeq \lambda_0 I$, every ideal and actual one-step contraction has spectrum in $[q_{\min}, \rho]$, and the ideal and actual source statistics share a fixed point θ^*. If the squared source endpoint errors are ϵ_i, then

$$\left|\widehat{\mathcal{C}}_{\mathrm{comp}} - \mathcal{C}_{\mathrm{struct}}\right| \le \sqrt{\frac{\mathrm{tr}(M_0)}{\lambda_0}} \left(\frac{\rho}{q_{\min}}\right)^{K-1} \sum_i \alpha_i \sqrt{\epsilon_i}. \tag{101}$$

Proof of Corollary F.2. Because both ideal and actual maps fix θ^*, their source endpoint difference is $R_i(\theta_0 - \theta^*)$. The preceding display bounds each one-step contraction difference. After weighted union,

$$\|\widehat{B} - B^\circ\|_F \le \sum_i \alpha_i \|X_i - Y_i\|_F. \tag{102}$$

Applying equation 87 and $\|AM_0^{1/2}\|_F \le \|A\|_2 \sqrt{\mathrm{tr}(M_0)}$ yields

$$\mathcal{C}_{P_0}(\widehat{B}^K, (B^\circ)^K) \le \sqrt{\frac{\mathrm{tr}(M_0)}{\lambda_0}} \left(\frac{\rho}{q_{\min}}\right)^{K-1} \sum_i \alpha_i \sqrt{\epsilon_i}. \tag{103}$$

The reverse triangle inequality in $L_2(P_0)$ proves equation 101. □

F.4 Proof of the sufficient detectability condition

Proof of Corollary 4.3. Write $E_i^H = \widehat{H}_i - H_i^\circ$. For $K = 1$, the source endpoint difference around the common fixed point is $-\eta E_i^H(\theta_0 - \theta^*)$, and therefore

$$\sqrt{\epsilon_i} = \eta \|E_i^H M_0^{1/2}\|_F. \tag{104}$$

The independently composed statistic perturbation is $\sum_i \alpha_i E_i^H$. Hence

$$\mathcal{C}_{P_0}(\widehat{\Phi}_{\mathrm{ind}}, \Phi_{\mathrm{ind}}^\circ) = \eta \left\| \sum_i \alpha_i E_i^H M_0^{1/2} \right\|_F \le \sum_i \alpha_i \sqrt{\epsilon_i} = \mathcal{U}_{\mathrm{src}}. \tag{105}$$

The reverse triangle inequality against the real-union map proves equation 9. When $\mathcal{C}_{\mathrm{struct}} > \mathcal{U}_{\mathrm{src}}$, the same reverse triangle inequality gives the displayed positive lower bound. Failure of this sufficient condition does not imply non-detectability, and Proposition F.3 rules out a converse based only on the ϵ_i. □

F.5 Why individual errors alone are insufficient

Proposition F.3 (No control from individual error alone)**.** No function $F : \mathbb{R}_+^m \to \mathbb{R}_+$ with $F(0) = 0$ can upper-bound all composition errors using only constituent endpoint errors. Exact scalar constituents can have unbounded composition error as their optima separate.

Proof of Proposition F.3. There is no function $F : \mathbb{R}_+^m \to \mathbb{R}_+$, with $F(0, \ldots, 0) = 0$, that upper-bounds every composition error using only individual endpoint errors. Take the one-dimensional

construction with $\eta = 1/2$, $T = 2$, $K = 1$, curvatures $h_1 = 1/2$, $h_2 = 3/2$, equal masses, and minimizers $u_1 = a$, $u_2 = -a$. Individually exact synthetic curvatures are $7/8$ and $15/8$, yet

$$\epsilon_1 = \epsilon_2 = 0, \qquad \mathcal{E}_{\text{ind}} = \frac{\sigma_0^2}{256} + \frac{a^2}{64}. \tag{106}$$

Here $\mathbb{E}[\theta_0] = 0$. For a general initialization mean μ_0, the right-hand side is $\sigma_0^2/256 + (\mu_0/16 + a/8)^2$. The first term gives a positive structural gap at $a = 0$, and the second makes the error unbounded while all constituent errors remain zero. □

G Local Nonlinear Remainder

Let

$$g = \nabla F(\theta^\circ), \qquad H = \nabla^2 F(\theta^\circ), \tag{107}$$

and

$$Q(\theta) = F(\theta^\circ) + g^\top(\theta - \theta^\circ) + \frac{1}{2}(\theta - \theta^\circ)^\top H(\theta - \theta^\circ). \tag{108}$$

Theorem G.1 (Hessian-Lipschitz local remainder)**.** Suppose $\nabla^2 F$ is ν-Lipschitz in operator norm on a radius-r ball, the nonlinear K-step gradient-descent trajectory remains in that ball, and $\|I - \eta H\|_2 \le \rho$. Then

$$\|\Phi_F^{K,\eta}(\theta_0) - \Phi_Q^{K,\eta}(\theta_0)\|_2 \le \frac{\eta\nu r^2}{2} \sum_{t=0}^{K-1} \rho^t.$$

Consequently, the nonlinear real-versus-synthetic RMS gap differs from its frozen-quadratic counterpart by at most the sum of the two remainders.

Throughout, the expansion point is the trajectory start, $\theta^\circ = \theta_0$. With gradients bounded by G along the path, the K-step displacement satisfies $r \le K\eta G = O(\eta)$, so the remainder is $O(\eta^3)$, which is the scaling assumed in Corollary G.2. For $K = 1$ the synthetic remainder vanishes identically, since the one-step update on the synthetic loss equals the one-step update on its quadratic expansion at θ_0, so $\delta_S = 0$ and the uncertainty in Corollary H.3 reduces to $\Delta_{\text{aff}} + \delta_R$.

Proof of Theorem G.1. Hessian Lipschitzness and Taylor's theorem imply, throughout the radius-r ball,

$$\|\nabla F(\theta) - g - H(\theta - \theta^\circ)\|_2 \le \frac{\nu}{2}\|\theta - \theta^\circ\|_2^2 \le \frac{\nu r^2}{2}. \tag{109}$$

Let e_t be the difference between nonlinear and frozen-quadratic iterates initialized at the same point. Their recursions give

$$\|e_{t+1}\|_2 \le \rho\|e_t\|_2 + \frac{\eta\nu r^2}{2}, \qquad e_0 = 0. \tag{110}$$

Unrolling gives equation 10. For real and synthetic maps, the reverse triangle inequality in $L_2(P_0)$ and the ordinary triangle inequality bound the difference between their nonlinear and quadratic RMS gaps by $\delta_R + \delta_S$. □

G.1 Proof of the finite-step detectability window

Corollary G.2 (Finite-step detectability window)**.** Suppose on a valid local interval that the observed RMS error satisfies $|\mathcal{C}_{\text{obs}} - \mathcal{C}_{\text{struct}}| \le \mathcal{U}_{\text{src}} + \delta_S + \delta_R$, as in Corollary 4.3 extended by the nonlinear remainders of Theorem G.1, and that $\mathcal{C}_{\text{struct}} = c_1\eta^2$, $\mathcal{U}_{\text{src}} \le c_2\eta$, and $\delta_S + \delta_R \le c_3\eta^3$, where $c_1, c_2, c_3 > 0$. Then $\mathfrak{D}_{\text{loc}}(\eta) = c_1\eta/(c_2 + c_3\eta^2)$ is maximized at $\eta^* = \sqrt{c_2/c_3}$, with value $c_1/(2\sqrt{c_2c_3})$. A nonempty certified interval exists iff $c_1^2 > 4c_2c_3$, before intersection with the stability interval.

Proof of Corollary G.2. Under the assumed bound, a positive structural signal is certified whenever $c_1\eta^2 > c_2\eta + c_3\eta^3$ (Corollary H.3 gives the same form with Δ_{aff} in place of $\mathcal{U}_{\text{src}}$). Dividing by

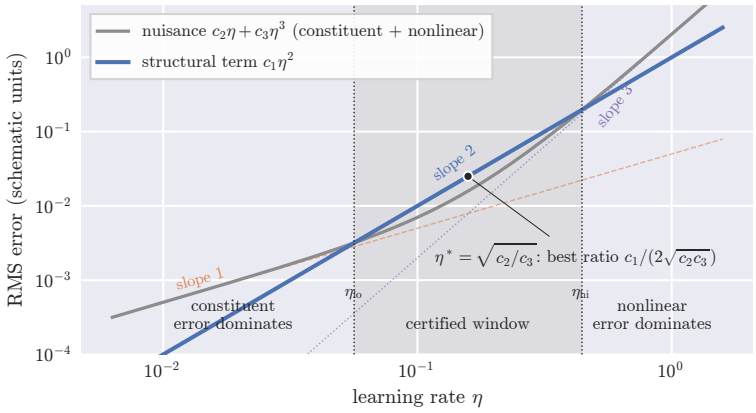

Figure 8: The detectability window (schematic). The structural term grows as η^2, constituent error as η, nonlinear error as η^3; the structural term is certifiably visible only where it exceeds their sum, the shaded interval of Corollary G.2.

the positive factor η gives $c_3\eta^2 - c_1\eta + c_2 < 0$, which has a nonempty solution interval iff its discriminant is positive. The two roots are

$$\eta_\pm = \frac{c_1 \pm \sqrt{c_1^2 - 4c_2c_3}}{2c_3}, \tag{111}$$

and the certified interval is (η_-, η_+). Direct differentiation of $c_1\eta/(c_2 + c_3\eta^2)$ gives the unique maximizer $\eta^* = \sqrt{c_2/c_3}$ and the stated maximum. The resulting interval must still be intersected with the positive-branch and local-trajectory conditions. □

The theorem requires the trajectories to stay in the local ball. This condition can be verified retrospectively for the local dynamics test, or enforced by restricting the step budget and initialization radius. The theorem does not cover a trajectory that crosses multiple curvature regimes.

Limits of the mechanism. The mechanism vanishes in the flow limit (Proposition E.3), can be masked when constituent error exceeds it (Corollary 4.3 then gives no conclusion), and is local, since equation 10 needs a slowly varying Hessian (Example 5.1). With terms scaling as $c_1\eta^2, c_2\eta, c_3\eta^3$, a certified interval in η exists iff $c_1^2 > 4c_2c_3$ (Corollary G.2, Appendix Figure 8). Which components a scalar retiming can repair is treated in Appendix H.

H Calibration and Components That Scalar Retiming Cannot Repair

Propositions H.1 and H.2 below are stated in the continuous gradient-flow limit, where retiming acts through matrix exponentials. The rank bound of Proposition H.2 carries over to discrete steps by the same argument with $(I - \eta H)^T$ in place of e^{-TH}, and its discrete form is given at the end of Appendix H.4. The exact scalar calibration of Proposition H.1 is specific to the flow limit: in discrete time a single step count generally cannot compensate a curvature scaling on every mode at once, as equation 112 shows.

H.1 Mode-specific retiming

For commuting real and synthetic Hessians, let λ_j and μ_j denote the real and synthetic curvatures on common mode j. When $0 < 1 - \eta\lambda_j < 1$ and $0 < 1 - \eta\mu_j < 1$, the continuous effective synthetic step count that exactly repairs that mode is

$$K_j^* = T\frac{\log(1 - \eta\lambda_j)}{\log(1 - \eta\mu_j)}. \tag{112}$$

A single scalar retiming repairs all task-relevant modes only if their K_j^* values coincide. If the protocol requires an integer number of GD steps, the common value must additionally be an integer or be approximated.

H.2 Uniform source clocks

Proposition H.1 (Exact calibration of uniform source clocks)**.** Let $Z = \sum_i \alpha_i/c_i$. Choosing $\beta_i = (\alpha_i/c_i)/Z$ and $\tau = TZ$ gives $e^{-\tau \sum_i \beta_i c_i H_i} = e^{-T \sum_i \alpha_i H_i}$. If the H_i are linearly independent, exactness for fixed β_i requires $\beta_i c_i/\alpha_i$ to be constant over sources.

Proof. For the stated weights, $H_S = \sum_i \beta_i c_i H_i = Z^{-1} \sum_i \alpha_i H_i$. Hence $\tau H_S = T \sum_i \alpha_i H_i$, and the two matrix exponentials agree. If the H_i are linearly independent, equality of their generators requires $\tau \beta_i c_i = T\alpha_i$ for every source. Thus $\beta_i c_i/\alpha_i$ is constant across sources. □

H.3 Repairability taxonomy

The formal clock and rank results separate defects that admit a scalar calibration from those that require a mode-dependent or higher-rank repair (Table 3). Only the semantic-mass and clock interventions are tested empirically in this paper. The other entries summarize theoretical consequences; a nonzero mean affine discrepancy is not by itself a criterion of non-repairability, since a common optimum with $\mu_0 \neq \theta^*$ gives $V_H \mu_0 - w \neq 0$ while a single scalar still repairs the map whenever the linear part and the affine offset are matched by the same factor.

Table 3: Theory-derived repairability taxonomy.

Component	Signature	Diagnostic	Minimum repair required
Global clock	common ratio	endpoint speed	scalar retiming
Source clock	coefficients c_i	source HVP scale	per-source reweighting
Curvature variance	$V_H \neq 0$	P_{comp}	mode-dependent if the required retimings differ (H.1)
Affine conflict	$V_H \mu_0 - w \neq 0$	mean shift	scalar retiming only if it matches the linear part and the affine offset simultaneously (the scalar example of Section 4 is repaired by the factor $3.6/3.5$); otherwise mode-dependent
Missing subspace	rank deficit	persistent residual	not repairable by retiming

H.4 Rank-budget proof

Proposition H.2 (Rank-budget non-repairability)**.** Let $s_1 \geq \cdots \geq s_d$ be singular values of $B = (I - e^{-TH_R}) M_0^{1/2}$. Every rank-$r$ synthetic Hessian and scalar retiming τ obey

$$\mathcal{E}(\tau) \geq \sum_{j=r+1}^{d} s_j(B)^2. \tag{113}$$

Proof of Proposition H.2. Let Q be the orthogonal projector onto $\ker(H_S)$. Since $\operatorname{rank}(H_S) \leq r$, $\operatorname{rank}(Q) \geq d - r$, and

$$Q e^{-\tau H_S} = Q. \tag{114}$$

Projecting the endpoint difference gives

$$Q(e^{-\tau H_S} - e^{-TH_R}) = Q(I - e^{-TH_R}). \tag{115}$$

The norm of an orthogonal projection lower-bounds the full endpoint error, so

$$\mathcal{E}(\tau) \geq \|QB\|_F^2, \qquad B = (I - e^{-TH_R}) M_0^{1/2}. \tag{116}$$

Among all projectors of rank at least $d - r$, the Ky Fan minimum principle gives

$$\|QB\|_F^2 = \operatorname{tr}(QBB^\top) \geq \sum_{j=r+1}^{d} s_j(B)^2, \tag{117}$$

which proves Proposition H.2. A scalar squared-loss example contributes a rank-one positive-semidefinite outer product to its Hessian, proving the sample-budget interpretation. In discrete time the projector satisfies $Q(I-\eta H_S)^K = Q$, so the same argument gives $\mathcal{E} \geq \sum_{j=r+1}^{d} s_j(B_{\text{disc}})^2$ with $B_{\text{disc}} = (I - (I - \eta H_R)^T) M_0^{1/2}$, which unifies the bound with the finite-step setting of Theorems 4.2 and E.2. □

H.5 Proof of the observable-gap certificate

Corollary H.3 (Two-to-one observable-gap certificate)**.** Assume $T = 2$ for each real source and union map, $K = 1$ for synthetic maps, and the hypotheses of Theorems F.1 and G.1. With P_{comp} evaluated at the frozen-model reference and nonlinear RMS bounds δ_S, δ_R,

$$\left| \mathcal{C}_{\text{obs}} - \sqrt{P_{\text{comp}}} \right| \leq \zeta, \qquad \zeta := \Delta_{\text{aff}} + \delta_S + \delta_R. \tag{118}$$

For partitions a, b, a proxy separation larger than $\zeta_a + \zeta_b$ certifies the corresponding observed ordering.

Proof of Corollary H.3. Theorem 4.2 and equation 11 give the ideal quadratic RMS gap $\sqrt{P_{\text{comp}}}$, without a common-optimum assumption. Theorem F.1 changes this norm by at most Δ_{aff}. Replacing the actual and reference quadratic dynamics by their nonlinear counterparts changes it by at most $\delta_S + \delta_R$, by two applications of the triangle inequality. This proves equation 118. For partitions a, b,

$$\mathcal{C}_{\text{obs}}^{(a)} \geq \sqrt{P_{\text{comp}}^{(a)}} - \zeta_a, \tag{119}$$

$$\mathcal{C}_{\text{obs}}^{(b)} \leq \sqrt{P_{\text{comp}}^{(b)}} + \zeta_b. \tag{120}$$

The displayed separation condition therefore certifies the ordering. □

I HVP Estimator and Concentration

Under a common optimum, the structural statistic is

$$P_H := \eta^4 \operatorname{tr}(V_H^2 M_0) = \eta^4 \mathbb{E}_{z \sim (0, M_0)} \|V_H z\|_2^2. \tag{121}$$

For general (p, K), its small-step leading coefficient is

$$P_H^{(p,K)} := \left[K \binom{p}{2} \eta^2 \right]^2 \operatorname{tr}(V_H^2 M_0). \tag{122}$$

The positive-branch factor in Theorem E.2 calibrates a lower bound and is not part of this leading coefficient. Two evaluations should be kept apart. The experiments evaluate $r_{\text{comp}}(\theta_0) = V_H \theta_0 - w$ directly for each initialization draw with one HVP per source from the source gradients (Appendix J.2), which includes the affine term and needs no probes. This appendix concerns the curvature-only trace P_H, estimated with random probes at $3m$ HVPs per probe; the concentration result of Theorem I.1 applies to that estimator and not to the sample mean of $\|r_{\text{comp}}(\theta_0)\|_2^2$ over $\theta_0 \sim P_0$.

For probes z_r, define

$$\widehat{P}(L) = \frac{\eta^4}{R} \sum_{r=1}^{R} \|V_H z_r\|_2^2, \qquad z_r = L g_r. \tag{123}$$

With $L = I$, this estimates $P_{\text{iso}} = \eta^4 \operatorname{tr}(V_H^2)$, and shaped probes with $LL^\top = M_0$ estimate equation 121.

I.1 HVP-ONLY ACTION

For a probe v, compute

$$q_i = H_i v, \qquad s_i = H_i q_i, \qquad \bar{q} = \sum_i \alpha_i q_i. \tag{124}$$

Then

$$V_H v = \sum_i \alpha_i s_i - \sum_i \alpha_i H_i \bar{q}. \tag{125}$$

This requires $3m$ HVP calls per probe vector v for m sources and no explicit Hessian.

I.2 GAUSSIAN CONCENTRATION

Theorem I.1 (Finite-probe concentration)**.** For independent $g_r \sim \mathcal{N}(0, I)$, $z_r = Lg_r$, and $A = L^\top V_H^2 L$, the estimator in equation 123 is unbiased and, with probability at least $1 - 2e^{-t}$,

$$|\widehat{P}_H - P_H| \le 2\eta^4 \|A\|_F \sqrt{t/R} + 2\eta^4 \|A\|_2 t/R. \tag{126}$$

For $M_0 = I$, relative error at most ε with probability $1 - \delta$ is ensured by

$$R \ge \max\left\{ \frac{16 \log(2/\delta)}{\varepsilon^2 r_4(V_H)}, \frac{4 \log(2/\delta)}{\varepsilon r_s(V_H)} \right\}, \tag{127}$$

where $r_4(V_H) = \operatorname{tr}(V_H^2)^2 / \operatorname{tr}(V_H^4)$ and $r_s(V_H) = \operatorname{tr}(V_H^2)/\|V_H\|_2^2$.

Proof of Theorem I.1. Since $\|V_H L g_r\|_2^2 = g_r^\top A g_r$, unbiasedness follows from $\mathbb{E}[g_r g_r^\top] = I$. Diagonalize $A = U \operatorname{diag}(\lambda_1, \dots, \lambda_d) U^\top$. Gaussian rotational invariance gives

$$\frac{1}{R} \sum_{r=1}^{R} g_r^\top A g_r - \operatorname{tr}(A) = \sum_{j=1}^{d} \sum_{r=1}^{R} \frac{\lambda_j}{R} (Z_{rj}^2 - 1), \tag{128}$$

where the Z_{rj} are independent standard Gaussians. Weighted chi-square tail bounds give

$$\Pr\left[X \ge 2\|A\|_F \sqrt{t/R} + 2\|A\|_2 t/R \right] \le e^{-t}, \tag{129}$$

$$\Pr\left[X \le -2\|A\|_F \sqrt{t/R} \right] \le e^{-t}. \tag{130}$$

A union bound and multiplication by η^4 prove equation 126. For $M_0 = I$, $A = V_H^2$, so

$$\|A\|_F^2 = \operatorname{tr}(V_H^4), \qquad \|A\|_2 = \|V_H\|_2^2. \tag{131}$$

Divide by $P_H = \eta^4 \operatorname{tr}(V_H^2)$ (for $M_0 = I$), take $t = \log(2/\delta)$, and make both relative-error terms at most $\varepsilon/2$ to obtain the sample complexity. □

I.3 RADEMACHER PROBES

The estimator is a Hutchinson-type stochastic trace estimator (Hutchinson, 1989). For independent raw Rademacher probes $\xi_r \in \{-1, +1\}^d$, let $B = A - \operatorname{diag}(A)$. The estimator remains unbiased and has exact variance

$$\operatorname{Var}(\widehat{P}_H) = \frac{2\eta^8}{R} \|B\|_F^2. \tag{132}$$

Indeed,

$$\xi^\top A \xi - \operatorname{tr}(A) = 2 \sum_{j<k} A_{jk} \xi_j \xi_k, \tag{133}$$

and distinct Rademacher monomials are orthogonal in L_2. Chebyshev gives

$$\Pr\left(|\widehat{P}_H - P_H| \ge \eta^4 u \right) \le \frac{2\|B\|_F^2}{R u^2}. \tag{134}$$

If $B = 0$, every raw Rademacher probe is exact.

I.4 Implementation safeguards

For the estimator to match the theory: all H_i are evaluated at the same parameter; each H_i is the Hessian of the complete empirical-mean source loss (summing batch losses with reduction $\texttt{sum}/N_i$ is valid, whereas averaging $H_b(H_b v)$ over minibatches does not equal $H_i^2 v$); the first HVP $H_i v$ is detached before the second, since automatic differentiation would otherwise introduce third derivatives; source masses α_i reflect the complete real partition even when an HVP subset is used; raw probes estimate $\mathrm{tr}(V_H^2)$ and probes divided by $\sqrt{d}$ estimate the trace divided by d; and probes used to tune a calibration are kept separate from hold-out probes used to report the predictor.

J Additional Experimental Results

All experiments were conducted on a single GPU with PyTorch. E9, E10 and E11 were run on an NVIDIA RTX A6000 with PyTorch 1.12.1.

J.1 Shared protocol

The experiment codes E1–E15 used in the text and in the supplementary code and result files name the subsections below (E5, quadratic endpoint matching, is in Appendix C.1; E15, the CIFAR-100 replication, is in Appendix J.10; the numbering has gaps at E4 and E6). We use the CIFAR-10 training split with deterministic tensor conversion and channel normalization. Every two-source partition is a disjoint cover of all 50,000 training examples. Heterogeneity is controlled through IID assignment or class-wise Dirichlet allocation (Hsu et al., 2019) with concentration 1.0, 0.5, or 0.1; the 0.5 level appears only in the official-schedule independent-versus-joint sweep. Source masses are computed from complete source sizes before any HVP subset is drawn. We store partition seeds, label histograms, source sizes, and hashes of sorted index sets.

The local evaluation network is a three-block ConvNet with GroupNorm and average pooling. The local-law validation uses GroupNorm(8); the independent-versus-joint comparison follows the official implementation with GroupNorm(4). Unless activation smoothness is itself the intervention, the network uses ReLU. For each sampled initialization θ_0, all source gradients and Hessians are evaluated at that same θ_0, and the frozen local model uses $\theta^\circ = \theta_0$. "The same checkpoint" therefore means consistency across sources within one initialization, not one parameter vector shared across initialization draws. We use exact loss Hessians rather than a generalized Gauss–Newton approximation. The main endpoint measurements use two real steps against one synthetic step, the smallest schedule in which time compression is non-affine and the defect has an exact closed form with no remainder ($R_2 = 0$ in Theorem E.2); for larger T/K or K the leading term is the same V_H up to the coefficient $K\binom{p}{2}\eta^2$ (Appendix I), so the two-step measurement isolates the mechanism rather than restricting it. The primary response is the RMS endpoint-map distance over a fixed initialization distribution, which the theorems bound and which is measurable without training an evaluation network. Test accuracy is evaluated separately as a downstream consequence, because decision margins can hide parameter-map differences.

The independent uncertainty unit is a generated source partition. Initialization seeds are repeated measurements, aggregated within partition before any between-partition statistic, and confidence intervals use a bootstrap over partitions. Within a partition, $\mathcal{C}_{\mathrm{ind}}$, $\mathcal{C}_{\mathrm{joint}}$ and $\mathcal{C}_{\mathrm{loc}}$ are root-mean-square endpoint distances over the initialization seeds, as Definition 3.1 specifies, and $\widehat{\Gamma}_{\mathrm{RMS}}$ is the difference of the two RMS values. The no-split control, the budget sweep, the SVHN replication and the short-schedule tables were aggregated as the mean over seeds of the per-seed endpoint-distance difference $\|\widehat{A}_{\mathrm{ind}} - A_R\| - \|\widehat{A}_{\mathrm{joint}} - A_R\|$, which we denote $\widehat{\Gamma}_{\mathrm{mean}}$. On the CIFAR-10 main grid the change from $\widehat{\Gamma}_{\mathrm{mean}}$ to $\widehat{\Gamma}_{\mathrm{RMS}}$ is below 3% at the label-skewed levels. For IID partitions, the absolute change is 7.8×10^{-6}. The endpoint measurements show a systematic seed effect (a variance decomposition attributes 75% of the variation to the partition, 19% to the seed and 7% to their interaction), which is why seeds are aggregated within partition; the level means change by at most 7% between three and five seeds, and the main text reports the three-seed values. The evaluation stage applies differentiable augmentation without a fixed seed, so repeated evaluation of the same synthetic set moves test accuracy by up to ± 0.35 points; accuracies are reported to one decimal.

J.2 Verification of the local prediction on the network (E1)

Configuration. The original sweep contains 15 partitions and three initialization seeds at $\eta = 0.4 \times 2^{-j}$ for $j = 0, \ldots, 9$. An extension adds 15 independent partitions at the shared points 0.0125 and 0.00078125. Scaling fits therefore use the original 15-partition sweep, whereas predictor comparisons at $\eta = 0.00078125$ use all 30 partitions. This split-input analysis avoids both discarding independent partitions and imputing missing learning rates.

For each initialization we compute

$$P_{\mathrm{comp}} = \eta^4 \left\| \sum_i \alpha_i H_i (g_i - \bar{g}) \right\|_2^2 . \tag{135}$$

The probe diagnostic $P_{\mathrm{iso}} = \eta^4 \operatorname{tr}(V_H^2)$ uses isotropic HVP probes and is reported separately. We compare both quantities with label divergence and gradient disagreement (Table 4). The estimator suite contains 15 checks, including explicit-matrix identities, minibatch invariance, source-order invariance, singular M_0, Gaussian and Rademacher probes, and spectral-norm estimation on H^2 for nearly tied positive/negative extreme eigenvalues. All checks pass.

Table 4: E1: which statistic predicts the observed discrepancy.

Predictor	Spearman ρ	partial ρ (adj. label div.)	partial ρ (adj. grad. dis.)
P_{comp}	0.991 [0.960, 0.999]	0.869 [0.568, 0.989]	0.701 [0.210, 0.971]
P_{iso}	0.955 [0.857, 0.984]	0.316 [−0.316, 0.785]	−0.001 [−0.262, 0.496]
Label divergence	0.964 [0.907, 0.985]	N/A	N/A
Gradient disagreement	0.983 [0.937, 0.995]	N/A	N/A

$n = 30$ partitions at η_{ref}; rank correlation with $\mathcal{C}_{\mathrm{loc}}$; brackets are 95% bootstrap intervals over partitions. N/A: a baseline cannot be adjusted for itself.

The high marginal correlations partly reflect the three pre-specified heterogeneity strata. The partial correlations measure residual rank association after adjusting for one baseline; they are not a complete within-level or causal decomposition. The weak adjusted result for P_{iso} shows that the isotropic curvature-only diagnostic loses useful task- or initialization-dependent information, without showing which affine component dominates.

The calibration cells (10^4-example sources, float32, three partitions per cell, one initialization seed) use a single-initialization form of the statistic. For the network initialization θ_0, the endpoint defect of Theorem 4.2 is $\eta^2 r_{\mathrm{comp}}(\theta_0)$, where $r_{\mathrm{comp}}(\theta_0) = \sum_i \alpha_i H_i (g_i - \bar{g}) = V_H \theta_0 - w$ (Eq. 11) is computed exactly by one Hessian-vector product per source, without random probes; the affine term is included, because it enters through the source gradients. The calibration ratio is the measured discrepancy $\|\Delta\Phi(\theta_0)\|$ divided by $\eta^2 \|r_{\mathrm{comp}}(\theta_0)\|$. It coincides with $\mathcal{C}_{\mathrm{obs}}/\sqrt{P_{\mathrm{comp}}}$ when the initialization distribution is concentrated at θ_0, and it is averaged over the three partitions of a cell rather than over P_0. Levels are never pooled, because the ratio can differ between levels.

Analysis. The local-law validation tests whether the local reference $\mathcal{C}_{\mathrm{loc}}$ of Section 6, the RMS distance between the weighted average of the exact source-wise two-step endpoints and the real-union two-step endpoint on the actual network (no distillation involved), follows the quadratic law. Its square should scale as η^4, and its size should match $\sqrt{P_{\mathrm{comp}}}$ (Eq. 11). On the 10^4-example calibration sources (three partitions per cell, one seed, eight rates from 7.8×10^{-4} to 0.1, each activation at its own measured L), the squared discrepancy inside $\eta L < 1$ scales with exponent 3.71 ± 0.08 for SiLU and 3.86 ± 0.06 for tanh at Dir-0.1, and 3.96 ± 0.02 and 3.93 ± 0.04 under IID (mean $\pm$ SD of per-partition fits). ReLU gives 3.49 ± 0.02 at Dir-0.1 and 3.01 under IID. Over the full sweep the exponents fall to 3.25–3.50 as the largest rates leave the perturbative window. At η_{ref} the single-initialization calibration ratios (Appendix J.2) are 0.996/0.969 for SiLU and 0.992/0.991 for tanh at IID/Dir-0.1 (SD over partitions at most 0.004), within approximately 3% of one, and they drift below one as ηL grows (Figure 9). The lower ReLU exponent is expected: a frozen Hessian misses activation crossings, whose gradient discrepancy is of lower order than the quadratic remainder, so reducing η does not restore it over the tested step-size range. Its calibration ratio is 1.00 at Dir-0.1 but 14.9 ± 1.7 at IID, where the structural signal ($\mathcal{C}_{\mathrm{loc}} \approx 10^{-5}$) approaches zero

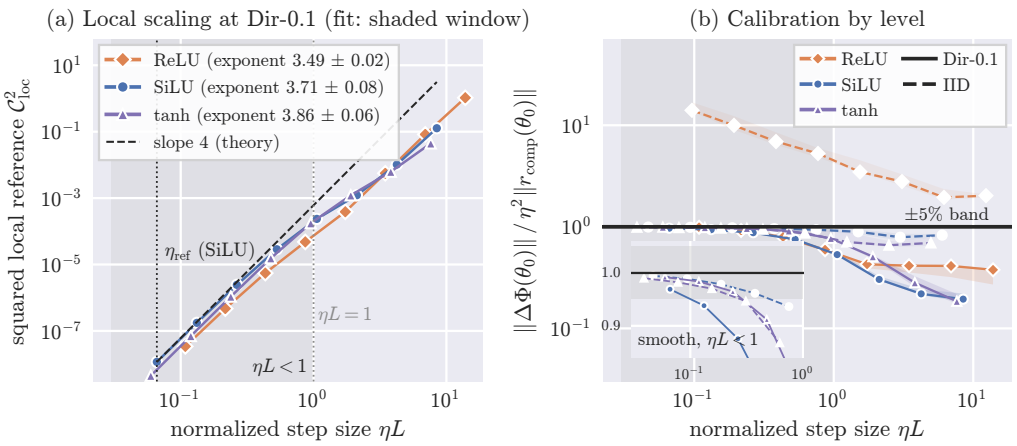

Figure 9: Local-law validation with all activations and both levels. (a) Squared discrepancy $\|\Delta\Phi(\theta_0)\|^2$ against ηL at Dir-0.1, with exponents fitted inside the shaded window $\eta L < 1$ and dotted lines at η_{ref} and $\eta L = 1$. (b) Calibration ratio $\|\Delta\Phi(\theta_0)\|/(\eta^2\|r_{\text{comp}}(\theta_0)\|)$ against ηL (solid: Dir-0.1, dashed: IID, bands over partitions), with the small-step regime enlarged in the inset. At IID the ReLU structural term is at numerical zero, so its ratio is not meaningful.

faster than the non-smooth remainder and the ratio is not meaningful; ReLU therefore serves as a boundary case for the smoothness assumption. The curvature disagreement has low effective rank ($r_H = 208.5 \pm 5.2$ in 320,010 dimensions; Appendix J.13).

A denser sweep inside the perturbative window. On the full sources, a second ReLU sweep uses nine rates from 3.9×10^{-4} to 6.25×10^{-3} in steps of $\sqrt{2}$, eight of which lie in the strict window $\eta L < 1$ for every partition and level, with five partitions per level and three initializations per partition (Figure 10). Fitting $\log \mathcal{C}^2_{\text{loc}}$ against $\log \eta$ on this fixed subset gives exponents of 3.878 $[3.867, 3.890]$ at Dir-1.0 and 3.754 $[3.689, 3.837]$ at Dir-0.1. The slope is fitted per initialization, averaged within partition, and the 95% interval comes from a bootstrap over the five partitions; on the common grid of step sizes this equals a single fit pooled over all initializations. The estimates are within 1% of the original three-point estimates, and both intervals exclude four. The 3–6% shortfall is the direction expected when the two expansion conditions, $\eta L \ll 1$ and a slowly varying Hessian, hold only marginally (Appendix J.13). The IID exponent is 3.00 $[2.97, 3.02]$; its discrepancy is dominated by terms other than the structural one and is not a test of the η^4 law. Dir-0.5 appears only in the distillation grid.

J.3 Endpoint-matching distillates on the network (E10)

Configuration. E10 applies the endpoint-matching objective of E5 to the E1 network, so that learned synthetic sets, a real network and the local law can be compared in one setting. The network is the three-block ConvNet with SiLU activation, the sources are the E1 sources (10^4 examples each) on the IID and Dir-0.1 partitions (three partitions per level), and the main setting is $T = 2$, $K = 1$ at $\eta_{\text{ref}} = 7.8 \times 10^{-4}$. For each source, a synthetic set with one, five or ten images per represented class (labels fixed, images initialized from random real images of the class) is optimized to minimize $\mathbb{E}_{\theta_0}\|\Phi^{K,\eta}_{S_i}(\theta_0) - \Phi^{T,\eta}_{D_i}(\theta_0)\|^2$ with Adam (learning rate 0.03, chosen once on the IID partition 0 at five images per class from $\{0.01, 0.03, 0.1\}$ and then frozen) for 1,100 iterations, the point at which the ten-image pilot reached within 3% of its loss after 3,000 iterations. Under Dir-0.1 each source allocates its budget to its represented classes as in E2. The joint set is trained with the same objective against $\Phi^{T,\eta}_{D_\cup}$ at the same total number of images. Training draws four initializations per iteration from a fixed bank of 64 (32 for the step-size and $(4, 2)$ supplements); evaluation uses 32 fresh initializations never used in training. The bank leaves a generalization gap over θ_0 (IID, ten images per class: matching loss 0.034 on the bank against 0.063 on the held-out initializations). To attribute the source-error plateau reported below, the Dir-0.1 cells at ten images per class were retrained on all three partitions under the four combinations of bank size (64 or 512 initializations)

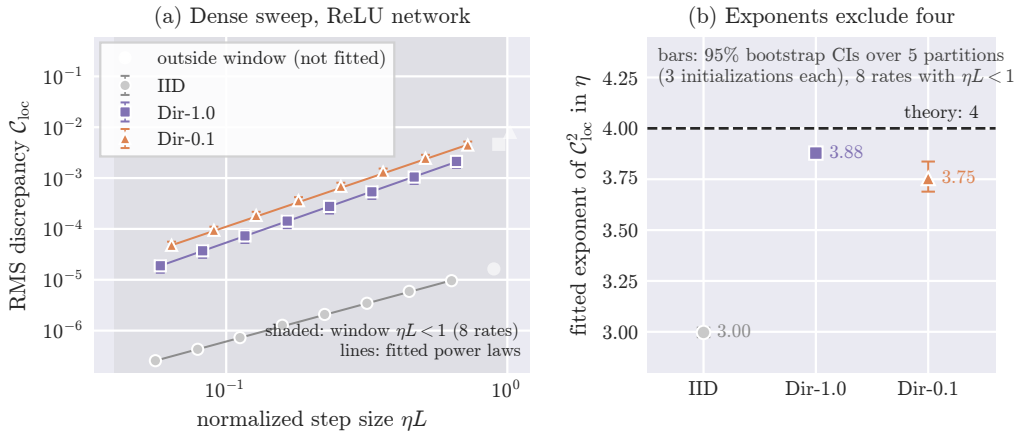

Figure 10: Dense E1 sweep on the ReLU network. (a) Partition-level mean RMS discrepancy against ηL at nine step sizes (five partitions per level with three initializations each, error bars are SDs over partitions); filled markers are the eight rates inside the strict window and lines are the fitted power laws. (b) Fitted exponents of $\mathcal{C}^2_{\text{loc}}$ in η with 95% intervals from a bootstrap over the five partitions; both skewed levels lie 3–6% below four, and the IID discrepancy is not governed by the structural term.

and iteration count (1,100 or 3,000), scored on the same 32 held-out initializations and the same real targets. The supplements vary the step size ($4\eta_{\text{ref}}$, $16\eta_{\text{ref}}$), the ratio ($T = 4$, $K = 2$) and the activation (tanh in place of SiLU) at five images per class. For $(4, 2)$ two predictions are reported: the composed-synthetic prediction with coefficient $K\binom{p}{2} = 2$ (Theorem E.2) and the real four-step local discrepancy, whose second-order coefficient is $\binom{T}{2} = 6$.

Analysis. Four findings follow (Table 6 lists the per-cell values, Figure 12 plots the Dir-0.1 rows and Figure 11 shows learned sets). First, at η_{ref} the structural prediction $\eta^2 r_{\text{comp}}(\theta_0)$ captures the local endpoint discrepancy $\Delta\Phi(\theta_0)$ in magnitude and direction on both activations: cosine 1.000 (SiLU) and 0.999 (tanh) with relative residual 0.02–0.04 on every partition, and for $(4, 2)$ the real four-step discrepancy is approximated by $6\eta^2 r_{\text{comp}}$ with cosine 0.999 and norm ratio 0.985 (IID) and 0.905 (Dir-0.1). Second, the learned sets retain a large source error: the relative source error is 0.50 at one image per class and 0.25–0.39 at five and ten (0.21–0.24 for tanh), so at Dir-0.1 the aggregate residual $\|r_{\text{src}}\|_{L^2(P_0)}$ is 23.6, 13.3 and 12.7 times $\mathcal{C}_{\text{struct}}$ at the three budgets (the bound $\mathcal{U}_{\text{src}}$ is 39–50 times), the composed discrepancy is nearly orthogonal to the prediction (cosine at most 0.05), and the inequality of Corollary 4.3, evaluated initialization by initialization, is satisfied for none of 96 initializations at η_{ref}, for $(4, 2)$ and for tanh. The count is the numerical inequality only, since the corollary's common-optimum assumption is not established on the network. The joint set has a lower error than the union in every cell, and the excess shrinks with the budget. Third, the local prediction degrades as the step size grows: at $16\eta_{\text{ref}}$ it overshoots $\Delta\Phi$ by a factor of 1.9 and $\mathcal{C}_{\text{ind}}$ falls below $\mathcal{C}_{\text{struct}}$, so the 3 of 96 initializations that satisfy the inequality there reflect the failing prediction rather than a smaller source error. Fourth, a larger initialization bank and longer training do not remove the source error (Table 5): 3,000 instead of 1,100 iterations changes the held-out relative source error by at most 0.002, a bank of 512 instead of 64 initializations lowers it from 0.380 to 0.370 while narrowing the gap between bank and held-out matching loss (0.116 against 0.144 to 0.130 against 0.137), $\mathcal{U}_{\text{pw}}/\mathcal{C}_{\text{struct}}$ moves from 38.5 to 37.5, and the inequality holds for none of 96 initializations under every condition. Other optimization choices and the objective itself were not varied.

J.4 Independent versus joint distillation (E2)

The independent-versus-joint results of Section 6 (Table 7; Figure 13) use distribution matching (DM) at its official schedule: 20,000 iterations, image learning rate 1.0, real batch 256, differentiable Siamese augmentation enabled, and IPC 10 per independently distilled source, with snapshots at iterations $\{0, 1000, 2500, 5000, 10000, 15000, 20000\}$. The joint set receives the sum of the in-

Table 5: E10: attribution of the source-error plateau. Dir-0.1, ten images per class per source, $(2, 1)$ at η_{ref}, SiLU, three partitions, all conditions scored on the same 32 held-out initializations. Bank loss is the relative matching loss of the source sets on their training bank, held-out is the relative source error on the evaluation initializations (mean and SD over partitions), errors are in units of 10^{-3}, and $\widehat{\Gamma}_{\mathrm{RMS}} = \mathcal{C}_{\mathrm{ind}} - \mathcal{C}_{\mathrm{joint}}$ is computed from the unrounded RMS values. These runs were archived with $\mathcal{U}_{\mathrm{pw}}$, the RMS over initializations of $\sum_i \alpha_i \|e_i(\theta_0)\|$, in place of $\mathcal{U}_{\mathrm{src}}$ of Table 6; on the cell common to both tables the two differ by 0.2% (3.808 against 3.814). Ineq. counts initializations as in Table 6.

Bank	Iter.	Bank loss	Held-out	$\mathcal{U}_{\mathrm{pw}}$	$\mathcal{C}_{\mathrm{ind}}$	$\mathcal{C}_{\mathrm{joint}}$	$\widehat{\Gamma}_{\mathrm{RMS}}$	$\mathcal{U}_{\mathrm{pw}}/\mathcal{C}_{\mathrm{struct}}$	ineq.
64	1,100	0.116	0.380 ± 0.008	3.81	1.25	0.99	0.26	38.5	0/96
64	3,000	0.114	0.379 ± 0.008	3.79	1.25	0.99	0.26	38.4	0/96
512	1,100	0.132	0.372 ± 0.008	3.73	1.20	0.95	0.24	37.7	0/96
512	3,000	0.130	0.370 ± 0.008	3.71	1.20	0.95	0.25	37.5	0/96

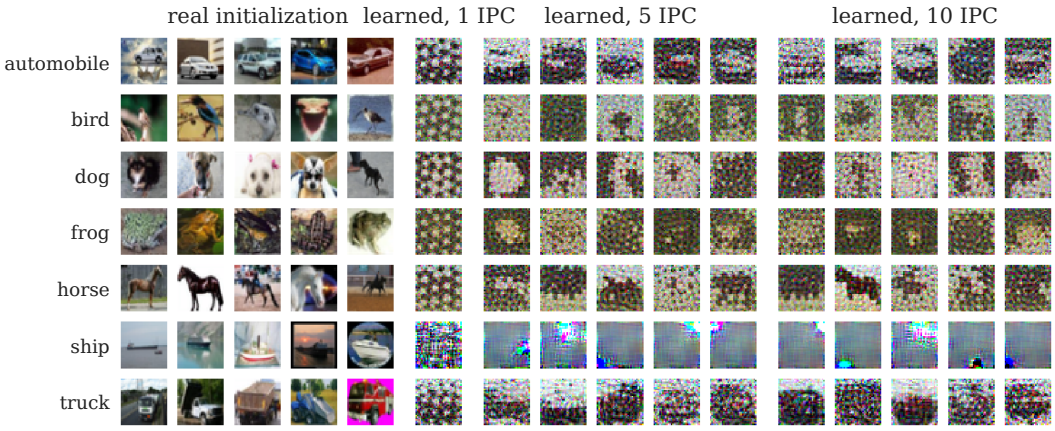

Figure 11: Endpoint-matching distillates on the SiLU network (E10; Dir-0.1 partition 0 of the E1 subset, source 0, which holds the seven classes shown). Images are initialized from real images of the class and optimized for 1,100 Adam iterations on the one-step-versus-two-step endpoint discrepancy; the columns show the initialization and the learned sets at one, five and ten images per class (up to five per class displayed). The learned images move far from their initialization (median relative displacement 1.1), since the objective constrains the induced training step and not the appearance.

dependent image budgets, which is IPC 15–20 over the union depending on how many classes each source represents, and stores the same semantic mass (Figure 14 shows the sets on one Dir-0.1 partition). Its accuracy is therefore comparable to published IPC-10 numbers only up to that budget difference. Evaluation follows DM's official protocol (1000 epochs, augmentation enabled), with three evaluation seeds averaged within each partition.

Gradient-based condensation (DC) is used only as a short-schedule optimization control (Appendix J.14). At the official 1000-iteration schedule on the full CIFAR-10 training set our DC runs reach 41.4% at IPC 10, below the published 44.9%, so they are not part of the official-schedule evidence. The short-schedule DC results are retained in the tables marked "short schedule"; their endpoint-level behaviour matches DM's, with a positive gap at Dir-0.1, and the canonical and matched-step choices of DC's IPC-dependent loop counts for the joint set agree at the endpoint level to within 0.3%. DM performs one image update per iteration and has no IPC-dependent loop structure.

Independent and joint endpoints use the same θ_0 and $\eta = 0.00625$. The endpoint protocol compares $\Phi^{1,\eta}_{S_i}$ with $\Phi^{2,\eta}_{D_i}$ for every source and $\Phi^{1,\eta}_{S_1 \oplus S_2}$ with $\Phi^{2,\eta}_{D_1 \oplus D_2}$ for the union. Hence $T = 2$ governs both real targets, while $K = 1$ governs only the synthetic maps. Each synthetic set enters with its stored class masses as loss weights and with a scalar step multiplier fitted at each evaluation initialization: for source i the measured map is $\Phi^{1,c_i\eta}_{S_i}$ with $c_i = \arg\min_c \|\Phi^{1,c\eta}_{S_i}(\theta_0) - \Phi^{2,\eta}_{D_i}(\theta_0)\|$,

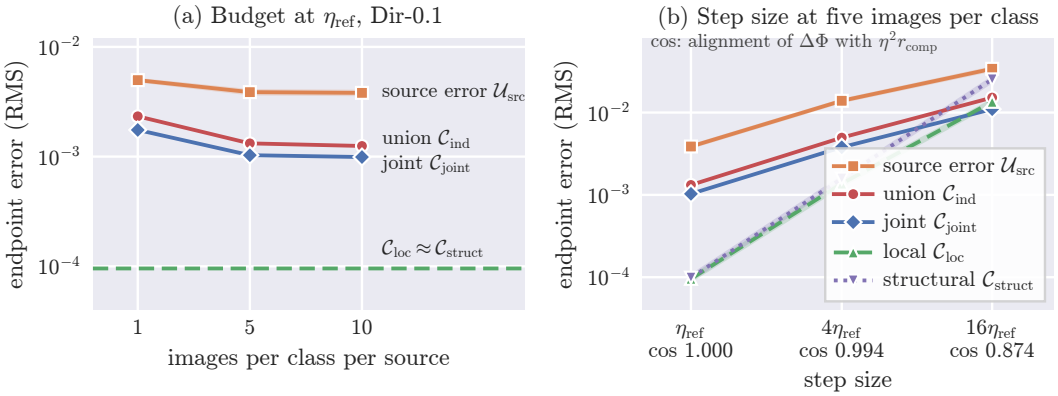

Figure 12: At small step sizes the local prediction is accurate, but the endpoint error of learned synthetic sets lies far above the local discrepancy. Endpoint-matching distillates on the SiLU network at Dir-0.1 (three partitions, RMS over 32 held-out initializations, mean and SD over partitions). (a) Weighted source error, error of the union of the learned sets and error of the jointly learned set against the image budget at η_{ref}; the local reference $\mathcal{C}_{\text{loc}}$ and the structural composition error $\mathcal{C}_{\text{struct}}$ closely agree and lie far below. (b) The same quantities against the step size at five images per class, with the cosine between $\Delta\Phi$ and $\eta^2 r_{\text{comp}}$ at each step size.

Table 6: E10: endpoint-matching distillates on the network (SiLU unless stated). Per cell, RMS over 32 held-out initializations; mean over three partitions. Errors in units of 10^{-3}. $\mathcal{U}_{\text{src}} = \sum_i \alpha_i \|\Phi_{S_i}^{K,\eta} - \Phi_{D_i}^{T,\eta}\|_{L^2(P_0)}$ is the weighted source error of Section 6 at the row's (T, K), rel. is the weighted mean over sources of each set's endpoint error divided by its own real T-step displacement, $\hat{s} = K\binom{p}{2}\eta^2 r_{\text{comp}}$ is the structural prediction (coefficient 1 for $(2,1)$ and 2 for $(4,2)$, Theorem E.2), and d_{ind} is the per-initialization error of the union of the learned sets against the real-union endpoint. The last column (ineq.) is evaluated before aggregation: it counts initializations θ_0 at which $\|\hat{s}(\theta_0)\| > \sum_i \alpha_i \|e_i(\theta_0)\|$, with $e_i(\theta_0) = \Phi_{S_i}^{K,\eta}(\theta_0) - \Phi_{D_i}^{T,\eta}(\theta_0)$.

Level	Setting	IPC	$\mathcal{U}_{\text{src}}$	rel.	$\mathcal{C}_{\text{ind}}$	$\mathcal{C}_{\text{joint}}$	$\mathcal{C}_{\text{loc}}$	$\mathcal{C}_{\text{struct}}$	$\cos(\Delta\Phi, \hat{s})$	$\cos(d_{\text{ind}}, \hat{s})$	ineq.
IID	$(2,1)$, η_{ref}	1	2.03	0.51	1.84	1.51	0.00009	0.00009	1.000	0.05	0/96
IID	$(2,1)$, η_{ref}	5	1.10	0.27	1.04	1.01	0.00009	0.00009	1.000	0.04	0/96
IID	$(2,1)$, η_{ref}	10	1.02	0.25	0.99	0.99	0.00009	0.00009	1.000	0.04	0/96
IID	$(2,1)$, $4\eta_{\text{ref}}$	5	4.11	0.26	3.86	3.71	0.0014	0.0014	0.997	0.03	0/96
IID	$(2,1)$, $16\eta_{\text{ref}}$	5	12.9	0.23	11.6	10.6	0.021	0.022	0.955	−0.04	0/96
IID	$(4,2)$, η_{ref}	5	2.16	0.27	2.04	1.97	0.0005	0.0002	0.999	0.04	0/96
Dir-0.1	$(2,1)$, η_{ref}	1	4.98	0.50	2.33	1.75	0.096	0.099	1.000	0.01	0/96
Dir-0.1	$(2,1)$, η_{ref}	5	3.87	0.39	1.32	1.03	0.096	0.099	1.000	0.04	0/96
Dir-0.1	$(2,1)$, η_{ref}	10	3.81	0.38	1.25	0.99	0.096	0.099	1.000	0.04	0/96
Dir-0.1	$(2,1)$, $4\eta_{\text{ref}}$	5	13.9	0.36	4.96	3.79	1.37	1.58	0.994	0.04	0/96
Dir-0.1	$(2,1)$, $16\eta_{\text{ref}}$	5	34.0	0.28	15.2	11.0	13.5	25.3	0.874	0.02	3/96
Dir-0.1	$(4,2)$, η_{ref}	5	7.50	0.39	2.61	2.01	0.54	0.20	0.998	0.04	0/96
IID, tanh	$(2,1)$, η_{ref}	5	1.06	0.21	0.93	0.85	0.00013	0.00013	0.999	0.01	0/96
Dir-0.1, tanh	$(2,1)$, η_{ref}	5	1.98	0.24	1.24	0.90	0.054	0.054	0.999	0.01	0/96

which has the closed form $c_i = \langle d_i, \delta_i \rangle / \|d_i\|^2$ for the raw synthetic displacement d_i and the real two-step displacement δ_i, and the joint set is calibrated in the same way against the union target. This is the per-set form of the scalar retiming of Appendix H: it uses the real two-step endpoint of the set's own target, so $\mathcal{C}_{\text{ind}}$, $\mathcal{C}_{\text{joint}}$ and every quantity derived from them in E2, E7, E9, E11 and E12 are endpoint errors after per-set scalar calibration, whereas the endpoint-matching sets of E10 are trained and evaluated at the fixed step η. The composed union is the mass-weighted sum of the calibrated source displacements. Appendix J.5 reports the same decomposition without the calibration (class-mass weights only, $c_i = 1$). The reported constituent error, including ϵ_{rel}, uses this one-step versus two-step comparison. The learning rate supports ranking but is not used for the asymptotic calibration claim.

Table 7: Separate-then-union versus joint distillation on CIFAR-10 (distribution matching, official schedule, 10 images per class per source). Operator quantities are RMS endpoint distances to the real-union two-step map, in units of 10^{-3}. Accuracies in percent.

	Endpoint errors				Test accuracy		
Source split	reference $\mathcal{C}_{\mathrm{loc}}$	separate $\mathcal{C}_{\mathrm{ind}}$	joint $\mathcal{C}_{\mathrm{joint}}$	excess $\widehat{\Gamma}_{\mathrm{RMS}}$	separate	joint	gain
Heterogeneity sweep (five partitions per level)							
IID	0.016 ± 0.001	7.41 ± 0.04	7.34 ± 0.08	0.06 ± 0.10	46.3 ± 0.9	51.6 ± 0.6	$+5.3 \pm 0.9$
Dir-1.0	4.85 ± 0.99	8.72 ± 0.29	7.34 ± 0.08	1.38 ± 0.25	46.2 ± 0.6	51.9 ± 0.5	$+5.7 \pm 0.6$
Dir-0.5	5.58 ± 0.76	8.71 ± 0.31	7.33 ± 0.06	1.38 ± 0.27	45.8 ± 1.1	51.7 ± 0.5	$+5.9 \pm 1.2$
Dir-0.1	8.09 ± 1.17	9.85 ± 1.33	7.50 ± 0.26	2.35 ± 1.44	45.8 ± 0.7	51.1 ± 1.1	$+5.3 \pm 0.7$
No-split control (three replicates)							
No split	0 (exact)	–	–	–	46.0 ± 0.4	52.6 ± 0.4	$+6.6 \pm 0.4$

Mean (SD) over partitions. Within a partition, endpoint errors are RMS over three initialization seeds and accuracies average three evaluation seeds. Gain: joint minus separate accuracy in percentage points. 0 (exact): $V_H \equiv 0$ by construction. The no-split endpoint errors were aggregated as $\widehat{\Gamma}_{\mathrm{mean}}$ and are reported in Appendix J.7 and Table 14, hence the dashes.

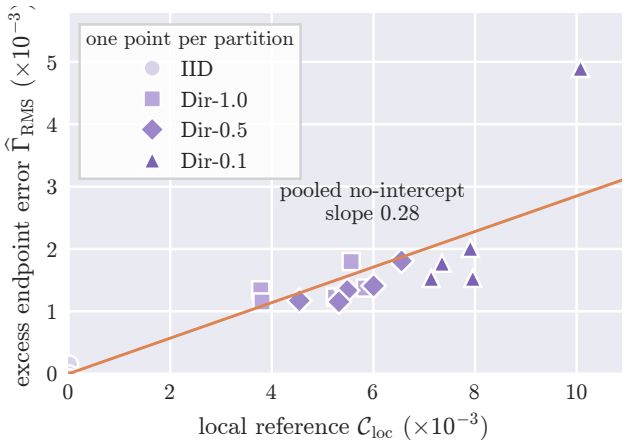

Figure 13: Independent versus joint distillation: excess endpoint error versus the local reference $\mathcal{C}_{\mathrm{loc}}$ on CIFAR-10. Each point represents one partition (five per level, three seeds each); the line is a descriptive pooled no-intercept fit across the skewed levels (slope 0.28). Level means appear in Table 7.

Paired analysis. The official-schedule analysis uses five partitions at each of four levels (IID, Dir-1.0, Dir-0.5, Dir-0.1), with the same five partition seeds at every level so that level comparisons can be paired within partition. Initialization repeats are aggregated within partition before any between-partition statistic is computed. Results from the earlier short schedule (500 DM iterations, 50-epoch evaluation) and the feature-skew family are collected in Appendix J.14.

J.5 Exact decomposition of the observed discrepancy (E9)

Configuration. For the E2 sets, E9 records the full endpoint vectors for each initialization seed (100–102) and decomposes the discrepancy of the separately distilled union exactly. With the union trained under a single mass-weighted loss, the one-step update on $S_1 \cup S_2$ is the weighted sum of the one-step updates on the parts, so

$$d_{\mathrm{ind}}(\theta_0) = \underbrace{\sum_i \alpha_i \Phi^{2,\eta}_{D_i}(\theta_0) - \Phi^{2,\eta}_{D_\cup}(\theta_0)}_{\Delta\Phi(\theta_0)} + \underbrace{\sum_i \alpha_i \big(\Phi^{1,\eta}_{S_i}(\theta_0) - \Phi^{2,\eta}_{D_i}(\theta_0)\big)}_{r_{\mathrm{src}}(\theta_0)}$$

holds without any quadratic or common-optimum assumption, with $\Phi^{1,\eta}_{S_i}$ the calibrated one-step maps of the E2 protocol (Appendix J.4). The identity was verified numerically on every cell (residual

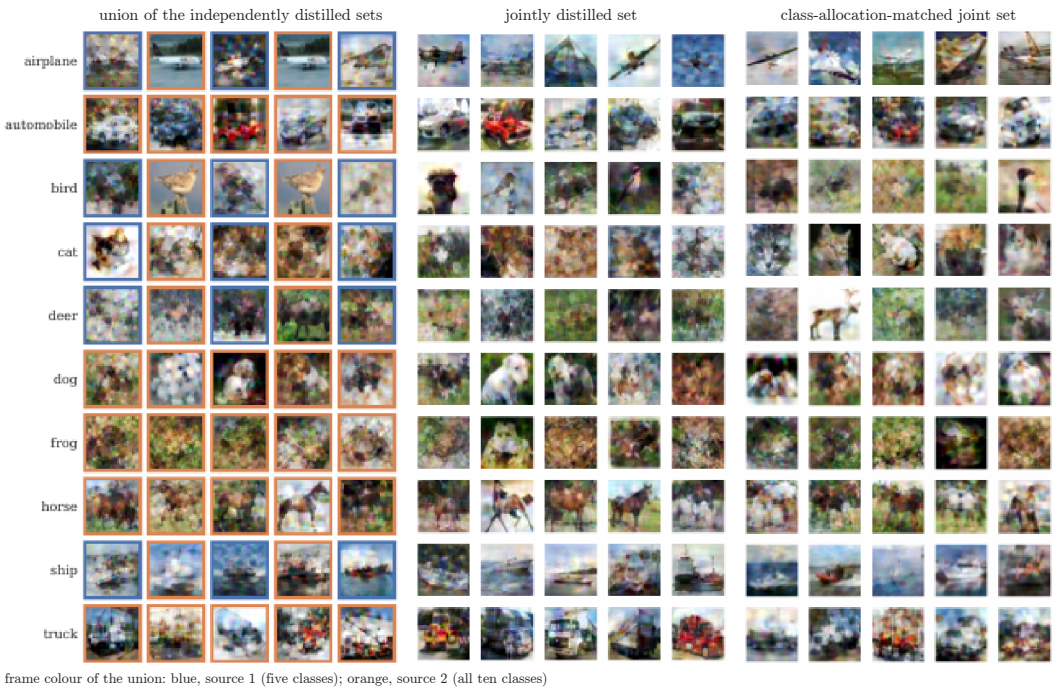

Figure 14: Synthetic images on one Dir-0.1 partition of CIFAR-10 (partition 0, DM at the official schedule; up to five images per class per set are shown). Left: the union of the two independently distilled sets (E2), with the frame colour marking the source of each image (blue, source 1, which holds five classes; orange, source 2, which holds all ten); each source stores ten images per class it holds, so the union has twenty images in the shared classes and ten in the others. Middle: the jointly distilled set at the same total budget of 150 images, 15 per class. Right: the class-allocation-matched joint set of E11 (Appendix J.8), which places 20 images in the five shared classes and 10 in the five classes held by one source, the per-class counts of the union.

at most 6×10^{-8}, relative 9×10^{-6}, the float32 accumulation error), and the norms $\|d_{\mathrm{ind}}\|$, $\|d_{\mathrm{joint}}\|$ and $\|\Delta\Phi\|$ reproduce $\mathcal{C}_{\mathrm{ind}}$, $\mathcal{C}_{\mathrm{joint}}$ and $\mathcal{C}_{\mathrm{loc}}$ of Table 7 to 3×10^{-6}. The excess of Section 6 subtracts the joint error, $\widehat{\Gamma}_{\mathrm{RMS}} = \|\Delta\Phi + r_{\mathrm{src}}\|_{L^2(P_0)} - \|d_{\mathrm{joint}}\|_{L^2(P_0)}$, so the decomposition bears on its first term. The local approximation residual $\Delta\Phi - \eta^2 r_{\mathrm{comp}}$ separates the network nonlinearity from the source distillation error. All 108 cells were processed (CIFAR-10 at four levels and the no-split control, SVHN at two levels and its no-split control), and the 18 CIFAR-100 cells of E15 (Appendix J.10) were added with the same script.

Analysis. At the three skewed CIFAR-10 levels $\Delta\Phi$ and r_{src} are anti-aligned: the cosine is -0.35, -0.48 and -0.46 at Dir-1.0, Dir-0.5 and Dir-0.1 (-0.78 on SVHN at Dir-0.1), and none of the 45 skewed CIFAR-10 cells or the 15 skewed SVHN cells has a positive cosine, and the measured $\|d_{\mathrm{ind}}\|$ is smaller than $\|r_{\mathrm{src}}\|$ although the two terms are of the same order. At the E2 step size ($\eta L \approx 1$) the prediction $\eta^2 r_{\mathrm{comp}}$ has a magnitude comparable to $\Delta\Phi$ (ratio 1.0–1.5) but only partly its direction (cosine 0.41–0.60, relative residual above one), and it is nearly orthogonal to d_{ind} (cosine 0.01–0.06), so the fraction of $\|d_{\mathrm{ind}}\|^2$ it explains is negative (-0.25, -0.69 and -1.55). Under IID partitioning $\|\Delta\Phi\| = 1.6 \times 10^{-5}$, so $d_{\mathrm{ind}} \approx r_{\mathrm{src}}$ (cosine 1.000); in the no-split control $\Delta\Phi$ is exactly zero and $d_{\mathrm{ind}} = r_{\mathrm{src}}$. Table 8 lists the per-level values.

Without the scalar calibration. Table 9 repeats the decomposition with the class-mass weights only, without the per-set step multiplier of the E2 protocol; $\Delta\Phi$ and $\eta^2 r_{\mathrm{comp}}$ do not depend on the synthetic sets and are unchanged. The fitted multipliers are 1.39–1.42 at the skewed CIFAR-10 levels, 1.56 at IID, 1.25–1.29 on SVHN and 1.9 on CIFAR-100, so without them the source errors are dominated by a scale mismatch common to all levels: $\|d_{\mathrm{ind}}\|$ rises to 13–14×10^{-3} on CIFAR-10 and varies little with heterogeneity, and the excess shrinks to between one eighth and one eighteenth

Table 8: E9: exact decomposition of the observed discrepancy on the E2 sets. Norms in units of 10^{-3}, RMS over seeds within partition and mean over partitions; cosines are means over seeds and partitions. $\hat{s} = \eta^2 r_{\mathrm{comp}}$.

Dataset	Level	$\|d_{\mathrm{ind}}\|$	$\|d_{\mathrm{joint}}\|$	$\|\Delta\Phi\|$	$\|r_{\mathrm{src}}\|$	$\|\hat{s}\|$	$\cos(\Delta\Phi, r_{\mathrm{src}})$	$\cos(\Delta\Phi, \hat{s})$	$\cos(d_{\mathrm{ind}}, \hat{s})$
CIFAR-10	IID	7.41	7.34	0.016	7.41	0.002	−0.01	0.10	−0.04
CIFAR-10	Dir-1.0	8.72	7.34	4.85	9.15	4.84	−0.35	0.41	0.06
CIFAR-10	Dir-0.5	8.71	7.33	5.58	9.84	7.26	−0.48	0.51	0.01
CIFAR-10	Dir-0.1	9.85	7.50	8.09	10.4	12.3	−0.46	0.60	0.04
CIFAR-10	no split	7.43	7.34	0	7.43	0	–	–	–
SVHN	IID	6.08	6.13	0.013	6.08	0.002	−0.02	0.08	0.09
SVHN	Dir-0.1	7.65	6.18	10.5	12.1	12.5	−0.78	0.32	−0.06
SVHN	no split	6.21	6.12	0	6.21	0	–	–	–
CIFAR-100	IID	1.95	1.86	0.018	1.95	0.002	−0.01	0.11	0.00
CIFAR-100	Dir-0.1	2.51	1.96	1.35	2.87	1.32	−0.48	0.86	−0.02

of its calibrated value on CIFAR-10, to a quarter on SVHN and to about one fiftieth on CIFAR-100. The heterogeneity-dependent part of the decomposition is present in both readings: $\Delta\Phi$ and r_{src} remain anti-aligned at every skewed level (cosine -0.11 to -0.56) and the excess remains positive at every skewed level and indistinguishable from zero at IID.

Table 9: E9 without the per-set scalar calibration. Each entry gives the calibrated value of Table 8 followed by the uncalibrated value (class-mass weights, $c_i = 1$). Norms in units of 10^{-3}; the excess is $\widehat{\Gamma}_{\mathrm{mean}}$, the signed per-initialization difference $\|d_{\mathrm{ind}}\| - \|d_{\mathrm{joint}}\|$ averaged within partition, mean $\pm$ SD over partitions, which differs from $\widehat{\Gamma}_{\mathrm{RMS}} = \mathcal{C}_{\mathrm{ind}} - \mathcal{C}_{\mathrm{joint}}$ of Table 7 by a few percent.

Dataset	Level	$\|d_{\mathrm{ind}}\|$	$\|d_{\mathrm{joint}}\|$	$\|r_{\mathrm{src}}\|$	$\cos(\Delta\Phi, r_{\mathrm{src}})$	excess
CIFAR-10	IID	7.41 / 13.4	7.34 / 13.4	7.41 / 13.4	−0.01 / +0.02	+0.07 ± 0.09 / −0.00 ± 0.09
CIFAR-10	Dir-1.0	8.72 / 13.4	7.34 / 13.4	9.15 / 13.6	−0.35 / −0.20	+1.42 ± 0.25 / +0.08 ± 0.08
CIFAR-10	Dir-0.5	8.71 / 13.5	7.33 / 13.3	9.84 / 13.9	−0.48 / −0.27	+1.42 ± 0.27 / +0.17 ± 0.11
CIFAR-10	Dir-0.1	9.85 / 13.7	7.50 / 13.4	10.4 / 14.4	−0.46 / −0.36	+2.38 ± 1.45 / +0.28 ± 0.08
SVHN	Dir-0.1	7.65 / 11.0	6.18 / 10.7	12.1 / 12.8	−0.78 / −0.56	+1.47 ± 0.95 / +0.37 ± 0.16
CIFAR-100	Dir-0.1	2.51 / 7.83	1.96 / 7.82	2.87 / 7.85	−0.48 / −0.11	+0.56 ± 0.03 / +0.01 ± 0.03

J.6 Post-hoc class reweighting (E12)

Configuration. E12 asks how much of the composed error of the saved E2 sets can be removed by changing only the class weights of the synthetic loss, with the images fixed. The one-step synthetic update is linear in the class weights, $d_{\mathrm{syn}}(w) = -\eta \sum_{(i,c)} w_{i,c}\, g_{i,c}(\theta_0)$, where $g_{i,c}(\theta_0)$ is the cross-entropy gradient of the class-c images of set S_i at θ_0, so the weights that best match the real-union two-step displacement $r_{\mathrm{real}}(\theta_0) = \Phi^{2,\eta}_{D_\cup}(\theta_0) - \theta_0$ solve a non-negative least-squares problem in at most twenty variables. Three weightings are compared for the independent union and, with the same procedure, for the joint set: the paper's weighting (class-mass weights with one scalar step multiplier per set fitted at each initialization, as in E2); weights fitted on a bank of 16 initializations (seeds 300–315, disjoint from the evaluation seeds) by accumulating the normal equations over the bank; and an oracle fitted on the evaluation initialization itself, which bounds what any class reweighting can achieve. All three are evaluated on the paper's initialization seeds 100–102, the endpoint error is $\|d_{\mathrm{syn}}(w) - r_{\mathrm{real}}\|$ as in E2, and the fraction of squared error removed is $1 - (\mathcal{C}_w/\mathcal{C}_{\mathrm{paper}})^2$. Under the bank weights the accuracy of the Dir-0.1 sets was re-evaluated with the official protocol and a class-weighted sampler (weights normalized to mean one), against the same sampler with uniform weights.

Analysis. Table 10 lists the results. Two quantities move differently. The squared error of the independent union falls by 15–25% at the skewed levels under weights fitted on held-out initializations and by 45–53% under the oracle on CIFAR-10 (41% on SVHN), so at least 47% of $\|d_{\mathrm{ind}}\|^2$ at these levels is not removed by any class reweighting, in line with the components that scalar retiming cannot repair (Appendix H); at IID the held-out weights do not transfer. The mean excess over

the joint set falls much more, from 1.38, 1.38 and 2.35 to 0.10, 0.21 and 0.67 ($\times 10^{-3}$) at Dir-1.0, Dir-0.5 and Dir-0.1 under the held-out weights, partly because the same fitting procedure increases the squared endpoint error of the joint set on CIFAR-10 by 12–14%; it remains positive at every skewed level under all three weightings. At IID it is $+0.06 \pm 0.10$ under the paper's weighting and -0.01 ± 0.12 under the bank weights, neither distinguishable from zero, and $+0.07 \pm 0.04$ under the oracle ($t = 4.16$, $df = 4$, $p = 0.014$, uncorrected two-sided). On accuracy, the held-out weights give no detectable change on the five Dir-0.1 partitions: 45.66 ± 0.73 against 45.50 ± 0.70 for the union (paired difference $+0.16 \pm 0.49$) and 50.36 ± 1.01 against 50.28 ± 0.91 for the joint set ($+0.08 \pm 0.41$).

Table 10: E12: endpoint error of the saved DM sets under three class weightings. RMS over the three evaluation initializations within partition, mean ± SD over five partitions, in units of 10^{-3}. Removed is the fraction of squared error removed relative to the paper's weighting; excess is $\mathcal{C}_{\mathrm{ind}} - \mathcal{C}_{\mathrm{joint}}$ under the same weighting.

Dataset	Level	Weighting	$\mathcal{C}_{\mathrm{ind}}$	$\mathcal{C}_{\mathrm{joint}}$	removed (ind / joint)	excess
CIFAR-10	IID	paper	7.41 ± 0.04	7.34 ± 0.08	–	$+0.06 \pm 0.10$
		bank	7.82 ± 0.05	7.83 ± 0.08	-0.11 / -0.14	-0.01 ± 0.12
		oracle	5.83 ± 0.05	5.76 ± 0.04	0.38 / 0.38	$+0.07 \pm 0.04$
CIFAR-10	Dir-1.0	paper	8.72 ± 0.29	7.34 ± 0.08	–	$+1.38 \pm 0.25$
		bank	7.93 ± 0.08	7.83 ± 0.08	0.17 / -0.14	$+0.10 \pm 0.08$
		oracle	6.00 ± 0.10	5.76 ± 0.04	0.53 / 0.38	$+0.24 \pm 0.10$
CIFAR-10	Dir-0.5	paper	8.71 ± 0.31	7.33 ± 0.06	–	$+1.38 \pm 0.27$
		bank	8.03 ± 0.28	7.81 ± 0.07	0.15 / -0.14	$+0.21 \pm 0.29$
		oracle	6.26 ± 0.32	5.78 ± 0.05	0.48 / 0.38	$+0.47 \pm 0.32$
CIFAR-10	Dir-0.1	paper	9.85 ± 1.33	7.50 ± 0.26	–	$+2.35 \pm 1.44$
		bank	8.62 ± 0.27	7.95 ± 0.24	0.20 / -0.12	$+0.67 \pm 0.13$
		oracle	7.15 ± 0.36	6.10 ± 0.36	0.45 / 0.34	$+1.04 \pm 0.09$
SVHN	IID	paper	6.08 ± 0.05	6.13 ± 0.06	–	-0.05 ± 0.03
		bank	5.83 ± 0.07	5.91 ± 0.05	0.08 / 0.07	-0.08 ± 0.06
		oracle	4.98 ± 0.07	5.09 ± 0.09	0.33 / 0.31	-0.11 ± 0.11
SVHN	Dir-0.1	paper	7.65 ± 0.91	6.18 ± 0.09	–	$+1.46 \pm 0.98$
		bank	6.54 ± 0.24	5.96 ± 0.11	0.25 / 0.07	$+0.58 \pm 0.31$
		oracle	5.80 ± 0.24	5.15 ± 0.09	0.41 / 0.31	$+0.64 \pm 0.30$

The removed fractions are means of the per-partition values; the Dir-0.1 union value under bank weights varies across partitions (SD 0.18 on CIFAR-10, 0.17 on SVHN).

J.7 Downstream accuracy and the no-split control (E7)

Each final synthetic set is evaluated with the distilling method's official protocol. For DM, a ConvNet is trained for 1000 epochs with differentiable augmentation, and three evaluation seeds are averaged within each partition. The earlier short schedule used 50 epochs without augmentation, which under-trained both sets and compressed the gain to about two points. Image budgets are matched within every partition. Because some sources omit classes under Dir-0.1, the joint budget varies between 150 and 200 images across partitions, so absolute accuracy should be compared only within partition, not across levels. The accuracies are those of Table 7.

The gain from joint distillation is $+5.3$, $+5.7$, $+5.9$, and $+5.3$ points at IID, Dir-1.0, Dir-0.5, and Dir-0.1. The within-partition paired differences from IID are $+0.44 \pm 0.94$, $+0.58 \pm 1.67$, and $+0.06 \pm 1.23$ ($t = 1.03$, 0.77, and 0.12 on four degrees of freedom). The same partition seed gives the largest gain at three of the four levels, so partition identity carries part of the variance and paired comparisons are the appropriate ones. Even when paired, no level effect is detectable. Repeated evaluation of a fixed synthetic set moves accuracy by up to ± 0.35 points, because the augmentation stage is not seeded (Appendix J). In the no-split control (two distillations of the full training set with different seeds, unioned, against one distillation at the combined budget), the gain is $+6.55 \pm 0.42$ over three replicates, with per-class effective rank 8.9 vs. 10.3 and cross/within ratio 0.941, the IID values. The mean endpoint-distance difference is $\widehat{\Gamma}_{\mathrm{mean}} = (0.08 \pm 0.21) \times 10^{-3}$, negative in one of the three replicates, and the measured $\mathcal{C}_{\mathrm{loc}}$ is exactly zero, as the identity requires.

Warm-started joint distillation (E13). To test whether the no-split gain is recovered by continuing distillation from the union, DM was run on the full training set at the official settings (image learning rate 1.0, differentiable augmentation, real batch 256) starting from the union of the two independent sets of each replicate, for 500 and 2,000 iterations, and, as a control for the short schedule, from random real images for the same 2,000 iterations with the same seeds (Table 11). The union and the official joint set were re-evaluated in the same run (46.44 and 52.59 against 46.04 and 52.59 in E7, within the evaluation noise noted above), so the differences in Table 11 are paired within the run. Warm-starting recovers $+0.44 \pm 0.47$ of the 6.15-point gap after 2,000 iterations, whereas the cold start reaches $+1.83 \pm 0.65$ in the same number of iterations with a higher per-class effective rank. Under the same 2,000-iteration budget, initialization from real images therefore yields the larger gain, and the warm-started images change little over this interval (median relative displacement 0.035). Whether a longer continuation or other optimization settings would close the gap from the union was not tested.

Table 11: E13: joint distillation warm-started from the union of the no-split sets (CIFAR-10, DM, 200 images). Accuracy in percent, mean $\pm$ SD over three replicates; paired differences are within replicate; effective rank is the per-class participation ratio of the frozen embedding.

Set	Iterations	Accuracy	vs. union	effective rank
union of the independent sets	0	46.44 ± 0.50	–	8.91
warm start from the union	500	46.16 ± 0.51	-0.28 ± 0.43	9.00
warm start from the union	2,000	46.88 ± 0.67	$+0.44 \pm 0.47$	9.15
cold start from real images	2,000	48.27 ± 0.54	$+1.83 \pm 0.65$	9.66
joint set (official schedule)	20,000	52.59 ± 0.36	$+6.15 \pm 0.66$	10.29

J.8 Class-allocation-matched joint set (E11)

Configuration. Under Dir-0.1 the canonical joint set of E2 places images in every class, whereas the composed set stores only the classes each source holds, so the two sets differ in their per-class allocation as well as in how they were distilled. E11 removes the first difference. On the three Dir-0.1 partitions of CIFAR-10 the joint set was re-distilled with DM at the official schedule (20,000 iterations, differentiable augmentation, real initialization) with its per-class counts fixed to those of the composed set (Figure 14, right): on partition 0, 20 images in the five classes both sources hold and 10 in the five held by one source (150 in all), on partition 1, 20 in seven classes and 10 in three (170), and on partition 2, where both sources hold all ten classes, 20 in every class (200), which is the canonical design and serves as a replicate with a different distillation draw. The composed set, the matched joint set and the canonical joint set were evaluated with five evaluation seeds under the 1,000-epoch protocol, and the endpoint quantities with the three initialization seeds of E2.

Analysis. Matching the class allocation leaves the excess endpoint error positive in all nine (partition, seed) cells, unchanged on partition 0 (1.55 against 1.53×10^{-3}) and reduced by about a third on partition 1 (1.06 against 1.53×10^{-3}; Table 12), while the accuracy gain on these two partitions falls from 4.5 to 1.0 and from 5.4 to 3.4 points. On the replicate partition both quantities agree with the canonical set to within 0.3 points and 10^{-5}. The class allocation therefore accounts for part of the accuracy gain at Dir-0.1 and can change the size of the excess endpoint error without removing it, so the two quantities respond differently to the same intervention (Appendix J.7). The canonical values reproduce the rows of Table 7 to 10^{-4} relative in the endpoint quantities and to within 0.5 points in the five-seed re-evaluation of accuracy.

J.9 Image budget and the local reference

The constituent-error regime. Corollary 4.3 guarantees the structural gap only when it exceeds the weighted constituent error, and the independent-versus-joint comparison lies on the other side of that threshold. Along the official DM schedule (Figure 15) the constituent error falls from 0.33 to 0.15 within 5,000 iterations and then plateaus, the composed error tracks it with elasticities of 0.93–1.08 over the four training intervals, and both settle with $\mathcal{C}_{\mathrm{ind}}/\mathcal{C}_{\mathrm{loc}} = 1.22 \pm 0.03$ at the end of the schedule. Additional iterations therefore do not bring ten images per class closer to the two-step map of 25,000 images; the budget sweep below varies the budget directly.

Table 12: E11: class-allocation-matched joint set at Dir-0.1 on CIFAR-10. The image count is the budget of both joint sets on that partition. Endpoint quantities are RMS over three initialization seeds in units of 10^{-3}, accuracy is the mean and SD over five evaluation seeds in percent, and the gain is joint minus independent.

		Matched joint		Canonical joint			Matched		Canonical	
Partition (images)	$\mathcal{C}_{\text{ind}}$	$\mathcal{C}_{\text{joint}}$	$\widehat{\Gamma}_{\text{RMS}}$	$\mathcal{C}_{\text{joint}}$	$\widehat{\Gamma}_{\text{RMS}}$	Acc. indep.	Acc. joint	Gain	Acc. joint	Gain
0 (150)	9.48	7.92	1.55	7.95	1.53	45.4 ± 0.7	46.4 ± 0.6	$+1.0$	49.9 ± 0.5	$+4.5$
1 (170)	8.88	7.82	1.06	7.35	1.53	46.1 ± 1.0	49.5 ± 0.5	$+3.4$	51.4 ± 0.4	$+5.4$
2 (200)	12.18	7.29	4.90	7.29	4.90	46.5 ± 0.8	52.5 ± 0.8	$+6.0$	52.8 ± 0.7	$+6.3$

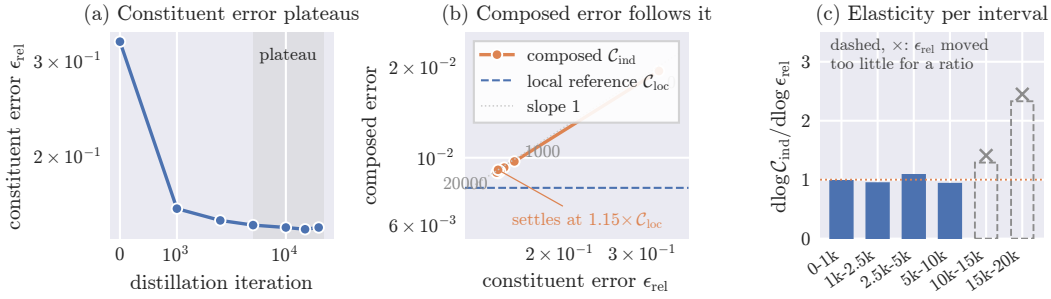

Figure 15: Composed error follows constituent error (DM, Dirichlet-0.1, partition p0). (a) Relative constituent error at the snapshot checkpoints {0,1000,2500,5000,10000,15000,20000} of the official schedule; it plateaus after 5,000 iterations. (b) The composed error $\mathcal{C}_{\text{ind}}$ tracks it with unit elasticity and settles above the local reference $\mathcal{C}_{\text{loc}}$ (dashed), at $1.15\times$ for this partition and about $1.2\times$ over the five partitions of the official sweep.

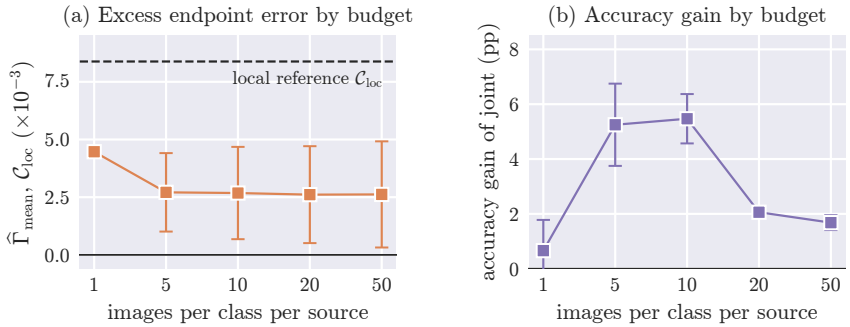

Figure 16: Image budget at Dir-0.1 (the same three partitions at every budget). (a) Excess endpoint error $\widehat{\Gamma}_{\text{mean}}$ of the separately distilled union, aggregated as the mean over seeds of the per-seed endpoint-distance difference as in Table 13, with the local reference $\mathcal{C}_{\text{loc}}$, which the partitions fix. (b) Accuracy gain of joint over separate-then-union distillation.

Varying the image budget. Table 13 varies the images per class (IPC) of each source at Dir-0.1 on the same three partitions (p0–p2) and three initialization seeds at every budget (Figure 16). Three findings follow. First, the budget moves the constituent error where iterations could not: ϵ_{rel} falls from 0.42 at IPC 1 to 0.15 at IPC 50, the composed error follows it with elasticities between 0.83 and 1.09, and $\mathcal{C}_{\text{ind}}/\mathcal{C}_{\text{loc}}$ descends from 2.60 to 1.02 ± 0.08, while $\mathcal{C}_{\text{loc}}$, which depends only on the real partition and the initialization, is identical at every budget. Second, whether $\mathcal{C}_{\text{ind}}$ settles at a persistent plateau near $\mathcal{C}_{\text{loc}}$ is not resolved. The criteria were fixed before the two largest budgets were run: a plateau is a ratio stabilizing at or above one with elasticity falling clearly below one, a crossing is a partition mean below 0.95 with most measurements below one, and anything between is undetermined. At IPC 50 the partition mean is 1.02 ± 0.08, the three partitions give 0.95, 1.02 and 1.11, four of the nine partition–seed measurements lie below one and the last elasticity is 0.95, so

the data fall in the undetermined region; no general bound requires a plateau, because approximate constituents can cancel part of the structural term (Appendix C.1). Third, the excess endpoint error and the accuracy gain respond differently to the budget: the excess is about $1.7\times$ its IPC-10 value at IPC 1 and essentially constant for IPC ≥ 5, whereas the gain is near zero at IPC 1, about five points at IPC 5–10 and about two points at IPC 20–50. Together with the heterogeneity axis, which changes the excess but not the gain on CIFAR-10, the budget is a second manipulation on which the two quantities separate; Appendix J.12 discusses one reading of the inverted-U profile.

Table 13: Image budget at Dir-0.1 (DM, official schedule; partitions p0–p2 $\times$ three initialization seeds at every budget).

IPC	ϵ_{rel}	$\mathcal{C}_{\mathrm{ind}}$	$\mathcal{C}_{\mathrm{ind}}/\mathcal{C}_{\mathrm{loc}}$	$\widehat{\Gamma}_{\mathrm{mean}}$	acc. diff. (pp)	elasticity to next
1	0.416 ± 0.023	21.3	2.60 ± 0.45	4.47 ± 0.07	$+0.66 \pm 1.12$	0.83
5	0.206 ± 0.021	11.9	1.44 ± 0.10	2.71 ± 1.70	$+5.25 \pm 1.50$	0.88
10	0.172 ± 0.019	10.2	1.22 ± 0.03	2.68 ± 2.00	$+5.47 \pm 0.90$	1.09
20	0.157 ± 0.019	9.20	1.09 ± 0.06	2.61 ± 2.10	$+2.06 \pm 0.14$	0.95
50	0.147 ± 0.017	8.63	1.02 ± 0.08	2.62 ± 2.30	$+1.68 \pm 0.27$	N/A

Each entry is the mean $\pm$ SD over the three partitions of the partition-level value (itself the mean over three seeds); $\mathcal{C}_{\mathrm{loc}} = 8.38 \times 10^{-3}$ at every IPC. $\mathcal{C}_{\mathrm{ind}}$ and $\widehat{\Gamma}_{\mathrm{mean}}$ in units of 10^{-3}. Elasticity is $\mathrm{d}\log \mathcal{C}_{\mathrm{ind}}/\mathrm{d}\log \epsilon_{\mathrm{rel}}$ between consecutive rows, computed from the unrounded partition means (recomputing from the rounded entries gives 0.83, 0.85, 1.13, 0.97). The joint set's budget is matched to the sum of the two sources' budgets (IPC 2 to 100 over the union). The IPC-10 row uses the three sweep partitions and therefore differs slightly from the five-partition values of Table 7.

J.10 Further datasets (SVHN and CIFAR-100)

SVHN (Netzer et al., 2011) uses the standard training split (73,257 images, no *extra* split) with the same tensor conversion and per-channel normalization as CIFAR-10, five partitions and three initialization seeds per level at IID and Dir-0.1, and the same DM schedule, budget and evaluation protocol. CIFAR-100 (E15) uses three partitions per level, the Dirichlet draw being per class over the 100 classes, and ten images per represented class per source, so the union stores 1,660–2,000 images and the joint set the same budget; the E9 decomposition was applied to every CIFAR-100 cell (Tables 8 and 9). Table 14 collects the results. The endpoint layer replicates on both datasets: the IID excess is small on both (-0.04 ± 0.04 on SVHN, 0.09 ± 0.02 on CIFAR-100, $\times 10^{-3}$) and the paired increase from IID to Dir-0.1 is positive on every partition (five of five on SVHN, three of three on CIFAR-100), and the CIFAR-100 decomposition shows the CIFAR-10 pattern, with the two terms anti-aligned (cosine -0.48, negative in nine of nine cells) and the prediction aligned with $\Delta\Phi$ (cosine 0.86) but orthogonal to d_{ind}. The accuracy layer depends on the dataset. On SVHN the gain rises with skew (paired difference $+2.24 \pm 1.21$, five of five, $t = 4.14$), because the composed set loses more accuracy under skew ($70.8 \to 66.4$, against $46.3 \to 45.8$ on CIFAR-10); on CIFAR-100 it is smaller at Dir-0.1 than at IID (paired difference -1.10 ± 1.00, $p = 0.20$), a difference three partitions do not resolve. The SVHN no-split control gives $+2.16 \pm 1.27$ ($t = 2.95$, $p \approx 0.10$) with $\mathcal{C}_{\mathrm{loc}}$ exactly zero, smaller than on CIFAR-10, and SVHN's frozen-feature effective ranks are lower throughout (4.5 against 4.9 per class, compared with 8.9 against 10.3 on CIFAR-10), consistent with the smaller within-class variation of digit images.

J.11 Cross-architecture evaluation (E14)

Configuration. The saved E2 sets (four levels, five partitions), the E7 no-split sets and the E11 class-allocation-matched sets were re-evaluated on ResNet-18 (the instance-normalized variant of the dataset-condensation code) with the DM evaluation protocol otherwise unchanged (1000 epochs, learning rate 0.01, batch 256, differentiable augmentation, three evaluation seeds). No set was re-distilled.

Analysis. Table 15 compares the two evaluation networks. On ResNet-18 the joint set remains more accurate than the union on every CIFAR-10 partition, but the gain is smaller than on the ConvNet ($+1.8$ to $+3.4$ against $+5.3$ to $+5.9$ points) and its dependence on skew differs. With the paired design of Appendix J.7 (the same partition seeds at every level), the within-partition

Table 14: Cross-dataset comparison (DM, official schedule, IPC 10 per source; five partitions per split level on CIFAR-10 and SVHN, three on CIFAR-100, three replicates for the no-split controls).

Dataset	Level	n	$\mathcal{C}_{\mathrm{loc}}$	$\widehat{\Gamma}_{\mathrm{mean}}$	accuracy gain (pp)
CIFAR-10	IID	5	0.016	0.08 ± 0.06	$+5.28 \pm 0.92$
CIFAR-10	Dir-0.1	5	7.72	2.33 ± 1.40	$+5.35 \pm 0.75$
SVHN	IID	5	0.013	-0.04 ± 0.04	$+3.24 \pm 1.10$
SVHN	Dir-0.1	5	10.5	1.47 ± 0.95	$+5.48 \pm 1.45$
CIFAR-10	no split	3	0 (exact)	0.08 ± 0.21	$+6.55 \pm 0.42$
SVHN	no split	3	0 (exact)	0.09 ± 0.16	$+2.16 \pm 1.27$
CIFAR-100	IID	3	0.018	0.09 ± 0.02	$+4.98 \pm 0.14$
CIFAR-100	Dir-0.1	3	1.35	0.56 ± 0.03	$+3.88 \pm 1.14$
Paired difference Dir-0.1 − IID within partition (t on 4 degrees of freedom, 2 for CIFAR-100)					
CIFAR-10		5		$+2.25 \pm 1.39$ ($t = 3.62$)	$+0.06 \pm 1.23$ ($t = 0.12$)
SVHN		5		$+1.51 \pm 0.94$ ($t = 3.58$)	$+2.24 \pm 1.21$ ($t = 4.14$)
CIFAR-100		3		$+0.47 \pm 0.02$ ($t = 37$)	-1.10 ± 1.00 ($t = -1.91$)

$\mathcal{C}_{\mathrm{loc}}$ and $\widehat{\Gamma}_{\mathrm{mean}}$ in units of 10^{-3}; mean ± SD over partitions, each partition averaging its initialization seeds (five seeds for the CIFAR-10 excess error here, three for SVHN and CIFAR-100). Two-sided p: 0.022, 0.91 (CIFAR-10), 0.023, 0.014 (SVHN) and 0.001, 0.20 (CIFAR-100).

difference of the ResNet-18 gain between Dir-0.1 and IID is $+1.64 \pm 1.39$ points, positive on five of five partitions ($t = 2.63$ on four degrees of freedom, $p = 0.06$); between Dir-0.5 and IID it is $+1.50 \pm 1.50$ (five of five, $p = 0.09$) and between Dir-1.0 and IID $+0.24 \pm 0.74$ (three of five). The corresponding ConvNet differences are $+0.06 \pm 1.23$, $+0.58 \pm 1.67$ and $+0.44 \pm 0.94$. The no-split gain on ResNet-18 is $+1.13 \pm 0.79$ (positive in all three replicates, $p = 0.13$ on two degrees of freedom), and the class-allocation-matched joint set gains $+1.37$ against $+2.81$ for the canonical joint set on the same three Dir-0.1 partitions. On SVHN the ResNet-18 gain is $+0.39 \pm 1.89$ at IID and $+4.12 \pm 2.01$ at Dir-0.1, a paired difference of $+3.72 \pm 3.41$ (four of five partitions, $p = 0.07$). Source heterogeneity is therefore not necessary for the gain on either network; its size depends on the evaluation network, and on ResNet-18 the gain at the highest skew exceeds the IID gain on every partition, a pattern that the ConvNet shows on SVHN but not on CIFAR-10.

Table 15: E14: accuracy of the same synthetic sets on the ConvNet and on ResNet-18 (DM, official evaluation protocol, three evaluation seeds). Percent, mean ± SD over partitions; gain is joint minus union within partition.

			ConvNet		ResNet-18		
Dataset	Level	n	joint	gain	union	joint	gain
CIFAR-10	IID	5	51.6 ± 0.6	$+5.28 \pm 0.92$	40.3 ± 0.6	42.1 ± 0.5	$+1.81 \pm 0.85$
CIFAR-10	Dir-1.0	5	51.9 ± 0.5	$+5.72 \pm 0.64$	40.2 ± 0.6	42.3 ± 0.6	$+2.05 \pm 0.77$
CIFAR-10	Dir-0.5	5	51.7 ± 0.5	$+5.86 \pm 1.24$	39.2 ± 1.1	42.6 ± 0.4	$+3.31 \pm 0.84$
CIFAR-10	Dir-0.1	5	51.1 ± 1.1	$+5.35 \pm 0.75$	37.8 ± 1.1	41.2 ± 0.8	$+3.44 \pm 0.95$
CIFAR-10	no split	3	52.6 ± 0.4	$+6.55 \pm 0.42$	41.1 ± 0.8	42.2 ± 0.1	$+1.13 \pm 0.79$
SVHN	IID	5	73.7 ± 0.9	$+3.24 \pm 1.10$	67.0 ± 1.2	67.4 ± 1.1	$+0.39 \pm 1.89$
SVHN	Dir-0.1	5	72.8 ± 0.7	$+5.48 \pm 1.45$	61.9 ± 2.4	66.0 ± 0.7	$+4.12 \pm 2.01$

ConvNet accuracies are those of Tables 7 and 14 (three evaluation seeds); ResNet-18 accuracies are from this experiment (three evaluation seeds). The one-sample t of the ResNet-18 gain against zero is 4.8, 6.0, 8.9, 8.1 at the four CIFAR-10 levels and 0.5, 4.6 on SVHN.

J.12 Feature-space diagnostics of independently and jointly distilled sets (E8)

For each synthetic set (Table 16) we embed every image with a frozen, randomly initialized ConvNet and compute, per class, the participation ratio of the Gram spectrum of the embedded images, that is, the number of directions the class's synthetic images effectively occupy, averaged over classes. For the two independently distilled sets we also compute the ratio of the mean cross-set distance to the mean within-set distance. A ratio below one is consistent with overlap between the two sets, and

a ratio above one with the sets occupying different regions. These are descriptive statistics of the images and are unrelated to r_H, which is a property of the real sources' Hessians.

Table 16: Effective rank per class and cross/within distance ratio in a frozen feature space. Mean ± SD over partitions (n as listed; two IID partitions, five Dir-0.1 partitions, three no-split replicates); image budgets matched within partition. Dir-1.0 and Dir-0.5 were not measured.

Level	n	eff. rank, separate (unioned)	eff. rank, joint	cross/within
IID	2	8.91 ± 0.11	10.27 ± 0.02	0.942 ± 0.000
Dirichlet-0.1	5	7.39 ± 0.42	9.79 ± 0.40	1.288 ± 0.097
No split (E7)	3	8.91 ± 0.05	10.29 ± 0.11	0.941 ± 0.001

Under IID the two independent sets have a cross/within ratio below one, consistent with overlapping coverage in the frozen feature space; under label skew the ratio is above one, consistent with each set covering its own skewed subset. In both cases the joint set occupies more directions per class at the same budget, a statement about the chosen embedding rather than about coverage of the real data. The pattern accompanies the accuracy gain and is present when the curvature-variance term is identically zero, as in the no-split control. Class omission under Dir-0.1 cannot account for the gain under IID and with no split, where every class is present in both sources at identical budgets; its contribution at Dir-0.1 is isolated by the class-allocation-matched control of Appendix J.8. One reading of the inverted-U budget profile of Appendix J.9 is that redundancy between the two sets matters most at intermediate budgets, where each set is large enough to overlap the other and small enough for the overlap to cost accuracy; the statistics here are consistent with that reading but do not measure class coverage directly.

J.13 Assumption diagnostics (E3)

Table 17: Assumption diagnostics.

Network	Level	ηL	r_H	δ_H
Local-law network at $\eta_{\text{ref}} = 7.8 \times 10^{-4}$				
ReLU, 2,000-sample subset	IID / Dir-1.0 / Dir-0.1	0.112 / 0.116 / 0.126	—	0.29 ± 0.08
ReLU, 10^4 subset	IID / Dir-0.1	0.096 / 0.108	—	—
SiLU, 10^4 subset	IID / Dir-0.1	0.047 / 0.066	—	0.14 ± 0.02
tanh, 10^4 subset	IID / Dir-0.1	0.039 / 0.060	—	0.20 ± 0.02
Distillation network (ReLU) at $\eta = 6.25 \times 10^{-3}$				
	IID	0.80 ± 0.00	N/A	—
	Dir-1.0	0.85 ± 0.02	204.0 ± 8.4	—
	Dir-0.5	0.86 ± 0.02	201.0 ± 4.0	—
	Dir-0.1	1.03 ± 0.06	208.5 ± 5.2	0.28 ± 0.08

$L = \max_i \|H_i\|_2$ at the shared initialization; r_H is the participation ratio $\operatorname{tr}(V_H)^2 / \operatorname{tr}(V_H^2)$ from Hutchinson probes (Hutchinson, 1989), out of 320,010 parameters; $\delta_H = \|H(\theta_2) - H(\theta_0)\|_2 / \|H(\theta_0)\|_2$ along the two-step real-union trajectory at Dir-0.1, mean ± SD over five partitions. Local-law rows are partition means; distillation rows are mean ± SD over five partitions (three seeds at Dir-0.1, one otherwise). Dashes: not measured. N/A: under IID V_H is at numerical zero and its participation ratio is not meaningful. L differs by activation (ReLU 123–139, SiLU 60–85, tanh 50–76), so the activations sit at different ηL at a given η.

The main-text assumptions are stated in Section 3. Table 17 reports the diagnostics available for them: the normalized step size ηL (which enters the positive-branch condition and the Theorem E.2 bracket) for the local-law network at its calibration rate and for the distillation network at its operating rate, and the effective rank of V_H (the $\sqrt{r_H}$ factor in the same bracket) for the distillation network. It also reports the relative Hessian change δ_H along the two-step trajectory, the empirical proxy for the slowly-varying-Hessian condition of equation 10. Two of the three diagnostics sit at their boundaries. $\eta L \approx 1$ at the distillation rate, and δ_H lies between 0.14 (SiLU) and 0.29 (ReLU), meaning that the Hessian moves by 14–29% within two steps. ReLU has the largest and most variable δ_H although its activation has zero second derivative almost everywhere. The loss Hessian still varies through the products of layer parameters, the normalization layers, and the cross-entropy loss,

and it also changes non-smoothly when the trajectory crosses activation boundaries, so the measured δ_H should not be attributed to the crossings alone. The two ReLU constructions (local-law and official networks) agree (0.29 vs. 0.28), so the activation comparison transfers to the network used for distillation. Only the low effective rank holds with a wide margin, consistent with the concentrated Hessian spectra reported for trained and initialized networks (Sagun et al., 2017; Ghorbani et al., 2019). The exponent measured in the local-law validation, 3–6% below four (Appendix J.2), is consistent with both boundary conditions. r_H on the local-law network was not measured.

J.14 Earlier short-schedule and preliminary diagnostics

The results in this subsection were obtained with the earlier DM schedule (500 iterations, 50-epoch evaluation) or are single-condition preliminaries. They support the main results but do not carry them; the structural quantities P_{comp} and $\mathcal{C}_{\mathrm{loc}}$ in them depend only on the real partitions and are unchanged by the schedule.

Type of heterogeneity: the feature-skew family. A label-balanced feature-skew family complements the label-skewed partitions. A frozen, untrained ConvNet embeds every image, and within each class the examples are sorted by a projection whose direction is derived from the partition seed and split at the median, so each source receives half of every class ($\alpha_1 = \alpha_2 = 1/2$, label total variation zero) and only the within-class inputs differ. Random directions give eight partitions; a single PC1 split serves as a higher-contrast diagnostic and is not treated as an independent sample. The structural term of this family is $60\times$ below Dirichlet-1.0 in P_{comp} (about $8\times$ in RMS), and under the short schedule its excess endpoint error is significantly negative (Table 18), so it is a second level at which the structural signal is small relative to the observed residuals.

Table 18: Type of heterogeneity changes the empirical outcome.

Shift	label TV	n	P_{comp}	$\sqrt{P_{\mathrm{comp}}}$	$\mathcal{C}_{\mathrm{loc}}$	$\mathcal{C}_{\mathrm{loc}}/\sqrt{P_{\mathrm{comp}}}$	$\widehat{\Gamma}_{\mathrm{mean}}{}^{\dagger}$
IID	0.000	5	4.13×10^{-12}	0.002	0.016	(8.1)	-0.377
Feature skew[f]	0.000	8	3.95×10^{-7}	0.63	0.652	(1.04)	-0.478
Dirichlet-1.0	0.227	5	2.30×10^{-5}	4.61	4.54	1.01	$+0.928$
Dirichlet-0.5	0.300	5	5.25×10^{-5}	7.12	5.48	0.78	not run
Dirichlet-0.1	0.392	5	1.42×10^{-4}	11.7	7.72	0.67	$+1.96$

P_{comp} and $\mathcal{C}_{\mathrm{loc}}$ from one protocol (E2 network, $\eta = 6.25 \times 10^{-3}$, full sources, five partitions $\times$ five seeds); $\sqrt{P_{\mathrm{comp}}}$, $\mathcal{C}_{\mathrm{loc}}$, and $\widehat{\Gamma}_{\mathrm{mean}}$ in units of 10^{-3}. Each column is the mean of the per-measurement values over partitions and seeds, so the ratio column is the mean of the per-measurement ratios and is not the quotient of the two column means, and $\sqrt{P_{\mathrm{comp}}}$ is not the square root of the P_{comp} column. In the quadratic model $\mathcal{C}_{\mathrm{loc}} = \mathcal{C}_{\mathrm{struct}} = \sqrt{P_{\mathrm{comp}}}$. On the network, the ratio $\mathcal{C}_{\mathrm{loc}}/\sqrt{P_{\mathrm{comp}}}$ is the calibration of the quadratic prediction against the measured local reference at the E2 rate, where $\eta L \approx 1$. It is 1.01 at Dir-1.0 and drifts to 0.67 at Dir-0.1, the same downward drift with step size that the local-law validation shows above η_{ref} (Figure 9b). At IID both quantities are at numerical zero, and the ratio (in parentheses) is not meaningful. The same P_{comp} values measured at η_{ref} and rescaled by $(\eta/\eta_{\mathrm{ref}})^4$ agree with these to within 2.4%. [f]Feature skew is from the earlier short-schedule run (P_{comp} at the E2 rate, eight partitions), so its ratio, in parentheses, is not directly comparable with the others. [†]$\widehat{\Gamma}_{\mathrm{mean}}$ is from the earlier 500-iteration schedule (Table 7 reports $\widehat{\Gamma}_{\mathrm{RMS}}$ at the official schedule); a negative value means the composed set's map was closer to the real-union map than the joint set's.

For random-projection feature skew, P_{comp} is positive in all eight partitions, with mean 3.95×10^{-7} at the E2 rate and 95% CI $[2.78, 5.16] \times 10^{-7}$. Nevertheless, $\widehat{\Gamma}_{\mathrm{mean}} = -4.78 \times 10^{-4}$ with CI $[-5.92, -3.65] \times 10^{-4}$, and all eight gaps are negative. The PC1 diagnostic increases P_{comp} to 4.70×10^{-6} and $\mathcal{C}_{\mathrm{loc}}$ to 2.14×10^{-3}, but its single measured gap remains negative at -6.59×10^{-4}. This single-partition probe argues against an explanation by lack of signal. In this family the observed composed error is 13.4 times the local reference and its cosine with the quadratic prediction is 0.011, so the measured endpoint is dominated by a residual nearly orthogonal to the structural term.

Semantic-mass intervention. Figure 17 compares stored semantic mass with uniform synthetic class weights on the original three partitions per heterogeneous level, with clock calibration in all

cells (IID is exactly one by construction). Improved constituent fidelity does not necessarily transfer to the union.

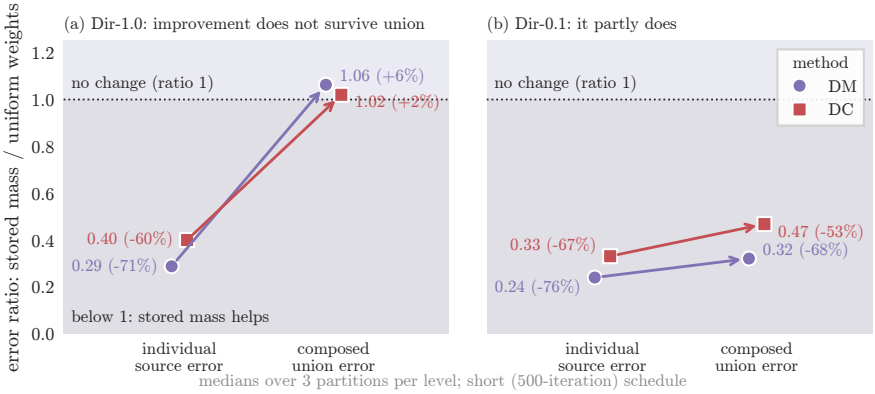

Figure 17: Better constituents do not always yield a better union. Ratio of the error with stored semantic mass to the error with uniform weights, for the individual sources and for their composed union (percentages: change relative to uniform weights). (a) At Dir-1.0 the individual error falls by 60–71% while the composed error is unchanged. (b) At Dir-0.1 the improvement partly survives union.

Five sources. One condition with $m = 5$ sources was run at Dir-0.1 with the same DM schedule and IPC 10 per source, on five partitions with three initialization seeds each; the joint set received the sum of the five budgets (Table 19). The local reference, the quadratic prediction, the excess endpoint error and the accuracy gain all grow from two to five sources, so the mechanism is not restricted to two constituents, although V_H is not monotone in the number of sources in general and this is an observation rather than a prediction. With IPC fixed per source the composed set's image count roughly doubles, so $\mathcal{C}_{\text{ind}}$ and its ratio to $\mathcal{C}_{\text{loc}}$ confound source count with budget and are not interpreted.

Table 19: Two versus five sources at Dir-0.1 (DM, official schedule, IPC 10 per source). Mean $\pm$ SD over $n = 5$ partitions; within a partition the endpoint quantities are RMS over the three initialization seeds 100–102 and the accuracy gain is their mean, the convention of Table 7, whose Dir-0.1 row the $m = 2$ column reproduces. The first block depends only on the real partition and the network; the second describes the distilled sets, whose total budget is not matched between $m = 2$ and $m = 5$, so $\mathcal{C}_{\text{ind}}$ and its ratio to $\mathcal{C}_{\text{loc}}$ confound source count with budget and are not interpreted.

	$m = 2$	$m = 5$	ratio
Real partition and network only			
$\mathcal{C}_{\text{loc}}$ ($\times 10^{-3}$)	8.09 ± 1.17	14.1 ± 2.3	1.75
$\sqrt{P_{\text{comp}}}$ ($\times 10^{-3}$)	12.3 ± 2.1	25.8 ± 2.2	2.09
Distilled sets (IPC 10 per source; total budget grows with m)			
$\widehat{\Gamma}_{\text{RMS}}$ ($\times 10^{-3}$)	2.35 ± 1.44	5.65 ± 1.56	2.41
accuracy gain (pp)	$+5.35 \pm 0.75$	$+7.34 \pm 1.34$	1.37
composed-set images	176 ± 19	376 ± 32	2.14
$\mathcal{C}_{\text{ind}}$ ($\times 10^{-3}$)	9.85 ± 1.33	12.2 ± 1.5	1.24

K Limitations and Future Work

Scope of the theory. The exact results concern full-batch, fixed-step gradient descent on quadratic objectives (A1). They apply to a neural network only locally, through a frozen quadratic model that requires a slowly varying Hessian along the training trajectory (A4); Appendix J.13 reports the measured Hessian variation. The remainder constant ν cannot be estimated reliably for a deep

network, so the calibration of P_{comp} is checked empirically rather than certified. The arbitrary-ratio expansion (Theorem E.2) additionally assumes a common optimum and $\eta L < 1$, whereas the distillation experiments run at $\eta L \approx 1$ and rely only on the exact two-to-one identity. For two-to-one compression, Lemma D.2 shows that heavy-ball momentum and independent minibatches (in expectation) leave the discrepancy $\eta^2(V_H\theta_0 - w)$ unchanged; general ratios under momentum, adaptive preconditioners such as Adam, whose effective step depends on the gradient history, and long horizons, over which the curvature is path dependent, lie outside the model.

Scope of the evidence. The main evidence is the official-schedule comparison on two-source CIFAR-10 partitions with its controls and the SVHN and CIFAR-100 replications (Appendices J.4–J.10); the short-schedule runs, the feature-skew family and the five-source condition are preliminary (Appendix J.14). One distillation method was used at the official schedule, and the accuracy gain was evaluated on two networks (Appendix J.11). Several controls rest on few units: three SVHN no-split replicates, three CIFAR-100 partitions per level, two partitions in the class-allocation-matched control and in the IID feature-space diagnostics.

Open questions. Three questions remain open. First, how the theory extends to longer horizons and other optimizers, where the linear superposition of source statistics and the frozen-quadratic approximation both fail. Second, whether a distillation method can bring the source error on a network into the regime of the law: the inequality $\mathcal{C}_{\mathrm{struct}} > \mathcal{U}_{\mathrm{src}}$ of Corollary 4.3 was not satisfied at the reference settings with either method (Appendices J.3 and J.5), the few initializations that satisfy it at $16\eta_{\mathrm{ref}}$ do so where the local prediction itself degrades, and the budget sweep does not settle whether the composed error stabilizes near $\mathcal{C}_{\mathrm{loc}}$ (Appendix J.9). Third, what produces the accuracy advantage of joint distillation: its dependence on heterogeneity differs between datasets and networks (Appendices J.10 and J.11), warm-starting joint distillation from the union recovers only a small part of it and class reweighting gives no detectable accuracy change (Appendices J.7 and J.6), and the feature-space diagnostics of Appendix J.12 are consistent with a coordination deficit between sets optimized without knowledge of each other without identifying the cause.

Future work. Four directions follow: the endpoint-level dissociation at ImageNet scale; a composition-aware objective that adds the penalty $\mathcal{R}_{\mathrm{comp}} = \mathbb{E}_{\theta_0}\|\Phi^{K,\eta}_{\oplus_i \alpha_i S_i}(\theta_0) - \Phi^{T,\eta}_{\oplus_i \alpha_i D_i}(\theta_0)\|_2^2$ to source-wise fidelity through a surrogate in the synthetic samples; a budget-matched study of the number of sources (Appendix J.14); and a coordination mechanism that does not pool data.