# WIDTH OR INTERACTION FIDELITY? ALLOCATING A FIXED STATE BUDGET IN ANALYTIC CONTINUAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Analytic continual learners fit a closed-form ridge readout on frozen features and carry only its sufficient statistics, the continuation state, across tasks. Expanding features into many random coordinates improves accuracy, but the Gram matrix in this state grows quadratically with width. We ask how a fixed continuation-state budget should be split between representation width and the fidelity of stored feature interactions. Because the feature–label statistic and the Gram diagonal grow only linearly with width, we keep them exact and compress only the interactions with Diagonal-Preserving Spectral Ridge (DPSR), which merges per-task spectra into a rank-bounded factor and solves a full-width readout in closed form. On three class-incremental benchmarks with ViT-B/16 features, DPSR stays within 0.1 points of a full-Gram readout of the same expansion in stage-average accuracy while storing 88–89% less continuation state. At equal budgets of 20–100 MiB with penalties chosen by generalized cross-validation (GCV), trading interaction rank for width beats exact ridge on a narrower expansion in 17 of 18 configurations by 0.17–2.26 points (three seeds, paired 95% intervals above zero) and loses 0.68 points where too little rank remains. With validation-and-refit selection, gains are smaller, clearest on CIFAR-100 and shared by other corrected sketches, and a penalty chosen before truncation can collapse accuracy after refitting. Under GCV, width thus outweighs interaction fidelity down to a dataset-dependent rank; the savings concern persistent state, not training time or peak memory.

## 1 INTRODUCTION

Analytic continual learners keep a pretrained backbone frozen and fit a ridge readout in closed form as tasks arrive (Zhuang et al., 2022; McDonnell et al., 2023; Zhuang et al., 2024). Because the readout depends on past data only through two additive statistics, the Gram matrix $G_t$ and the feature–label statistic $S_t$, these learners need neither a replay buffer nor gradient updates. Expanding the frozen features into $m \gg D$ random coordinates makes classes more linearly separable and improves accuracy further (McDonnell et al., 2023; Tran et al., 2025; Zou et al., 2026). Width, however, is paid for in the state that the learner carries from task to task, which we call the *continuation state*: the Gram matrix has $m^2$ entries, about 381 MiB in FP32 at $m = 10{,}000$, whereas $S_t$ has only $mK$ entries for $K$ classes. Whenever this state must be kept small, for instance when many task streams keep their own readouts on one shared frozen backbone, its quadratic term limits the width a learner can afford.

A fixed state budget therefore poses a choice that the literature has not isolated: keep exact interactions among fewer coordinates, or keep more coordinates and store how they interact only approximately. Existing analytic learners either store the full Gram of one representation (McDonnell et al., 2023; Zou et al., 2026) or compress it at a fixed width (Peng et al., 2025), and streaming sketches such as Frequent Directions (FD) and Robust FD (RFD) bound the error of a compressed Gram at a given size (Liberty, 2013; Ghashami et al., 2016; Luo et al., 2019). Comparisons across these methods confound training recipes with state sizes, so it remains open whether, at equal state, width is worth more than interaction fidelity, and when.

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.

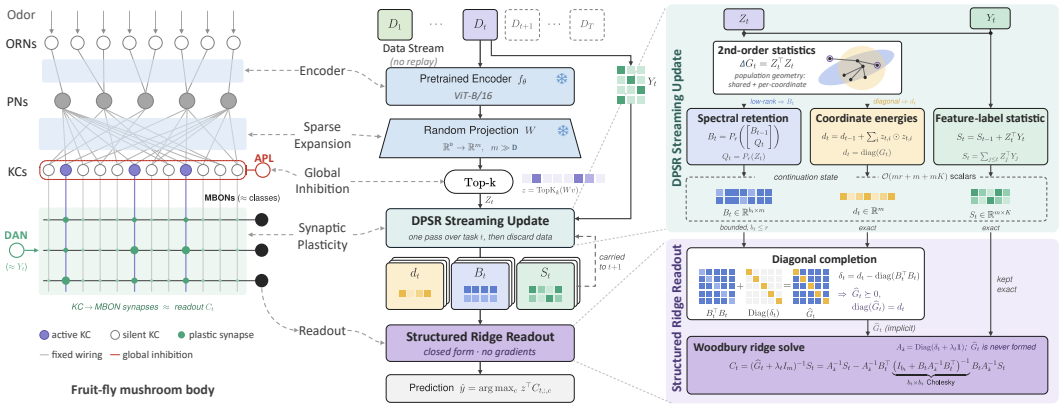



Figure 1: **DPSR keeps the per-coordinate statistics exact and compresses only the interactions.** Middle: frozen features pass through a fixed sparse random expansion with Top-$k$ (Zou et al., 2026). Right: for each task, DPSR updates the exact coordinate energies $d_t$ and label statistic $S_t$ and a rank-bounded interaction factor $B_t$, i.e., $b_t m + m + mK$ continuation scalars instead of $m^2 + mK$; diagonal completion restores the discarded energies, and a Woodbury solve yields a full-width classifier. Left: the fruit-fly mushroom body, a conceptual analogy only, not a model of the circuit (ORN, PN, KC: olfactory receptor, projection, Kenyon cells; APL, DAN, MBON: anterior paired lateral, dopaminergic, output neurons; Appendix B.6).

We study this question with a controlled design on CIFAR-100, CUB-200 and VTAB-50 with a frozen ViT-B/16. The cost structure of the ridge state (Section 2.3) suggests an allocation: the label statistic and the coordinate energies $d_t = \mathrm{diag}(G_t)$ grow linearly with width and can be kept exactly, and only the off-diagonal interactions, which grow quadratically, need compression. *Diagonal-Preserving Spectral Ridge* (DPSR; Figure 1) is the simplest realization of this allocation. It keeps $S_t$ and $d_t$ exact, merges each task's spectrum into a rank-bounded factor, restores every discarded coordinate energy by diagonal completion, and solves a full-width ridge readout with the Woodbury identity. Its components are established; DPSR contributes exact per-coordinate bookkeeping across the two-stage continual update, which makes it a transparent probe of the allocation rather than a new sketch. Standard identities show that its completed system stays well posed after every truncation and that isotropic completion only reparameterizes low-rank ridge (LR; Section 3.3). We compare it and four other compression rules with exact ridge on a narrower expansion of the same features at matched continuation budgets, sharing encoder, data, class orders, seeds and penalty selection, so that only width and interaction rank differ, and at fixed width we follow how much of each coordinate's energy the factor captures along the task stream.

Our answer is conditional: spending state on width pays while enough interaction rank remains, how much rank is enough depends on the data, and the size of the gain depends on how penalties are selected (Figures 2 and 3). Our contributions are:

- **A controlled matched-budget study of width versus interaction fidelity.** With penalties selected by generalized cross-validation (GCV), trading interaction rank for width beats narrow exact ridge in 17 of 18 width–rank configurations, by up to 2.26 points, and in 9 of 9 encoder and task-stream extensions. With validation and refit, the gains are smaller and mostly inconclusive outside CIFAR-100, where RFD and FD with diagonal completion (FD+diag) gain comparably to DPSR; since completion also cannot be separated from penalty selection, we claim no DPSR-specific gain.
- **Interaction fidelity along the task stream.** At fixed width, a rank-bounded factor captures a shrinking share of each coordinate's energy as tasks accumulate. With too little rank (256 on CIFAR-100), DPSR's deficit to the full Gram grows to 8.38 points by the last stage, whereas at rank 1024 DPSR stays within 0.1 points of the full Gram with 88.0–89.3% less continuation state, although several public methods remain more accurate.
- **A failure mode of full-width compressed readouts.** Select-before-truncation, characterized exactly for LR and, for DPSR, through the curvature left outside the retained span,

collapses first-task accuracy on CUB-200 in two of three seeds for DPSR, LR and FD, but not for RFD or FD+diag at the same penalty.

## 2 Problem: Where the Continuation State Grows

### 2.1 Analytic Readouts and Their State

We consider class-incremental learning in which the representation is fixed and only a linear readout is learned continually. Task $t$ supplies $n_t$ examples from new classes; past examples and task identities are unavailable. A frozen encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ is followed by a fixed sparse expansion $\phi_m(v) = \mathrm{TopK}_k(Wv)$, where each row of $W \in \mathbb{R}^{m\times D}$ has $p$ randomly placed Gaussian entries (Zou et al., 2026). Let $Z_t \in \mathbb{R}^{n_t\times m}$ stack the expanded features of task $t$ and $Y_t \in \{0,1\}^{n_t\times K}$ their one-hot labels; all $K$ classes are allocated from the start, and prediction is an argmax over all classes. For ridge loss with penalty $\lambda_t > 0$, the exact readout after task $t$ depends on past data only through two additive statistics (Zhuang et al., 2022; McDonnell et al., 2023; Zhuang et al., 2024):

$$G_t = \sum_{j\le t} Z_j^\top Z_j, \quad S_t = \sum_{j\le t} Z_j^\top Y_j, \quad C_t^\star = (G_t + \lambda_t I)^{-1} S_t. \tag{1}$$

We distinguish two parts of this state by how they scale with width. The *per-coordinate statistics* are $S_t$ ($mK$ scalars) and the coordinate energies $d_t = \mathrm{diag}(G_t)$ ($m$ scalars); both grow linearly in $m$. The *population interaction structure* is the off-diagonal part of $G_t$, which grows as $m^2$.

The state that must persist for learning to continue, $(G_t, S_t)$ or $m^2 + mK$ scalars for exact ridge, is the *continuation state*; the *total state* also counts the projection $W$ and the materialized $m \times K$ classifier, both of which can be regenerated (Appendix C). We budget continuation state because the learner must carry it between tasks and it grows quadratically with width. Total state still differs: at 43 MiB, narrow exact ridge carries 53.93, 55.01 and 53.38 MiB on CIFAR-100, CUB-200 and VTAB-50, every wide rule 76.10, 79.92 and 74.20 MiB (Table 27); peak GPU memory and runtime are reported separately (Section 4.4). States are counted in FP32 (1 MiB $= 2^{20}$ bytes). Exact readouts store the dense $m \times m$ Gram, as the full-Gram readouts we compare with do (Fly-CL: 385.3 MiB at $m = 10{,}000$, $K = 100$); storing only its upper triangle would let narrow exact ridge be about 1.4 times wider at every budget (e.g., $m_n = 4{,}647$ instead of 3,307 at 43 MiB on CIFAR-100; Appendix E), a stronger exact baseline that we did not evaluate.

### 2.2 The Allocation Problem

Fix a continuation budget $\beta$. A *narrow exact* learner spends it on a narrower expansion of the same frozen features, of width $m_n$ with $m_n^2 + m_n K \approx \beta$, and solves Eq. (1) exactly. A *wide compressed* learner keeps width $m > m_n$ and an approximation $\widehat{G}_t$ whose storage, together with any statistics kept exactly, fits in $\beta$. In a *matched-budget* comparison, the budget, encoder, data, class order, seeds, penalty selection and storage precision are shared, and only width and interaction rank differ. *Exact wide* ridge at the full width $m$ serves as an unbudgeted reference.

### 2.3 Only the Interactions Scale Quadratically

Three observations motivate the allocation. First, width helps when state is free: exact wide ridge ($m = 10{,}000$) has a higher mean stage-average accuracy than narrow exact ridge in all nine dataset–budget cells, by 0.35 to 2.00 points (Table 1, last row), yet a full-Gram readout at this width needs 383.4–389.1 MiB of continuation state (Fly-CL in Table 5). Second, nearly all of this cost lies in the interaction structure: at the same width, the per-coordinate statistics occupy only 1.9–7.7 MiB (Table 17). Third, spectral truncation discards cheap information: a rank-$r$ factor $B_t$ with $B_t^\top B_t \approx G_t$ costs $mr$ scalars, yet after the task stream the median coordinate retains only 0.675–0.790 of its energy at $r = 256$ and 0.808–0.958 at $r = 1024$ (Table 22; Figure 3a), although restoring all of it would cost only $m$ scalars.

Existing readouts occupy different points of this design space (Table 7; Appendix B): LoRanPAC truncates the spectrum and confines the classifier to the retained span (Peng et al., 2025), FD shrinks the spectrum and RFD adds an isotropic correction (Liberty, 2013; Luo et al., 2019), streaming ridge

sketches bound regression error (Shi & Phillips, 2021), MVP keeps within-view Gram blocks (Tran et al., 2025), and prompt- and adapter-based learners adapt the representation (Wang et al., 2022b;a; Zhou et al., 2024; Liang & Li, 2024). These works fix one representation or compare different state sizes; we fix the state and vary how it is spent.

# 3 DIAGONAL-PRESERVING SPECTRAL RIDGE

DPSR instantiates the allocation principle of Section 2.3 in its simplest form: its continuation state $(B_t, d_t, S_t)$ holds the per-coordinate statistics exactly and summarizes the population interaction structure by a rank-bounded factor. The individual components are established: per-task truncated SVD followed by a merge is the kind of continual spectral update used by LoRanPAC and Frequent Directions (Peng et al., 2025; Liberty, 2013; Ghashami et al., 2016), low-rank-plus-diagonal approximations date back to sparse Gaussian processes (Snelson & Ghahramani, 2005), and the solve is the Woodbury identity (Henderson & Searle, 1981). What DPSR adds is exact per-coordinate bookkeeping across the two-stage continual update (Figure 1).

## 3.1 STREAMING STATE UPDATES

For each task, DPSR accumulates $S_t = S_{t-1} + Z_t^\top Y_t$ and $d_t = d_{t-1} + \sum_i z_{t,i} \odot z_{t,i}$ exactly. For a matrix $X \in \mathbb{R}^{\ell \times m}$ with SVD $X = U\Sigma V^\top$, let $\mathcal{P}_r(X) = \Sigma_{1:b,1:b} V_{:,1:b}^\top \in \mathbb{R}^{b \times m}$ with $b = \min(r, \ell, m)$, so that $\mathcal{P}_r(X)^\top \mathcal{P}_r(X)$ is the best rank-$b$ approximation of $X^\top X$ (Eckart & Young, 1936; Mirsky, 1960). DPSR compresses each task and merges it with the previous factor:

$$Q_t = \mathcal{P}_r(Z_t), \qquad B_t = \mathcal{P}_r(M_t), \quad M_t = [B_{t-1}; Q_t], \qquad B_0 = 0_{0 \times m}. \tag{2}$$

The factor $B_t$ has $b_t \le r$ rows. The task step discards the tail $L_t = Z_t^\top Z_t - Q_t^\top Q_t$ and the merge step discards $E_t = M_t^\top M_t - B_t^\top B_t$; both are positive semidefinite (PSD). Interaction storage is therefore independent of the number of tasks, but directions discarded early cannot be recovered, so $B_t^\top B_t$ need not be the best rank-$r$ approximation of $G_t$. In total, DPSR stores $b_t m + m + mK$ continuation scalars instead of $m^2 + mK$: the width $m$ enters linearly, and the rank $r$ controls how much of the budget goes to interactions.

## 3.2 FULL-WIDTH READOUT BY DIAGONAL COMPLETION

Let $R_t = G_t - B_t^\top B_t$ denote the discarded geometry, $\delta_t = \mathrm{diag}(R_t)$ its diagonal, $\Delta_t = R_t - \mathrm{Diag}(\delta_t)$ its off-diagonal part and $\tau_t = \mathrm{tr}(R_t)$ its trace. DPSR completes the factor with the discarded coordinate energies, which it computes from its own state:

$$\widehat{G}_t = B_t^\top B_t + \mathrm{Diag}(\delta_t), \qquad \delta_t = d_t - \mathrm{diag}(B_t^\top B_t). \tag{3}$$

In exact arithmetic, $\mathrm{diag}(\widehat{G}_t) = d_t$ and $G_t - \widehat{G}_t = \Delta_t$: every coordinate energy is restored, whereas no discarded interaction is. Rounding-level negative entries of $\delta_t$ are clipped after a tolerance check (Appendix D.10). With $A_\lambda = \mathrm{Diag}(\delta + \lambda \mathbf{1})$, the Woodbury identity gives the ridge solution on the completed Gram without forming any $m \times m$ matrix; only a $b \times b$ Cholesky system is solved:

$$C = (\widehat{G} + \lambda I)^{-1} S = A_\lambda^{-1} S - A_\lambda^{-1} B^\top (I + B A_\lambda^{-1} B^\top)^{-1} B A_\lambda^{-1} S. \tag{4}$$

Because all of $S$ enters $C$ through $A_\lambda^{-1} S$, including its component outside the row span of $B$, the readout is not confined to the retained span, unlike principal-component regression (PCR) or LoRanPAC-style truncation. This full-width property is a double-edged design choice: directions outside the span receive curvature only $\lambda + \sum_i \delta_i v_i^2$ (Lemma 2), so when $\lambda$ is small, a small label component there can dominate $C$.

The penalty is selected either by task-local generalized cross-validation (GCV) (Golub et al., 1979) on the current task's spectrum, or by current-task validation: DPSR builds a temporary state from the pre-task state and 90% of the task, scores the held-out 10%, discards the temporary state and refits on the full task at the selected $\lambda$ (Algorithm 1, Appendix C). The structured solve costs $O(mr^2 + r^3 + mrK)$ instead of $O(m^3)$, but training time is dominated by the task SVD and by penalty selection.

### 3.3 Properties of the Completed Readout

The following statements are standard identities adapted to the two-stage update; each clarifies one aspect of the allocation. They assume fixed features, at least two classes, $\lambda > 0$ and exact arithmetic; proofs are in Appendix D. First, restoring the cheap statistics keeps the ridge system well posed: under the updates (2) with exactly accumulated $d_t$, $R_t \succeq 0$, $\delta_t \geq 0$, $\widehat{G}_t \succeq 0$ and $\mathrm{diag}(\widehat{G}_t) = d_t$ after every truncation, and for a fixed factor $B$, $u = \delta$ is the unique Frobenius-optimal diagonal completion $B^\top B + \mathrm{Diag}(u)$ of $G$ (Proposition 3, Appendix D.2). This optimality concerns approximating $G$ with $B$ held fixed, not classifier error. The next statement shows how completion enters the readout: as a penalty rather than as geometry.

**Proposition 1** (Regularization view)**.** *In exact arithmetic with $\lambda > 0$ and $B, \delta$ as in Proposition 3, $C = (\widehat{G} + \lambda I)^{-1} S$ is the unique minimizer of $\|BC\|_F^2 + \sum_i (\lambda + \delta_i)\|C_{i,:}\|_2^2 - 2\langle S, C\rangle$. At fixed $B, S, \delta$, isotropic completion $B^\top B + \bar{\delta} I$ satisfies $C_{\mathrm{iso}}(\lambda) = C_{\mathrm{LR}}(\lambda + \bar{\delta})$, where $C_{\mathrm{LR}}(\gamma) = (B^\top B + \gamma I)^{-1} S$ is low-rank ridge (LR); hence the positive-$\lambda$ isotropic path lies in the LR path.*

Completion thus acts as a coordinate-dependent penalty $\lambda + \delta_i$. The reparameterization is proved only for isotropic completion versus LR, where any difference after selection is a difference in the selected penalty; for DPSR versus LR no such identity holds, and that their difference also reflects penalty selection is a hypothesis, consistent with the fine-grid control of Section 4.2. The same coupling between penalty and truncation underlies the failure mode described next.

**Lemma 2** (Select-before-truncation)**.** *Let $B \in \mathbb{R}^{b \times m}$, $S \in \mathbb{R}^{m \times K}$ and $\lambda > 0$ (exact arithmetic), let $\Pi$ be the orthogonal projector onto the row span of $B$, and let $S_\parallel = \Pi S$, $S_\perp = (I - \Pi)S$. Then $C_{\mathrm{LR}}(\lambda) = (B^\top B + \lambda I)^{-1} S_\parallel + \lambda^{-1} S_\perp$, and the two terms are column-wise orthogonal. A label component along a right singular vector of $B$ with singular value $\sigma$ is scaled by $(\sigma^2 + \lambda)^{-1}$, and one orthogonal to the span by $\lambda^{-1}$: a relative amplification of $(\sigma^2 + \lambda)/\lambda$. If the state is exact, i.e., $B^\top B = Z^\top Z$ and $S = Z^\top Y$ for the stacked data $(Z, Y)$, then $S_\perp = 0$. For DPSR, only a Rayleigh-quotient statement holds: for a unit vector $v$ orthogonal to the span, $v^\top (\widehat{G} + \lambda I) v = \lambda + \sum_i \delta_i v_i^2$.*

The lemma explains why validation can miss the problem: if the entire validation state is exact, as on the first CUB-200 task, then $S_\perp = 0$ and a small $\lambda$ looks safe, but after a truncating refit the same $\lambda$ amplifies $S_\perp \neq 0$. For DPSR we claim no orthogonal decomposition. Appendix D.6 bounds how discarded interactions change predictions; the bound needs a common penalty and the exact-wide classifier, so it is retrospective and we do not use it to explain our results.

## 4 Experiments

We ask whether width is worth more than interaction fidelity at equal state (Section 4.1), what the state must retain (Section 4.2), when the allocation fails (Section 4.3), and how DPSR compares with the full Gram and public methods, at what cost (Section 4.4).

We use CIFAR-100 and CUB-200 split into ten tasks and VTAB-50 split into five (Krizhevsky, 2009; Wah et al., 2011; Zhai et al., 2019); CUB-200 (9,430/2,358 train/test images) and VTAB-50 (1,796/8,619) use local partitions rather than official splits. Unless stated otherwise, features come from a frozen ViT-B/16 (ImageNet-21K pretraining, ImageNet-1K fine-tuning), expanded to $m = 10{,}000$ coordinates with $p = 300$ nonzeros per row and $k/m = 0.3$. Narrow exact ridge competes with five wide rules, LR on the leading spectrum, FD, RFD, FD+diag and DPSR, where LR, FD and RFD receive one extra factor row to offset the $m$ energies that DPSR stores; the unbudgeted exact-wide reference uses an FP64 Gram, and the public comparison adds prompt-, adapter- and analytic learners, among them Fly-CL, which keeps the full Gram of the same expansion as DPSR. We report stage-average accuracy $\bar{A} = T^{-1} \sum_t t^{-1} \sum_{j \leq t} \mathrm{acc}_{t,j}$, where $\mathrm{acc}_{t,j}$ is the accuracy on task $j$ after stage $t$, final accuracy $A_T$ and backward transfer as means$\pm$SD over seeds, and paired effects as "mean [95% paired $t$ interval]" over seed-level differences. An interval that contains zero is inconclusive, not evidence of equivalence; with three seeds and no correction for multiplicity, we read counts of intervals descriptively.

*Protocol A (public comparison)* runs every method with its own recipe over five seeds; Fly-CL and DPSR ($r = 1024$) select $\lambda$ by task-local GCV over $\{10^6, \ldots, 10^9\}$ and store FP32 state. Protocols B and C are both matched-budget evaluations; they differ in penalty selection, so we never pool their

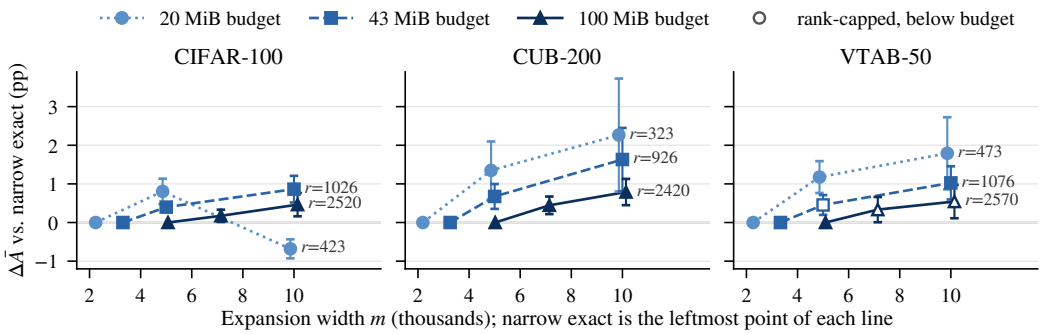



Figure 2: **Spending a fixed continuation budget on width pays off until the interaction rank becomes too small.** Paired change in $\bar{A}$ (points) of DPSR relative to narrow exact ridge at the same continuation budget, against width; Protocol C, three seeds. Each line is one budget and starts at narrow exact ridge (zero by construction); other points use a wider expansion and the rank $r$ that fills the budget. Bars: 95% paired intervals; lines join measured points only; hollow: rank-capped VTAB-50 points below budget.

Table 1: **Under validation and refit, matched-budget gains are clearest on CIFAR-100 and mostly inconclusive elsewhere.** Protocol B, three seeds. First row: $\bar{A}$ (%) of narrow exact; other rows: paired change (points), colored bold if the 95% paired interval excludes zero (Table 14); last row: unbudgeted exact wide (FP64 Gram) minus narrow exact.

| | CIFAR-100 | | | CUB-200 | | | VTAB-50 | | |
|---|---|---|---|---|---|---|---|---|---|
| Budget (MiB) | 20 | 43 | 100 | 20 | 43 | 100 | 20 | 43 | 100[§] |
| Narrow exact $\bar{A}$ (%) | 91.33 | 91.66 | 91.88 | 91.28 | 91.47 | 91.66 | 92.88 | 93.84 | 94.47 |
| *Spectral factor only* | | | | | | | | | |
| Low rank (LR) | **↑0.379** | ↑0.055 | **↑0.355** | ↓0.456 | ↓5.481 | ↑0.344 | ↑0.106 | ↑0.401 | ↑0.402 |
| FD | **↑0.534** | ↑0.179 | **↑0.392** | ↓0.220 | ↓5.284 | ↑0.270 | ↑0.447 | ↑0.182 | ↑0.402 |
| *Spectral factor with isotropic or diagonal correction* | | | | | | | | | |
| RFD | **↑0.510** | **↑0.578** | **↑0.449** | ↓0.164 | **↑0.472** | ↑0.502 | **↑1.768** | **↑0.737** | ↑0.402 |
| FD+diag | **↑0.184** | **↑0.547** | **↑0.424** | ↓0.157 | ↑0.079 | ↑0.260 | ↑1.199 | ↑0.498 | ↑0.407 |
| DPSR (ours) | ↑0.293 | **↑0.556** | ↑0.405 | **↓0.579** | ↓2.933 | ↓0.019 | **↑0.691** | ↑0.548 | ↑0.407 |
| Exact wide − narrow | 1.01 | 0.68 | 0.46 | 0.73 | 0.54 | 0.35 | 2.00 | 1.04 | 0.41 |

[§]On VTAB-50 at 100 MiB, the rank cap leaves the wide rules at 70.42–70.46 MiB.

numbers. In *Protocol B (matched budget, validation and refit)*, at 20, 43 and 100 MiB of continuation state, narrow exact ridge takes the largest width that fits and each wide rule keeps $m = 10{,}000$ with the rank that fills the budget (Appendix E); every readout selects $\lambda \in \{10^2, \ldots, 10^9\}$ by current-task validation and refit (Section 3.2), over three seeds, with FP32 states and FP64 solves, and a fine-grid control uses 40 penalty values. *Protocol C (matched budget, GCV)* selects every penalty, including that of narrow exact ridge, by task-local GCV, and varies DPSR's width–rank split at the same budgets (three seeds) and the encoder (ResNet-50, DINOv2 ViT-g/14) and stream length (five seeds; three for 50 tasks).

### 4.1 Is Width Worth More Than Interaction Fidelity at Equal State?

Under GCV, trading interaction rank for width pays off while enough rank remains (Figure 2; means in Table 16). Seventeen of the 18 wider DPSR allocations exceed narrow exact ridge at the same budget, with paired intervals above zero. The exception marks the limit: at 20 MiB on CIFAR-100, full width leaves rank 423 and loses 0.68 $[-0.93, -0.43]$ points, whereas width 5,000 with rank 947 gains $+0.81$ $[0.48, 1.14]$. The trade also holds across encoders and task streams (Table 2): against exact ridge on a narrower expansion whose continuation state matches DPSR's, DPSR is

Table 2: **Across encoders and task streams, DPSR beats state-matched narrow exact ridge.** Protocol C, five seeds (three for 50 tasks); mean $\bar{A}$ (%). Fly-CL: full Gram at $m = 10{,}000$; Narrow: exact ridge at widths 3,304/3,401/3,254, matching DPSR's continuation state (42.90/46.72/41.01 MiB); intervals in Table 8.

| | CIFAR-100 | | | CUB-200 | | | VTAB-50 | | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | Fly-CL | Narrow | **DPSR** | Fly-CL | Narrow | **DPSR** | Fly-CL | Narrow | **DPSR** |
| ResNet-50 | 84.12 | 82.62 | 83.24 | 78.27 | 75.14 | 78.60 | 93.29 | 92.87 | 93.26 |
| DINOv2 ViT-g/14 | 96.27 | 95.87 | 96.23 | 93.65 | 92.73 | 93.70 | 96.70 | 96.08 | 96.70 |
| ViT-B/16, 20/20/10 tasks | 93.01 | 92.29 | 93.01 | 91.63 | 90.27 | 91.71 | 95.59 | 94.75 | 95.59 |
| ViT-B/16, 50 tasks | 92.78 | – | 92.74 | 92.00 | – | 92.12 | 95.70 | – | 95.70 |

Table 3: **Without interactions the state fails under a shared penalty grid; span-restricted PCR loses on VTAB-50.** $\bar{A}$ (%), Protocol A, five seeds, mean±SD, $r = 1024$, $m = 10{,}000$, one penalty grid; bold: highest mean. DPSR: separate run of the Table 5 setting (one CUB-200 seed differs by 0.004).

| Retained state | CIFAR-100 | CUB-200 | VTAB-50 |
|---|---|---|---|
| Diagonal only | $83.58_{\pm 1.74}$ | $63.50_{\pm 2.96}$ | $36.66_{\pm 2.95}$ |
| PCR (in span) | $92.17_{\pm 0.90}$ | $91.44_{\pm 0.57}$ | $89.92_{\pm 1.90}$ |
| Low rank (LR) | $92.17_{\pm 0.85}$ | $91.54_{\pm 0.56}$ | $95.05_{\pm 0.47}$ |
| FD | $91.77_{\pm 0.95}$ | $\mathbf{91.67}_{\pm 0.57}$ | $\mathbf{95.09}_{\pm 0.46}$ |
| RFD | $92.63_{\pm 0.88}$ | $91.28_{\pm 0.59}$ | $95.07_{\pm 0.47}$ |
| DPSR (ours) | $\mathbf{92.78}_{\pm 0.88}$ | $91.46_{\pm 0.57}$ | $95.05_{\pm 0.47}$ |

more accurate in all nine encoder and stream cells, with paired gains of $+0.36$ to $+3.45$ points and intervals above zero.

Under validation and refit, the gains are smaller and concentrated on CIFAR-100 (Table 1). There, at 43 MiB, DPSR gains $+0.556$ $[0.423, 0.689]$ over narrow exact ridge, RFD $+0.578$ $[0.414, 0.742]$ and FD+diag $+0.547$ $[0.279, 0.815]$, and paired contrasts do not separate DPSR from either (DPSR−RFD: $-0.022$ $[-0.108, 0.064]$; DPSR−FD+diag: $+0.008$ $[-0.334, 0.351]$); plain LR and FD, without any correction, have intervals containing zero, and DPSR exceeds LR by $+0.501$ $[0.276, 0.727]$ on the 8-value grid. The gain is thus not specific to DPSR's rule, and some correction of the spectral factor appears to matter, although three seeds cannot establish that the corrected rules are equivalent and a finer grid does not separate DPSR from LR (Section 4.2). Across all 45 cells, 15 intervals lie above zero (11 on CIFAR-100; 6 for RFD), one lies below zero (DPSR on CUB-200 at 20 MiB: $-0.579$ $[-1.030, -0.129]$), and 29 contain zero; on CUB-200 at 43 MiB, LR, FD and DPSR fall below narrow exact with very wide intervals (e.g., LR: $-5.481$ $[-17.041, 6.079]$), a reversal traced in Section 4.3. The headroom under this protocol is small: even unbudgeted exact wide ridge exceeds narrow exact ridge by only 0.35–2.00 points.

The size of the gain thus depends on penalty selection. On CUB-200, narrow exact scores 0.98–1.92 points lower under GCV than under validation and refit, and the full-width DPSR gains of Protocol B are smaller than those of Protocol C in eight of nine cells.

### 4.2 What Must the State Retain?

Under a shared penalty grid, a state without interactions fails, and the benefit of keeping the diagonal exactly cannot be separated from penalty selection (Table 3). Keeping only the per-coordinate statistics yields $\bar{A}$ of 83.58, 63.50 and 36.66, far below every rank-1024 rule; since the penalty was not tuned for this state, this does not show that interactions are necessary in general. Confining the classifier to the retained span (PCR) lowers VTAB-50 $\bar{A}$ by 5.13 points relative to LR on the same factor, yet raises final accuracy (92.84 vs. 92.50 for LR and 92.49 for DPSR; Table 17), so keeping all of $S$ helps mainly at intermediate stages; on CIFAR-100 and CUB-200, PCR is close to LR. Among rules with a factor, an equal 40-value penalty grid at ranks 256 and 1024 (Protocol B) leaves all 12 paired intervals for DPSR versus LR or isotropic completion containing zero (mean differences from $-0.138$ to $+2.033$ points; Tables 18 and 19), which also qualifies the 8-value DPSR–LR contrast above.

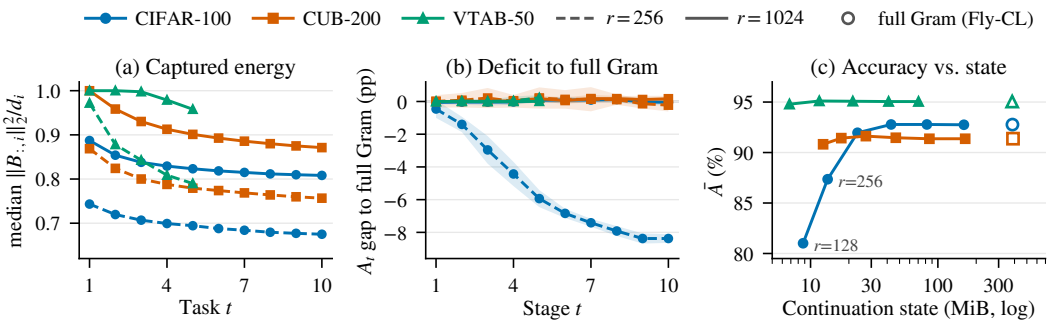



Figure 3: **Interaction fidelity along the task stream: a fixed rank captures less of each coordinate's energy as tasks accumulate, at a dataset-dependent cost in accuracy.** (a) Median of $\|B_{:,i}\|_2^2/d_i$ over coordinates after each task (fine-grid-control factors, three seeds; independent of the penalty). (b) Paired gap in $A_t$ (mean accuracy over seen tasks after stage $t$) between DPSR and the full Gram; bands: 95% paired intervals. (c) $\bar{A}$ against continuation state for ranks 128–4096 and the full Gram (hollow). (b, c): Protocol A, five seeds. All: $m = 10{,}000$; lines join measured points.

Table 4: **A penalty selected on an exact validation state collapses DPSR after a truncating refit.** First-task accuracy (%) on CUB-200, Protocol B, 43 MiB (capacity $r = 926$), one row per seed; $n_{\text{fit}}$, $n_{\text{refit}}$: examples in the validation fit and the refit. Projected $S$ is a diagnostic counterfactual, not a remedy.

| Seed | $\lambda$ | $n_{\text{fit}}$ | $n_{\text{refit}}$ | FP32 state | FP64 state | Exact wide | Projected $S$ (FP64) |
|---|---|---|---|---|---|---|---|
| 1993 | $10^7$ | 868 | 966 | 92.27 | 92.27 | 92.27 | 92.27 |
| 1994 | $10^2$ | 849 | 943 | 55.02 | 43.23 | 99.13 | 99.13 |
| 1995 | $10^2$ | 843 | 935 | 42.15 | 28.93 | 97.11 | 97.11 |

### 4.3 When Does the Allocation Fail?

The allocation fails in two identifiable ways. The first is too little interaction rank, whose cost builds up along the stream: at fixed rank, the factor captures a shrinking share of each coordinate's energy as tasks accumulate on every dataset (Figure 3a), and where that share is lowest, CIFAR-100 at rank 256, DPSR's deficit to the full Gram grows from 0.46 points after the first stage to 8.38 points after the last, whereas in the other five series the mean gap stays within 0.23 points of zero (Figure 3b). Accordingly, CIFAR-100 needs rank 1024 to match the full Gram, whereas CUB-200 and VTAB-50 peak at ranks 512 and 256 (Figure 3c; Table 15) and still gain at ranks 323 and 473 in Figure 2, where CIFAR-100 loses at rank 423. Captured energy and accuracy are associated across three datasets, not causally linked: CUB-200 keeps its accuracy with a median share of 0.76 at rank 256. The best split can therefore lie between the extremes.

The second is select-before-truncation, which explains the CUB-200 reversal: under Protocol B at 43 MiB, DPSR's CUB-200 mean falls below narrow exact ($-2.933$ $[-9.650, 3.783]$), as do LR and FD but not RFD and FD+diag (Table 12), whereas under Protocol C the same allocation gains $+1.63$ $[0.81, 2.45]$ (Figure 2). Table 4 traces DPSR's collapse to the first task: the validation fit subsets lie below the capacity $r = 926$, so the temporary state is exact, whereas the refits exceed it and truncate. Two seeds select the smallest candidate, $\lambda = 10^2$, and collapse after refit; the third selects $10^7$ and matches exact wide. At the same penalty, LR and FD collapse further (14.0–23.1%), whereas RFD and FD+diag stay at 96.7–98.7% (Appendix F.7). The collapse persists with an FP64 state, so FP32 storage is not its sole cause, and in the collapsed refits a label component outside the retained span, with a norm fraction below 0.034%, dominates the classifier. This is consistent with Lemma 2: on an exact validation state $S_\perp = 0$, so validation cannot detect the risk that truncation creates. Projecting $S$ onto the span restores exact-wide first-task accuracy in both seeds; it is a diagnostic counterfactual, not a remedy. The evidence is limited to the first CUB-200 task in two of three seeds; we have not traced the mechanism at later stages or on other datasets, explained why the RFD and FD+diag corrections protect them, or evaluated a refit-aware selection rule.

Table 5: **DPSR matches the full Gram of the same expansion with 88–89% less continuation state, but is not the most accurate method.** Protocol A, five seeds; $\bar{A}$ (%), mean±SD; MiB: continuation state (LoRanPAC: measured state of its implementation). Bold/underline: best/second best shown; all methods in Table 10.

| | CIFAR-100 | | CUB-200 | | VTAB-50 | |
|---|---|---|---|---|---|---|
| Method | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ |
| ACIL | $92.17_{\pm 1.16}$ | 518.2 | $89.06_{\pm 0.42}$ | 524.5 | $94.97_{\pm 0.54}$ | 515.1 |
| RanPAC | $\underline{92.90}_{\pm 0.90}$ | <u>385.3</u> | $\underline{92.58}_{\pm 0.46}$ | <u>389.1</u> | $\underline{95.21}_{\pm 0.46}$ | <u>383.4</u> |
| LoRanPAC | $\mathbf{93.33}_{\pm 0.82}$ | 4218.3 | $\mathbf{92.79}_{\pm 0.27}$ | 3177.7 | $\mathbf{95.39}_{\pm 0.41}$ | 879.3 |
| Fly-CL (full Gram) | $92.77_{\pm 0.89}$ | <u>385.3</u> | $91.37_{\pm 0.60}$ | <u>389.1</u> | $95.05_{\pm 0.46}$ | <u>383.4</u> |
| DPSR (ours) | $92.78_{\pm 0.88}$ | **42.9** | $91.47_{\pm 0.56}$ | **46.7** | $95.05_{\pm 0.47}$ | **41.0** |

### 4.4 How Close Is DPSR to the Full Gram, and at What Cost?

At fixed width, DPSR reaches the accuracy of the full Gram with a fraction of its state (Table 5; Figure 3c). Against Fly-CL, which keeps the full Gram of the same expansion, its mean $\bar{A}$ is within 0.1 points on all three datasets and its mean $A_T$ within 0.18 points (Table 10), with 88.0–89.3% less continuation state; the agreement extends to other encoders and longer streams, except with ResNet-50 on CIFAR-100 (Table 2). DPSR is not the most accurate method: RanPAC and LoRanPAC have higher means on all three datasets at 383.4–4,218.3 MiB (for LoRanPAC, its implementation's measured state), and F-OAL$^{\dagger}$, which fuses features from all encoder blocks, has the highest VTAB-50 mean with 8.0 MiB (Table 10); since recipes and state sizes differ, these comparisons do not isolate the allocation.

A compact continuation state does not make training cheaper. At 43 MiB on CIFAR-100, a task takes 0.9 s to fit with narrow exact ridge, 12.8 s with DPSR and 28.0 s with RFD, mostly for the task SVD and validation (Table 25); DPSR's peak GPU allocation exceeds narrow exact's only on CIFAR-100 (1226.4 vs. 645.2 MiB; 525.0 vs. 527.0 on CUB-200, 366.3 vs. 531.2 on VTAB-50; Table 27), although the structured solve itself is 13.0–15.0 times faster than a dense solve (Table 26). The savings are thus in persistent memory, not compute, and DPSR's total state (76.10 MiB on CIFAR-100) exceeds that of narrow exact ridge (53.93 MiB).

## 5 Conclusion

We asked how a fixed continuation-state budget should be split between representation width and interaction fidelity. Keeping the cheap per-coordinate statistics exact and compressing only the interactions, DPSR matches a full-Gram readout of the same expansion with 88–89% less continuation state. Spending the budget on width pays while enough interaction rank remains: under GCV it beats narrow exact ridge in 17 of 18 width–rank configurations and in all nine encoder and stream extensions. How much rank is enough depends on the data: a fixed rank captures less of each coordinate's energy as tasks accumulate, and the accuracy deficit builds up where it captures least. The benefit also depends on penalty selection: under validation and refit it is smaller and clearest on CIFAR-100, where differently corrected rules gain comparably, so we do not attribute it to DPSR's particular correction, and a penalty selected on an untruncated state can collapse a full-width compressed readout after refitting. Refit-aware penalty selection, packed or reduced-precision exact baselines, and trainable representations are natural next steps.

**Limitations.** The matched-budget studies use three seeds without multiplicity correction, and their conclusions depend on penalty selection: gains are consistent under GCV but mostly inconclusive under validation and refit outside CIFAR-100. The budget counts continuation state with a dense exact Gram; packed (about 1.4 times wider) and reduced-precision exact baselines are not reported, wide rules carry 20.8–24.9 MiB more total state at 43 MiB, and the exact-wide reference uses an FP64 Gram. Completion and penalty selection are not separated, and the diagonal-only control shares one penalty grid. Select-before-truncation is traced only on the first CUB-200 task, with no remedy evaluated. All results concern frozen features, one expansion family and local CUB-200/VTAB-50 partitions.

## REPRODUCIBILITY STATEMENT

Section 4 defines the three evaluation protocols, their seeds, penalty grids, selection rules and storage precision. Appendix C specifies training, task-local GCV, current-task validation with refit, and inference. Appendix D states every assumption and gives all proofs. Appendix E lists datasets, local partitions, encoder checkpoints, baseline adaptations, the recorded Protocol B widths and ranks, and state accounting. Our anonymized code release recomputes every table from the saved per-seed results and includes numerical checks of the stated identities and bounds against dense FP64 references.

## ETHICS STATEMENT

This work uses public image-classification benchmarks and pretrained encoders and involves no human subjects. Compact sufficient statistics are derived from labeled training data and may still encode information about it; they are not a privacy guarantee. The biological illustrations in Appendix B are adapted from published figures under CC BY licenses.

## AI USE STATEMENT

Generative AI tools were used to assist with mathematical understanding, manuscript editing, LaTeX formatting, and code development. All theoretical results and experimental findings were independently verified by the authors, who take full responsibility for the manuscript.

## REFERENCES

Yoshinori Aso, Daisuke Hattori, Yang Yu, Rebecca M. Johnston, Nirmala A. Iyer, Teri-T. B. Ngo, Heather Dionne, L. F. Abbott, Richard Axel, Hiromu Tanimoto, and Gerald M. Rubin. The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife*, 3: e04577, 2014. doi: 10.7554/eLife.04577.

Zhiwen Cao, Yanfeng Li, Shu-Dong Huang, Yalan Ye, Shuyin Xia, Yi Wang, and Jiancheng Lv. On the power of statistics in class-incremental learning with pretrained models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

N. Alex Cayco-Gajic and R. Angus Silver. Re-evaluating circuit mechanisms underlying pattern separation. *Neuron*, 101(4):584–602, 2019.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016. doi: 10.1137/15M1009718.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23 (1):53–60, 1981.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Carlos Hernández-Pérez, Eduardo Weruaga, and David Díaz. Lobe X of the cerebellum: A natural neuro-resistant region. *Anatomia*, 2(1):43–62, 2023. doi: 10.3390/anatomia2010005. URL https://www.mdpi.com/2813-0545/2/1/5.

Toshihide Hige, Yoshinori Aso, Mehrab N. Modi, Gerald M. Rubin, and Glenn C. Turner. Heterosynaptic plasticity underlies aversive olfactory learning in Drosophila. *Neuron*, 88(5):985–998, 2015. doi: 10.1016/j.neuron.2015.11.003.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3):489–501, 2006.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pp. 189–206. American Mathematical Society, 1984.

Ron A. Jortner. Network architecture underlying maximal separation of neuronal representations. *Frontiers in Neuroengineering*, 5:19, 2013. doi: 10.3389/fneng.2012.00019. URL https://www.frontiersin.org/journals/neuroengineering/articles/10.3389/fneng.2012.00019/full.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Edo Liberty. Simple and deterministic matrix sketching. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 581–588, 2013.

Andrew C. Lin, Alexei M. Bygrave, Alix de Calignon, Tzumin Lee, and Gero Miesenböck. Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination. *Nature Neuroscience*, 17(4):559–568, 2014.

Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and L. F. Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

Luo Luo, Cheng Chen, Zhihua Zhang, Wu-Jun Li, and Tong Zhang. Robust frequent directions with application in online learning. *Journal of Machine Learning Research*, 20(45):1–41, 2019.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

Saleh Momeni, Changnan Xiao, and Bing Liu. AnaCP: Toward upper-bound continual learning via analytic contrastive projection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/699b19e638a086ed3a6d1710c5aea504-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Maria Papadopoulou, Stijn Cassenaer, Thomas Nowotny, and Gilles Laurent. Normalization for sparse encoding of odors by a wide-field interneuron. *Science*, 332(6030):721–725, 2011.

Liangzu Peng, Juan Elenter Litwin, Joshua Agterberg, Alejandro Ribeiro, and René Vidal. Lo-RanPAC: Low-rank random features and pre-trained models for bridging theory and practice in continual learning. In *International Conference on Learning Representations*, 2025.

Ameya Prabhu, Shiven Sinha, Ponnurangam Kumaraguru, Philip H. S. Torr, Ozan Sener, and Puneet K. Dokania. RanDumb: Random representations outperform online continually learned representations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Benwei Shi and Jeff M. Phillips. A deterministic streaming sketch for ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 586–594, 2021.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, pp. 1257–1264, 2005. URL https://mlg.eng.cam.ac.uk/zoubin/papers/nips05spgp.pdf.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999.

Quyen Tran, Lam Tran, Khanh Doan, Toan Tran, Dinh Phung, Khoat Than, and Trung Le. Boosting multiple views for pretrained-based continual learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/480e563034dbb6d1dd622d8eab7d129b-Paper-Conference.pdf.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Shusen Wang, Alex Gittens, and Michael W. Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *International Conference on Machine Learning*, pp. 3608–3616, 2017.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

Hongwei Yan, Guanglong Sun, Kanglei Zhou, Qian Li, Liyuan Wang, and Yi Zhong. Flyprompt: Brain-inspired random-expanded routing with temporal-ensemble experts for general continual learning. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2602.01976`.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Da-Wei Zhou, Qi-Wei Wang, and Han-Jia Ye. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*, 2023.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564, 2024.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11602–11615, 2022.

Huiping Zhuang, Yuchen Liu, Run He, Kai Tong, Ziqian Zeng, Cen Chen, Yi Wang, and Lap-Pui Chau. F-OAL: Forward-only online analytic learning with fast training and low memory footprint in class-incremental learning. *arXiv preprint arXiv:2403.15751*, 2024.

Heming Zou, Yunliang Zang, Wutong Xu, and Xiangyang Ji. Fly-CL: A fly-inspired framework for enhancing efficient decorrelation and reduced training time in pre-trained model-based continual representation learning. In *International Conference on Learning Representations*, 2026.

# A NOTATION

Table 6: **Notation.** Each symbol has one meaning throughout the paper.

| Symbol | Meaning |
|---|---|
| $t, T, n_t$ | Current task, number of tasks, current examples |
| $D, m, m_n, p, k$ | Encoder, expansion and narrow widths; projection degree; active coordinates |
| $K, r, b_t, q_t$ | Class count, rank capacity, actual factor rows ($b_t \le r$), $\min(n_t, m)$ |
| $b(X)$ | Rows kept by $\mathcal{P}_r$ for a matrix $X$ with $\ell$ rows: $\min(r, \ell, m)$ |
| $f_\theta, W, Z_t, Y_t$ | Frozen encoder, sparse projection, expanded features, one-hot labels |
| $G_t, S_t, d_t$ | Cumulative Gram, feature–label statistic, coordinate energies $\mathrm{diag}(G_t)$ |
| $B_t, Q_t, M_t, \mathcal{P}_r$ | Shared factor, task factor, stacked factor, spectral truncation operator |
| $R_t, \delta_t, \Delta_t, \tau_t$ | Discarded geometry (PSD), its diagonal, its off-diagonal part, its trace |
| $L_t, E_t, \ell_t, e_t$ | Task and merge tails and their largest eigenvalues |
| $\widehat{G}_t$ | Completed Gram $B_t^\top B_t + \mathrm{Diag}(\delta_t)$ |
| $A_\lambda$ | $\mathrm{Diag}(\delta + \lambda \mathbf{1})$ in the Woodbury solve (Eq. 4, Algorithm 1) |
| $\widehat{H}, H^\star$ | System matrices $\widehat{G} + \lambda I$ and $G + \lambda I$ |
| $F_t, N_t, J_t$ | Intermediate matrices of the Woodbury solve (Algorithm 1) |
| $u$ | Generic diagonal completion vector (Proposition 3) |
| $\Pi, S_\parallel, S_\perp$ | Projector onto the row span of $B$, and the split of $S$ (Lemma 2) |
| $C_t, C_t^\star, C_n$ | DPSR, exact-wide and narrow-exact classifiers |
| $C_{\mathrm{LR}}, C_{\mathrm{iso}}$ | Low-rank ridge and isotropic-completion classifiers |
| $\lambda, \Lambda, \bar{\delta}$ | Ridge penalty, candidate grid, mean residual diagonal |
| $\varepsilon_{\mathrm{cap}}$ | State-computable coefficient cap (Proposition 5) |
| $\mu_w, \gamma_h, \zeta$ | Exact-wide top-one margin, signed label margin, allowance $2\|s_D - s_w\|_\infty$ |
| $\Xi, \widehat{\Xi}$ | Violation event of Proposition 5 and its realized-perturbation version |
| $\kappa$ | Heterogeneity index of the discarded diagonal (Appendix D.2) |
| $U, \Sigma, V$ | Factors of a singular value decomposition |
| $\nu_i, \Psi$ | GCV leverage and projected labels $U^\top Y$ (Algorithm 3) |
| $P$ | Expanded class-prototype matrix (Appendix F.8) |
| $h, a$ | Parameters of the synthetic generator only (Appendix F.12) |
| $c_{\mathrm{enc}}$ | Encoder cost per batch (Algorithms 1 and 2) |
| $\mathrm{acc}_{t,j}, \bar{A}, A_T, \mathrm{BWT}$ | Task accuracy after stage $t$, stage-average and final accuracy, backward transfer |
| subscripts $D, w, n$ | DPSR, exact-wide and narrow-exact quantities (Proposition 5, Lemma 6) |

$D$ is the encoder output dimension, whereas the subscript $D$ in $\mathrm{Acc}_D$, $\gamma_D$ and $s_D$ labels DPSR. All class-score statements assume at least two configured classes; with one class, decisions agree automatically.

# B DETAILED RELATED WORK

The design space in Table 7 separates width from interaction fidelity. Exact coordinate energy also appears in block-diagonal, diagonal-only and FD+diagonal states; DPSR combines it with shared spectral retention and the full label statistic. This appendix extends the positioning paragraph of Section 2.3.

## B.1 CONTINUAL LEARNING WITH PRETRAINED MODELS

Pretraining changes the representation problem in class-incremental learning. Earlier approaches address forgetting through parameter regularization, output distillation or exemplar-supported classification, including EWC, Learning without Forgetting and iCaRL (Kirkpatrick et al., 2017; Li & Hoiem, 2018; Rebuffi et al., 2017). These methods illuminate the distinction between preserving an existing function and accommodating new classes. The wider taxonomy includes rehearsal, architectural expansion and parameter-isolation approaches (Wang et al., 2024). DPSR addresses a narrower setting in which the feature map remains fixed and accumulated readout geometry is the continual resource.

Prompt-based methods use a pretrained model as a basis for task-dependent adaptation. L2P learns a prompt pool, DualPrompt separates complementary prompt roles, and CODA-Prompt constructs

Table 7: **Coordinate energy and interaction fidelity are separate choices.** Gram-component cost excludes $S$, projection and classifier; $k_v$ denotes views. Full-width means no retained-span constraint.

| Construction | Representative | Cost | Exact energy | Interactions | Full width |
|---|---|---|---|---|---|
| Full Gram | Fly-CL/RanPAC | $m^2$ | ✓ | All | ✓ |
| Leading spectrum | Low-rank ridge | $mr$ | × | Rank $r$ | ✓ |
| Truncated inverse | LoRanPAC-style | $mr$ | × | Rank $r$ | × |
| Shrunk spectrum | FD / RFD | $mr$ | × | Rank $r$ / isotropic | ✓ |
| Within-view blocks | MVP | $k_v d_a^2$ | ✓ | Within block | ✓ |
| Diagonal only | Control | $m$ | ✓ | None | ✓ |
| FD + diagonal | Control | $mr + m$ | ✓ | Shrunk + diagonal | ✓ |
| Shared factor + diagonal | DPSR (Ours) | $mr + m$ | ✓ | Rank $r$ + energies | ✓ |

decomposed attention-based prompts (Wang et al., 2022b;a; Smith et al., 2023). Their retained parameters and selection machinery influence the representation presented to the classifier. Comparing them with an analytic readout therefore involves more than changing a matrix approximation, even with shared class orders and preprocessing. DPSR compresses accumulated statistics of one fixed expanded representation rather than allocating trainable prompts.

Subspace and low-rank adaptation provide another way to use pretrained features. EASE expands an ensemble of subspaces, InfLoRA limits interference during low-rank adaptation, and the model-versus-exemplars analysis emphasizes how much pretrained structure can replace stored examples (Zhou et al., 2024; Liang & Li, 2024; Zhou et al., 2023). RanDumb further demonstrates the relevance of random representations as a continual-learning reference (Prabhu et al., 2024). These studies motivate strong simple baselines and explicit accounting for adaptation parameters. DPSR instead keeps the encoder and expansion fixed, allowing its continual-state trade to be studied through sufficient statistics.

Encoder families and evaluation collections affect that comparison. Residual networks, vision transformers and DINOv2 supply different feature geometries (He et al., 2016; Dosovitskiy et al., 2021; Oquab et al., 2024); ImageNet pretraining, CIFAR-100, CUB and VTAB define different sources of semantic structure and task granularity (Deng et al., 2009; Krizhevsky, 2009; Wah et al., 2011; Zhai et al., 2019). ImageNet-R and ImageNet-A add distinct distribution shifts (Hendrycks et al., 2021a;b). Our additional settings retain their actual encoder, partition and seed group. DPSR's contribution is the allocation of retained geometry within these settings, not a new encoder or dataset protocol.

### B.2 Analytic continual learning

Analytic learners make the cumulative quadratic objective explicit. ACIL develops a class-incremental analytic formulation, while F-OAL emphasizes forward-only online updates with low memory (Zhuang et al., 2022; 2024). Their sufficient-statistic or recursive updates can preserve the solution of a specified regression problem, conditional on feature and regularization choices. That algebraic property differs from a guarantee of unchanged classification accuracy as classes arrive. DPSR approximates the Gram component while retaining the complete feature–label statistic, making approximation of the cumulative objective its central object.

Random-feature analytic learners show why width matters in practice. RanPAC combines pretrained features with random projection and analytic classification, and Fly-CL introduces sparse expansion with a ridge readout (McDonnell et al., 2023; Zou et al., 2026). LoRanPAC studies low-rank random features and continual spectral approximation (Peng et al., 2025). These methods are especially relevant because they expose interactions between width, spectrum and linear readout quality. DPSR separates the cost of retaining a wide coordinate system from the cost of retaining its unrestricted second-order interactions.

Analytic components also appear inside broader pretrained systems. FlyPrompt uses a random-expanded analytic router, class-statistics methods examine the information retained by compact statistics, and AnaCP develops analytic contrastive projection (Yan et al., 2026; Cao et al., 2026; Momeni et al., 2025). Feature drift, routing targets and class sets determine whether a cumulative

Gram replacement is valid in such systems. A statistic substitution would require reproducing the native update and evaluation protocol. DPSR's present evaluation isolates frozen-feature readouts and does not treat these broader systems as automatic plug-in applications.

### B.3 Random features, expansion and separability

Random features provide explicit coordinates for nonlinear prediction. Random-feature kernels approximate kernel computations, extreme learning machines use fixed hidden features with analytic output fitting, and dimension-reduction theory explains when random linear maps preserve specified distances (Rahimi & Recht, 2007; Huang et al., 2006; Johnson & Lindenstrauss, 1984). These statements concern different maps and targets: distance preservation does not itself establish a ridge classification margin. DPSR conditions on the chosen expanded features and studies the geometry that must persist after those features have been observed.

Expansion-based continual analyses use separability, prototype correlation and margins as complementary currencies. RanPAC and Fly-CL study prototype geometry; MVP studies a Bayes margin under an expansive nonlinear random map and retains within-view Gram structure (McDonnell et al., 2023; Zou et al., 2026; Tran et al., 2025). The Bayes margin belongs to a hypothesis-family analysis, whereas the signed margins in our transfer result belong to fixed realized score vectors. The distinction matters for a noninvertible Top-$k$ feature map. Our transfer statement (Proposition 5) is instead a deterministic perturbation bound on realized scores at a common penalty, and it does not reuse those assumptions.

### B.4 Streaming sketches and ridge sketches

Frequent Directions gives a deterministic streaming summary by repeatedly shrinking singular values, and its analysis relates the sketch to covariance approximation (Liberty, 2013; Ghashami et al., 2016). Robust Frequent Directions supplements the spectral factor with an isotropic component (Luo et al., 2019). A shrunk spectrum can distribute error differently from hard truncation even when the storage cost is similar. DPSR uses task truncation followed by merge truncation and preserves the residual diagonal, while our equal-budget experiments retain FD, RFD and FD+diagonal as competing realizations.

Ridge-specific sketching results connect matrix approximation to the downstream objective. Deterministic streaming ridge sketches and analyses of classical versus Hessian sketching distinguish optimization error from statistical behavior (Shi & Phillips, 2021; Wang et al., 2017). Randomized matrix decompositions provide complementary tools for reducing factorization costs, and Nyström methods approximate kernel structure using selected components (Halko et al., 2011; Williams & Seeger, 2001). A good matrix approximation need not be the best classifier after separate parameter selection. We keep this distinction explicit: our identities concern the approximation, while the allocation experiments tune every method independently.

Spectral optimality is also conditional on the approximation class. Classical low-rank approximation characterizes optimal truncation for a single matrix, while unitarily invariant norm results clarify the role of its singular values (Eckart & Young, 1936; Mirsky, 1960). Sequential task/merge truncation need not equal the global best rank-$r$ factor of the final Gram. GCV and inverse-update identities then concern selection and solution of the resulting quadratic problem (Golub et al., 1979; Henderson & Searle, 1981). Proposition 3 fixes the spectral factor when stating optimal diagonal completion, and Appendix D.2 accounts separately for sequential spectral tails.

### B.5 Low-rank-plus-diagonal models

Low-rank-plus-diagonal structure predates its use in this continual state. Probabilistic PCA provides an isotropic-noise latent-variable model, and dimensionality-reduction work discusses related latent structure in neural recordings (Tipping & Bishop, 1999; Cunningham & Yu, 2014). Low-rank and sparse matrix decomposition additionally studies when structured components can be distinguished (Chandrasekaran et al., 2011). These probabilistic and identification questions are different from accumulating an uncentered feature Gram. DPSR uses the matrix form as a sufficient-statistic approximation and makes no likelihood-fitting claim.

**Algorithm 1** DPSR over a task stream; no example is retained across tasks

**Require:** Tasks $\mathcal{D}_{1:T}$; frozen $f_\theta, W$; width $m$, capacity $r$, sparsity $k$; penalty grid $\Lambda$

$B_0 \leftarrow 0_{0\times m}$; $d_0 \leftarrow 0_m$; $S_0 \leftarrow 0_{m\times K}$ ▷ continuation state

**for** $t = 1, \dots, T$ **do**

  $Z_t \leftarrow \mathrm{TopK}_k(W f_\theta(X_t))$ ▷ $c_{\mathrm{enc}}(X_t) + O(n_t mp + n_t m \log m)$

  $S_t \leftarrow S_{t-1} + Z_t^\top Y_t$; $d_t \leftarrow d_{t-1} + \sum_i z_{t,i}^{\odot 2}$ ▷ $O(n_t m)$

  $U_t \Sigma_t V_t^\top \leftarrow \mathrm{SVD}(Z_t)$; $Q_t \leftarrow \mathcal{P}_r(Z_t)$ ▷ $O(n_t m q_t)$, $q_t = \min(n_t, m)$

  $\lambda_t \leftarrow \textsc{Select}(\Lambda)$ from the pre-task state and current task ▷ GCV or validation; App. C

  $B_t \leftarrow \mathcal{P}_r([B_{t-1}; Q_t])$; $\delta_t \leftarrow [d_t - \mathrm{diag}(B_t^\top B_t)]_+$ ▷ $O(mr^2)$; tolerance check before clipping

  $A_t \leftarrow \mathrm{Diag}(\delta_t + \lambda_t \mathbf{1})$; $F_t \leftarrow A_t^{-1} B_t^\top$; $N_t \leftarrow A_t^{-1} S_t$ ▷ $O(mr + mK)$

  Solve $(I + B_t F_t) J_t = B_t N_t$ by Cholesky; $C_t \leftarrow N_t - F_t J_t$; discard $Z_t, Y_t$ ▷ $O(mr^2 + r^3 + mrK)$

**end for**

**return** $(B_T, d_T, S_T)$ and $C_T$ ▷ $b_T m + m + mK$ continuation scalars, $b_T \le r$

Sparse Gaussian-process constructions offer a particularly close algebraic precedent for diagonal correction. Pseudo-input approximations preserve information beyond a low-rank covariance through an input-dependent correction (Snelson & Ghahramani, 2005). The diagonal term also acts like a coordinate-dependent penalty in the corresponding linear system; the isotropic special case simply shifts ridge regularization. These precedents rule out novelty claims for the matrix form or for Woodbury. DPSR uses them as the simplest realization of an allocation, with exact per-coordinate bookkeeping across a continual update.

### B.6 Biological motivation

Expansion coding recodes inputs in a larger population (Figure 4; the left panel of Figure 1 sketches the mushroom body). Projection neurons feed Kenyon cells in the insect mushroom body, and mossy fibers feed cerebellar granule cells before Purkinje-cell readout (Litwin-Kumar et al., 2017; Cayco-Gajic & Silver, 2019). In the mushroom body, feedback inhibition supports sparse odor responses (Lin et al., 2014) and wide-field inhibition contributes global normalization (Papadopoulou et al., 2011); learning acts at the Kenyon-cell-to-output synapses (Aso et al., 2014; Hige et al., 2015). In *Drosophila*, about 2,000 Kenyon cells expand the olfactory input, so the number of plastic output synapses grows only linearly with the expansion (Aso et al., 2014; Litwin-Kumar et al., 2017). This motivated the resource question of the paper: what must be stored to learn a readout on top of a wide code?

The analogy has a tension that we state explicitly. If the label statistic $S$ is likened to the Kenyon-cell-to-output synapses, the circuit changes those synapses by local plasticity at a cost linear in the number of Kenyon cells. Our results point the other way for a ridge readout: a state with only per-coordinate statistics fails under the shared penalty selection (Section 4.2), and the readout needs some interaction structure. Biology motivates the question, not the solution; DPSR is not a model of the circuit.

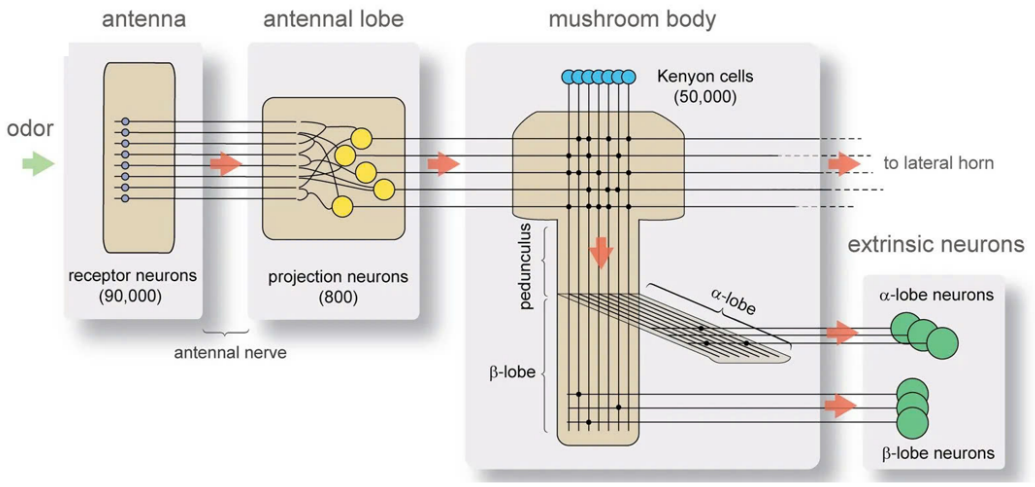
## C Detailed algorithms and workspace

Figure 1 (Section 1) summarizes the pipeline and the state that DPSR carries between tasks; Algorithms 1–2 give the exact procedures.

### C.1 Training and spectral updates

The method retains $(B, d, S)$ at task boundaries. A task with $n$ samples has feature matrix $Z \in \mathbb{R}^{n\times m}$, rank capacity $r$, actual factor rows $b \le r$ and $q = \min(n, m)$. The step Select does not change the retained state; with current-task validation it builds and discards a temporary state (Algorithm 4). Algorithm 1 gives the full-task refit; its costs assume dense economy SVD after sparse feature construction. The current implementation materializes expanded features, so a sparse expansion does not imply sparse SVD storage. For one-hot $Y$, the cross statistic is accumulated by class in $O(nm)$; constructing a dense one-hot product would be a different implementation cost.

**(A)**




**(B)**

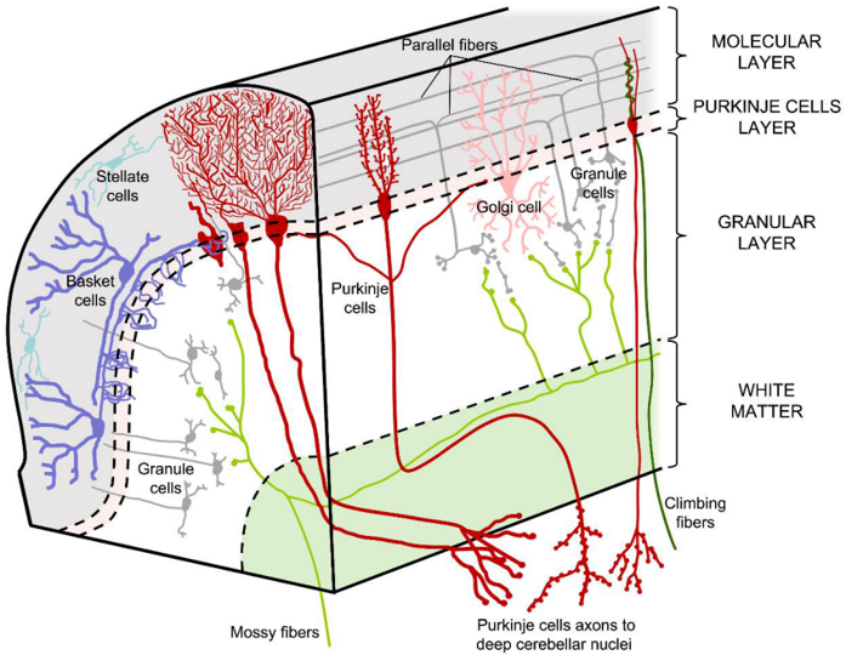



Figure 4: **Biological expansion coding (conceptual illustration).** (A) Insect mushroom body, adapted from Jortner (2013), Fig. 1A (CC BY). The cell counts printed in the panel are those of the source figure and differ from the *Drosophila* counts cited above; the panel should not be read as *Drosophila*. (B) Cerebellar cortex, adapted from Hernández-Pérez et al. (2023), Fig. 2 (CC BY 4.0). Biology motivates the resource question behind this work, not the algorithm.

The task SVD costs $O(nmq)$; merging a factor with at most $2r$ rows costs $O(mr^2)$ when $r \leq m$. The solve constructs $F \in \mathbb{R}^{m\times b}$, $N \in \mathbb{R}^{m\times K}$, a $b \times b$ system and a $b \times K$ right-hand side. Its arithmetic is $O(mb^2 + b^3 + mbK)$ and temporary storage $O(mb + mK + b^2 + bK)$ beyond the retained factors, excluding backend workspace. The current task's $O(nm)$ expanded features and economy-SVD outputs can dominate this temporary memory.

**Algorithm 2** Inference in the full expanded representation

**Require:** Frozen $f_\theta, W$, classifier $C \in \mathbb{R}^{m\times K}$, input $x$

$v \leftarrow f_\theta(x)$ ▷ $c_{\mathrm{enc}}(x)$
$z \leftarrow \mathrm{TopK}_k(Wv)$ ▷ $O(mp + m\log m)$
$s \leftarrow z^\top C$ ▷ $O(mK)$ dense; $O(kK)$ with sparse accumulation
**return** smallest-index maximizer of $s$ ▷ $O(K)$

**Algorithm 3** Task-local GCV selection (Protocols A and C)

**Require:** Current $Z, Y$, economy SVD $U\Sigma V^\top$, grid $\Lambda$

$\Psi \leftarrow U^\top Y$ ▷ $O(nqK)$; smaller with one-hot accumulation
**for** $\lambda \in \Lambda$ **do**
  $\nu_i \leftarrow \sigma_i^2/(\sigma_i^2+\lambda)$; $\mathrm{df} \leftarrow \sum_i \nu_i$ ▷ $O(q)$
  $\widehat{Y} \leftarrow U\,\mathrm{Diag}(\nu)\Psi$ ▷ $O(nqK)$
  $g_\lambda \leftarrow (\|Y-\widehat{Y}\|_F^2/n)/(1-\mathrm{df}/n)^2$ ▷ $O(nK)$
**end for**
**return** smallest $\lambda$ minimizing $g_\lambda$ ▷ $O(|\Lambda|)$

**Algorithm 4** Protocol B selection from the pre-task state: current-task validation and refit

**Require:** Pre-task state $(B_-, d_-, S_-)$; current $Z, Y$; grid $\Lambda$

Form stratified fit/validation indices ▷ $O(n)$
Update a temporary state using fit examples and the pre-task state ▷ $O(n_f m q_f + mr^2 + n_f m)$
**for** $\lambda \in \Lambda$ **do**
  Solve the temporary system at $\lambda$ ▷ $O(mr^2 + r^3 + mrK)$
  Score held-out current examples and record accuracy ▷ $O(n_v mK)$
**end for**
Choose highest validation accuracy; ties favor smaller $\lambda$ ▷ $O(|\Lambda|)$
Discard temporary state; update $(B_-, d_-, S_-)$ with all current examples ▷ $O(nmq + mr^2 + nm)$
Solve the refitted system at the selected nominal $\lambda$ ▷ $O(mr^2 + r^3 + mrK)$
Discard current examples and return the full-task state ▷ $O(1)$ logical release

### C.2 Inference

The classifier can be regenerated from continuation state by Eq. (4). A retained projection contributes its materialized nonzeros and indices; deterministic regeneration is an alternative only when its seed and construction recipe remain available. The reported continuation state excludes the projection and deployed classifier, while total non-backbone state counts both without double counting.

### C.3 Task-local GCV

GCV uses the current task's full economy spectrum, including zero singular values. Its loss is unnormalized ridge, so changing representation scale changes the effective role of the same nominal grid. Selection depends only on current-task data; it does not jointly optimize past and future task accuracy. For $\lambda > 0$, $\mathrm{df} < n$ in exact arithmetic, giving a positive denominator.

### C.4 Current-task validation and refit

The retained-state update is performed once on the full task; feature reuse and temporary validation updates are additional current-task work. A capacity crossing can make the temporary factor exact while the refit factor is truncated; Lemma 2 describes the consequence. Isotropic completion can also change its effective penalty across these two states.

### C.5 Workspace accounting

Continuation storage is $bm + m + mK$ scalars for DPSR, $m^2 + mK$ for exact ridge, and the corresponding declared factors/statistics for each comparator. During selection, the pre-task and temporary states coexist; SVD and Cholesky add workspace and the current examples remain accessible until refit completes. Logical tensor payload differs from allocator reservation and host-process

RSS. The isolated measurements in Appendix F.13 include training, selection and refit workspace but exclude encoder extraction and saved test caches. No past validation examples are retained.

# D THEORETICAL STATEMENTS AND PROOFS

## D.1 ASSUMPTIONS AND CONVENTIONS

All statements concern exact arithmetic, a fixed feature map, a fixed finite class set and $\lambda > 0$; the stored floating-point statistics approximate these ideal quantities. A positive margin means a unique maximizer; deterministic argmax ties use the smallest class index, and ties count as violations in every agreement condition. Singular data matrices and empty factors are allowed. Every statement is a standard identity or perturbation bound (orthogonal projection, Loewner order, Eckart–Young–Mirsky (Eckart & Young, 1936; Mirsky, 1960), Woodbury (Henderson & Searle, 1981), the resolvent identity) adapted to the two-stage update. For a matrix $X$ with $\ell$ rows we write $b(X) = \min(r, \ell, m)$ for the number of rows kept by $\mathcal{P}_r$, so $b_t = b(M_t)$.

## D.2 INVARIANTS AND FIXED-FACTOR COMPLETION

**Proposition 3** (Invariants and fixed-factor completion)**.** *In exact arithmetic, under the updates* (2) *with exactly accumulated $d_t$, for every $t$: $R_t \succeq 0$, $\delta_t \geq 0$, $\widehat{G}_t \succeq 0$ and $\operatorname{diag}(\widehat{G}_t) = d_t$. For any fixed $B$ and $u \in \mathbb{R}^m$, $\|G - B^\top B - \operatorname{Diag}(u)\|_F^2 = \|\Delta\|_F^2 + \|\delta - u\|_2^2$, so $u = \delta$ is the unique minimizer; its improvements over $u = 0$ and over the best isotropic $u = c\mathbf{1}$ are $\|\delta\|_2^2$ and $\|\delta - \bar{\delta}\mathbf{1}\|_2^2$, with $\bar{\delta} = m^{-1}\sum_i \delta_i$.*

The invariants hold after every truncation. The optimality concerns approximating $G$ with $B$ held fixed; it is not optimality in classifier error, and it does not hold jointly over $(B, u)$ (Appendix D.8).

*Proof of the invariants.* For $X = U\Sigma V^\top$ with right singular vectors $v_i$, $X^\top X - \mathcal{P}_r(X)^\top \mathcal{P}_r(X) = \sum_{i>b(X)} \sigma_i(X)^2 v_i v_i^\top \succeq 0$, so $0 \preceq \mathcal{P}_r(X)^\top \mathcal{P}_r(X) \preceq X^\top X$. Stacking adds Grams: $M_t^\top M_t = B_{t-1}^\top B_{t-1} + Q_t^\top Q_t$. Assume $B_{t-1}^\top B_{t-1} \preceq G_{t-1}$ (true for $t = 1$ since $B_0^\top B_0 = G_0 = 0$). Because the Loewner order is preserved under addition, $B_t^\top B_t \preceq M_t^\top M_t = B_{t-1}^\top B_{t-1} + Q_t^\top Q_t \preceq G_{t-1} + Z_t^\top Z_t = G_t$. Hence $R_t \succeq 0$, and its diagonal $\delta_t$ is nonnegative. $\widehat{G}_t = B_t^\top B_t + \operatorname{Diag}(\delta_t)$ is a sum of PSD matrices. Since $d_t$ is accumulated exactly, $d_t = \operatorname{diag}(G_t)$ and $\operatorname{diag}(\widehat{G}_t) = \operatorname{diag}(B_t^\top B_t) + d_t - \operatorname{diag}(B_t^\top B_t) = d_t$. □

All matrices above are $m \times m$; $B_t \in \mathbb{R}^{b_t \times m}$, $Q_t \in \mathbb{R}^{b(Z_t) \times m}$ and $M_t \in \mathbb{R}^{(b_{t-1} + b(Z_t)) \times m}$. With the tails $L_t = Z_t^\top Z_t - Q_t^\top Q_t$ and $E_t = M_t^\top M_t - B_t^\top B_t$, adding and subtracting $M_t^\top M_t$ gives $R_t = R_{t-1} + L_t + E_t$, hence

$$R_t = \sum_{j \leq t} (L_j + E_j), \qquad \tau_t = \operatorname{tr}(R_t) = \sum_{j \leq t} \Big( \sum_{i > b(Z_j)} \sigma_i(Z_j)^2 + \sum_{i > b(M_j)} \sigma_i(M_j)^2 \Big). \tag{5}$$

With $\ell_j$ and $e_j$ the largest eigenvalues of $L_j$ and $E_j$ (zero when nothing is discarded), the triangle inequality for PSD matrices gives $\|R_t\|_2 \leq \sum_{j \leq t} (\ell_j + e_j)$. The trace is computable from $(B_t, d_t)$ as $\tau_t = \sum_i \delta_{t,i}$; the tail sum needs per-update records that the state does not keep. Since $B_t^\top B_t$ has rank at most $r < m$, Eckart–Young–Mirsky in spectral norm gives $\lambda_{r+1}(G_t) \leq \|R_t\|_2$.

*Proof of fixed-factor completion.* Let $R = G - B^\top B$, $\delta = \operatorname{diag}(R)$ and $\Delta = R - \operatorname{Diag}(\delta)$. For any $u \in \mathbb{R}^m$, $R - \operatorname{Diag}(u) = \Delta + \operatorname{Diag}(\delta - u)$. $\Delta$ has zero diagonal and $\operatorname{Diag}(\delta - u)$ is diagonal, so $\langle \Delta, \operatorname{Diag}(\delta - u) \rangle_F = 0$ and $\|R - \operatorname{Diag}(u)\|_F^2 = \|\Delta\|_F^2 + \|\delta - u\|_2^2$. The unique minimizer is $u = \delta$. Its gap to $u = 0$ is $\|\delta\|_2^2$. For $u = c\mathbf{1}$, $\sum_i (\delta_i - c)^2$ is minimized at $c = \bar{\delta}$, with gap $\|\delta - \bar{\delta}\mathbf{1}\|_2^2$. □

This is an orthogonal-projection identity for the Gram approximation with $B$ fixed. It says nothing about classifier error, which depends on $\Delta C^\star$ (Proposition 5), and it does not optimize $B$ and $u$ jointly.

**Heterogeneity index.** Let $\kappa = \|\delta - \bar{\delta}\mathbf{1}\|_2^2/\|R\|_F^2$, with $\kappa = 0$ when $R = 0$. Its numerator is computable from the state; its denominator generally is not. For $R \succeq 0$ with eigenvalues $\rho_i \geq 0$, $\|\delta\|_2^2 \leq \|R\|_F^2 = \sum_i \rho_i^2 \leq (\sum_i \rho_i)^2 = \tau^2$, so for $R \neq 0$,

$$\frac{\|\delta - \bar{\delta}\mathbf{1}\|_2^2}{\tau^2} \leq \kappa \leq \frac{\|\delta - \bar{\delta}\mathbf{1}\|_2^2}{\|\delta\|_2^2}.$$

If $\tau = 0$ or $\|\delta\|_2 = 0$, positive semidefiniteness implies $R = 0$, and all three expressions are set to zero.

### D.3 Operator sandwich

**Proposition 4** (Operator sandwich and spectral tails)**.** *Under the assumptions of Proposition 3,* $-\operatorname{Diag}(\delta) \preceq \Delta \preceq R$ *and* $\|\Delta\|_2 \leq \|R\|_2 \leq \tau$. *For the task/merge update,* $R_t = \sum_{j \leq t}(L_j + E_j)$ *with PSD tails, and* $\|R_t\|_2 \leq \sum_{j \leq t}(\ell_j + e_j)$.

*Proof.* $\Delta = R - \operatorname{Diag}(\delta) \preceq R$ because $\operatorname{Diag}(\delta) \succeq 0$, and $\Delta + \operatorname{Diag}(\delta) = R \succeq 0$ gives $-\operatorname{Diag}(\delta) \preceq \Delta$. Hence $\lambda_{\max}(\Delta) \leq \lambda_{\max}(R) = \|R\|_2$ and $\lambda_{\min}(\Delta) \geq -\max_i \delta_i$. Since $0 \leq \delta_i = \mathbf{e}_i^\top R\,\mathbf{e}_i \leq \|R\|_2$, every eigenvalue of $\Delta$ lies in $[-\|R\|_2, \|R\|_2]$, so $\|\Delta\|_2 \leq \|R\|_2$. For PSD $R$, $\|R\|_2 \leq \operatorname{tr}(R) = \tau$. The tail statements were shown in Appendix D.2. □

*Remark.* The triangle inequality alone gives the weaker $\|\Delta\|_2 \leq 2\tau$. Both bounds use $\tau$, the *total* discarded energy, whereas $\|R\|_2$ can be much smaller; bounds in terms of $\tau$ are therefore loose by construction. Neither bound compares $\|\Delta C^\star\|_F$ with $\|RC^\star\|_F$, and neither implies a prediction advantage over low-rank ridge.

### D.4 Proposition 1: regularization view

*Proof.* The objective equals $\operatorname{tr}(C^\top \widehat{G} C) + \lambda\|C\|_F^2 - 2\langle S, C\rangle$, because $\|BC\|_F^2 + \sum_i \delta_i \|C_{i,:}\|_2^2 = \operatorname{tr}(C^\top(B^\top B + \operatorname{Diag}\delta)C)$. Its gradient is $2(\widehat{G} + \lambda I)C - 2S$, and $\widehat{G} + \lambda I \succ 0$ by Proposition 3, so the objective is strictly convex with unique minimizer $(\widehat{G} + \lambda I)^{-1}S$. With $C_{\mathrm{LR}}(\gamma) = (B^\top B + \gamma I)^{-1}S$ and $C_{\mathrm{iso}}(\lambda) = (B^\top B + (\lambda + \bar{\delta})I)^{-1}S$, $C_{\mathrm{iso}}(\lambda) = C_{\mathrm{LR}}(\lambda + \bar{\delta})$ by definition; since $\bar{\delta} \geq 0$, every $\lambda > 0$ maps to $\gamma = \lambda + \bar{\delta} > 0$, proving the path inclusion. □

The inclusion concerns one fixed state. A nominal $\lambda$ selected during validation uses the effective penalty $\lambda + \bar{\delta}_{\mathrm{val}}$ and, after refit, $\lambda + \bar{\delta}_{\mathrm{full}}$, so mapping a validation-time isotropic choice to LR need not reproduce the isotropic refit. For DPSR the penalty $\lambda + \delta_i$ varies across coordinates, so DPSR's path is not a reparameterization of the LR path; the proposition proves no statement about DPSR versus LR after selection. The penalty acts in the original coordinate basis, and although $B^\top B + \operatorname{Diag}(\delta)$ has a factor-analysis form, $G$ is an uncentered second-moment sum and no likelihood fit follows.

### D.5 Lemma 2: select-before-truncation

*Proof.* Let $B \in \mathbb{R}^{b \times m}$ have economy SVD $B = U_B \Sigma_B V_B^\top$ with positive singular values $\sigma_1, \ldots, \sigma_q$, $q = \operatorname{rank}(B)$, so that the row span of $B$ is $\operatorname{range}(V_B)$ and $\Pi = V_B V_B^\top$. The matrix $B^\top B + \lambda I = V_B \Sigma_B^2 V_B^\top + \lambda I$ is symmetric positive definite, has eigenvector $v_i$ (the $i$th column of $V_B$) with eigenvalue $\sigma_i^2 + \lambda$, and acts as $\lambda I$ on the orthogonal complement of $\operatorname{range}(V_B)$, since $Bw = 0$ there. Its inverse therefore maps $v_i$ to $v_i/(\sigma_i^2 + \lambda)$ and every $w \perp \operatorname{range}(V_B)$ to $w/\lambda$. Applying this column by column, $(B^\top B + \lambda I)^{-1}S = (B^\top B + \lambda I)^{-1}S_\parallel + \lambda^{-1}S_\perp$, where the first term lies in $\operatorname{range}(V_B)$ and the second in its orthogonal complement; corresponding columns are orthogonal. The ratio of the scalings of a unit component along $v_i$ and one orthogonal to the span is $(\sigma_i^2 + \lambda)/\lambda$. In particular, if $S_\parallel \neq 0$ and $\sigma_{\min}$ is the smallest positive singular value, $\|(B^\top B + \lambda I)^{-1}S_\parallel\|_F \leq \|S_\parallel\|_F/(\sigma_{\min}^2 + \lambda)$, so the $S_\perp$ term has Frobenius norm at least $\frac{\sigma_{\min}^2 + \lambda}{\lambda}\frac{\|S_\perp\|_F}{\|S_\parallel\|_F}$ times that of the $S_\parallel$ term.

If the state is exact, $B^\top B = Z^\top Z$ for the stacked features $Z$. Then $\operatorname{range}(B^\top) = \operatorname{range}(B^\top B) = \operatorname{range}(Z^\top Z) = \operatorname{range}(Z^\top)$, and $S = Z^\top Y \in \operatorname{range}(Z^\top)$ column by column, so $S_\perp = 0$. For

DPSR and a unit $v$ orthogonal to the row span, $Bv = 0$, hence $v^\top(\widehat{G} + \lambda I)v = v^\top \mathrm{Diag}(\delta)v + \lambda = \lambda + \sum_i \delta_i v_i^2 \in [\lambda, \lambda + \max_i \delta_i]$. □

For DPSR, $\mathrm{Diag}(\delta)$ generally mixes the row span of $B$ with its complement, so $(\widehat{G}+\lambda I)^{-1}$ does not map $S_\parallel$ and $S_\perp$ to orthogonal parts. The decomposition $C = \widehat{H}^{-1}S_\parallel + \widehat{H}^{-1}S_\perp$, with $\widehat{H} = \widehat{G} + \lambda I$, used in Section 4.3 and Appendix F.7, is linear but not orthogonal; the ratio $\|\widehat{H}^{-1}S_\perp\|_F / \|C\|_F$ is therefore not a fraction of energy. On the first task of a stream the pre-task state is empty; if the fit subset has at most $r$ examples, the temporary factor keeps every nonzero singular value, the temporary state is exact, and the lemma gives $S_\perp = 0$ at validation.

### D.6 Prediction transfer

Write $\widehat{H} = \widehat{G} + \lambda I$ and $H^\star = G + \lambda I$ (both $m \times m$), $C = \widehat{H}^{-1}S$ and $C^\star = (H^\star)^{-1}S$ (both $m \times K$).

**Proposition 5** (Prediction transfer)**.** *Assume exact arithmetic, fixed features, a fixed class set with at least two classes, and that DPSR and exact wide ridge are fitted on the same data at the same stage with the* same $\lambda > 0$*. Then*

$$C - C^\star = \widehat{H}^{-1}\Delta C^\star, \qquad \|C - C^\star\|_F \le \frac{\|\Delta C^\star\|_F}{\lambda + \min_i \delta_i} \le \varepsilon_{\mathrm{cap}} := \frac{\tau \|S\|_F}{\lambda(\lambda + \min_i \delta_i)}. \tag{6}$$

*Let $\mu_w(z)$ be the exact-wide top-one margin of input $z$ and $\Xi = \{\mu_w(z) \le 2\|z\|_2 \|C - C^\star\|_F\}$. For any narrow classifier $C_n$ and any evaluation distribution shared by the three classifiers,*

$$\mathrm{Acc}_D - \mathrm{Acc}_n \ge \mathrm{Acc}_w - \mathrm{Acc}_n - \Pr(\Xi). \tag{7}$$

*The statement is retrospective: $\Xi$ requires the exact-wide classifier.*

**Scope in our experiments.** Proposition 5 gives $\mathrm{Acc}_D - \mathrm{Acc}_n \ge \mathrm{Acc}_w - \mathrm{Acc}_n - \Pr(\Xi)$ for DPSR ($D$), exact wide ($w$) and any narrow classifier ($n$), where $\Xi$ is the event that the exact-wide top-one margin is at most the induced score change. It requires a common $\lambda$ for DPSR and exact wide ridge at the same stage, exact arithmetic and the exact-wide reference, so it is retrospective and does not apply directly to our results, in which every method selects its own $\lambda$. Its state-computable certificate covers no test point at any recorded stage, and in common-penalty diagnostics the realized-perturbation violation rate is 0.143 and 0.062 on CIFAR-100 and CUB-200 (Table 23), more than ten times the exact-wide gains at 43 MiB (0.68 and 0.54 points; a cross-protocol, order-of-magnitude comparison). The proposition therefore explains the structure of the trade-off but neither predicts nor certifies it.

*Proof.* $H^\star - \widehat{H} = G - \widehat{G} = \Delta$. The resolvent identity gives $C - C^\star = \widehat{H}^{-1}(H^\star - \widehat{H})(H^\star)^{-1}S = \widehat{H}^{-1}\Delta C^\star$. Since $\widehat{G} \succeq \mathrm{Diag}(\delta) \succeq (\min_i \delta_i)I$, $\lambda_{\min}(\widehat{H}) \ge \lambda + \min_i \delta_i$ and $\|\widehat{H}^{-1}\|_2 \le (\lambda + \min_i \delta_i)^{-1}$, which gives the first inequality via $\|XY\|_F \le \|X\|_2 \|Y\|_F$. Proposition 4 gives $\|\Delta C^\star\|_F \le \|\Delta\|_2 \|C^\star\|_F \le \tau \|C^\star\|_F$, and $H^\star \succeq \lambda I$ gives $\|C^\star\|_F \le \|S\|_F / \lambda$.

For each class $k$, $|z^\top(C - C^\star)_{:,k}| \le \|z\|_2 \|(C - C^\star)_{:,k}\|_2 \le \|z\|_2 \|C - C^\star\|_F$ by Cauchy–Schwarz. Outside $\Xi$, the exact-wide winner's score exceeds every competitor's by more than $2\|z\|_2\|C - C^\star\|_F$, and each score moves by at most half of that, so DPSR predicts the same class. Hence $\{\text{DPSR correct}\} \supseteq \{\text{exact wide correct}\} \setminus \Xi$ and $\mathrm{Acc}_D \ge \mathrm{Acc}_w - \Pr(\Xi)$; subtracting $\mathrm{Acc}_n$ proves Eq. (7). □

Accuracies are fractions; for percentage points multiply $\Pr(\Xi)$ by 100. For stage-average accuracy, the stagewise bounds are averaged with the metric's weights. The narrow classifier may be tuned separately, but DPSR and exact wide ridge must share $\lambda$, features, stage, class set and evaluation distribution. Our reported experiments select $\lambda$ separately for every method, so the proposition does not apply to them directly.

**Size of the cap.** At any stage with a zero-energy coordinate, $d_i = 0$ forces the $i$th column of $Z$ and of $B$ to vanish, so $\delta_i = 0$, $\min_i \delta_i = 0$ and $\varepsilon_{\mathrm{cap}} = \tau\|S\|_F/\lambda^2$; Table 22 records such coordinates. The denominator can be replaced by $\lambda + \lambda_{\min}(\widehat{G})$, because $\|\widehat{H}^{-1}\|_2 = 1/\lambda_{\min}(\widehat{H}) = 1/(\lambda + \lambda_{\min}(\widehat{G}))$ and $\lambda_{\min}(\widehat{G}) \geq \min_i \delta_i$. This does not help in the same case, since $\mathbf{e}_i^\top \widehat{G}\mathbf{e}_i = d_i = 0$ gives $\lambda_{\min}(\widehat{G}) = 0$.

**State-computable agreement certificate.** Let $j$ be DPSR's unique predicted class and $\mu(z) = (z^\top C)_j - \max_{j' \neq j}(z^\top C)_{j'}$ its own margin. If $\mu(z) > 2\|z\|_2\varepsilon_{\mathrm{cap}}$, the argument above with the roles of $C$ and $C^\star$ exchanged shows that exact wide ridge at the same $\lambda$ also predicts $j$. Every term is available from $(B, d, S, \lambda, z)$. The certificate concerns agreement, not correctness. If $R = 0$, $\varepsilon_{\mathrm{cap}} = 0$ and every positive margin passes.

### D.7 Pointwise margin exchange

**Lemma 6** (Realized margin exchange)**.** *Fix an input with label $y$, a stage and a class set with at least two classes. Let $s_D = z_w^\top C$, $s_w = z_w^\top C^\star$ and $s_n = z_n^\top C_n$ be the DPSR, exact-wide and narrow score vectors, $\gamma_h = s_{h,y} - \max_{j \neq y} s_{h,j}$ the signed label margins, and $\zeta = 2\|s_D - s_w\|_\infty$ the allowance. Then*

$$\gamma_D - \gamma_n \geq (\gamma_w - \gamma_n) - \zeta, \qquad \zeta \leq 2\|z_w\|_2\|C - C^\star\|_F, \tag{8}$$

*and $\gamma_w - \zeta > 0 > \gamma_n$ implies that DPSR is correct where narrow exact is wrong.*

*Proof.* Let $\epsilon = \|s_D - s_w\|_\infty$. Then $s_{D,y} \geq s_{w,y} - \epsilon$ and $\max_{j \neq y} s_{D,j} \leq \max_{j \neq y} s_{w,j} + \epsilon$, so $\gamma_D \geq \gamma_w - 2\epsilon$; subtracting $\gamma_n$ gives Eq. (8). This step holds for any finite score vectors. Cauchy–Schwarz gives $\epsilon \leq \|z_w\|_2\|C - C^\star\|_F$, and under the assumptions of Proposition 5 further $\epsilon \leq \|z_w\|_2\varepsilon_{\mathrm{cap}}$. A positive signed margin implies a correct unique prediction and a negative one an incorrect prediction. □

Margins are unrescaled ridge score gaps, not geometric distances or MVP's Bayes margins.

**A larger top-one margin does not imply pointwise dominance.** Take $Z = I_2$, $Y = I_2$, $\lambda = 1$ and $B = [1, 0]$. Then $B^\top B = \mathrm{Diag}(1, 0)$, $\delta = (0, 1)$ and $\widehat{G} = I_2 = G$, so $C = C^\star = I_2/2$. Narrow exact ridge on the first feature has coefficients $[1/2, 0]$. At $z = [1/2, 2]$ with label 1, the exact-wide and DPSR scores are $[1/4, 1]$ and the narrow scores $[1/4, 0]$. The wide top-one margin, $3/4$, exceeds the narrow one, $1/4$, and compression changes no score, yet wide/DPSR are wrong and narrow is correct. Top-one margin inequalities alone therefore cannot imply pointwise accuracy dominance. Our anonymized code release checks this witness and 300 random score configurations.

### D.8 Tightness and what was measured

The trace-based certificate has zero observed coverage at every recorded stage of the fine-grid control. The matched-effective identity of Proposition 1 holds with maximum relative discrepancy $4.39 \times 10^{-9}$ across these checks. Saved factors do not recover the full Gram or every discarded tail, so the bounds cannot be evaluated exactly on the large final states; small dense streams permit direct verification.

**Which violation event Table 23 measures.** Appendix F.10 describes these diagnostics as using the realized logit perturbation and a top-one margin rather than the trace bound. The natural realized event is $\widehat{\Xi} = \{\mu_w(z) \leq 2\|s_D(z) - s_w(z)\|_\infty\}$. Because $\|s_D - s_w\|_\infty \leq \|z\|_2\|C - C^\star\|_F$, $\widehat{\Xi} \subseteq \Xi$ and $\Pr(\widehat{\Xi}) \leq \Pr(\Xi)$ whenever the margin is the exact-wide margin $\mu_w$. The saved summaries do not state the norm of the perturbation or whose margin is used, so we report the column as a realized-perturbation violation rate and do not identify it with $\Pr(\Xi)$.

Equality cases explain the limits. When $R = 0$, the diagonal gap, the cap and the perturbation vanish. A diagonal residual gives $\Delta = 0$ although $\tau$ may be positive, making the cap loose; a rank-one PSD residual can have $\|R\|_2 = \tau$. A smaller Gram error can coexist with a classification loss, because the perturbation depends on $\Delta C^\star$ and on the data margins; our code release includes a sparse-feature example.

### D.9 Relation to LoRanPAC, MVP, FD/RFD and Fly-CL

**Full second-order geometry.** RanPAC connects expanded-feature separability and prototype decorrelation to an analytic readout, and Fly-CL studies sparse expansion and prototype correlations along its pipeline (McDonnell et al., 2023; Zou et al., 2026). Fly-CL's rank and Top-$k$ results rest on their own assumptions (energy concentration, a Lipschitz loss); we do not use them as proofs for our streams. DPSR changes the retained state after the map is fixed.

**Leading spectrum and continual truncation.** LoRanPAC bounds training and test MSE under its own recurrence and data/noise assumptions (Peng et al., 2025). Its cumulative tail term sums the largest discarded eigenvalue at each update; the closest DPSR analogue is $\sum_j(\ell_j+e_j)$, not $\tau_t$, which sums all discarded eigenvalues. Our two-stage update and full-$S$ objective differ, so its bounds do not transfer automatically.

**Shrunk spectrum and diagonal completion.** FD/RFD and deterministic streaming ridge already separate a wide feature map from compressed interaction statistics (Liberty, 2013; Luo et al., 2019; Shi & Phillips, 2021). RFD adds an isotropic correction, and FD+diag keeps exact coordinate energies just as DPSR does, provided its uncorrected residual is PSD. These are other realizations of the same allocation; our comparison is between realizations, not between the presence and absence of geometry-preserving sketches.

**Within-view geometry.** MVP studies a Bayes margin over measurable classifiers for an independent Gaussian expansion with an invertible, expansive elementwise nonlinearity, and decomposes the readout after replacing the Gram by within-view blocks (Tran et al., 2025). Its assumptions do not cover the noninvertible Top-$k$ map, and its blocks retain the original diagonal. DPSR instead summarizes interactions globally.

**Low-rank-plus-diagonal geometry.** The matrix form and the diagonal approximation principle have precedents, including sparse Gaussian-process approximations (Snelson & Ghahramani, 2005). DPSR's specific elements are its sequential update with exact per-coordinate bookkeeping and its evaluation at matched continual state.

### D.10 Numerical verification and comparator updates

The implementation computes $d - \operatorname{diag}(B^\top B)$ before clipping. Negative entries exceeding an absolute-plus-relative tolerance invalidate the PSD-based interpretation; for the stored FP32 states the tolerance is $256\epsilon_{32}\max(1, \|d\|_\infty, \|\operatorname{diag}(B^\top B)\|_\infty)$. Passing this diagonal check alone does not prove that the full residual is PSD. Large-state operator norms estimated by power iteration are estimates. The normalized solve residual is

$$\eta = \frac{\|\widehat{H}C - S\|_F}{\|\widehat{H}\|_2\|C\|_F + \|S\|_F}.$$

The small dense checks use the exact largest eigenvalue. Small backward error does not guarantee small forward error for an ill-conditioned matrix, and independent dense solves address a different question from reconstructing ideal statistics in higher precision.

Our anonymized code release evaluates the hard-truncation and FD updates against independent dense FP64 references (absolute tolerance $10^{-10}$, relative $10^{-8}$), over varied spectra and ranks, singular and zero streams, deterministic class ties, the Frobenius gaps, the operator and tail bounds, the resolvent identity and the agreement implications, and searches for counterexamples to "better Gram approximation implies better classification". These finite checks are not proofs, and we have not implemented rigorous floating-point certification.

For a merge factor with singular values $\sigma_1 \geq \cdots$, the FD implementation retains capacity $r$ and, when the merged row count exceeds $r$, subtracts $\sigma_r^2$ from the first $r$ squared singular values; the last retained row then has zero energy. Without compression it returns the exact factor. RFD additionally accumulates half this shrinkage in an isotropic correction $\alpha$. The uncorrected FD residual is PSD, but $G - (B_{\mathrm{FD}}^\top B_{\mathrm{FD}} + \alpha I)$ need not be, so a PSD-residual statement does not apply to the corrected RFD matrix. FD+diag uses its own exact coordinate energies and the residual diagonal of the FD factor. RFD's scalar has an eight-byte numeric payload.

# E EXPERIMENTAL DETAILS

## E.1 DATASETS, ENCODERS AND PREPROCESSING

Unless stated otherwise, comparisons use the stored ViT-B/16 checkpoint `augreg2_in21k_ft_in1k`, with ImageNet-21K pretraining and ImageNet-1K fine-tuning. This identity was confirmed by strict checkpoint loading and sampled reconstruction of cached test features. The feature cache is fixed within each paired group, and training labels and test ordering are checked. Reconstruction checks cover the first and last test feature of every task/seed rather than a complete re-extraction of all cached training features.

CIFAR-100 uses 50,000 training and 10,000 test examples. The local CUB-200 split contains 9,430/2,358 examples; the local VTAB-50 collection contains 1,796/8,619. Identical image-byte hashes are absent across the local CUB/VTAB training and test folders. Their results should be interpreted on these local partitions rather than as official-split leaderboard claims. The standard sequences contain ten, ten and five tasks, respectively. The 20/20/10-task extension uses its recorded finer class partitions; the 50-task extension uses one, two or four classes per task according to the dataset. All matrix-based metrics weight tasks uniformly.

The encoder extensions use ResNet-50 and DINOv2 ViT-g/14, with the latter producing 1536-dimensional features from 518-pixel inputs. Stored run configurations retain their feature preprocessing, expansion width and rank. Their five seeds preserve the paired class orders; the 50-task diagnostic has three seeds. These groups are displayed separately from the three-seed matched-budget evaluations.

## E.2 BASELINES AND ADAPTATIONS

The public table contains L2P, DualPrompt, EASE, InfLoRA, ACIL, F-OAL, RanPAC, LoRanPAC, Fly-CL and DPSR. Shared task boundaries, class orders, normalization and evaluation do not make the methods' training recipes identical: prompt/adaptation methods update trainable components, while analytic methods use their specified fixed-feature recipes. RanPAC and LoRanPAC retain the recorded no-PETL/no-adapter configurations. LoRanPAC's materialized implementation can retain substantially more state than a compact theoretical rank-factor formula; the table reports the measured per-component state.

InfLoRA$^\dagger$ retains its released CIFAR optimizer/rank/epoch recipe with the shared checkpoint and freezes those hyperparameters for CUB/VTAB. F-OAL$^\dagger$ retains all-block feature fusion, sigmoid projection and recursive least squares but uses the shared task boundaries, normalization and task-level evaluation; CUB/VTAB are supported by the local adapter. ACIL similarly uses the official random-buffer and analytic modules in the shared frozen-ViT, zero-base setting. The daggers identify adaptations rather than published native-regime replications.

## E.3 SELECTION, METRICS AND COMPARISON GROUPS

Protocol A (public comparison) uses five seeds and method-specific recipes; Fly-CL/DPSR select task-local GCV over four nominal powers of ten. Protocol C (matched budget, GCV) uses task-local GCV with three seeds for the width–rank sweep and the margin diagnostics, and five seeds (three for 50 tasks) for the encoder and stream extensions. Task-local GCV scores only the current task, whereas the selected penalty is applied to the cumulative Gram. Protocol B (matched budget, validation and refit) uses three seeds (1993, 1994, 1995) and current-task stratified validation followed by full-task refit, over eight powers of ten. Both break selection ties toward smaller nominal penalties. At 43 MiB, the recorded Protocol B configurations of all three seeds use narrow exact widths 3,307/3,258/3,332 on CIFAR-100/CUB-200/VTAB-50, DPSR ranks 1,026/926/1,076, and low-rank, FD and RFD ranks 1,027/927/1,077; the extra row of the latter three equals the $m$-entry energy vector that DPSR stores. The width sweep in Table 16 lists similar widths and ranks but belongs to Protocol C and is not a source for the Protocol B configuration. Fine-grid controls use 40 logarithmically spaced penalties with equal search budget for LR, isotropic completion and DPSR. All three seeds and both ranks are retained in that control.

The reported $\bar{A}$, $A_T$ and BWT are recomputed from task-by-stage accuracy matrices. Sample SD measures between-seed dispersion; paired $t$ intervals use seed-level differences, not tasks or class pairs as independent replicates. A confidence interval crossing zero is inconclusive rather than proof of equivalence. Rank and width grids are fixed in advance; we do not select a test-optimal operating point.

### E.4 Numerical reliability and state accounting

The matched-budget states are stored in FP32 with FP64 solves. All-class versus seen-class predictions were compared across selection and refit stages, with zero changes in the checked group. Raw residual diagonals are checked before clipping. The CUB/43 stored systems additionally agree with independent dense FP64 solves; separately reconstructed FP64 states address state-construction error. Operator norms in large-system normalized residuals use a power estimate, so these are numerical diagnostics rather than rigorous certificates.

The exact-wide reference uses FP64 Gram accumulation because an earlier FP32 full-Gram group produced nonfinite low-penalty solutions. Its label statistic and expanded inputs remain FP32, and its larger Gram payload is included in the table. This reference matches the representation and selection procedure but differs in statistic precision from the compact competitors. Consequently recovered width-gain ratios are descriptive comparisons of the recorded methods, not a precision-controlled instantiation of Proposition 5.

Continuation MiB counts unique logical state required for the next task, in FP32 scalars with $1\,\text{MiB} = 2^{20}$ bytes; exact readouts count the dense $m \times m$ Gram, as stored by the implementations. Packed storage of the upper triangle, $m(m+1)/2 + mK$ scalars, would allow narrow exact widths of 3,138/4,647/7,140 (CIFAR-100), 3,042/4,550/7,041 (CUB-200) and 3,186/4,697/7,190 (VTAB-50) at 20/43/100 MiB, instead of the dense widths 2,240/3,307/5,070, 2,191/3,258/5,020 and 2,264/3,332/5,095 used here. These widths follow from the formula; the accuracy of packed narrow exact ridge was not measured. total MiB adds available projection and deployed classifier state while excluding one shared backbone. DPSR includes $S, d, B$; RFD includes its correction scalar. In the matched-budget implementation this scalar has eight numeric bytes. Diagnostic exact Grams, saved evaluation caches and offline state snapshots are excluded from deployment state. The isolated profiling group uses RTX A5000 and PyTorch 1.13.1+cu117; the same-system structured/dense solve benchmark uses RTX 3090, so its latencies remain a separate measurement block.

### E.5 Additional encoder, sequence and domain-shift metrics

Table 8 gives the complete $\bar{A}/A_T$/BWT records, and Table 9 the domain-shift extension. Each reported value is linked to its per-seed result file, class-order hash, statistic definition and aggregation, so it can be rechecked without new training or hyperparameter search.

## F Additional experiments

### F.1 Public comparison: accuracy, forgetting and state

Table 10 reports the complete public-method metric set under Protocol A. All values are recomputed over the five seeds; F-OAL and InfLoRA retain their adaptation markers. LoRanPAC's state is the measured state of its implementation, which can exceed the state of a compact rank-factor formulation of the method.

### F.2 Allocation envelope at three budgets

Tables 11–13 report $\bar{A}$ and state for every budget and rule under Protocol B, and Table 14 all 45 paired intervals summarized in main Table 1; Figure 8 plots the same effects. We show these data once in the main text (Table 1) and once graphically here. Exact wide uses the same feature map with an FP64 Gram. The recovered-gain ratio is computed from group means; it is unstable near a zero denominator (CUB-200: $-79.3\%$, $-541.9\%$, $-5.5\%$ at 20/43/100 MiB, with denominators of 0.35–0.73 points), is confounded by the precision difference between the exact-wide reference and the budgeted rows, and carries no paired uncertainty. We therefore report it only descriptively.

Table 8: **Encoder and sequence extensions.** Protocol C; five seeds (three for 50 tasks); mean±SD. State is continuation MiB; –: not reported. ‡: narrow exact ridge at width 3,304/3,401/3,254 (CIFAR-100/CUB-200/VTAB-50; recorded run configuration), whose FP32 continuation state $m^2 + mK$ is 42.90/46.72/41.01 MiB by the accounting of Section 2.1, matching DPSR's 42.92/46.73/41.01 MiB; the MiB column lists only measured states. Paired DPSR−narrow differences (five seeds, 95% paired $t$ intervals): ResNet-50 $+0.62$ $[0.46, 0.78]$, $+3.45$ $[3.23, 3.68]$, $+0.39$ $[0.14, 0.64]$; DINOv2 $+0.36$ $[0.26, 0.46]$, $+0.97$ $[0.79, 1.15]$, $+0.62$ $[0.49, 0.75]$; 20/20/10 tasks $+0.73$ $[0.56, 0.90]$, $+1.44$ $[1.13, 1.75]$, $+0.84$ $[0.57, 1.12]$ (CIFAR-100, CUB-200, VTAB-50; `tables/generality_paired_effects.json`).

| Dataset | Setting | Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | BWT↑ | MiB↓ |
|---|---|---|---|---|---|---|
| CIFAR-100 | ResNet-50 | Fly-CL | $84.12_{\pm 1.08}$ | $76.71_{\pm 0.12}$ | $-7.86_{\pm 0.49}$ | – |
| CUB-200 | ResNet-50 | Fly-CL | $78.27_{\pm 1.23}$ | $69.54_{\pm 0.57}$ | $-9.49_{\pm 0.77}$ | – |
| VTAB-50 | ResNet-50 | Fly-CL | $93.29_{\pm 0.60}$ | $89.88_{\pm 0.39}$ | $-4.67_{\pm 0.89}$ | – |
| CIFAR-100 | ResNet-50 | Narrow‡ | $82.62_{\pm 1.16}$ | $74.00_{\pm 0.18}$ | $-9.28_{\pm 0.48}$ | – |
| CUB-200 | ResNet-50 | Narrow‡ | $75.14_{\pm 1.17}$ | $65.92_{\pm 0.35}$ | $-10.14_{\pm 0.30}$ | – |
| VTAB-50 | ResNet-50 | Narrow‡ | $92.87_{\pm 0.76}$ | $89.17_{\pm 0.50}$ | $-5.03_{\pm 1.14}$ | – |
| CIFAR-100 | ResNet-50 | DPSR (Ours) | $83.24_{\pm 1.12}$ | $75.79_{\pm 0.18}$ | $-7.98_{\pm 0.18}$ | – |
| CUB-200 | ResNet-50 | DPSR (Ours) | $78.60_{\pm 1.17}$ | $69.69_{\pm 0.31}$ | $-10.03_{\pm 0.78}$ | – |
| VTAB-50 | ResNet-50 | DPSR (Ours) | $93.26_{\pm 0.59}$ | $89.78_{\pm 0.42}$ | $-4.76_{\pm 0.83}$ | – |
| CIFAR-100 | DINOv2 g/14 | Fly-CL | $96.27_{\pm 0.66}$ | $94.17_{\pm 0.07}$ | $-2.49_{\pm 0.28}$ | – |
| CUB-200 | DINOv2 g/14 | Fly-CL | $93.65_{\pm 0.46}$ | $91.21_{\pm 0.14}$ | $-2.75_{\pm 0.11}$ | – |
| VTAB-50 | DINOv2 g/14 | Fly-CL | $96.70_{\pm 0.55}$ | $94.70_{\pm 0.35}$ | $-2.85_{\pm 0.38}$ | – |
| CIFAR-100 | DINOv2 g/14 | Narrow‡ | $95.87_{\pm 0.71}$ | $93.51_{\pm 0.07}$ | $-2.82_{\pm 0.38}$ | – |
| CUB-200 | DINOv2 g/14 | Narrow‡ | $92.73_{\pm 0.61}$ | $89.78_{\pm 0.20}$ | $-3.23_{\pm 0.16}$ | – |
| VTAB-50 | DINOv2 g/14 | Narrow‡ | $96.08_{\pm 0.64}$ | $93.78_{\pm 0.45}$ | $-3.39_{\pm 0.52}$ | – |
| CIFAR-100 | DINOv2 g/14 | DPSR (Ours) | $96.23_{\pm 0.67}$ | $94.08_{\pm 0.08}$ | $-2.55_{\pm 0.21}$ | – |
| CUB-200 | DINOv2 g/14 | DPSR (Ours) | $93.70_{\pm 0.47}$ | $91.34_{\pm 0.14}$ | $-2.74_{\pm 0.12}$ | – |
| VTAB-50 | DINOv2 g/14 | DPSR (Ours) | $96.70_{\pm 0.55}$ | $94.69_{\pm 0.37}$ | $-2.85_{\pm 0.37}$ | – |
| CIFAR-100 | ViT, 20/20/10 tasks | Fly-CL | $93.01_{\pm 0.84}$ | $88.79_{\pm 0.12}$ | $-4.39_{\pm 0.35}$ | – |
| CUB-200 | ViT, 20/20/10 tasks | Fly-CL | $91.63_{\pm 0.63}$ | $87.03_{\pm 0.22}$ | $-4.68_{\pm 0.16}$ | – |
| VTAB-50 | ViT, 20/20/10 tasks | Fly-CL | $95.59_{\pm 0.37}$ | $92.78_{\pm 0.23}$ | $-3.42_{\pm 0.41}$ | – |
| CIFAR-100 | ViT, 20/20/10 tasks | Narrow‡ | $92.29_{\pm 0.92}$ | $87.58_{\pm 0.21}$ | $-4.97_{\pm 0.28}$ | – |
| CUB-200 | ViT, 20/20/10 tasks | Narrow‡ | $90.27_{\pm 0.76}$ | $85.22_{\pm 0.22}$ | $-5.08_{\pm 0.48}$ | – |
| VTAB-50 | ViT, 20/20/10 tasks | Narrow‡ | $94.75_{\pm 0.57}$ | $91.46_{\pm 0.46}$ | $-4.01_{\pm 0.56}$ | – |
| CIFAR-100 | ViT, 20/20/10 tasks | DPSR (Ours) | $93.01_{\pm 0.84}$ | $88.72_{\pm 0.11}$ | $-4.45_{\pm 0.27}$ | – |
| CUB-200 | ViT, 20/20/10 tasks | DPSR (Ours) | $91.71_{\pm 0.60}$ | $87.17_{\pm 0.33}$ | $-4.70_{\pm 0.34}$ | – |
| VTAB-50 | ViT, 20/20/10 tasks | DPSR (Ours) | $95.59_{\pm 0.37}$ | $92.77_{\pm 0.24}$ | $-3.44_{\pm 0.44}$ | – |
| CIFAR-100 | ViT, 50 tasks | Fly-CL | $92.78_{\pm 0.73}$ | $88.72_{\pm 0.09}$ | $-4.58_{\pm 0.10}$ | 385.284 |
| CUB-200 | ViT, 50 tasks | Fly-CL | $92.00_{\pm 0.66}$ | $87.11_{\pm 0.32}$ | $-4.81_{\pm 0.24}$ | 389.099 |
| VTAB-50 | ViT, 50 tasks | Fly-CL | $95.70_{\pm 0.46}$ | $92.49_{\pm 0.16}$ | $-3.02_{\pm 0.68}$ | 383.377 |
| CIFAR-100 | ViT, 50 tasks | DPSR (Ours) | $92.74_{\pm 0.73}$ | $88.68_{\pm 0.08}$ | $-4.53_{\pm 0.12}$ | 42.915 |
| CUB-200 | ViT, 50 tasks | DPSR (Ours) | $92.12_{\pm 0.60}$ | $87.38_{\pm 0.18}$ | $-4.77_{\pm 0.50}$ | 46.730 |
| VTAB-50 | ViT, 50 tasks | DPSR (Ours) | $95.70_{\pm 0.45}$ | $92.48_{\pm 0.10}$ | $-3.09_{\pm 0.62}$ | 41.008 |

Table 9: **Under domain shift, DPSR's means stay close to Fly-CL's with less continuation state.** Protocol A extension, ViT-B/16, five seeds; accuracy in %, continuation in MiB.

| | ImageNet-R | | | ImageNet-A | | |
|---|---|---|---|---|---|---|
| Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | $A_T\uparrow$ | MiB↓ |
| RanPAC | $76.55_{\pm 0.98}$ | $70.93_{\pm 0.67}$ | 389.1 | $70.22_{\pm 1.18}$ | $62.73_{\pm 0.54}$ | 389.1 |
| LoRanPAC | $\mathbf{77.71}_{\pm 0.68}$ | $\mathbf{73.12}_{\pm 0.15}$ | 3177.7 | $\mathbf{70.87}_{\pm 1.16}$ | $\mathbf{64.07}_{\pm 0.41}$ | 3177.7 |
| Fly-CL | $76.77_{\pm 0.70}$ | $71.16_{\pm 0.36}$ | 389.1 | $69.88_{\pm 1.22}$ | $62.12_{\pm 0.56}$ | 389.1 |
| DPSR (Ours) | $76.36_{\pm 0.77}$ | $70.83_{\pm 0.39}$ | **46.7** | $69.79_{\pm 1.34}$ | $61.89_{\pm 0.59}$ | **46.7** |

### F.3 Rank sensitivity

Table 15 reports the complete rank sweep under Protocol A (task-local GCV, five seeds); rank is not selected by test performance.

### F.4 Intermediate-width allocation

Table 16 reports the intermediate allocations at 20, 43 and 100 MiB; these Protocol C runs are separate from Protocol B.

Table 10: **Public methods trade accuracy, forgetting and state differently.** Protocol A, ViT-B/16, five seeds; all metrics are recomputed from task matrices. MiB excludes the shared backbone; LoRanPAC's MiB is the measured state of its implementation, not the theoretical state of the method.

| Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | BWT↑ | Cont. MiB↓ | Total MiB↓ | Seeds |
|---|---|---|---|---|---|---|
| **CIFAR-100** | | | | | | |
| L2P | $88.03_{\pm 2.04}$ | $81.88_{\pm 1.55}$ | $-10.30_{\pm 1.11}$ | 327.84 | 327.84 | 5 |
| DualPrompt | $88.69_{\pm 1.47}$ | $83.26_{\pm 0.75}$ | $-6.47_{\pm 0.32}$ | 328.56 | 328.56 | 5 |
| EASE | $91.77_{\pm 0.90}$ | $86.95_{\pm 0.20}$ | $-6.47_{\pm 0.61}$ | 48.31 | 48.31 | 5 |
| InfLoRA† | $90.88_{\pm 1.16}$ | $85.09_{\pm 0.51}$ | $-5.08_{\pm 0.37}$ | 102.09 | 102.09 | 5 |
| ACIL | $92.17_{\pm 1.16}$ | $89.12_{\pm 0.04}$ | $\mathbf{-3.37}_{\pm 0.48}$ | 518.25 | 566.25 | 5 |
| F-OAL† | $92.06_{\pm 0.93}$ | $87.39_{\pm 0.13}$ | $-5.26_{\pm 0.29}$ | **8.39** | **11.70** | 5 |
| RanPAC | $92.90_{\pm 0.90}$ | $89.02_{\pm 0.09}$ | $-4.32_{\pm 0.25}$ | 385.28 | 418.40 | 5 |
| LoRanPAC | $\mathbf{93.33}_{\pm 0.82}$ | $\mathbf{89.83}_{\pm 0.11}$ | $-3.76_{\pm 0.18}$ | 4218.34 | 4549.46 | 5 |
| Fly-CL | $92.77_{\pm 0.89}$ | $88.79_{\pm 0.12}$ | $-4.38_{\pm 0.29}$ | 385.28 | 423.44 | 5 |
| DPSR (Ours) | $92.78_{\pm 0.88}$ | $88.74_{\pm 0.10}$ | $-4.45_{\pm 0.28}$ | 42.92 | 81.07 | 5 |
| **CUB-200** | | | | | | |
| L2P | $77.79_{\pm 1.90}$ | $66.96_{\pm 1.87}$ | $-10.86_{\pm 2.75}$ | 328.13 | 328.13 | 5 |
| DualPrompt | $79.54_{\pm 2.15}$ | $69.73_{\pm 1.57}$ | $-9.54_{\pm 2.26}$ | 328.85 | 328.85 | 5 |
| EASE | $86.97_{\pm 1.32}$ | $81.68_{\pm 0.17}$ | $-5.87_{\pm 0.48}$ | 51.83 | 51.83 | 5 |
| InfLoRA† | $79.43_{\pm 1.71}$ | $67.46_{\pm 0.68}$ | $-9.71_{\pm 1.94}$ | 100.21 | 100.21 | 5 |
| ACIL | $89.06_{\pm 0.42}$ | $83.17_{\pm 0.26}$ | $-6.03_{\pm 0.36}$ | 524.50 | 572.50 | 5 |
| F-OAL† | $90.38_{\pm 0.61}$ | $85.17_{\pm 0.27}$ | $-5.26_{\pm 0.54}$ | **9.16** | **12.85** | 5 |
| RanPAC | $92.58_{\pm 0.46}$ | $89.01_{\pm 0.68}$ | $\mathbf{-3.58}_{\pm 0.49}$ | 389.10 | 426.03 | 5 |
| LoRanPAC | $\mathbf{92.79}_{\pm 0.27}$ | $\mathbf{89.47}_{\pm 0.13}$ | $-3.60_{\pm 0.28}$ | 3177.67 | 3546.93 | 5 |
| Fly-CL | $91.37_{\pm 0.60}$ | $87.06_{\pm 0.25}$ | $-4.50_{\pm 0.32}$ | 389.10 | 431.07 | 5 |
| DPSR (Ours) | $91.47_{\pm 0.56}$ | $87.24_{\pm 0.31}$ | $-4.50_{\pm 0.38}$ | 46.73 | 88.70 | 5 |
| **VTAB-50** | | | | | | |
| L2P | $83.88_{\pm 3.38}$ | $73.48_{\pm 4.33}$ | $-11.93_{\pm 4.53}$ | 327.69 | 327.69 | 5 |
| DualPrompt | $84.16_{\pm 1.58}$ | $75.77_{\pm 2.83}$ | $-9.06_{\pm 2.21}$ | 328.41 | 328.41 | 5 |
| EASE | $90.72_{\pm 2.21}$ | $85.00_{\pm 2.37}$ | $-8.01_{\pm 2.75}$ | **6.54** | **6.54** | 5 |
| InfLoRA† | $86.55_{\pm 1.89}$ | $77.78_{\pm 3.17}$ | $-5.14_{\pm 1.80}$ | 91.05 | 91.05 | 5 |
| ACIL | $94.97_{\pm 0.54}$ | $92.43_{\pm 0.42}$ | $-3.45_{\pm 0.59}$ | 515.12 | 563.12 | 5 |
| F-OAL† | $\mathbf{95.97}_{\pm 0.33}$ | $\mathbf{94.10}_{\pm 0.29}$ | $-2.80_{\pm 0.56}$ | 8.01 | 11.13 | 5 |
| RanPAC | $95.21_{\pm 0.46}$ | $93.05_{\pm 0.32}$ | $-3.03_{\pm 0.46}$ | 383.38 | 414.58 | 5 |
| LoRanPAC | $95.39_{\pm 0.41}$ | $93.37_{\pm 0.26}$ | $\mathbf{-2.77}_{\pm 0.36}$ | 879.29 | 1191.33 | 5 |
| Fly-CL | $95.05_{\pm 0.46}$ | $92.49_{\pm 0.32}$ | $-3.50_{\pm 0.59}$ | 383.38 | 419.62 | 5 |
| DPSR (Ours) | $95.05_{\pm 0.47}$ | $92.49_{\pm 0.33}$ | $-3.49_{\pm 0.65}$ | 41.01 | 77.25 | 5 |

Table 11: **At 20 MiB, every wide rule has a higher mean than narrow exact on CIFAR-100 and VTAB-50 and a lower mean on CUB-200.** Protocol B, ViT-B/16, three seeds; $\bar{A}$ mean±SD and continuation MiB. Bold/underline: largest/second-largest $\bar{A}$ mean (MiB is matched by design and not ranked). Exact wide uses an FP64 Gram, the budgeted rows FP32 states. Recovered gain $(\bar{A}_{\mathrm{DPSR}} - \bar{A}_{\mathrm{narrow}})/(\bar{A}_{\mathrm{wide}} - \bar{A}_{\mathrm{narrow}})$ is a ratio of means, unstable when the denominator is small (CUB-200); it is descriptive only.

| | CIFAR-100 | | CUB-200 | | VTAB-50 | |
|---|---|---|---|---|---|---|
| Method | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ |
| Narrow exact | $91.33_{\pm 0.74}$ | 20.00 | $\mathbf{91.28}_{\pm 0.74}$ | 19.98 | $92.88_{\pm 0.81}$ | 19.98 |
| Low rank | $91.71_{\pm 0.71}$ | 19.99 | $90.82_{\pm 0.48}$ | 19.99 | $92.99_{\pm 1.62}$ | 19.99 |
| FD | $\mathbf{91.86}_{\pm 0.82}$ | 19.99 | $91.06_{\pm 0.74}$ | 19.99 | $93.33_{\pm 0.97}$ | 19.99 |
| RFD | $91.84_{\pm 0.76}$ | 19.99 | $91.11_{\pm 0.83}$ | 19.99 | $\mathbf{94.65}_{\pm 0.42}$ | 19.99 |
| FD+diag | $91.51_{\pm 0.70}$ | 19.99 | $91.12_{\pm 0.68}$ | 19.99 | $94.08_{\pm 0.48}$ | 19.99 |
| DPSR (Ours) | $91.62_{\pm 0.62}$ | 19.99 | $90.70_{\pm 0.63}$ | 19.99 | $93.58_{\pm 0.93}$ | 19.99 |
| *Exact wide* | $92.34_{\pm 0.80}$ | 766.8 | $92.01_{\pm 0.48}$ | 770.6 | $94.89_{\pm 0.14}$ | 764.8 |
| Width gain recovered | 28.9% | | -79.3% | | 34.5% | |

### F.5 State ablations

Table 17 reports the five-seed spectral controls and their state under Protocol A. Diagonal-only spends less state; PCR restricts the classifier span. Every rule, including diagonal only, uses the same task-local GCV grid $\{10^6, \ldots, 10^9\}$; the penalty was not tuned separately for the diagonal-only state, whose curvature differs greatly from that of the rules with a factor, so its low accuracy is a result under this shared selection rule only. The matched-budget FD+diagonal comparison is in Tables 1 and 12; main Table 3 shows the $\bar{A}$ column of this table. The DPSR row of this

Table 12: **At 43 MiB, every wide rule has a higher CIFAR-100 mean than narrow exact; LR, FD and DPSR have lower CUB-200 means.** Protocol B, ViT-B/16, three seeds; $\bar{A}$ mean±SD and continuation MiB. Bold/underline: largest/second-largest $\bar{A}$ mean (MiB is matched by design and not ranked). Exact wide uses an FP64 Gram, the budgeted rows FP32 states. Recovered gain as in Table 11; descriptive only.

| | CIFAR-100 | | CUB-200 | | VTAB-50 | |
|---|---|---|---|---|---|---|
| Method | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ |
| Narrow exact | $91.66_{\pm 0.81}$ | 42.98 | $91.47_{\pm 0.63}$ | 42.98 | $93.84_{\pm 0.22}$ | 42.99 |
| Low rank | $91.71_{\pm 0.76}$ | 42.99 | $85.99_{\pm 4.25}$ | 42.99 | $94.25_{\pm 0.86}$ | 42.99 |
| FD | $91.84_{\pm 0.86}$ | 42.99 | $86.18_{\pm 4.30}$ | 42.99 | $94.03_{\pm 1.04}$ | 42.99 |
| RFD | $\mathbf{92.23}_{\pm 0.77}$ | 42.99 | $\mathbf{91.94}_{\pm 0.62}$ | 42.99 | $\mathbf{94.58}_{\pm 0.36}$ | 42.99 |
| FD+diag | $92.20_{\pm 0.83}$ | 42.99 | $\underline{91.55}_{\pm 0.37}$ | 42.99 | $94.34_{\pm 0.56}$ | 42.99 |
| DPSR (Ours) | $\underline{92.21}_{\pm 0.76}$ | 42.99 | $88.54_{\pm 2.39}$ | 42.99 | $\underline{94.39}_{\pm 0.70}$ | 42.99 |
| *Exact wide* | $92.34_{\pm 0.80}$ | 766.8 | $92.01_{\pm 0.48}$ | 770.6 | $94.89_{\pm 0.14}$ | 764.8 |
| Width gain recovered | 81.4% | | -541.9% | | 52.6% | |

Table 13: **At 100 MiB, the wide rules have means close to exact wide on CIFAR-100 and VTAB-50.** Protocol B, ViT-B/16, three seeds; $\bar{A}$ mean±SD and continuation MiB. Bold/underline: largest/second-largest $\bar{A}$ mean (MiB is matched by design and not ranked). Exact wide uses an FP64 Gram, the budgeted rows FP32 states. Recovered gain as in Table 11; descriptive only. On VTAB-50 the rank cap leaves the wide rules at 70.42–70.46 MiB, below budget.

| | CIFAR-100 | | CUB-200 | | VTAB-50 | |
|---|---|---|---|---|---|---|
| Method | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ | $\bar{A}\uparrow$ | MiB↓ |
| Narrow exact | $91.88_{\pm 0.89}$ | 99.99 | $91.66_{\pm 0.65}$ | 99.96 | $94.47_{\pm 0.28}$ | 100.00 |
| Low rank | $92.24_{\pm 0.80}$ | 99.98 | $\underline{92.00}_{\pm 0.54}$ | 99.98 | $\underline{94.87}_{\pm 0.16}$ | 70.42 |
| FD | $92.27_{\pm 0.81}$ | 99.98 | $91.93_{\pm 0.40}$ | 99.98 | $\underline{94.87}_{\pm 0.16}$ | 70.42 |
| RFD | $\mathbf{92.33}_{\pm 0.79}$ | 99.98 | $\mathbf{92.16}_{\pm 0.54}$ | 99.98 | $\underline{94.87}_{\pm 0.16}$ | 70.42 |
| FD+diag | $\underline{92.30}_{\pm 0.75}$ | 99.98 | $91.92_{\pm 0.58}$ | 99.98 | $\mathbf{94.88}_{\pm 0.17}$ | 70.46 |
| DPSR (Ours) | $92.29_{\pm 0.77}$ | 99.98 | $91.64_{\pm 0.48}$ | 99.98 | $\mathbf{94.88}_{\pm 0.17}$ | 70.46 |
| *Exact wide* | $92.34_{\pm 0.80}$ | 766.8 | $92.01_{\pm 0.48}$ | 770.6 | $94.89_{\pm 0.14}$ | 764.8 |
| Width gain recovered | 88.2% | | -5.5% | | 98.3% | |

Table 14: **Allocation effects vary with budget, dataset and sketch rule.** Paired $\bar{A}$ difference from narrow exact (pp), with 95% paired $t$ intervals, three seeds; Protocol B, ViT-B/16. Means are summarized in Table 1 and plotted in Figure 8.

| Dataset | Method | 20 MiB | 43 MiB | 100 MiB |
|---|---|---|---|---|
| CIFAR-100 | Low rank | +0.379 [+0.150, +0.609] | +0.055 [-0.254, +0.363] | +0.355 [+0.013, +0.697] |
| | FD | +0.534 [+0.285, +0.784] | +0.179 [-0.009, +0.366] | +0.392 [+0.135, +0.649] |
| | RFD | +0.510 [+0.312, +0.707] | +0.578 [+0.414, +0.742] | +0.449 [+0.113, +0.785] |
| | FD+diag | +0.184 [+0.072, +0.296] | +0.547 [+0.279, +0.815] | +0.424 [+0.051, +0.797] |
| | DPSR (Ours) | +0.293 [-0.152, +0.738] | +0.556 [+0.423, +0.689] | +0.405 [-0.043, +0.854] |
| CUB-200 | Low rank | -0.456 [-1.124, +0.212] | -5.481 [-17.041, +6.079] | +0.344 [-0.516, +1.205] |
| | FD | -0.220 [-0.560, +0.119] | -5.284 [-16.957, +6.388] | +0.270 [-0.573, +1.113] |
| | RFD | -0.164 [-0.502, +0.174] | +0.472 [+0.068, +0.876] | +0.502 [-0.043, +1.046] |
| | FD+diag | -0.157 [-0.574, +0.260] | +0.079 [-0.873, +1.032] | +0.260 [-0.551, +1.071] |
| | DPSR (Ours) | -0.579 [-1.030, -0.129] | -2.933 [-9.650, +3.783] | -0.019 [-1.050, +1.011] |
| VTAB-50 | Low rank | +0.106 [-2.080, +2.292] | +0.401 [-1.421, +2.222] | +0.402 [-0.105, +0.910] |
| | FD | +0.447 [-0.905, +1.799] | +0.182 [-1.993, +2.358] | +0.402 [-0.105, +0.910] |
| | RFD | +1.768 [+0.151, +3.386] | +0.737 [+0.261, +1.214] | +0.402 [-0.105, +0.910] |
| | FD+diag | +1.199 [-0.405, +2.802] | +0.498 [-0.403, +1.398] | +0.407 [-0.098, +0.912] |
| | DPSR (Ours) | +0.691 [+0.366, +1.016] | +0.548 [-0.886, +1.983] | +0.407 [-0.098, +0.912] |

table and the DPSR row of Tables 5 and 10 are separate executions of the same configuration and seeds; their per-seed $\bar{A}$ agree except on CUB-200 seed 1995 (91.298 vs. 91.302), which moves the rounded CUB-200 mean±SD from $91.47_{\pm 0.56}$ to $91.46_{\pm 0.57}$. The rank sweep of Table 15 is a third execution.

Table 15: **Rank sensitivity.** Protocol A (task-local GCV), ViT-B/16, five seeds; mean±SD. State is continuation MiB; on VTAB-50 the rank saturates at 70.457 MiB for $r \geq 2048$.

| Dataset | Setting | Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | BWT↑ | MiB↓ |
|---|---|---|---|---|---|---|
| CIFAR-100 | 128 | DPSR | $81.01_{\pm 1.15}$ | $77.75_{\pm 0.39}$ | $-6.10_{\pm 0.84}$ | 8.736 |
| CIFAR-100 | 256 | DPSR | $87.35_{\pm 1.02}$ | $80.42_{\pm 0.13}$ | $-8.73_{\pm 0.22}$ | 13.618 |
| CIFAR-100 | 512 | DPSR | $91.96_{\pm 0.93}$ | $86.99_{\pm 0.24}$ | $-5.58_{\pm 0.48}$ | 23.384 |
| CIFAR-100 | 1024 | DPSR | $92.78_{\pm 0.88}$ | $88.74_{\pm 0.09}$ | $-4.46_{\pm 0.28}$ | 42.915 |
| CIFAR-100 | 2048 | DPSR | $92.78_{\pm 0.88}$ | $88.78_{\pm 0.07}$ | $-4.36_{\pm 0.27}$ | 81.978 |
| CIFAR-100 | 4096 | DPSR | $92.75_{\pm 0.87}$ | $88.79_{\pm 0.05}$ | $-4.36_{\pm 0.22}$ | 160.103 |
| CUB-200 | 128 | DPSR | $90.82_{\pm 0.60}$ | $85.56_{\pm 0.15}$ | $-6.08_{\pm 0.30}$ | 12.550 |
| CUB-200 | 256 | DPSR | $91.43_{\pm 0.58}$ | $86.84_{\pm 0.27}$ | $-5.16_{\pm 0.39}$ | 17.433 |
| CUB-200 | 512 | DPSR | $91.62_{\pm 0.48}$ | $87.46_{\pm 0.17}$ | $-4.61_{\pm 0.39}$ | 27.199 |
| CUB-200 | 1024 | DPSR | $91.46_{\pm 0.56}$ | $87.23_{\pm 0.29}$ | $-4.50_{\pm 0.38}$ | 46.730 |
| CUB-200 | 2048 | DPSR | $91.36_{\pm 0.59}$ | $87.04_{\pm 0.28}$ | $-4.61_{\pm 0.18}$ | 85.793 |
| CUB-200 | 4096 | DPSR | $91.37_{\pm 0.60}$ | $87.01_{\pm 0.22}$ | $-4.54_{\pm 0.33}$ | 163.918 |
| VTAB-50 | 128 | DPSR | $94.79_{\pm 0.51}$ | $91.80_{\pm 0.40}$ | $-4.20_{\pm 0.44}$ | 6.828 |
| VTAB-50 | 256 | DPSR | $95.11_{\pm 0.44}$ | $92.71_{\pm 0.25}$ | $-3.38_{\pm 0.51}$ | 11.711 |
| VTAB-50 | 512 | DPSR | $95.08_{\pm 0.45}$ | $92.58_{\pm 0.30}$ | $-3.46_{\pm 0.55}$ | 21.477 |
| VTAB-50 | 1024 | DPSR | $95.05_{\pm 0.47}$ | $92.49_{\pm 0.33}$ | $-3.49_{\pm 0.64}$ | 41.008 |
| VTAB-50 | 2048 | DPSR | $95.05_{\pm 0.46}$ | $92.49_{\pm 0.32}$ | $-3.50_{\pm 0.59}$ | 70.457 |
| VTAB-50 | 4096 | DPSR | $95.05_{\pm 0.46}$ | $92.49_{\pm 0.32}$ | $-3.50_{\pm 0.59}$ | 70.457 |

Table 16: **Intermediate-width allocation.** Protocol C, ViT-B/16, three seeds; mean±SD. Setting: budget MiB / width $m$ / rank (exact: rank $=$ $m$); "large" is full width $m = 10{,}000$; Figure 2 plots paired differences from narrow exact (data in `figures/src/width_split/width_split_data.json`). State is continuation MiB; VTAB-50 rows at 35.229, 49.320 and 70.457 MiB are below budget because the rank is capped by 1,796 training examples.

| Dataset | Setting | Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | BWT↑ | MiB↓ |
|---|---|---|---|---|---|---|
| CIFAR-100 | 20 / 2240 / 2240 | exact | $91.12_{\pm 0.64}$ | $86.93_{\pm 0.27}$ | $-5.35_{\pm 0.44}$ | 19.995 |
| CUB-200 | 20 / 2191 / 2191 | exact | $89.36_{\pm 1.11}$ | $84.41_{\pm 0.38}$ | $-5.25_{\pm 0.31}$ | 19.984 |
| VTAB-50 | 20 / 2264 / 2264 | exact | $92.97_{\pm 0.43}$ | $89.80_{\pm 0.30}$ | $-4.84_{\pm 0.64}$ | 19.985 |
| CIFAR-100 | 20 / 5000 / 947 | mid | $91.93_{\pm 0.73}$ | $88.14_{\pm 0.18}$ | $-4.87_{\pm 0.37}$ | 19.989 |
| CUB-200 | 20 / 5000 / 847 | mid | $90.71_{\pm 0.81}$ | $86.00_{\pm 0.19}$ | $-4.83_{\pm 0.26}$ | 19.989 |
| VTAB-50 | 20 / 5000 / 997 | mid | $94.15_{\pm 0.27}$ | $91.48_{\pm 0.24}$ | $-4.17_{\pm 0.55}$ | 19.989 |
| CIFAR-100 | 20 / 10000 / 423 | large | $90.44_{\pm 0.64}$ | $85.21_{\pm 0.11}$ | $-6.50_{\pm 0.26}$ | 19.989 |
| CUB-200 | 20 / 10000 / 323 | large | $91.62_{\pm 0.52}$ | $87.15_{\pm 0.24}$ | $-4.75_{\pm 0.36}$ | 19.989 |
| VTAB-50 | 20 / 10000 / 473 | large | $94.76_{\pm 0.06}$ | $92.40_{\pm 0.18}$ | $-3.64_{\pm 0.47}$ | 19.989 |
| CIFAR-100 | 43 / 3307 / 3307 | exact | $91.49_{\pm 0.65}$ | $87.53_{\pm 0.21}$ | $-5.10_{\pm 0.33}$ | 42.980 |
| CUB-200 | 43 / 3258 / 3258 | exact | $89.98_{\pm 0.97}$ | $85.10_{\pm 0.35}$ | $-5.26_{\pm 0.46}$ | 42.977 |
| VTAB-50 | 43 / 3332 / 3332 | exact | $93.69_{\pm 0.24}$ | $90.87_{\pm 0.35}$ | $-4.43_{\pm 0.61}$ | 42.987 |
| CIFAR-100 | 43 / 5000 / 2153 | mid | $91.89_{\pm 0.64}$ | $88.09_{\pm 0.20}$ | $-4.85_{\pm 0.37}$ | 42.992 |
| CUB-200 | 43 / 5000 / 2053 | mid | $90.65_{\pm 0.86}$ | $85.77_{\pm 0.04}$ | $-4.97_{\pm 0.42}$ | 42.992 |
| VTAB-50 | 43 / 5000 / 2203 | mid | $94.14_{\pm 0.27}$ | $91.47_{\pm 0.24}$ | $-4.21_{\pm 0.52}$ | 35.229 |
| CIFAR-100 | 43 / 10000 / 1026 | large | $92.36_{\pm 0.75}$ | $88.76_{\pm 0.11}$ | $-4.57_{\pm 0.31}$ | 42.992 |
| CUB-200 | 43 / 10000 / 926 | large | $91.61_{\pm 0.64}$ | $87.45_{\pm 0.25}$ | $-4.46_{\pm 0.20}$ | 42.992 |
| VTAB-50 | 43 / 10000 / 1076 | large | $94.72_{\pm 0.07}$ | $92.28_{\pm 0.21}$ | $-3.72_{\pm 0.56}$ | 42.992 |
| CIFAR-100 | 100 / 5070 / 5070 | exact | $91.88_{\pm 0.66}$ | $88.09_{\pm 0.23}$ | $-4.85_{\pm 0.37}$ | 99.990 |
| CUB-200 | 100 / 5020 / 5020 | exact | $90.68_{\pm 0.87}$ | $85.89_{\pm 0.17}$ | $-4.87_{\pm 0.28}$ | 99.962 |
| VTAB-50 | 100 / 5095 / 5095 | exact | $94.17_{\pm 0.25}$ | $91.50_{\pm 0.22}$ | $-4.14_{\pm 0.56}$ | 99.998 |
| CIFAR-100 | 100 / 7000 / 3643 | mid | $92.06_{\pm 0.71}$ | $88.47_{\pm 0.12}$ | $-4.66_{\pm 0.33}$ | 99.976 |
| CUB-200 | 100 / 7000 / 3543 | mid | $91.12_{\pm 0.79}$ | $86.52_{\pm 0.25}$ | $-4.76_{\pm 0.42}$ | 99.976 |
| VTAB-50 | 100 / 7000 / 3693 | mid | $94.51_{\pm 0.14}$ | $91.92_{\pm 0.13}$ | $-3.95_{\pm 0.56}$ | 49.320 |
| CIFAR-100 | 100 / 10000 / 2520 | large | $92.34_{\pm 0.76}$ | $88.71_{\pm 0.02}$ | $-4.59_{\pm 0.27}$ | 99.983 |
| CUB-200 | 100 / 10000 / 2420 | large | $91.47_{\pm 0.74}$ | $87.13_{\pm 0.31}$ | $-4.60_{\pm 0.16}$ | 99.983 |
| VTAB-50 | 100 / 10000 / 2570 | large | $94.72_{\pm 0.08}$ | $92.29_{\pm 0.24}$ | $-3.70_{\pm 0.47}$ | 70.457 |

### F.6 Regularization controls and isotropic identity

Tables 18 and 19 report the complete regularization comparison. The nominal grid is $\lambda_i = 10^{2+7i/39}$ for $i = 0, \ldots, 39$. It is identical for all three methods. The rank is fixed in advance, not selected using either test performance or additional parallel validation states. Each independent rank experiment follows the same current-task split and refit rule; sharing the task SVD across experiments only saves evaluation computation. The isotropic implementation retains the diagonal vector to recompute its residual mean, so its reported state includes that vector rather than assuming a scalar-only implementation.

Table 17: **The state controls separate diagonal energy from retained interactions.** Protocol A, ViT-B/16, five seeds, mean±SD; all rules share one penalty grid and selection rule; the diagonal-only allocation is smaller. On VTAB-50, PCR has the lowest $\bar{A}$ of the spectral rules but the highest $A_T$.

| Method | $\bar{A}\uparrow$ | $A_T\uparrow$ | Cont. MiB↓ | Total MiB↓ |
|---|---|---|---|---|
| **CIFAR-100** | | | | |
| Diagonal only | $83.58_{\pm 1.74}$ | $77.83_{\pm 0.30}$ | 3.853 | 42.006 |
| Low rank | $92.17_{\pm 0.85}$ | $87.06_{\pm 0.14}$ | 42.915 | 81.068 |
| FD | $91.77_{\pm 0.95}$ | $85.76_{\pm 0.18}$ | 42.915 | 81.068 |
| RFD | $92.63_{\pm 0.88}$ | $88.47_{\pm 0.15}$ | 42.915 | 81.068 |
| PCR | $92.17_{\pm 0.90}$ | $87.72_{\pm 0.13}$ | 42.915 | 81.068 |
| DPSR (Ours) | $92.78_{\pm 0.88}$ | $88.74_{\pm 0.10}$ | 42.915 | 81.068 |
| **CUB-200** | | | | |
| Diagonal only | $63.50_{\pm 2.96}$ | $52.85_{\pm 0.43}$ | 7.668 | 49.635 |
| Low rank | $91.54_{\pm 0.56}$ | $87.37_{\pm 0.34}$ | 46.730 | 88.698 |
| FD | $91.67_{\pm 0.57}$ | $87.46_{\pm 0.05}$ | 46.730 | 88.698 |
| RFD | $91.28_{\pm 0.59}$ | $86.69_{\pm 0.16}$ | 46.730 | 88.698 |
| PCR | $91.44_{\pm 0.57}$ | $86.45_{\pm 0.25}$ | 46.730 | 88.698 |
| DPSR (Ours) | $91.46_{\pm 0.57}$ | $87.23_{\pm 0.29}$ | 46.730 | 88.698 |
| **VTAB-50** | | | | |
| Diagonal only | $36.66_{\pm 2.95}$ | $31.64_{\pm 2.60}$ | 1.945 | 38.191 |
| Low rank | $95.05_{\pm 0.47}$ | $92.50_{\pm 0.33}$ | 41.008 | 77.253 |
| FD | $95.09_{\pm 0.46}$ | $92.64_{\pm 0.30}$ | 41.008 | 77.253 |
| RFD | $95.07_{\pm 0.47}$ | $92.57_{\pm 0.33}$ | 41.008 | 77.253 |
| PCR | $89.92_{\pm 1.90}$ | $92.84_{\pm 0.32}$ | 41.008 | 77.253 |
| DPSR (Ours) | $95.05_{\pm 0.47}$ | $92.49_{\pm 0.33}$ | 41.008 | 77.253 |

Table 18: **Fine-grid selection does not establish a DPSR-specific gain.** Protocol B, ViT-B/16, with 40 nominal penalties; stage-average accuracy (%), mean±SD, three paired seeds. Bold/underline: largest/second-largest mean. Paired intervals: Table 19; all twelve include zero.

| Rank | Method | CIFAR-100 | CUB-200 | VTAB-50 |
|---|---|---|---|---|
| 256 | Low rank | $\underline{91.38}_{\pm 0.71}$ | $\mathbf{91.08}_{\pm 0.57}$ | $88.59_{\pm 5.49}$ |
| | Isotropic | $\mathbf{91.39}_{\pm 0.81}$ | $\underline{91.07}_{\pm 0.51}$ | $\underline{90.10}_{\pm 2.97}$ |
| | DPSR (Ours) | $91.30_{\pm 0.76}$ | $90.94_{\pm 0.35}$ | $\mathbf{90.62}_{\pm 2.33}$ |
| 1024 | Low rank | $92.17_{\pm 0.77}$ | $91.58_{\pm 0.19}$ | $92.90_{\pm 2.57}$ |
| | Isotropic | $\underline{92.19}_{\pm 0.84}$ | $\mathbf{91.90}_{\pm 0.48}$ | $\underline{93.66}_{\pm 1.33}$ |
| | DPSR (Ours) | $\mathbf{92.24}_{\pm 0.84}$ | $\underline{91.81}_{\pm 0.53}$ | $\mathbf{93.83}_{\pm 1.13}$ |

Table 19: **All twelve fine-grid paired intervals contain zero.** Protocol B with 40 nominal penalties, three paired seeds; DPSR minus comparator in $\bar{A}$ points with 95% paired $t$ intervals. Values are the stored paired statistics behind Table 18. Diagonal completion and penalty selection cannot be separated at this resolution.

| Rank | Comparator | CIFAR-100 | CUB-200 | VTAB-50 |
|---|---|---|---|---|
| 256 | LR | $-0.075$ [$-0.242$, $+0.092$] | $-0.138$ [$-0.819$, $+0.544$] | $+2.033$ [$-6.297$, $+10.364$] |
| | Isotropic | $-0.087$ [$-0.244$, $+0.070$] | $-0.130$ [$-0.563$, $+0.304$] | $+0.525$ [$-1.071$, $+2.121$] |
| 1024 | LR | $+0.070$ [$-0.105$, $+0.244$] | $+0.234$ [$-1.285$, $+1.753$] | $+0.926$ [$-3.150$, $+5.002$] |
| | Isotropic | $+0.042$ [$-0.056$, $+0.139$] | $-0.084$ [$-0.499$, $+0.332$] | $+0.161$ [$-0.478$, $+0.801$] |

The matched-effective diagnostic uses the same refitted $B, S, \delta$ and compares an FP64 Woodbury solve with an independently formed QR-subspace ridge solve. It checks an identity at matched effective penalty and is not a separately tuned baseline. Saved records contain every candidate's validation accuracy, all selected and boundary penalties, and residual means before and after refit. The current-task validation checks do not guarantee that the finite range contains a globally optimal penalty at every future stage.

### F.7 CUB REFIT MECHANISM

**Cache identity and arithmetic.** Every paired comparison reads the same hashed feature/label cache and class-order file. The local checkpoint loads strictly into the recorded ViT-B/16 archi-

Table 20: **The first-task collapse persists after FP64 state reconstruction.** Protocol B, CUB-200, 43 MiB, first task, one row per seed (three seeds), fixed validation-selected penalty; accuracy in %. Projected-$S$ is a diagnostic counterfactual, not a tuned competitor.

| Seed | $\lambda$ | State32 | State64 | Exact wide | Projected-$S$ (64) |
|---|---|---|---|---|---|
| 1993 | $1e+07$ | 92.27 | 92.27 | 92.27 | 92.27 |
| 1994 | $1e+02$ | 55.02 | 43.23 | 99.13 | 99.13 |
| 1995 | $1e+02$ | 42.15 | 28.93 | 97.11 | 97.11 |

tecture. Reconstruction checks compare the first and last test feature of every task and seed with extraction from that checkpoint; test label ordering and training label counts are checked separately. This is sampled checkpoint reconstruction, not re-extraction of every cached training feature. The original caches do not retain per-example provenance metadata. Local CUB and VTAB train/test folders have no identical image-byte hashes in common.

The unbudgeted exact-wide reference accumulates $G$ in FP64 and retains $S$ in FP32, with the same FP32 expanded inputs, split, nominal grid and refit procedure as the equal-byte group. Its larger Gram payload is included in Table 12. An earlier FP32 full-Gram reference produced non-finite solutions at the smallest validation penalty on CIFAR, despite the backend not raising an error. That complete reference group was excluded and rerun with FP64 Gram accumulation and explicit finite checks. This precision change applies to the unbudgeted reference; the matched-budget competitors retain their declared FP32 states. It therefore does not establish a precision-controlled accuracy decomposition between those competitors and the reference.

In CUB/43 the factor capacity is 926. The three first-task temporary sample counts are 868, 849 and 843, whereas refit uses 966, 943 and 935 examples. The selected nominal penalties are $10^7$, $10^2$ and $10^2$. Thus the latter two seeds select a weak penalty before any ideal-arithmetic truncation occurs in that task; refit then creates discarded directions while keeping that penalty fixed. This diagnostic stage was chosen because it is the observed collapse, not as an additional favorable benchmark. The per-seed counts and penalties above are taken from the Protocol B run records of each seed. At the same selected penalty $\lambda = 10^2$, the first-task accuracies of the other wide rules in seeds 1994/1995 are 23.14/17.77% for LR, 20.09/14.05% for FD, 98.69/96.69% for RFD and 98.25/96.69% for FD+diag, against 98.25/96.69% for narrow exact (which selects $10^2$/$10^5$); in seed 1993 all wide rules select $10^7$ and reach 92.27%. We have not tested why the RFD and FD+diag corrections prevent the collapse while DPSR's diagonal completion does not.

For a fixed refit state, decompose $S = S_{\|} + S_{\perp}$ with respect to the row span of $B$. A QR-orthonormalized basis defines this projection even when the stored singular vectors have lost orthogonality. With the completed system matrix $\widehat{H} = \widehat{G} + \lambda I$ (so that $C = \widehat{H}^{-1}S$), linearity gives $C = \widehat{H}^{-1}S_{\|} + \widehat{H}^{-1}S_{\perp}$. A small norm fraction in $S_{\perp}$ need not imply a small contribution to $C$: LR scales directions orthogonal to $B$ by $1/\lambda$ exactly (Lemma 2), and for DPSR the curvature along such directions is only $\lambda + \sum_i \delta_i v_i^2$. Replacing full-width ridge by PCR would therefore change the problem.

The follow-up keeps the originally selected penalty fixed, constructs both precisions from identical expanded inputs, and uses an independent FP64 full-task spectral solution as the exact reference. In the two collapsed seeds, the FP64 $S_{\perp}$ norm fractions are below $0.00034$, yet the norm ratios $\|\widehat{H}^{-1}S_{\perp}\|_F/\|C\|_F$ exceed $0.998$. Because $\widehat{H}^{-1}$ does not preserve the split for DPSR (Appendix D.5), the two parts are not orthogonal and these ratios are not fractions of classifier energy. Projecting $S$ into the retained span removes the observed first-task collapse in both precisions. We do not choose that projection, a new penalty, or a new rank using test accuracy; it remains a diagnostic counterfactual. Every stored system at every stage was also checked against an independent dense FP64 solve, but this targeted counterfactual does not establish the same mechanism at every stage or dataset.

The difference between State32 and State64 in Table 20 shows that constructing the sufficient statistics and factor is a material source of variation. Agreement of FP64 Woodbury and dense Cholesky on the same stored state cannot detect it. Conversely, the persistence of the collapse after FP64 state reconstruction rules out FP32 state error as its sole cause. The reported allocation results retain

Table 21: **Training-prototype correlations along the pipeline (descriptive).** Mean absolute off-diagonal Pearson correlation, fine-grid control states (Protocol B), three seeds (mean±SD). Class pairs are not independent replicates. Readout-output prototypes of a readout fit to one-hot targets are expected to be nearly decorrelated; as an arithmetic reference (not a measurement), centered one-hot vectors over $K$ classes have $|\mathrm{corr}| = 1/(K-1)$, i.e. about 0.010, 0.005 and 0.020 for CIFAR-100, CUB-200 and VTAB-50.

| Dataset | Rank | Backbone | Expansion | LR readout | DPSR readout |
|---|---|---|---|---|---|
| CIFAR-100 | 256 | $0.435_{\pm 0.000}$ | $0.398_{\pm 0.002}$ | $0.028_{\pm 0.002}$ | $0.032_{\pm 0.004}$ |
| CUB-200 | 256 | $0.353_{\pm 0.000}$ | $0.310_{\pm 0.003}$ | $0.040_{\pm 0.012}$ | $0.039_{\pm 0.015}$ |
| VTAB-50 | 256 | $0.166_{\pm 0.000}$ | $0.150_{\pm 0.001}$ | $0.036_{\pm 0.020}$ | $0.038_{\pm 0.019}$ |
| CIFAR-100 | 1024 | $0.435_{\pm 0.000}$ | $0.398_{\pm 0.002}$ | $0.021_{\pm 0.008}$ | $0.021_{\pm 0.008}$ |
| CUB-200 | 1024 | $0.353_{\pm 0.000}$ | $0.310_{\pm 0.003}$ | $0.021_{\pm 0.009}$ | $0.016_{\pm 0.001}$ |
| VTAB-50 | 1024 | $0.166_{\pm 0.000}$ | $0.150_{\pm 0.001}$ | $0.021_{\pm 0.001}$ | $0.021_{\pm 0.001}$ |

their original FP32 storage and FP64 solve protocol; they are not silently replaced by the follow-up precision.

### F.8 Coordinate energy, prototypes and spectrum

We evaluate the final saved states of the fine-grid control (Table 18) for both ranks and every dataset/seed, after checking their SHA256 against the original run records. No classifier fitting, feature learning or hyperparameter selection is repeated. The readout is reconstructed in FP64 from the stored FP32 sufficient statistics at its recorded penalty. Consequently these plots describe the saved-state geometry; they do not replace the original classification measurements.

**Prototype correlation (descriptive).** Let $n_c$ be the count of training examples of class $c$. The expanded prototype matrix is $P = S\,\mathrm{Diag}(1/n_c)$ and the readout-output prototype matrix is $C^\top P$. Backbone prototypes are computed from the corresponding hashed training caches. Pearson correlation centers each prototype over its coordinates and normalizes its Euclidean norm; zero-variance prototypes are excluded and counted. Figure 5 displays the class-pair matrices for the first seed; Table 21 summarizes absolute correlations over all unordered class pairs, with mean and sample SD across seeds. Class pairs are not independent trials. Columns of $C$ are classifier vectors; rows of $C$ are feature coordinates, so neither is silently substituted for the class-conditional output prototypes. Changing output dimensionality also changes the interpretation of these correlations. Low readout-output correlations are expected for any readout fit to one-hot targets: its output prototypes $C^\top p_c$ approximate the one-hot codes, which are nearly decorrelated. As an arithmetic reference, not a measurement, two distinct one-hot vectors over $K$ classes, centered over their coordinates, have Pearson correlation exactly $-1/(K-1)$, i.e. $|\mathrm{corr}| \approx 0.010$, $0.005$ and $0.020$ for $K = 100$, $200$ and $50$. The readout values in Table 21 should be read against this reference; they are not evidence that the representation itself becomes decorrelated.

**Retained spectrum and trace.** We compute the eigenvalues of $BB^\top$ in FP64; these are the nonzero eigenvalues of $B^\top B$. In Figure 6, the plotted cutoff is the capacity of the stored factor, not evidence that it is the global top-$r$ eigenspace of the cumulative Gram. The entropy effective rank is $\exp(-\sum_i p_i \log p_i)$ with $p_i = \lambda_i / \sum_j \lambda_j$. The raw outputs also report $\mathrm{tr}(B^\top B)/\sum_i d_i$. Coordinate medians and this trace fraction weight coordinates differently. Diagonal residual energy is defined in the coordinate basis and is not an orthogonal complement to the factor's row span after completion.

### F.9 Supplementary allocation and representation views

The following views are supplementary plots of recorded Protocol B summaries and saved-state prototype matrices; they involve no new training run.

**Width and state cost.** At fixed rank, DPSR grows linearly in the expansion width while exact ridge grows quadratically. For CIFAR-100 at 43 MiB, the narrow exact width is $m = 3{,}307$ and

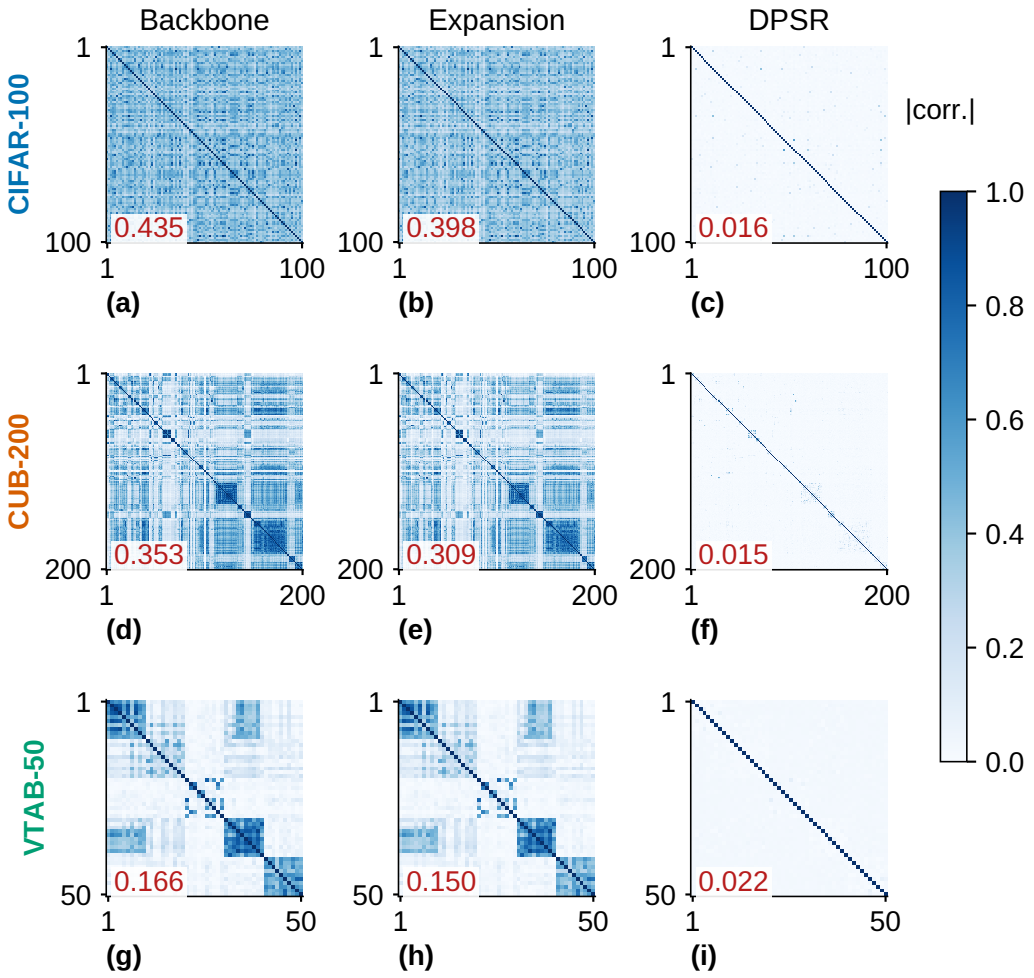



Figure 5: **Training-prototype correlations before and after the readout (descriptive).** Fine-grid control states (Protocol B), absolute class-pair Pearson matrices at rank 1024, final stage, first seed (1993) of three. Low readout-output correlations are expected for a readout fit to one-hot targets (Appendix F.8). All nine dataset/seed reconstructions match their saved correlation summaries; Table 21 shows three-seed dispersion. Matrices retain their within-seed class order and are not averaged across shuffled class orders.

the DPSR configuration uses $m = 10{,}000$ with $r = 1{,}026$. The accompanying retention summaries show the coordinate energy preserved by ranks 256 and 1024.

**Matched-budget allocation.** Figure 8 is the graphical form of Table 1. RFD and FD+diagonal match or exceed DPSR in several cells, and on CUB-200 at 43 MiB the intervals of LR, FD and DPSR are wide and mostly negative.

**Prototype correlations along the pipeline (descriptive).** The cached prototype matrices give a representation-level view; the one-hot caveat of Appendix F.8 applies to the readout values. On CIFAR-100, the mean absolute off-diagonal correlation falls from 0.435 in the backbone to 0.398 after expansion and 0.016 after the DPSR readout for seed 1993. Panel (b) summarizes the same pipeline over the three datasets at rank 1024, with the low-rank readout shown as a hollow diamond.

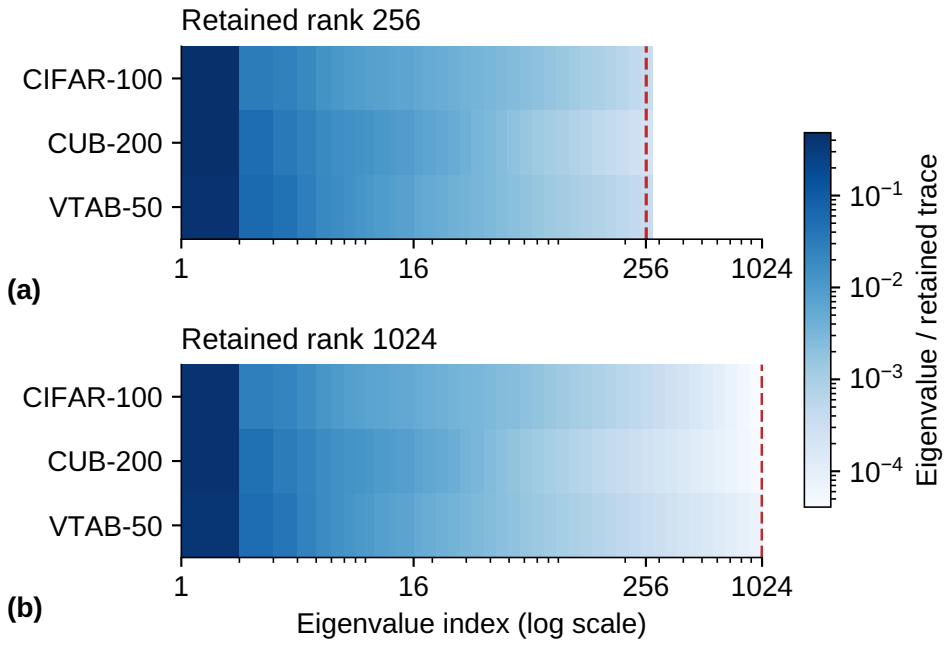



Figure 6: Spectra of the retained factors (fine-grid control, Protocol B, three seeds), normalized by their own trace. Heat-map cells show the mean normalized eigenvalues over three seeds, in logarithmic index bins; white denotes indices beyond retained capacity. Panels show capacities 256 and 1024. The shared logarithmic colour scale retains small eigenvalues. This is not the spectrum or effective rank of full $G$, whose eigenvalues cannot generally be reconstructed from $B, d, S$.

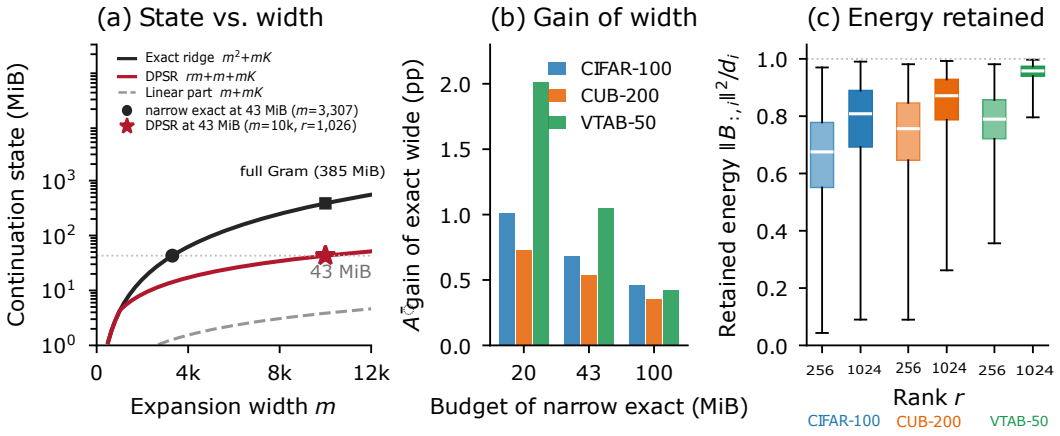



Figure 7: **Width, state cost and coordinate-energy retention.** (a) FP32 continuation state for exact ridge and DPSR as a function of expansion width (arithmetic); (b) unbudgeted exact-wide gain over narrow exact ridge at the three budgets (Protocol B, three seeds, difference of means); (c) final-stage coordinate-energy retention quantiles (fine-grid control, Protocol B, three seeds).

**Synthetic regularization control.** The synthetic comparison separates a common penalty from validation-tuned penalties. Large common-penalty gains over low-rank ridge largely disappear after method-specific tuning, while the isotropic comparison remains close to zero across most settings.

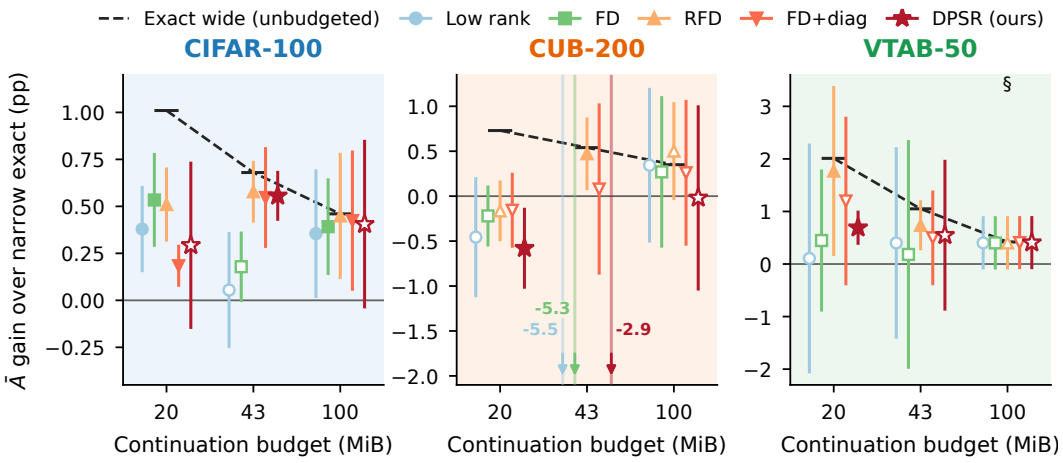



Figure 8: **Matched-budget allocation across datasets (same data as Table 1).** Protocol B, ViT-B/16, three seeds. Each marker is the paired $\bar{A}$ gain over narrow exact ridge at 20, 43 or 100 MiB; bars are 95% paired intervals and the dashed line is the unbudgeted exact-wide gain. Off-scale CUB-200 values are printed at the lower boundary.

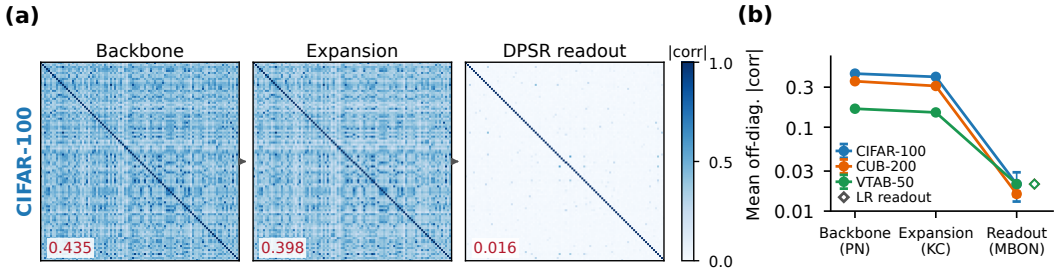



Figure 9: **Prototype correlations along the readout pipeline (descriptive).** Fine-grid control states, Protocol B. (a) Absolute Pearson correlations for CIFAR-100 class prototypes at the final stage, seed 1993 of three; red labels give the mean off-diagonal value. (b) Mean off-diagonal absolute correlation across datasets at rank 1024; error bars are the three-seed SD and hollow diamonds are low-rank ridge. The axis labels "(PN)", "(KC)" and "(MBON)" inside the image are a loose analogy only: the backbone, expansion and readout stages do not correspond to these cell types in any mechanistic sense. Low readout-output correlations are expected for a readout fit to one-hot targets.

**Coordinate-energy retention.** Table 22 reports per-seed coordinate-retention quantiles. For coordinate $i$ with $d_i > 0$, retention is $\|B_{:,i}\|_2^2/d_i$. Zero-energy coordinates are excluded from quantiles and counted separately. Seed-level trajectories retain the minimum, quartiles, median, and maximum at every stage; coordinates are not independent data replicates. The $\kappa$ bounds evaluate the PSD formulas on stored statistics and clipped diagonal residuals. They do not provide exact offline $\kappa$ or prove PSD of the stored full residual. The agreement certificate of Proposition 5 has zero observed coverage at all recorded stages; no stage is selected to make it look informative.

### F.10 Margin exchange, residual action and agreement

**Margin data and unresolved measurements.** The margin summaries in Table 23 use the realized logit perturbation and a top-one margin, not the retained-state trace bound; we call their violation column the realized-perturbation violation rate. If the margin is the exact-wide margin, this rate is $\Pr(\widehat{\Xi}) \leq \Pr(\Xi)$ (Appendix D.8). The summaries can therefore show low violation rates while the trace-based certificate has zero coverage. We verify arithmetic across the nine stored seed summaries and the stored zero disagreements outside the sufficient-condition violation event. These files

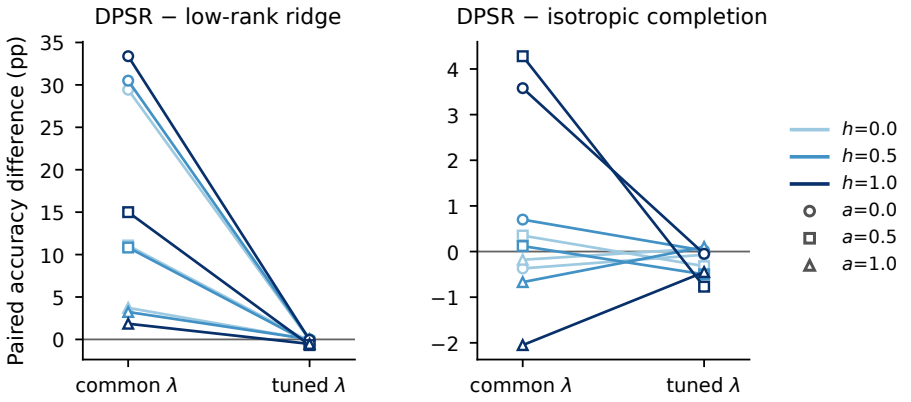



Figure 10: **Synthetic regularization control.** Synthetic generator (Appendix F.12), three seeds. Lines connect the mean DPSR-minus-comparator accuracy at a common penalty and after validation tuning for each setting $(h, a)$ of the generator. Tuning substantially reduces the apparent completion gain over low-rank ridge.

Table 22: **Low-rank truncation removes coordinate energy across seeds.** Fine-grid control states (Protocol B), three seeds, one row each; final-stage retention quantiles; the last column counts the maximum zero-energy coordinates over stages. Zero-energy coordinates are excluded from quantiles.

| Dataset | Rank | Seed | Min | Q1 | Median | Q3 | Max | Zero max |
|---|---|---|---|---|---|---|---|---|
| CIFAR-100 | 256 | 1993 | 0.059 | 0.552 | 0.676 | 0.779 | 0.967 | 0 |
| CIFAR-100 | 256 | 1994 | 0.023 | 0.551 | 0.675 | 0.779 | 0.973 | 1 |
| CIFAR-100 | 256 | 1995 | 0.048 | 0.550 | 0.675 | 0.777 | 0.970 | 0 |
| CIFAR-100 | 1024 | 1993 | 0.111 | 0.693 | 0.808 | 0.889 | 0.990 | 0 |
| CIFAR-100 | 1024 | 1994 | 0.052 | 0.693 | 0.808 | 0.890 | 0.991 | 1 |
| CIFAR-100 | 1024 | 1995 | 0.107 | 0.691 | 0.809 | 0.890 | 0.990 | 0 |
| CUB-200 | 256 | 1993 | 0.086 | 0.648 | 0.756 | 0.846 | 0.981 | 2 |
| CUB-200 | 256 | 1994 | 0.101 | 0.645 | 0.756 | 0.848 | 0.984 | 0 |
| CUB-200 | 256 | 1995 | 0.082 | 0.645 | 0.757 | 0.846 | 0.979 | 1 |
| CUB-200 | 1024 | 1993 | 0.235 | 0.789 | 0.871 | 0.928 | 0.992 | 2 |
| CUB-200 | 1024 | 1994 | 0.259 | 0.786 | 0.871 | 0.929 | 0.993 | 0 |
| CUB-200 | 1024 | 1995 | 0.292 | 0.787 | 0.872 | 0.928 | 0.993 | 1 |
| VTAB-50 | 256 | 1993 | 0.307 | 0.720 | 0.790 | 0.858 | 0.980 | 0 |
| VTAB-50 | 256 | 1994 | 0.376 | 0.721 | 0.790 | 0.857 | 0.980 | 0 |
| VTAB-50 | 256 | 1995 | 0.386 | 0.721 | 0.789 | 0.856 | 0.984 | 0 |
| VTAB-50 | 1024 | 1993 | 0.792 | 0.939 | 0.958 | 0.974 | 0.997 | 0 |
| VTAB-50 | 1024 | 1994 | 0.801 | 0.940 | 0.958 | 0.974 | 0.997 | 0 |
| VTAB-50 | 1024 | 1995 | 0.795 | 0.940 | 0.958 | 0.974 | 0.997 | 0 |

Table 23: **Small realized score changes preserve most exact-wide decisions.** Protocol C common-penalty diagnostics, ViT-B/16, three seeds, final stage. Margin and allowance are in score units; violation (the realized-perturbation violation rate, Appendix D.8) and disagreement are fractions.

| Dataset | Residual action | Median margin | Mean allowance | Violation | Disagreement |
|---|---|---|---|---|---|
| CIFAR-100 | $0.012_{\pm 0.000}$ | $0.604_{\pm 0.002}$ | $0.133_{\pm 0.002}$ | $0.143_{\pm 0.002}$ | $0.022_{\pm 0.001}$ |
| CUB-200 | $0.013_{\pm 0.000}$ | $0.308_{\pm 0.001}$ | $0.030_{\pm 0.001}$ | $0.062_{\pm 0.003}$ | $0.009_{\pm 0.001}$ |
| VTAB-50 | $0.003_{\pm 0.000}$ | $0.501_{\pm 0.002}$ | $0.008_{\pm 0.000}$ | $0.010_{\pm 0.002}$ | $0.001_{\pm 0.000}$ |

do not contain per-example logits or narrow-exact margins; they cannot reconstruct three-readout histograms, a joint signed-margin exchange distribution, or a width-margin gain. They also use a different selection path (common-penalty GCV) from Protocol B, so their CUB-200 values should not be combined with the Protocol B refit collapse.

A paired width/compression exchange measurement requires all three score vectors on the same examples and stage, with exact-wide evaluated at DPSR's penalty for the perturbation interpretation. Separately selected exact-wide penalties add another source of change. Such score vectors were not saved by the completed exact-wide run. Similarly, cumulative retained factors do not recover each

Table 24: **Dense-ReLU controls do not establish a DPSR-specific gain.** DPSR minus comparator in accuracy points, 95% paired intervals, three seeds. Gaussian projection with ReLU at width 10,000, Protocol B-style validation and refit, ViT-B/16 features.

| Dataset | Comparator | Paired difference [95% CI] |
|---|---|---|
| CIFAR-100 | LR | 0.025 [-0.179, 0.228] |
| CIFAR-100 | Isotropic | 0.048 [-0.179, 0.274] |
| CIFAR-100 | Narrow exact | 0.343 [-0.067, 0.753] |
| CUB-200 | LR | -0.252 [-0.897, 0.393] |
| CUB-200 | Isotropic | 0.026 [-0.068, 0.119] |
| CUB-200 | Narrow exact | 0.012 [-0.429, 0.454] |

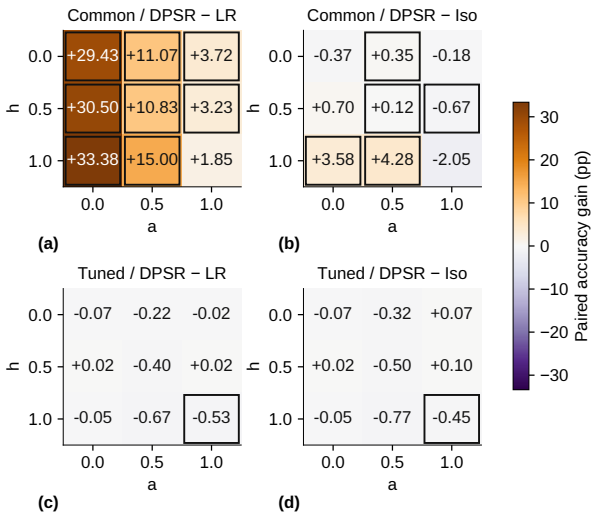



Figure 11: **Completion gains depend on the synthetic setting and tuning.** Cells show mean DPSR-minus-comparator accuracy in percentage points over three seeds. Outlines mark 95% paired intervals excluding zero; the shared colour scale includes negative outcomes.

task's uncompressed leading subspace; cross-task principal angles require additional task factors. We report these missing measurements explicitly rather than approximating them with cumulative-state similarity or invented histogram bins.

### F.11 Dense-ReLU map

The dense-ReLU follow-up shares Protocol B-style current-task validation and refit across its methods and retains three seeds. Table 24 reports paired differences; the complete control does not establish a DPSR-specific advantage on this map. Its feature scale differs from Top-$k$, so its nominal penalty range is interpreted separately.

### F.12 Synthetic study

The complete synthetic grid is shown in Figure 11. The generator uses dimension 128, rank 24, five classes and 3,000/1,000/2,000 training/validation/test examples. The common penalty scale is 0.01; tuned models search $10^{-4}, \ldots, 10^{2}$ using validation, with smaller-scale tie-breaking. Every $(h, a)$ cell and all three seeds are retained. These controls probe sensitivity rather than identify a causal explanation of the real datasets.

### F.13 Efficiency and structured solve

The same-system solve benchmark isolates the structured linear algebra from feature extraction and task SVD. Table 26 reports per-seed median latency and incremental workspace on RTX 3090.

Table 25: **Small continuation state does not imply lower fitting cost.** CIFAR-100: Protocol A (five seeds) and Protocol B at 43 MiB (three seeds); timing scopes differ by block (Appendix F.13).

| Method | Cont. | Total | Fit s/task | Solve ms | $\bar{A}\uparrow$ |
|---|---|---|---|---|---|
| Fly-CL | 385.3 | 423.4 | $19.0_{\pm 1.1}$ | – | $92.77_{\pm 0.89}$ |
| DPSR (Ours) | 42.9 | 81.1 | $18.2_{\pm 0.1}$ | $5.0_{\pm 0.1}$ | $92.78_{\pm 0.88}$ |
| *Matched-budget fit, feature extraction excluded* | | | | | |
| Narrow exact | 43.0 | 53.9 | $0.9_{\pm 0.0}$ | – | $91.66_{\pm 0.81}$ |
| RFD | 43.0 | 76.1 | $28.0_{\pm 0.0}$ | – | $92.23_{\pm 0.77}$ |
| DPSR (Ours) | 43.0 | 76.1 | $12.8_{\pm 0.0}$ | – | $92.21_{\pm 0.76}$ |

Table 26: **The structured solve reduces latency and workspace.** Same approximate system, three seeds, RTX 3090, separate from the Protocol A/B timing runs; milliseconds and incremental MiB.

| Dataset | Dense ms | DPSR ms | Speedup | Dense MiB | DPSR MiB |
|---|---|---|---|---|---|
| CIFAR-100 | $73.23_{\pm 0.86}$ | $5.03_{\pm 0.13}$ | $14.56_{\pm 0.53}$ | $1531.89_{\pm 0.00}$ | $104.04_{\pm 0.00}$ |
| CUB-200 | $74.08_{\pm 1.38}$ | $5.69_{\pm 0.10}$ | $13.02_{\pm 0.41}$ | $1535.71_{\pm 0.00}$ | $104.04_{\pm 0.00}$ |
| VTAB-50 | $72.02_{\pm 0.85}$ | $4.81_{\pm 0.02}$ | $14.98_{\pm 0.20}$ | $1529.99_{\pm 0.00}$ | $104.04_{\pm 0.00}$ |

Table 27: **Continuation savings and peak workspace are distinct.** Protocol B at 43 MiB, isolated profiling (RTX A5000), three seeds; continuation, total non-backbone state and peak GPU allocation in MiB.

| Dataset | Method | Continuation↓ | Total↓ | Peak GPU↓ |
|---|---|---|---|---|
| CIFAR-100 | Narrow exact | 42.98 | 53.93 | 645.2 |
| | DPSR (Ours) | 42.99 | 76.10 | 1226.4 |
| | LR ridge | 42.99 | 76.10 | 1226.3 |
| | FD | 42.99 | 76.10 | 1988.8 |
| | RFD | 42.99 | 76.10 | 1988.8 |
| | FD+diagonal | 42.99 | 76.10 | 1988.7 |
| CUB-200 | Narrow exact | 42.98 | 55.01 | 527.0 |
| | DPSR (Ours) | 42.99 | 79.92 | 525.0 |
| | LR ridge | 42.99 | 79.92 | 525.1 |
| | FD | 42.99 | 79.92 | 525.0 |
| | RFD | 42.99 | 79.92 | 525.0 |
| | FD+diagonal | 42.99 | 79.92 | 524.9 |
| VTAB-50 | Narrow exact | 42.99 | 53.38 | 531.2 |
| | DPSR (Ours) | 42.99 | 74.20 | 366.3 |
| | LR ridge | 42.99 | 74.20 | 366.3 |
| | FD | 42.99 | 74.20 | 366.3 |
| | RFD | 42.99 | 74.20 | 366.3 |
| | FD+diagonal | 42.99 | 74.20 | 366.3 |

These records use an earlier common-penalty reconstruction, so their accuracy fields are not pooled with the final public comparison. Maximum coefficient differences and rare prediction disagreements make this a floating-point benchmark, not an identity test at arbitrary conditioning.

Table 27 reports method-isolated matched-budget peaks on RTX A5000. The host RSS includes the interpreter, loaded libraries and current feature data; it is process context rather than deployment state. All selected penalties match the corresponding accuracy runs. Lower continuation bytes can coexist with larger peak workspace because validation, factorization and refit use temporary arrays.

### F.14 TASK-BY-STAGE ACCURACY MATRICES

Figures 12–13 display mean task accuracies across the five public-comparison seeds. Each lower-triangular row is a training stage; columns identify previously introduced tasks. The shared colour range permits comparison across methods and datasets. These matrices supply $\bar{A}$, $A_T$ and BWT rather than treating tasks as independent statistical replicates.

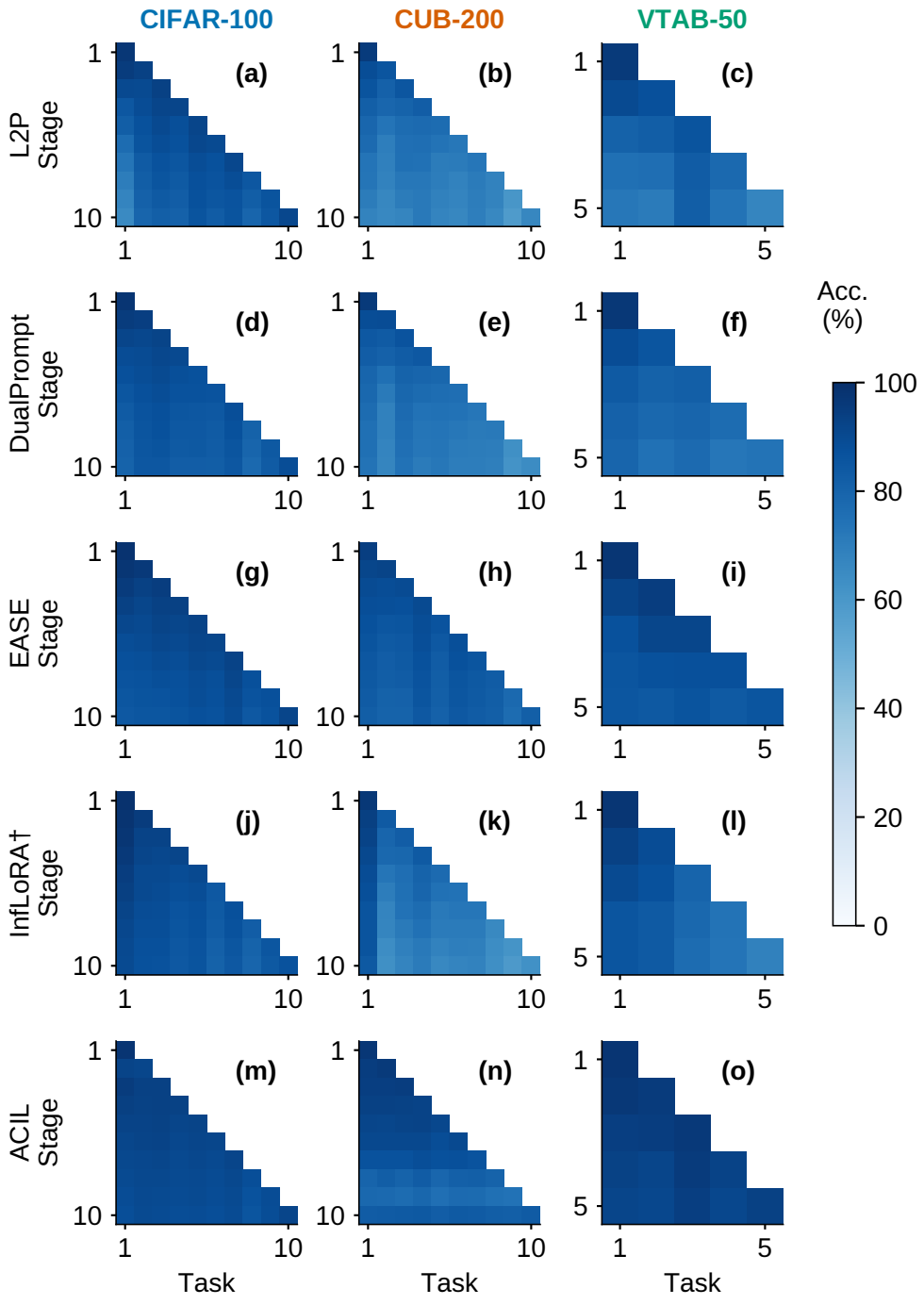



Figure 12: Public-comparison task-by-stage accuracy matrices: first five methods, mean over five seeds. White cells are future tasks.

## G Additional limitations and broader impact

This appendix complements the limitations paragraph of Section 5. The evidence concerns fixed representations, local partitions where stated, and fixed state allocations. The additional encoders and domain shifts broaden that scope without establishing a universally preferable compression rule.

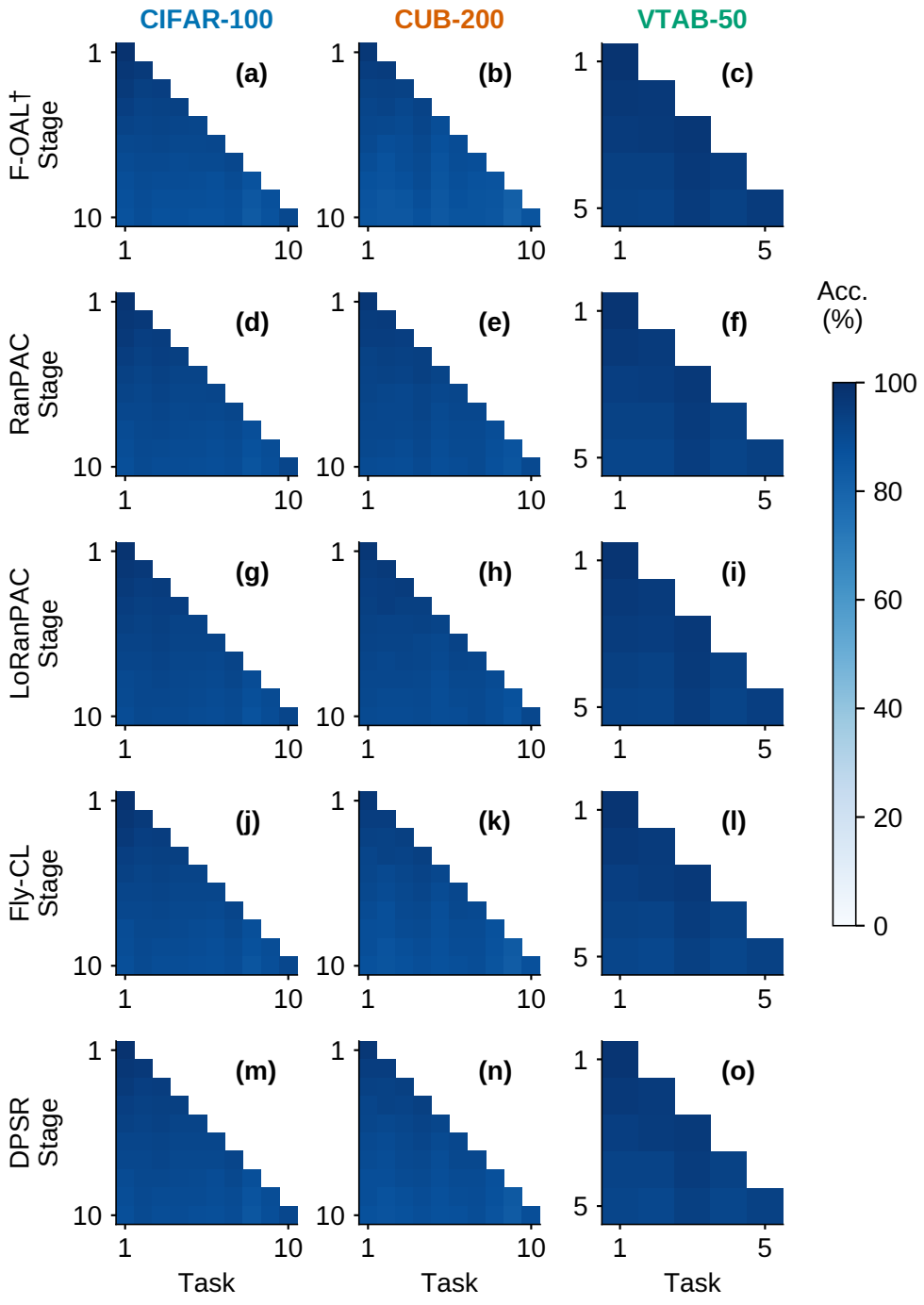



Figure 13: Public-comparison task-by-stage accuracy matrices: remaining five methods, mean over five seeds. All panels share the same accuracy colour range.

The representation measurements are descriptive: prototype correlation, trace retention and classifier agreement each answer a different question. Joint narrow/wide/compressed pointwise scores and full residual spectra were not saved for the main allocation runs, which limits quantitative margin-exchange and large-state tightness analyses.

The method can reduce retained feature statistics while leaving preprocessing, inference and temporary workspace substantial. Small seed counts limit uncertainty resolution; numerical precision and selection/refit behavior can materially affect an operating point. Statistical summaries retained from labeled data may still encode sensitive information, so compact state should not be interpreted as a privacy guarantee. Deployment should evaluate its own data, compute and reliability constraints.