SCORE-PRESERVING CONTINUAL LASSO UNDER TASK SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning (CL) aims to learn from sequential tasks while mitigating catastrophic forgetting when historical data are no longer accessible. We study continual learning for high-dimensional sparse linear regression, where ℓ_1 regularization creates a difficulty absent from continual ridge regression: that recursively reusing an ℓ_1-penalized estimator can propagate earlier shrinkage and accumulate regularization bias across tasks. We propose score-preserving continual Lasso (SP-CL), which separates sparse estimation from state transmission by preserving the historical squared-loss score. Under squared loss, its cumulative estimator exactly recovers a full-history Lasso minimizer. When task coefficients coincide, it inherits the full-history sparse rate. For related task-specific coefficients, SP-CL uses a regularized current-task correction that leaves the transmitted state unchanged and attains the offline source-assisted rate under an oracle related-task condition. We also derive Overall/BWT/FWT decompositions that distinguish historical information loss, task conflict, and current-task specialization, and validate the resulting regimes in simulation.

1 INTRODUCTION

Continual learning (CL) asks a model to learn from a sequence of tasks when data from earlier tasks are no longer fully accessible. A central difficulty is *catastrophic forgetting*: adapting to a new task can degrade performance on tasks learned earlier. At the same time, excessive stability can prevent useful adaptation. A successful continual learner must therefore maintain strong performance over observed tasks, preserve useful earlier knowledge, and exploit that knowledge on new tasks. These goals are commonly reflected by Overall performance, backward transfer (BWT), and forward transfer (FWT), respectively (Parisi et al., 2019; Wang et al., 2024).

Regularization-based methods address this stability–adaptation problem by constraining new solutions toward parameters learned from previous tasks. For two fixed-design regression tasks, Li et al. (2023) derived tight risk bounds for ℓ_2-regularized updates and characterized the forgetting–intransigence trade-off. Zhao et al. (2024) extended this perspective to sequences of tasks through generalized matrix-valued ℓ_2 regularization, obtaining offline-oracle-order estimation while balancing forward and backward transfer. In a complementary realizable setting, Levinstein et al. (2025) showed that fixed or increasing isotropic regularization can achieve near-optimal or optimal task-horizon convergence rates.

More recent work has examined genuinely high-dimensional regimes. Zhao et al. (2025) derived asymptotic prediction risks when dimension grows proportionally with task sample size, characterizing Overall performance, BWT, and FWT under a common coefficient and heterogeneous task covariances. Karpel et al. (2026) further allowed multiple task-specific linear teachers and characterized optimal isotropic ℓ_2 regularization in overparameterized continual regression. Thus, existing theory now covers multi-task, high-dimensional, and heterogeneous-task settings, but its continual regularization mechanism remains quadratic rather than sparse. Table 1 summarizes these representative theoretical settings and capabilities most relevant to our problem.

Empirical work uses sparse computation to reduce continual interference without replay or task information (Bricken et al., 2023). We study its statistical counterpart: high-dimensional predictors with sparse task-relevant coefficients, where ℓ_1 regularization creates a state-transmission problem absent from quadratic methods.

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.

Table 1: Comparison with representative theoretical work on regularization-based continual linear regression. "Task-specific coefficients" allows different population coefficients across tasks; "Current-task correction" denotes an explicit second-stage correction using the current task.

Work	Sparse	Task-specific coefficients	Current-task correction	Overall/BWT/FWT
Li et al. (2023)	×	×	×	×
Zhao et al. (2024)	×	×	×	✓
Zhao et al. (2025)	×	×	×	✓
Karpel et al. (2026)	×	✓	×	×
SP-CL	✓	✓	✓	✓

High-dimensional streaming methods support ℓ_1-regularized estimation from sequential summaries but do not study continual task shift (Luo et al., 2023), while sparse transfer learning permits coefficient adaptation but assumes that source samples remain accessible (Li et al., 2022).

Three issues remain for sparse continual regression.

First, sparsity changes what should be transmitted. Under squared loss, historical data are exactly represented by a quadratic objective through its cumulative curvature and score, so a quadratic state can preserve the information needed from inaccessible data. With ℓ_1 regularization, however, the fitted parameter also contains a penalty subgradient; recursively transmitting that penalized center can therefore propagate shrinkage that does not belong to the historical data objective.

Second, task variation creates a distinct adaptation problem. Existing high-dimensional ℓ_2 theory can accommodate task-specific coefficients by tuning regularization to balance historical retention and current adaptation (Karpel et al., 2026). Our question is different: can related history first reduce estimation variance through cumulative information, and can the resulting pooled bias then be corrected toward the current sparse coefficient without revisiting historical samples?

Third, task heterogeneity complicates the interpretation of forgetting. Old-task degradation may reflect genuine loss of historical information, but it may also arise because the predictor deliberately moves toward a different current-task optimum. Overall performance, BWT, and FWT must therefore be interpreted jointly when comparing cumulative retention with current-task specialization.

In this paper, we aim to answer the following question:

Can sparse continual learning preserve full-history information without accumulating past shrinkage, while adapting efficiently to related tasks?

We propose *score-preserving continual Lasso* (SP-CL), which separates sparse estimation from state transmission: ℓ_1 regularization produces a sparse cumulative estimator, while a corrected quadratic state preserves only the historical squared-loss score—it *regularizes locally but transmits the data score globally*. Under task shift, SP-CL uses historical aggregation to reduce estimation uncertainty and a sparse current-batch correction to adapt toward the current coefficient without modifying the transmitted state. For evaluation, we derive exact identities that decompose Overall, FWT, and BWT into current-task gain, historical cost, task conflict, and estimation error.

Our contributions are threefold.

- **Score-preserving sparse transmission.** We identify how recursively transmitting an ℓ_1-penalized estimate propagates past penalty subgradients, and develop SP-CL to remove this contamination from the transmitted state. Its cumulative estimator is exactly a full-history pooled-Lasso minimizer, yielding the pooled-Lasso sparse rate under a shared coefficient.
- **Efficient adaptation under task shift.** Under task shift, SP-CL uses a current-batch sparse correction to adapt the cumulative estimate without modifying the transmitted state. Under a related-task condition, the resulting specialist attains the offline two-stage rate using only the continual state and current batch.
- **Continual evaluation under heterogeneity.** We separate catastrophic forgetting from task conflict in standard CL evaluation. Exact decompositions of Overall performance, BWT, and FWT distinguish historical information loss, task conflict, and current-task specialization,

and quantify the current-gain versus historical-cost trade-off between cumulative and current-adapted prediction.

2 Problem Setup and Evaluation

Data and continual-learning protocol. We consider a standard continual-learning problem in which tasks $t = 1, \dots, T$ arrive sequentially. At task t, the learner observes $\mathcal{D}_t = \{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ generated from

$$y_{ti} = x_{ti}^\top \beta_t^\star + \varepsilon_{ti}, \qquad x_{ti} \in \mathbb{R}^p, \tag{1}$$

where $\beta_t^\star$ is the true parameter for task t and may vary across tasks. We focus on a high-dimensional sparse regime in which p may exceed the per-task sample size n_t, while the task-relevant coefficients are sparse. Write $X_t = (x_{t1}, \dots, x_{tn_t})^\top$, $Y_t = (y_{t1}, \dots, y_{tn_t})^\top$, and $N_t = \sum_{j=1}^t n_j$. When task t arrives, the learner may use the current batch $\mathcal{D}_t$ and its transmitted state, but cannot revisit raw observations from earlier tasks. Distributional and task-relatedness conditions are introduced later when needed for the statistical guarantees.

Task performance and continual-learning metrics. For regression, we measure task performance by the population mean-squared prediction error. For a coefficient $b \in \mathbb{R}^p$ and a fresh task-j observation (x, y), write $\Sigma_j = \mathbb{E}_j(xx^\mathsf{T})$, $\mathrm{MSE}_j(b) = \mathbb{E}_j[(y - x^\mathsf{T} b)^2]$ and $\mathcal{R}_j(b) = \mathrm{MSE}_j(b) - \mathrm{MSE}_j(\beta_j^\star) = \left\|b - \beta_j^\star\right\|_{\Sigma_j}^2$. Thus, same-task differences in MSE equal differences in excess prediction error. For tasks observed through time t, let $\pi_{j,t} = n_j / N_t$, define $\overline{\mathrm{MSE}}_t(b) = \sum_{j=1}^t \pi_{j,t} \mathrm{MSE}_j(b)$, and let $w_t^\star \in \arg\min_b \overline{\mathrm{MSE}}_t(b)$ denote the best single population coefficient for the tasks observed so far. With equal batch sizes, $\overline{\mathrm{MSE}}_t$ is the usual average task MSE.

For predictors $\{\hat{b}_t\}_{t=1}^T$ produced after successive tasks, we consider the standard Overall, backward-transfer (BWT), and forward-transfer (FWT) metrics. Here $\hat{\beta}_t^{\mathrm{cur}}$ denotes the current-only Lasso benchmark defined in Section 4.2:

$$\mathrm{BWT}_t(\hat{b}) = \frac{1}{t-1} \sum_{j<t} \{\mathrm{MSE}_j(\hat{b}_j) - \mathrm{MSE}_j(\hat{b}_t)\} \qquad \mathrm{Overall}_t(\hat{b}) = \frac{1}{t} \sum_{j=1}^t \mathrm{MSE}_j(\hat{b}_t)$$

$$\mathrm{FWT}_t(\hat{b}) = \frac{1}{t-1} \sum_{k=2}^t \{\mathrm{MSE}_k(\hat{\beta}_k^{\mathrm{cur}}) - \mathrm{MSE}_k(\hat{b}_k)\}$$

Overall evaluates the current predictor on the tasks seen so far; for $t \geq 2$, BWT compares old-task performance with its task-time reference, and FWT compares each predictor with current-only fitting. Positive BWT or FWT indicates improvement relative to its reference; the three metrics need not rank predictors in the same way when task parameters differ.

3 Score-preserving continual Lasso

3.1 From ℓ_2 Regularization to Sparse State Transmission

Regularization-based continual regression commonly updates the current parameter by penalizing its deviation from an estimate carried from earlier tasks. Generalized ℓ_2 regularization uses a matrix-weighted penalty of the schematic form

$$\widehat{w}_t^{\mathrm{GR}} \in \arg\min_w \left\{ \ell_t(w) + \frac{1}{2} \left\| w - \widehat{w}_{t-1}^{\mathrm{GR}} \right\|_{A_t}^2 \right\}, \qquad \|v\|_A^2 = v^\mathsf{T} A v, \quad A \succeq 0,$$

where ℓ_t is the current-task loss (Zhao et al., 2024). For squared loss, this quadratic penalty provides a natural way to carry information from earlier tasks. The following lemma identifies the curvature and score of the full historical least-squares objective.

Lemma 1 (Exact quadratic representation of historical squared loss)**.** *Let $H_t := \sum_{j=1}^t X_j^\mathsf{T} X_j$ and $G_t := \sum_{j=1}^t X_j^\mathsf{T} Y_j$. For any $u_t \in \mathbb{R}^p$ satisfying $H_t u_t = G_t$ (such a vector exists), and for every*

$w \in \mathbb{R}^p$,

$$\frac{1}{2}\sum_{j=1}^{t}\|Y_j - X_j w\|_2^2 = \frac{1}{2}\|w - u_t\|_{H_t}^2 + C_t, \tag{2}$$

where C_t does not depend on w.

The proof is given in Appendix A. Hence (H_t, u_t) represents the historical squared-loss objective up to a constant, and its score is $S_t(w) := G_t - H_t w = H_t(u_t - w)$. We call u_t a *score carrier*. This score contains only historical data information, not first-order contributions from taskwise regularization. A penalized estimate, however, need not satisfy this data-score equation.

For this subsection, write $\lambda_t := \lambda_{w,t}$. The most direct sparse analogue keeps the parameter-centered quadratic structure and adds an ℓ_1 penalty. We call it *direct transmission of the penalized center* (direct-center transmission):

$$\widehat{w}_t^{\mathrm{dir}} \in \arg\min_w \left\{ \frac{1}{2}\|Y_t - X_t w\|_2^2 + \frac{1}{2}\left\|w - \widehat{w}_{t-1}^{\mathrm{dir}}\right\|_{H_{t-1}}^2 + N_t\lambda_t\|w\|_1 \right\}, \tag{3}$$

with $H_0 = 0$. Let $z_t^{\mathrm{dir}} \in \partial\left\|\widehat{w}_t^{\mathrm{dir}}\right\|_1$ be a KKT subgradient.

Proposition 1 (Exact accumulation of penalty subgradients)**.** *Set $G_0 = H_0\widehat{w}_0^{\mathrm{dir}} = 0$. Then, deterministically, for every t,*

$$G_t - H_t\widehat{w}_t^{\mathrm{dir}} = \sum_{j=1}^{t} N_j\lambda_j z_j^{\mathrm{dir}}. \tag{4}$$

The proof is given in Appendix A. The left-hand side is the full-history data score, whereas the right-hand side consists entirely of accumulated ℓ_1-penalty subgradients. Thus, direct-center transmission mixes persistent data information with task-local regularization. The terms need not grow monotonically, but when they align the distortion can be severe, as the following example shows.

Example 1 (Support collapse under recursive penalization)**.** *Consider noiseless tasks with the same sparse coefficient $\beta^\star \neq 0$, equal batch size n_0, and a design that is orthogonal on the active coordinates. In this setting, recursively transmitting the penalized center eventually sets every coordinate in* $\mathrm{supp}(\beta^\star)$ *to zero. Appendix B gives a high-dimensional construction with $n_0 < p$. For every $k \in \mathrm{supp}(\beta^\star)$, the iterates obey*

$$\left|\widehat{w}_{t,k}^{\mathrm{dir}}\right| = \frac{1}{t}\max\left\{(t-1)\left|\widehat{w}_{t-1,k}^{\mathrm{dir}}\right| + |\beta_k^\star| - t\lambda_t,\, 0\right\}. \tag{5}$$

Each batch contributes the fixed signal $|\beta_k^\star|$, whereas inherited shrinkage enters through $t\lambda_t$. Hence, whenever $t\lambda_t$ eventually dominates the signal, every active coordinate reaches zero after finitely many tasks and remains zero thereafter, even though the coefficient is fixed and more data arrive. For example, $\lambda_t \asymp \sqrt{\log p/(tn_0)}$ gives $t\lambda_t \asymp \sqrt{t\log p/n_0} \to \infty$. Offline pooled Lasso instead applies only its current threshold, which decreases with cumulative sample size. The construction and derivation are in Appendix B.

Proposition 1 identifies the score discrepancy created by direct-center transmission; Figure 1 shows its statistical consequence under noisy, correlated, high-dimensional designs ($p > n_t$, Gaussian AR(1) covariates, $\sigma = 1$): substantial attenuation of the true active coefficients. By contrast, SP-CL remains aligned with offline pooled Lasso throughout. Appendix B shows that the score discrepancy is robust across the tested penalty constants, whereas the severity of signal attenuation depends on the penalty level.

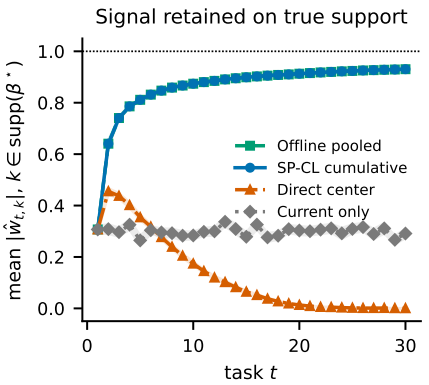

Figure 1: Signal retained on the true active support ($p = 200$, $s = 10$, $n_t = 40$, $T = 30$; 50 streams).

Persistent and task-local first-order information. Proposition 1 suggests separating estimation from transmission: a regularizer may be useful for the current estimate without belonging to the historical objective that the next state is intended to reconstruct. In the present Lasso setting, the ℓ_1 penalty remains essential for sparse estimation, but its KKT contribution is not part of the historical data loss. SP-CL therefore retains the ℓ_1 penalty in estimation but removes its KKT contribution before transmission – regularize locally, transmit the historical data score. Appendix G further confirms this exact-equivalence mechanism on a real chronological climate-prediction stream.

3.2 SCORE-PRESERVING CONTINUAL LASSO

Accordingly, SP-CL removes the current ℓ_1-KKT contribution before transmitting the state. Initialize $H_0 = 0$, $u_0 = 0$, and $N_0 = 0$.

Cumulative sparse fit. Given (H_{t-1}, u_{t-1}) satisfying $H_{t-1}u_{t-1} = G_{t-1}$, set $N_t = N_{t-1} + n_t$ and $H_t = H_{t-1} + X_t^\top X_t$, and solve

$$\widehat{w}_t \in \arg\min_w \left\{ \frac{1}{2} \|Y_t - X_t w\|_2^2 + \frac{1}{2} \|w - u_{t-1}\|_{H_{t-1}}^2 + N_t \lambda_{w,t} \|w\|_1 \right\}. \tag{6}$$

Score-preserving state correction. Choose $z_t \in \partial \|\widehat{w}_t\|_1$. The KKT condition for (6) gives

$$S_t(\widehat{w}_t) = G_t - H_t \widehat{w}_t = N_t \lambda_{w,t} z_t.$$

Unlike Proposition 1, this expression contains no penalty terms accumulated from earlier tasks: its only regularization contribution is the one used for the current sparse fit. Define

$$d_t = (H_t/N_t)^+ z_t, \qquad u_t = \widehat{w}_t + \lambda_{w,t} d_t. \tag{7}$$

The KKT equation also implies $z_t \in \mathrm{Col}(H_t)$, so $H_t d_t = N_t z_t$ and therefore

$$S_t(u_t) = G_t - H_t u_t = 0.$$

Thus, ℓ_1 regularization acts on the estimator $\widehat{w}_t$, while the transmitted quadratic state retains the historical curvature and data score without accumulated penalty contributions. The Moore–Penrose inverse selects the minimum-norm correction d_t when H_t is singular.

The full-history sparse benchmark is the offline pooled Lasso

$$\widehat{w}_t^{\mathrm{off}} \in \arg\min_w \left\{ \frac{1}{2N_t} \sum_{j=1}^{t} \|Y_j - X_j w\|_2^2 + \lambda_{w,t} \|w\|_1 \right\}. \tag{8}$$

It is unavailable to a continual learner because it requires the historical raw batches.

Theorem 1 (Exact pooled-Lasso equivalence)**.** *For every t, the cumulative estimator* $\widehat{w}_t$ *in* (6) *is a minimizer of the offline pooled Lasso* (8)*.*

The proof is given in Appendix A. Hence SP-CL exactly reproduces the pooled-Lasso objective without revisiting historical data; only $H_t u_t = G_t$ matters, so u_t represents the historical linear score.

Current-task correction under task shift. When task coefficients differ, the cumulative estimator targets a pooled coefficient rather than the current-task coefficient. Motivated by the two-stage correction principle in high-dimensional transfer learning (Li et al., 2022; Tian & Feng, 2023), we use the current batch to sparsely correct the pooled estimate without modifying the transmitted state:

$$\widehat{\Delta}_t = \arg\min_\Delta \left\{ \frac{1}{2n_t} \|Y_t - X_t(\widehat{w}_t + \Delta)\|_2^2 + \lambda_{\Delta,t} \|\Delta\|_1 \right\}, \qquad \widehat{\beta}_t = \widehat{w}_t + \widehat{\Delta}_t. \tag{9}$$

The predictor $\widehat{\beta}_t$ is used for current-task prediction but is not written back into the state. Section 4 analyzes when this correction improves on current-only and cumulative estimation.

Algorithm 1 Score-preserving continual Lasso (SP-CL)

Require: Sequential batches $\{(X_t, Y_t)\}_{t=1}^{T}$ and penalties $\{\lambda_{w,t}, \lambda_{\Delta,t}\}_{t=1}^{T}$
1: Initialize $H_0 = 0$, $u_0 = 0$, and $N_0 = 0$
2: **for** $t = 1, \ldots, T$ **do**
3: **Cumulative sparse fit:** Set $N_t = N_{t-1} + n_t$ and solve (6) for $\widehat{w}_t$
4: **Score-preserving state update:** Set $H_t = H_{t-1} + X_t^{\mathsf{T}} X_t$ and construct u_t by (7)
5: **Current-task correction (under task shift):** Solve (9) and output $\widehat{\beta}_t$
6: **end for**

Only (H_t, u_t, N_t) is transmitted to task $t+1$; neither $\widehat{\beta}_t$ nor $\widehat{\Delta}_t$ is written into the transmitted state. In unrestricted designs, this requires a $p \times p$ curvature state; under a fixed known common eigenbasis for pairwise commuting empirical task Hessians, Corollary 2 gives an exact $O(p)$-dimensional transmitted state without changing the original ℓ_1 penalty.

3.3 Consequence Under a Shared Coefficient

The simplest consequence of Theorem 1 arises when $\beta_1^\star = \cdots = \beta_t^\star = \beta^\star$: task mismatch disappears and the cumulative and current targets coincide.

Corollary 1 (Full-history sparse estimation under a shared coefficient)**.** *Suppose $\beta_1^\star = \cdots = \beta_t^\star = \beta^\star$, $\|\beta^\star\|_0 \le s$, and the pooled design and noise satisfy the standard sub-Gaussian conditions with well-conditioned covariance. If $N_t \gtrsim s \log p$ and $\lambda_{w,t} \asymp \sqrt{\log p / N_t}$, then, with high probability,*

$$\|\widehat{w}_t - \beta^\star\|_\Sigma^2 \lesssim \frac{s \log p}{N_t}, \qquad \|\widehat{w}_t - \beta^\star\|_1 \lesssim s\sqrt{\frac{\log p}{N_t}}. \tag{10}$$

Let $S = \mathrm{supp}(\beta^\star)$. Under the additional irrepresentability, beta-min, and sample-size conditions in Appendix D.1,

$$\mathbb{P}\{\mathrm{supp}(\widehat{w}_t) = S, \quad \mathrm{sgn}(\widehat{w}_{t,S}) = \mathrm{sgn}(\beta_S^\star)\} \longrightarrow 1. \tag{11}$$

Proofs are in Appendices A and D.1; unlike Table 2's rate comparison, Theorem 1 gives exact objective-level equivalence.

Table 2: Statistical comparison in the shared-coefficient setting. Rates show representative squared-estimation/excess-prediction scales under well-conditioned designs; dense and sparse rows correspond to different parameter classes and are not direct minimax comparisons.

Estimator	Raw history	Structure	Error scale	Signed support
Current-only Lasso	No	Sparse	$\frac{s \log p}{n_t}$	✓‡
Offline pooled Lasso	Yes	Sparse	$\frac{s \log p}{N_t}$	✓‡
Gen. ℓ_2 CL (Zhao et al., 2024)	No	Dense	$\frac{p}{N_t}$ †	–
SP-CL cumulative	No	Sparse	$\frac{s \log p}{N_t}$	✓‡

† Representative scaling of the oracle-order generalized-ℓ_2 error in a common, well-conditioned shared-coefficient specialization. ‡ Under the additional irrepresentability, beta-min, and sample-size conditions of Corollary 1, or their current-only analogue.

4 Current-Task Transfer under Related Task Shift

4.1 Design condition and the related-task class

We now allow task-specific coefficients and ask when related history improves current-task estimation under the following design condition.

Condition 1 (Design and noise)**.** *The rows x_{ji} are independent mean-zero sub-Gaussian vectors with common covariance Σ, whose eigenvalues lie in $[\kappa_l, \kappa_u]$ for fixed constants $0 < \kappa_l \le \kappa_u < \infty$; the errors ε_{ji} are independent mean-zero sub-Gaussian with uniformly bounded sub-Gaussian norm.*

For each task $t \ge 2$, consider the oracle related-task class

$$\Theta_t(s,h) = \Big\{(\beta_1^\star, \dots, \beta_t^\star) : \|\beta_t^\star\|_0 \le s,\ \max_{j<t} \big\|\beta_j^\star - \beta_t^\star\big\|_1 \le h\Big\}.$$

Here h measures the largest ℓ_1 discrepancy between the current task and a historical task; smaller h means more transferable history. We work in an oracle related-task regime in which all historical tasks are informative for the current target.

Under the common covariance in Condition 1, with $\pi_{j,t} = n_j/N_t$, the cumulative population target is $w_t^\star = \sum_{j=1}^t \pi_{j,t}\beta_j^\star$. Hence, writing $\Delta_t^\star = \beta_t^\star - w_t^\star$, the relatedness condition gives $\|\Delta_t^\star\|_1 \le h$.

4.2 Current-task transfer

We compare the cumulative and adapted predictors with the current-only Lasso $\widehat{\beta}_t^{\mathrm{cur}}$, obtained by minimizing $(2n_t)^{-1}\|Y_t - X_t b\|_2^2 + \lambda_{\mathrm{cur},t}\|b\|_1$ over b. The cumulative predictor $\widehat{w}_t$ uses all historical information through the score-preserving state, and the adapted predictor $\widehat{\beta}_t = \widehat{w}_t + \widehat{\Delta}_t$ uses the current batch to correct cumulative mismatch. Let $D_t = \|w_t^\star - \beta_t^\star\|_\Sigma$.

Theorem 2 (Anytime estimation under related task shift)**.** *Fix $\delta \in (0,1)$ and, for $t \ge 1$, let $L_{t,\delta} = \log(C_{\mathrm{u}} p\, t(t+1)/\delta)$, where C_{u} is sufficiently large, and take $\lambda_{\mathrm{cur},t}, \lambda_{\Delta,t} \asymp \sqrt{L_{t,\delta}/n_t}$ and $\lambda_{w,t} \asymp \sqrt{L_{t,\delta}/N_t}$. Under Condition 1 and the sparsity and sample-size conditions below, with probability at least $1-\delta$, the following statements hold simultaneously for every $t \ge 1$ for which their respective conditions hold. For* (a) *current-only fitting, suppose $\|\beta_t^\star\|_0 \le s$ and $n_t \gtrsim sL_{t,\delta}$. Then*

$$\Big\|\widehat{\beta}_t^{\mathrm{cur}} - \beta_t^\star\Big\|_\Sigma^2 \lesssim \frac{sL_{t,\delta}}{n_t}, \qquad \Big\|\widehat{\beta}_t^{\mathrm{cur}} - \beta_t^\star\Big\|_1 \lesssim s\sqrt{\frac{L_{t,\delta}}{n_t}}.$$

For (b) *the cumulative generalist, suppose $\|w_t^\star\|_0 \le s_{w,t}$ and $N_t \gtrsim s_{w,t}L_{t,\delta}$. Then*

$$\|\widehat{w}_t - w_t^\star\|_\Sigma^2 \lesssim \frac{s_{w,t}L_{t,\delta}}{N_t}, \qquad \|\widehat{w}_t - w_t^\star\|_1 \lesssim s_{w,t}\sqrt{\frac{L_{t,\delta}}{N_t}}.$$

Moreover, $\mathcal{R}_t(\widehat{w}_t) \lesssim s_{w,t}L_{t,\delta}/N_t + D_t^2$.

For (c) *the adapted specialist, let $t \ge 2$ and suppose $(\beta_1^\star, \dots, \beta_t^\star) \in \Theta_t(s,h)$ with $h\sqrt{L_{t,\delta}/n_t} = o(1)$ along the underlying high-dimensional sequence, $h \lesssim \sqrt{s}$, $n_t \gtrsim L_{t,\delta}$, and $N_{t-1} \gtrsim sL_{t,\delta}$. Then*

$$\Big\|\widehat{\beta}_t - \beta_t^\star\Big\|_\Sigma^2 \lesssim \frac{sL_{t,\delta}}{N_t} + \left(h\sqrt{\frac{L_{t,\delta}}{n_t}} \wedge h^2\right), \quad \Big\|\widehat{\beta}_t - \beta_t^\star\Big\|_1 \lesssim s\sqrt{\frac{L_{t,\delta}}{N_t}} + h. \tag{12}$$

For polynomially growing t and δ^{-1}, $L_{t,\delta} \asymp \log p$; stronger beta-min and support-alignment conditions yield signed-support recovery (Appendix D). By Theorem 1, the pooled first stage is exactly its offline counterpart, so the current correction inherits the offline two-stage rate without revisiting history; Appendix C gives the full constant-level statement.

Rate interpretation. Writing $q_t = \sqrt{L_{t,\delta}/n_t}$, current-only fitting has scale sq_t^2, whereas the specialist has scale $sL_{t,\delta}/N_t + (hq_t \wedge h^2)$. More history reduces the first term, while larger task discrepancy increases the second. Under common covariance, the generalist instead pays direct mismatch $D_t^2 \lesssim h^2$, so correction is most useful when pooling reduces variance but introduces an estimable sparse mismatch.

5 Continual Evaluation under Task Heterogeneity

We now compare the cumulative generalist $b_t^{\mathrm{G}} = \widehat{w}_t$ and current-adapted specialist $b_t^{\mathrm{S}} = \widehat{\beta}_t$ (superscripts G/S throughout) through Overall, FWT, and BWT under task heterogeneity.

Task conflict. Under the common covariance, define the unavoidable single-head conflict $\mathsf{TC}_t := \sum_{j=1}^{t} \pi_{j,t} \left\|\beta_j^\star - w_t^\star\right\|_\Sigma^2$. Then, for every $b \in \mathbb{R}^p$,

$$\overline{\mathrm{MSE}}_t(b) = \sum_{j=1}^{t} \pi_{j,t} \mathrm{MSE}_j(\beta_j^\star) + \mathsf{TC}_t + \|b - w_t^\star\|_\Sigma^2 . \tag{13}$$

Thus TC_t is the unavoidable loss from representing heterogeneous tasks by one coefficient, while the final term is the avoidable deviation from the best cumulative single head. The analogous historical decomposition is centered at $w_{t-1}^\star$ with conflict TC_{t-1}. Proposition 4 in Appendix B.6 gives a two-task lower bound.

Proposition 2 (Generalist–specialist metric relations). *Let $\overline{\mathrm{MSE}}_{<t}(b)$ denote the weighted average over tasks $j < t$; define the specialist's historical cost $\mathrm{BC}_t := \overline{\mathrm{MSE}}_{<t}(\widehat{\beta}_t) - \overline{\mathrm{MSE}}_{<t}(\widehat{w}_t)$ and current gain $\mathrm{CG}_t := \mathrm{MSE}_t(\widehat{w}_t) - \mathrm{MSE}_t(\widehat{\beta}_t)$.* Then, for the generalist and specialist, (a) $\mathrm{FWT}_t^{\mathrm{S}} - \mathrm{FWT}_t^{\mathrm{G}} = (t-1)^{-1} \sum_{k=2}^{t} \mathrm{CG}_k$ and (b) $\overline{\mathrm{MSE}}_t(\widehat{\beta}_t) - \overline{\mathrm{MSE}}_t(\widehat{w}_t) = N_{t-1}\mathrm{BC}_t/N_t - n_t \mathrm{CG}_t/N_t$. Under equal batch sizes, for any deployed sequence $\{\widehat{b}_j\}$, (c) $\mathrm{BWT}_t(\widehat{b}) = (t-1)^{-1} \sum_{j<t} \mathcal{R}_j(\widehat{b}_j) - \mathsf{TC}_{t-1} - \left\|\widehat{b}_t - w_{t-1}^\star\right\|_\Sigma^2$. The routine proof is given in Appendix E.

Part (a) links current-task correction directly to the specialist's FWT gain, while part (b) expresses Overall performance as a current-gain versus historical-cost trade-off. Part (c) additionally isolates task conflict: even with perfect historical information, $\widehat{b}_j = \beta_j^\star$ for $j < t$ and $\widehat{b}_t = w_{t-1}^\star$ give $\mathrm{BWT}_t(\widehat{b}) = -\mathsf{TC}_{t-1}$. Thus old-task degradation need not reflect loss of transmitted information; Appendix F gives a budgeted interpolation between the two outputs.

6 Experiments

Setup. We evaluate sparse estimation, task-shift transfer, and continual-learning performance under model (1) with Gaussian AR(1) covariates $\Sigma_{k\ell} = 0.5^{|k-\ell|}$, $\sigma = 1$, and sparse coefficients. We report analytic population risks over 50 streams; tuning, solver, and sensitivity details are in Appendix B.

Sparse high-dimensional estimation. We first consider a shared $s = 10$ sparse coefficient with $p = 1000$, $n_t = 40$, and $T = 20$, so that $N_T = 800 < p$ throughout the stream. Figure 2 compares sparse and dense estimators. At the final task, SP-CL attains excess risk 0.115, compared with 2.92 for pilot-tuned pooled Ridge, while using about 20 effective coordinates rather than roughly 950–990. In this globally underdetermined regime, SP-CL retains the statistical benefit of pooling while the dense pooled estimator remains substantially less accurate.

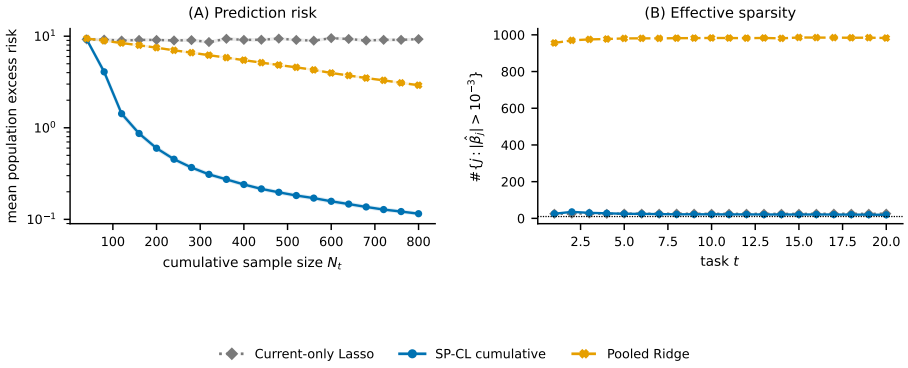

Figure 2: Sparse efficiency ($p = 1000$, $s = 10$, $n_t = 40$, $T = 20$, $N_T = 800 < p$; 50 streams). Excess risk and support size; pilot-tuned Ridge versus SP-CL (pooled Lasso matches SP-CL).

Transfer under task shift. We next vary the task mismatch h and the amount of historical information $m = N_{t-1}/n_t$ under a fixed-support coefficient path (Appendix B.5). Figure 3 reports

the gain from pooling, the additional gain from current-task correction, and their total gain over current-only fitting. Pooling is beneficial for small h; at intermediate mismatch, correction can rescue negative transfer from pooling; and with severe shift and little history, even the corrected estimator can underperform current-only fitting. The transition across h and m reflects the two terms in Theorem 2: more history reduces estimation error, whereas larger task mismatch increases the correction cost.

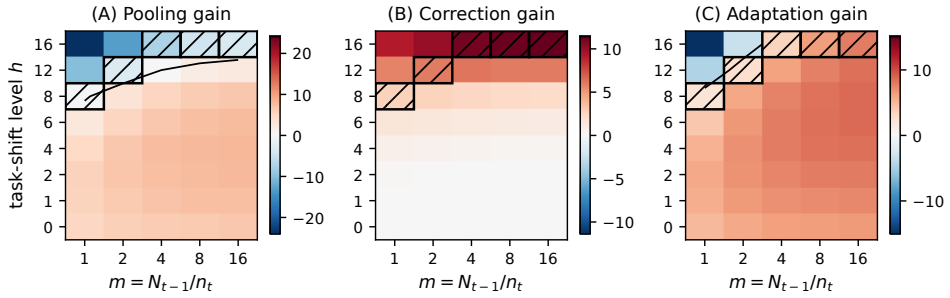

Figure 3: Transfer under task shift (50 streams). Mean gains as task mismatch h and history ratio $m = N_{t-1}/n_t$ vary: $G_{\text{pool}} = \mathcal{R}_t(\widehat{\beta}_t^{\text{cur}}) - \mathcal{R}_t(\widehat{w}_t)$, $G_{\text{corr}} = \mathcal{R}_t(\widehat{w}_t) - \mathcal{R}_t(\widehat{\beta}_t)$, and $G_{\text{adapt}} = G_{\text{pool}} + G_{\text{corr}}$; positive values = improvement and hatching = correction rescue (Appendix B.5).

Continual-learning performance. We fix $T = 20$ and vary $h \in \{0, 4, 8, 12\}$. Figure 4 compares the cumulative generalist $\widehat{w}_t$ with the current-adapted specialist $\widehat{\beta}_t$ on final-task Overall, FWT, and BWT. At $h = 12$, the specialist's forward transfer exceeds the generalist's ($\text{FWT}^{\text{S}} = 8.23$ vs. $\text{FWT}^{\text{G}} = 6.45$), while its backward transfer falls well below the generalist's ($\text{BWT}^{\text{S}} = -3.23$ vs. $\text{BWT}^{\text{G}} = +0.98$); Overall favors the generalist throughout. These results expose the generalist–specialist trade-off: increasing specialization improves forward transfer but incurs a growing historical cost, as decomposed in Proposition 2.

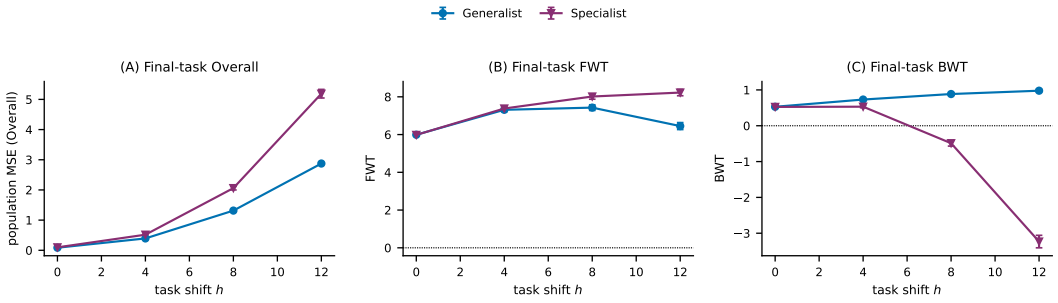

Figure 4: Continual performance under task heterogeneity ($p = 200$, $s = 10$, $n_t = 40$, $T = 20$; 50 streams). Final-task Overall, FWT, and BWT for the cumulative generalist and current-adapted specialist as h varies; error bars show one standard error.

7 Conclusion

We show that recursively transmitting an ℓ_1-penalized estimate propagates past shrinkage, and develop SP-CL to separate sparse estimation from state transmission. Its cumulative estimator exactly matches offline pooled Lasso, while a current-task correction provides statistically efficient adaptation under task shift. Overall/FWT/BWT decompositions further separate information retention, task conflict, and specialization. SP-CL therefore treats regularization used for estimation and information retained for future tasks as distinct objects.

REPRODUCIBILITY STATEMENT

The algorithm, assumptions, and complete proofs are provided in the main paper and appendices. The supplementary ZIP contains anonymized core implementation, fixed seeds, and scripts/configurations for reproducing the main experimental results. A cleaned and documented version of the full codebase will be released publicly after the review process.

AI USE STATEMENT

Generative AI tools were used to help develop the conceptual framework, formulate and organize mathematical claims, design the initial experimental protocol, and draft and edit this manuscript. The authors manually reviewed all AI-assisted content, checked the cited literature against primary sources, and take responsibility for the final text, mathematical claims, and empirical artifacts.

REFERENCES

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Trenton Bricken, Xander Davies, Deepak Singh, Dmitry Krotov, and Gabriel Kreiman. Sparse distributed memory is a continual learner. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JknGeelZJpHP.

Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.

Gilad Karpel, Edward Moroshko, Ran Levinstein, Ron Meir, Daniel Soudry, and Itay Evron. Optimal ℓ_2 regularization in high-dimensional continual linear regression. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–62, 2026.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232, pp. 513–533, 2023.

Sai Li, T. Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *Journal of the Royal Statistical Society: Series B*, 84:149–173, 2022.

Lan Luo, Ruijian Han, Yuanyuan Lin, and Jian Huang. Online inference in high-dimensional generalized linear models with streaming data. *Electronic Journal of Statistics*, 17:3443–3471, 2023.

Matthew J. Menne, Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910, 2012.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

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*, 46:5362–5383, 2024.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 61021–61039, 2024.

Yihan Zhao, Wenqing Su, and Ying Yang. High-dimensional asymptotics of generalization performance in continual ridge regression. *arXiv preprint arXiv:2508.15494*, 2025.

A Proofs for Score-Preserving Continual Lasso

A.1 Quadratic closure

Proof of Lemma 1. Let $\widetilde{X}_t = (X_1^\mathsf{T}, \dots, X_t^\mathsf{T})^\mathsf{T}$ and $\widetilde{Y}_t = (Y_1^\mathsf{T}, \dots, Y_t^\mathsf{T})^\mathsf{T}$. Then $H_t = \widetilde{X}_t^\mathsf{T}\widetilde{X}_t$ and $G_t = \widetilde{X}_t^\mathsf{T}\widetilde{Y}_t$. Since $\mathrm{Col}(\widetilde{X}_t^\mathsf{T}\widetilde{X}_t) = \mathrm{Col}(\widetilde{X}_t^\mathsf{T})$, we have $G_t \in \mathrm{Col}(H_t)$, so there exists u_t satisfying $H_t u_t = G_t$. Expanding the historical squared loss gives $\frac{1}{2} w^\mathsf{T} H_t w - w^\mathsf{T} G_t + C_t'$. Completing the square then yields (2). □

A.2 Penalty residual under direct transmission of the penalized center

Proof of Proposition 1. The KKT condition for (3) gives

$$H_t \widehat{w}_t^{\mathrm{dir}} = X_t^\mathsf{T} Y_t + H_{t-1} \widehat{w}_{t-1}^{\mathrm{dir}} - N_t \lambda_t z_t^{\mathrm{dir}}.$$

Subtracting this identity from $G_t = G_{t-1} + X_t^\mathsf{T} Y_t$ yields $D_t = D_{t-1} + N_t \lambda_t z_t^{\mathrm{dir}}$ for $D_t = G_t - H_t \widehat{w}_t^{\mathrm{dir}}$. Iterating proves (4). □

A.3 Exact pooled-Lasso equivalence

Proof of Theorem 1. The KKT equation for (6) and the induction hypothesis $H_{t-1} u_{t-1} = G_{t-1}$ give

$$H_t \widehat{w}_t = G_t - N_t \lambda_{w,t} z_t.$$

Because $(H_t/N_t) d_t = z_t$ on $\mathrm{Col}(H_t)$, $H_t(\lambda_{w,t} d_t) = N_t \lambda_{w,t} z_t$, proving $H_t u_t = G_t$. Lemma 1 then shows that the smooth part of (6) is, up to a constant, the full-history squared loss. Dividing by N_t yields (8). □

A.4 Exact compressed state under commuting empirical task Hessians

Corollary 2 (Linear-dimensional transmitted state under commuting empirical task Hessians)**.** *For each batch, let $\widehat{\Sigma}_j = n_j^{-1} X_j^\mathsf{T} X_j$. Suppose that the symmetric matrices $\{\widehat{\Sigma}_j\}_{j \geq 1}$ commute pairwise and have a fixed known common orthogonal eigenbasis U, so that*

$$\widehat{\Sigma}_j = U \Gamma_j U^\mathsf{T}, \qquad \Gamma_j \text{ diagonal}.$$

Set $D_0 = 0$, $a_0 = 0$, and, at task t, update

$$D_t = D_{t-1} + n_t \Gamma_t.$$

The cumulative fit and score-preserving update can be computed as

$$\widehat{w}_t \in \arg\min_w \left\{ \frac{1}{2} \|Y_t - X_t w\|_2^2 + \frac{1}{2} (U^\mathsf{T} w - a_{t-1})^\mathsf{T} D_{t-1} (U^\mathsf{T} w - a_{t-1}) + N_t \lambda_{w,t} \|w\|_1 \right\}, \tag{14}$$

$$a_t = U^\mathsf{T} \widehat{w}_t + N_t \lambda_{w,t} D_t^+ U^\mathsf{T} z_t, \qquad z_t \in \partial \|\widehat{w}_t\|_1 . \tag{15}$$

For every t, $D_t a_t = U^\mathsf{T} G_t$, and (14) *has the same minimizer set as the full-state SP-CL update and the offline pooled Lasso in* (8)*. Thus (D_t, a_t, N_t) is a $2p+1$-dimensional transmitted state. The ℓ_1 penalty remains $\|w\|_1$ in the original parameterization. This is an $O(p)$ dynamic-state representation conditional on the shared basis U; if a dense U must also be stored, total persistent storage remains $O(p^2)$.*

Proof. Because the empirical task Hessians are symmetric and commute, they are simultaneously diagonalizable, giving the displayed representation. Let $g_t = U^\mathsf{T} G_t$. Then

$$H_t = \sum_{j \le t} X_j^\mathsf{T} X_j = U D_t U^\mathsf{T}, \qquad g_t = g_{t-1} + U^\mathsf{T} X_t^\mathsf{T} Y_t.$$

We prove $D_t a_t = g_t$ by induction. It holds at $t = 0$. If it holds at $t-1$, then the quadratic historical term in (14) equals, up to a constant,

$$\frac{1}{2} w^\mathsf{T} U D_{t-1} U^\mathsf{T} w - w^\mathsf{T} U D_{t-1} a_{t-1} = \frac{1}{2} w^\mathsf{T} H_{t-1} w - w^\mathsf{T} G_{t-1}.$$

Hence (14) is exactly (6). Its KKT equation, after left multiplication by U^T, is

$$D_t U^\mathsf{T} \widehat{w}_t = g_t - N_t \lambda_{w,t} U^\mathsf{T} z_t.$$

As in the proof of Theorem 1, $z_t \in \mathrm{Col}(H_t)$; consequently $U^\mathsf{T} z_t \in \mathrm{Col}(D_t)$. Applying (15) therefore gives

$$D_t a_t = D_t U^\mathsf{T} \widehat{w}_t + N_t \lambda_{w,t} D_t D_t^{+} U^\mathsf{T} z_t = g_t.$$

The induction closes, so the reduced quadratic term is identical to the full-state one at every task. Theorem 1 then yields the stated offline pooled-Lasso equivalence. □

A.5 Proof of the estimation claim in Corollary 1

Proof. Theorem 1 identifies $\widehat{w}_t$ with an offline pooled Lasso minimizer targeting the s-sparse $\beta^\star$. The stated result is therefore the standard pooled Lasso bound under the displayed design, noise, and sample-size conditions, including its usual prediction and ℓ_1 error bounds. □

B Supporting results for direct transmission of the penalized center

B.1 Persistent-sign lower bound

This appendix gives detailed consequences of the penalty residual in Proposition 1, including the underdetermined construction used in Example 1. Suppose $n_j = n_0$ and $\beta_j^\star = \beta^\star$ for all j, write $\lambda_j = \lambda_{w,j}$, and define

$$q_j = \frac{1}{n_0} X_j^\mathsf{T} Y_j, \qquad \xi_j = q_j - \frac{1}{n_0} X_j^\mathsf{T} X_j \beta^\star, \qquad \xi_{1:t} = \sum_{j=1}^{t} \xi_j.$$

Then $G_t = H_t \beta^\star + n_0 \xi_{1:t}$. Let $S = \mathrm{supp}(\beta^\star)$, $s = |S|$, and $\overline{H}_t = H_t/(t n_0)$.

Proposition 3 (Persistent-sign lower bound)**.** *Suppose that, through time t,*

$$\mathrm{supp}(\widehat{w}_j^{\mathrm{dir}}) = S, \qquad z_{j,S}^{\mathrm{dir}} = \mathrm{sgn}(\beta_S^\star), \qquad 1 \le j \le t.$$

Then

$$\overline{H}_{t,SS}(\widehat{w}_{t,S}^{\mathrm{dir}} - \beta_S^\star) = \frac{1}{t} \xi_{1:t,S} - L_t \,\mathrm{sgn}(\beta_S^\star), \qquad L_t = \frac{1}{t} \sum_{j=1}^{t} j \lambda_j. \tag{16}$$

Consequently, if $\overline{H}_{t,SS} \succ 0$ and $\lambda_{\max}(\overline{H}_{t,SS}) \le C_{\max}$, then

$$\left\| \widehat{w}_t^{\mathrm{dir}} - \beta^\star \right\|_2 \ge \frac{\left[\sqrt{s} L_t - \| \xi_{1:t,S}/t \|_2 \right]_+}{C_{\max}}. \tag{17}$$

Proof. Divide (4) by tn_0, use $G_t = H_t\beta^\star + n_0\xi_{1:t}$, and restrict to S to obtain (16). The bound follows from the triangle inequality and the upper spectral bound on $\overline{H}_{t,SS}$. □

For the nominal high-dimensional schedule $\lambda_j \asymp \sqrt{\log p/(jn_0)}$, $L_t \asymp \sqrt{t \log p/n_0}$. On the usual score-concentration event, persistent signed-support recovery and ℓ_2 consistency are incompatible whenever $st \log p/n_0 \not\to 0$. This obstruction involves only the active block; the full cumulative Gram matrix may remain singular.

B.2 Support extinction

Example 2 (Support extinction under a batchwise underdetermined design)**.** *Suppose $s < n_0 < p$ and, after ordering coordinates by $S = \mathrm{supp}(\beta^\star)$, each noiseless batch satisfies*

$$\frac{1}{n_0} X_t^\top X_t = \begin{pmatrix} I_s & 0 \\ 0 & B_t \end{pmatrix}, \qquad B_t \succeq 0, \quad \mathrm{rank}(B_t) \le n_0 - s, \qquad Y_t = X_{t,S}\beta^\star_S.$$

Initialize $\widehat{w}_0^{\mathrm{dir}} = 0$. Then $\widehat{w}^{\mathrm{dir}}_{t,S^c} = 0$ for all t, and every $k \in S$ obeys

$$\widehat{w}^{\mathrm{dir}}_{t,k} = \mathcal{S}_{\lambda_t}\left\{ \frac{\beta^\star_k + (t-1)\widehat{w}^{\mathrm{dir}}_{t-1,k}}{t} \right\}, \qquad \mathcal{S}_a(u) = \mathrm{sgn}(u)(|u| - a)_+. \tag{18}$$

Equivalently, for $c_{t,k} = t|\widehat{w}^{\mathrm{dir}}_{t,k}|$,

$$c_{t,k} = \left[c_{t-1,k} + |\beta^\star_k| - t\lambda_t\right]_+ .$$

For $\lambda_t = c_\lambda\sqrt{\log(ep)/(tn_0)}$, $t\lambda_t \to \infty$. Hence, for every $k \in S$, there is a finite τ_k such that $\widehat{w}^{\mathrm{dir}}_{t,k} = 0$ for all $t \ge \tau_k$, and thus $\widehat{w}^{\mathrm{dir}}_t \to 0$. The construction remains batchwise underdetermined for arbitrarily large p, because only the s-dimensional active block is required to be orthogonal.

Proof. On the active coordinates, the current and accumulated quadratic terms in (3) reduce, after division by tn_0, to a squared loss centered at $(\beta^\star_k + (t-1)\widehat{w}^{\mathrm{dir}}_{t-1,k})/t$. Soft thresholding therefore gives (18), and multiplication by t gives the recursion for $c_{t,k}$. On S^c, the response is zero and the previous center is zero; the ℓ_1 penalty selects zero at every step. Since $t\lambda_t \to \infty$, the decrements $t\lambda_t - |\beta^\star_k|$ eventually become positive and have a divergent sum. Thus $c_{t,k}$ reaches zero in finite time and remains zero thereafter. □

B.3 Additional state-transmission diagnostics

This subsection reports the numerical diagnostics underlying Figure 1 and its penalty-constant sensitivity, and is the noisy, correlated-design counterpart to the deterministic construction above. Across the 50-seed noisy experiment ($p = 200$, $s = 10$, $n_t = 40$, $T = 30$, AR(1) $\rho = 0.5$, $\sigma = 1$, $c_\lambda = 1$), every solve converged (KKT residual: max 5.4×10^{-9}, median 1.6×10^{-9}, over 1500 rows and four estimators). The Proposition 1 identity holds to a relative error of at most 1.7×10^{-8} (median 4.7×10^{-15}), and the SP-CL/offline pooled-Lasso objective gap remains at floating-point precision throughout (max 8.9×10^{-16}, median exactly 0). The single largest value in each of these diagnostics traces to one seed at task 5, where the pooled design was exactly square ($N_5 = p = 200$) and produced a near-singular Gram matrix: this inflates only the carrier-reconstruction (pseudoinverse) residual, to 4.0×10^{-7} at that cell, not the pooled-objective equivalence, which for that same seed and task remains at 2.5×10^{-15}. An analogous but smaller effect (residuals up to about 10^{-7} in a small number of near-singular pooled-design cells) appears in the task-shift experiments of Section 6, for the same conditioning reason and with no loss of objective equivalence there either.

Figure 5 plots this discrepancy directly: SP-CL stays at floating-point precision at every task, while direct-center's remains persistently positive and does not shrink as more data arrive.

At the final task, direct-center's mean population MSE is 10.99, within 0.1% of the fully-shrunk null-predictor floor ($\sigma^2 + \|\beta^\star\|^2_\Sigma \approx 11$), against 1.06 for both SP-CL and offline pooled; its mean true-positive rate on the 10-coordinate active support falls to 0.046, against 1.000 for SP-CL and offline pooled.

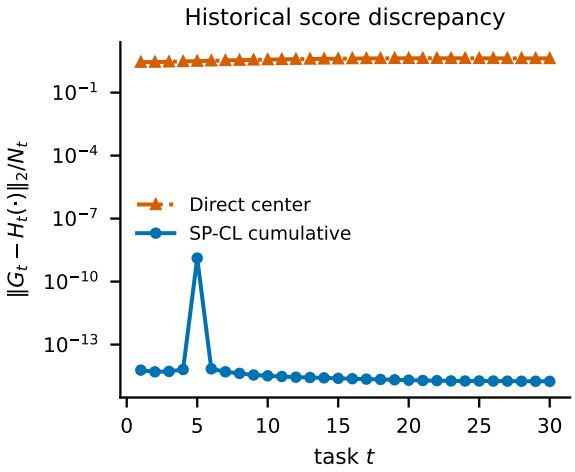

Figure 5: Historical score discrepancy $\|G_t - H_t(\cdot)\|_2/N_t$ under the same noisy, correlated design as Figure 1, for direct-center ($\widehat{w}_t^{\mathrm{dir}}$) and SP-CL ($u_t$), log scale.

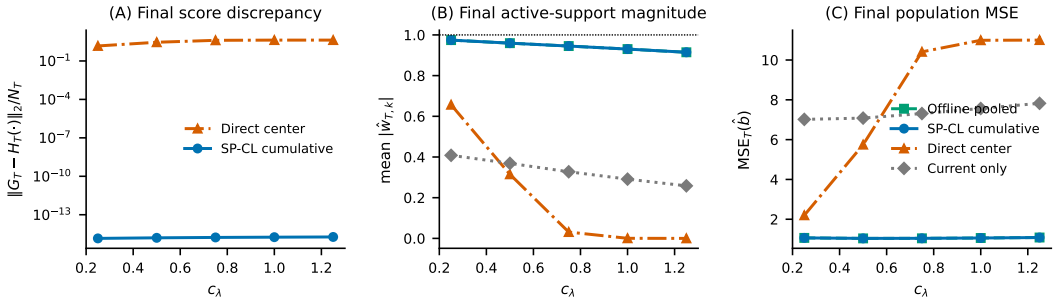

Figure 6: Final-task ($t = T$) sensitivity to the penalty constant c_λ, same 50 seeds reused across $c_\lambda \in \{0.25, 0.5, 0.75, 1, 1.25\}$. Panel (A): score discrepancy. Panel (B): active-support magnitude. Panel (C): population MSE.

Penalty-constant sensitivity. Repeating this experiment (same 50 seeds) for $c_\lambda \in \{0.25, 0.5, 0.75, 1, 1.25\}$, Figure 6 shows that SP-CL's score discrepancy remains at floating-point precision at every tested value, while direct-center's remains persistently positive throughout the same range: the mechanism itself does not depend on the penalty constant. The downstream severity does: at $c_\lambda = 0.25$, direct-center retains 66% of the true coefficient magnitude on average and its MSE (2.2) is far closer to SP-CL's (1.06) than to the null floor (≈ 11); support collapse and near-null MSE are specific to $c_\lambda \gtrsim 0.75$ in this design. We do not claim direct-center collapses for every statistically motivated penalty constant, only that it does so robustly at and above the order used as the main setting.

B.4 Additional sparse-efficiency results

Figure 7 gives the squared estimation-error trace for the high-dimensional experiment in Section 6. Its ordering agrees with the risk comparison in Figure 2. This experiment is not an exact variable-selection setting: the final-task false discovery rate of SP-CL is approximately 0.47. Its purpose is instead to quantify the finite-sample prediction benefit of sparse estimation when $p > N_T$.

B.5 Task-shift simulation construction

For Figure 3, we use $p = 200$, $s = 10$, and equal batch sizes $n_t = 40$. The fixed-support task path is

$$\beta_k^\star = \beta_0^\star + \frac{h(k-1)}{T-1} v,$$

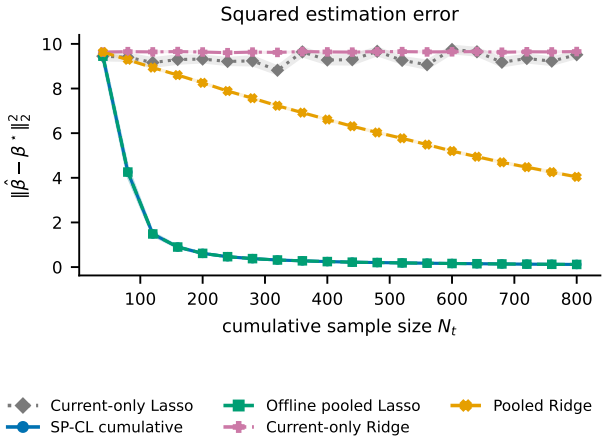

Figure 7: Squared ℓ_2 estimation error in the high-dimensional sparse-versus-dense experiment of Figure 2. Pooled Ridge uses the pilot-tuned task-specific penalty schedule.

where $\|v\|_1 = 1$ and v is supported on five of the ten active coordinates. Thus the final coefficient is within ℓ_1 distance h of every earlier coefficient. We vary $h \in \{0, 1, 2, 4, 6, 8, 12, 16\}$ and $m = N_{t-1}/n_t \in \{1, 2, 4, 8, 16\}$, using $T = m + 1$ equal-size batches. Each cell averages 50 common-random-number replicates, so the current-only fit is shared across values of m for a fixed (h, seed).

B.6 Unavoidable single-head conflict

For completeness, any one coefficient must incur nonzero loss on at least one of two tasks with distinct population optima:

Proposition 4 (Unavoidable single-head conflict)**.** *For any two tasks $j < t$,*

$$\inf_{b \in \mathbb{R}^p} \max\{\mathcal{R}_j(b), \mathcal{R}_t(b)\} = \frac{1}{4} \left\|\beta_j^\star - \beta_t^\star\right\|_\Sigma^2 . \tag{19}$$

Proof. The triangle inequality gives $\left\|\beta_j^\star - \beta_t^\star\right\|_\Sigma \le \left\|b - \beta_j^\star\right\|_\Sigma + \|b - \beta_t^\star\|_\Sigma$. The lower bound follows after squaring, and equality is attained at $b = (\beta_j^\star + \beta_t^\star)/2$. □

C Proof of Theorem 2

Parts (a) and (b) are the standard single-task and sparse-anchor Lasso bounds, proved at the end of this section. We prove part (c) by a confidence-level specialization of Tian & Feng (2023); c, C denote constants that may change between lines and depend only on κ_l, κ_u.

Step 1 (Exact offline reduction). By Theorem 1, $\widehat{w}_t$ is exactly a minimizer of the offline pooled Lasso (8) formed from $\mathcal{D}_1, \ldots, \mathcal{D}_t$, with no dependence on how the inaccessible batches $\mathcal{D}_j$, $j < t$, were summarized. Consequently the pair $(\widehat{w}_t, \widehat{\Delta}_t)$ defined by (6) and (9) is identical, as a random pair given $(\mathcal{D}_1, \ldots, \mathcal{D}_t)$, to the output of Algorithm 1 (A-Trans-GLM) of Tian & Feng (2023) specialized to $\psi(u) = u^2/2$, run with target sample $\mathcal{D}_t$ and source set $\{\mathcal{D}_j\}_{j<t}$: their transferring step is our (6) divided by N_t, and their debiasing step is our (9). This identification holds exactly and is the only streaming-specific step in the proof; everything below follows from their offline argument after making its confidence level explicit.

Step 2 (Notation map and population target). Their target parameter, target sample size, source total, and pooled total map to our $\beta_t^\star$, n_t, N_{t-1}, and N_t; their source parameters map to $\beta_j^\star$, $j < t$; their contrast bound maps to h. Under common covariance, the population version of the transferring step solves $\Sigma w = \sum_{j \le t} \pi_{j,t} \Sigma \beta_j^\star$, giving $w_t^\star = \sum_{j \le t} \pi_{j,t} \beta_j^\star$, and the related-task condition in Section 4.1 gives $\|\beta_t^\star - w_t^\star\|_1 \le \sum_{j<t} \pi_{j,t} \left\|\beta_t^\star - \beta_j^\star\right\|_1 \le h$, so $w_t^\star$ plays the role of their $w_{\mathcal{A}_h}$.

Step 3 (Covariance compatibility). Their Theorem 1 requires a covariance-compatibility condition (their Assumption 4). Under common covariance, $\widetilde{\Sigma}_h = \sum_k \alpha_k \Sigma = \Sigma$ and $\Sigma_h^{(k)} = \Sigma$ for every historical task k, so $\widetilde{\Sigma}_h^{-1}\Sigma^{(k)} = I_p$, whose induced ℓ_1-operator norm is $1 < \infty$: the condition holds automatically, with no separate assumption needed.

Step 4 (Confidence-level transfer bound). With $\psi(u) = u^2/2$, their Assumption 3(i) holds with $M_\psi = 1$. We retain a generic failure budget in the concentration argument underlying their Theorem 1. For $\xi \in (0,1)$, write

$$L(\xi) = \log\left(\frac{C'_{\mathrm{u}} p}{\xi}\right).$$

Under Condition 1, the finitely many sub-Gaussian score and empirical-covariance concentration events used in that proof hold jointly with probability at least $1-\xi$ after taking C'_{u} sufficiently large and replacing $\log p$ by $L(\xi)$ in the penalties and sample-size requirements. The remaining oracle inequalities in their proof are deterministic on this joint event. Thus their argument, translated via Step 2 ($n_0 \to n_t$, $n_{\mathcal{A}_h} \to N_{t-1}$, $n_{\mathcal{A}_h} + n_0 \to N_t$, and their target sparsity to $s = \|\beta_t^\star\|_0$), gives the following confidence-level form: if $h\sqrt{L(\xi)/n_t} = o(1)$ along the same asymptotic sequence, $h \le C\sqrt{s}$, $n_t \ge CL(\xi)$, and $N_{t-1} \ge CsL(\xi)$, then, with

$$\lambda_{w,t} = C_w\sqrt{L(\xi)/N_t}, \qquad \lambda_{\Delta,t} = C_\Delta\sqrt{L(\xi)/n_t},$$

$$\left\|\widehat{\beta}_t - \beta_t^\star\right\|_2 \lesssim \sqrt{\frac{sL(\xi)}{N_t}} + \left[\left(\frac{L(\xi)}{n_t}\right)^{1/4} h^{1/2} \wedge h\right],$$

$$\left\|\widehat{\beta}_t - \beta_t^\star\right\|_1 \lesssim s\sqrt{\frac{L(\xi)}{N_t}} + h.$$

with probability at least $1-\xi$. This is the confidence-level form of their inequalities (6)–(7) after the translation above; it retains the same joint high-probability event. In particular, the argument does not, and cannot, condition on $\widehat{w}_t$ and treat the current design as independent of it, because $\widehat{w}_t$ is fit using the current batch through the pooled objective.

Step 5 (Squaring). Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ to absorb the cross term into the sum of squared orders gives

$$\left\|\widehat{\beta}_t - \beta_t^\star\right\|_\Sigma^2 \lesssim \frac{sL(\xi)}{N_t} + \left(h\sqrt{\frac{L(\xi)}{n_t}} \wedge h^2\right),$$

on the same event, using the upper eigenvalue bound in Condition 1. The second line in Step 4 gives the ℓ_1 bound.

Parts (a) and (b). For every $\xi \in (0,1)$, the standard Lasso score and restricted-eigenvalue events under Condition 1 hold with probability at least $1-\xi$ when the penalties and sample sizes use $L(\xi)$ (Bickel et al., 2009). Thus part (a) gives $\left\|\widehat{\beta}_t^{\mathrm{cur}} - \beta_t^\star\right\|_\Sigma^2 \lesssim sL(\xi)/n_t$ and $\left\|\widehat{\beta}_t^{\mathrm{cur}} - \beta_t^\star\right\|_1 \lesssim s\sqrt{L(\xi)/n_t}$ once $n_t \ge CsL(\xi)$. Its sample-size requirement comes from the restricted-eigenvalue property over s-sparse directions, unlike Tian–Feng's $n_0 \ge C\log p$ condition, which controls their debiasing step. Part (b) additionally assumes $\|w_t^\star\|_0 \le s_{w,t}$: since $\widehat{w}_t$ is exactly the offline pooled Lasso targeting the $s_{w,t}$-sparse $w_t^\star$, the standard Lasso bound gives $\|\widehat{w}_t - w_t^\star\|_\Sigma^2 \lesssim s_{w,t}L(\xi)/N_t$ and $\|\widehat{w}_t - w_t^\star\|_1 \lesssim s_{w,t}\sqrt{L(\xi)/N_t}$ once $N_t \ge Cs_{w,t}L(\xi)$. Adding $D_t^2 = \|w_t^\star - \beta_t^\star\|_\Sigma^2$ by the triangle inequality (again via $(a+b)^2 \le 2a^2+2b^2$) gives the current-task risk bound in part (b). This ℓ_1 bound is used in Lemma 2 of Appendix D.

Step 6 (Anytime uniformization). For each $t \ge 1$, set $\xi_t = \delta/[3t(t+1)]$ and choose $C_{\mathrm{u}} = 3C'_{\mathrm{u}}$ in Theorem 2, so that $L(\xi_t) = L_{t,\delta}$. The three events establishing parts (a), (b), and (c) at time t then each fail with probability at most ξ_t. A union bound over all times and all three event classes gives

$$\mathbb{P}\left(\bigcup_{t\ge1}\bigcup_{r\in\{\mathrm{cur,cum,ad}\}} E_{t,r}^c\right) \le 3\sum_{t\ge1}\xi_t = \delta,$$

where $E_{t,r}$ denotes the event establishing assertion r at time t. Thus all displayed bounds hold simultaneously for every time at which their respective assumptions are satisfied. No independence across time is required.

D SUPPORT-RECOVERY CONSEQUENCES

This appendix records standard variable-selection consequences of the exact reduction. It gives the additional conditions and proof for the signed-support claim in Corollary 1, and records its related-task analogue. These stronger conditions make clear that score preservation transfers the relevant offline Lasso variable-selection guarantees.

D.1 CONDITIONS AND PROOF FOR SHARED-COEFFICIENT SIGNED-SUPPORT RECOVERY

Let $S = \mathrm{supp}(\beta^\star)$ and $\beta_{\min} = \min_{k\in S} |\beta_k^\star|$. In addition to the estimation conditions of Corollary 1, suppose there are constants $\kappa_S, M_S < \infty$ and $\gamma \in (0, 1]$ such that

$$\lambda_{\min}(\Sigma_{SS}) \geq \kappa_S, \qquad \left\|\Sigma_{SS}^{-1}\right\|_\infty \leq M_S, \qquad \left\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\right\|_\infty \leq 1-\gamma.$$

Assume also that $N_t \gtrsim s^2 \log p$ and $\beta_{\min} \geq C_{\mathrm{sel}}\lambda_{w,t}$ for a sufficiently large constant C_{sel}. These are the additional conditions for its signed-support conclusion. The constant implicit in $\lambda_{w,t} \asymp \sqrt{\log p/N_t}$ is taken sufficiently large to dominate the score noise.

Proof of the signed-support claim in Corollary 1. Theorem 1 identifies $\widehat{w}_t$ with the offline pooled Lasso. The stated irrepresentability, beta-min, and sample-size conditions therefore yield the usual primal–dual-witness signed-support argument (Wainwright, 2009). □

D.2 RELATED-TASK SELECTION

Write

$$S_{w,t} = \mathrm{supp}(w_t^\star), \qquad S_{\Delta,t} = \mathrm{supp}(\Delta_t^\star), \qquad S_{\beta,t} = \mathrm{supp}(\beta_t^\star),$$

and let $s_{\Delta,t} = |S_{\Delta,t}|$. In addition to Condition 1, $(\beta_1^\star, \ldots, \beta_t^\star) \in \Theta_t(s, h)$, and the sparse-anchor condition $\|w_t^\star\|_0 \leq s_{w,t}$, suppose the following hold for target time t:

(i) For each $A \in \{S_{w,t}, S_{\Delta,t}\}$,

$$\lambda_{\min}(\Sigma_{AA}) \geq \kappa_A, \qquad \left\|\Sigma_{AA}^{-1}\right\|_\infty \leq M_A, \qquad \left\|\Sigma_{A^cA}\Sigma_{AA}^{-1}\right\|_\infty \leq 1-\gamma_A$$

for positive constants κ_A, M_A and $\gamma_A \in (0, 1]$.

(ii) The covariance entries are uniformly bounded, $\max_{k,\ell\leq p} |\Sigma_{k\ell}| \leq M_\Sigma$, and, under the anytime calibration of Theorem 2, the sample sizes and penalties obey

$$N_t \gtrsim s_{w,t}^2 L_{t,\delta}, \qquad n_t \gtrsim s_{\Delta,t}^2 L_{t,\delta}, \qquad s_{w,t}\lambda_{w,t} \leq c_\Delta \lambda_{\Delta,t},$$

where c_Δ is sufficiently small relative to $\gamma_{S_{\Delta,t}}$.

(iii) The sparse anchor and contrast are support-compatible, $S_{\beta,t} = S_{w,t} \cup S_{\Delta,t}$, and satisfy the beta-min conditions

$$\min_{k\in S_{w,t}} |w_{t,k}^\star| \geq C_w\lambda_{w,t}, \qquad \min_{k\in S_{\Delta,t}} |\Delta_{t,k}^\star| \geq C_\Delta\lambda_{\Delta,t},$$

$$\min_{k\in S_{\beta,t}} |\beta_{t,k}^\star| \geq C_\beta(\lambda_{w,t} + \lambda_{\Delta,t})$$

for sufficiently large constants determined by the selection geometry.

Lemma 2 (First-stage score stability)**.** *Under Condition 1, the sparse-anchor condition $\|w_t^\star\|_0 \leq s_{w,t}$ used in Theorem 2(b), and the related-task selection conditions above, for a sufficiently large penalty constant in $\lambda_{\Delta,t}$,*

$$\left\|\frac{1}{n_t}X_t^\top X_t(\widehat{w}_t - w_t^\star)\right\|_\infty \leq \frac{\gamma_{S_{\Delta,t}}}{8}\lambda_{\Delta,t} \tag{20}$$

with probability $1 - o(1)$.

Proof. By Theorem 1, the first stage is the offline pooled Lasso targeting the $s_{w,t}$-sparse $w_t^\star$, whose standard ℓ_1 bound $\|\widehat{w}_t - w_t^\star\|_1 = O_{\mathbb{P}}(s_{w,t}\lambda_{w,t})$ is proved in Appendix C (the argument for Theorem 2(b)). The sub-Gaussian design and bounded covariance entries give $\max_{k,\ell\le p} |(X_t^\mathsf{T} X_t/n_t)_{k\ell}| = O_{\mathbb{P}}(1)$. The final condition in part (ii), with a sufficiently large current-stage penalty constant, therefore yields (20). □

Corollary 3 (Current-task signed-support recovery)**.** *Under Condition 1,* $(\beta_1^\star, \dots, \beta_t^\star) \in \Theta_t(s, h)$*, the sparse-anchor condition* $\|w_t^\star\|_0 \le s_{w,t}$*, and the related-task selection conditions above,*

$$\mathbb{P}\left\{\mathrm{supp}(\widehat{\beta}_t) = S_{\beta,t},\ \mathrm{sgn}(\widehat{\beta}_{t,S_{\beta,t}}) = \mathrm{sgn}(\beta^\star_{t,S_{\beta,t}})\right\} = 1 - o(1). \tag{21}$$

Proof. Theorem 1 turns the first stage into the offline pooled Lasso. The selection geometry and beta-min condition give its primal–dual-witness support result (Wainwright, 2009). Lemma 2 makes the residual first-stage offset in (9) smaller than the current-stage penalty, so the same argument applies to the contrast Lasso. Support compatibility and the final beta-min condition rule out cancellation when the two selected components are added. □

E Proof of the metric decomposition

Proof of Proposition 2. For part (a), subtract the two FWT definitions in Section 2 and use the definitions of BC_k and CG_k in Section 5 at each task k. Part (b) follows by reweighting the historical and current MSE terms in those same definitions. For part (c), substitute the excess-risk identity in Section 2 into the definition of BWT and apply the same decomposition as (13), restricted to tasks $j < t$; the irreducible MSE terms cancel under equal batch weights. □

F Optional partial correction under a historical budget

The two outputs define a simple continuum of deployment rules without changing the transmitted state. For $\eta \in [0, 1]$, let

$$\widehat{\beta}_t(\eta) = \widehat{w}_t + \eta\widehat{\Delta}_t. \tag{22}$$

Here $\eta = 0$ returns the generalist and $\eta = 1$ returns the fully adapted specialist. The stored state evaluates the historical squared loss, up to a constant independent of b, through

$$\widehat{L}^{\mathrm{hist}}_{<t}(b) = \frac{1}{2N_{t-1}}\left\{b^\mathsf{T} H_{t-1} b - 2b^\mathsf{T} H_{t-1} u_{t-1}\right\}. \tag{23}$$

One optional budgeted rule chooses the current-loss-optimal value over a fixed grid subject to

$$\widehat{L}^{\mathrm{hist}}_{<t}\{\widehat{\beta}_t(\eta)\} - \widehat{L}^{\mathrm{hist}}_{<t}(\widehat{w}_t) \le \varepsilon_t. \tag{24}$$

This is a deployment choice, not a new state-update algorithm.

Its population trade-off is explicit under the common-covariance model. Let $\rho_t = n_t/N_t$, write $d_t^\star = \beta_t^\star - w_{t-1}^\star$, and set $b_t^\eta = w_t^\star + \eta(\beta_t^\star - w_t^\star)$. Replacing the estimates in (22) by their population targets yields

$$\begin{aligned} \mathcal{R}_t(b_t^\eta) &= \{1 - \gamma_t(\eta)\}^2 \|d_t^\star\|_\Sigma^2, \\ \overline{\mathrm{MSE}}_{<t}(b_t^\eta) - \overline{\mathrm{MSE}}_{<t}(w_{t-1}^\star) &= \gamma_t(\eta)^2 \|d_t^\star\|_\Sigma^2, \end{aligned} \tag{25}$$

where $\gamma_t(\eta) = \rho_t + (1 - \rho_t)\eta$. Equation (25) separates the unavoidable conflict of a single head from the controllable choice of how far to adapt toward the new task.

G Real-Data Application: Climate-Stream Prediction

This appendix reports an independent real-data confirmation of Theorem 1 and Proposition 1 on a chronological climate-prediction stream, using a separate implementation of the same estimators written for this application. Because the true coefficient is unknown here, the analysis emphasizes future prediction, historical-block degradation, and the score-transmission diagnostic rather than estimation error against a ground truth.

G.1 Data and stream construction

The predictor is the NCEP–NCAR Reanalysis 1 sea-level-pressure field (Kalnay et al., 1996) over the North Pacific/North America window $0°$–$70°$N, $150°$E–$120°$W, on a $2.5°$ grid flattened to $p = 1073$ grid cells. The response is the one-week-ahead anomaly in weekly mean temperature at Los Angeles International Airport (NOAA station USW00023174, $33.938°$N, $118.387°$W) (Menne et al., 2012); missing daily TAVG values are filled by $(\mathrm{TMAX} + \mathrm{TMIN})/2$ before weekly averaging. We retain the extended winter months November–March and pair each week's pressure field with the following week's temperature anomaly within the same winter, so no pair spans two winters.

To avoid leakage, all preprocessing is fixed on a 1980–1984 reference period (week-of-year SLP and temperature climatology, per-cell SLP scaling, with reference standard deviations below 10^{-8} floored at 1) before the retrospective stream begins; these reference winters are never used for fitting or testing. Starting in winter 1985, each of $T = 8$ chronological tasks pools four training winters ($n_t \approx 81$–83 weekly pairs) and holds out a fifth winter as a task-specific test block (20–21 pairs); the test winters are 1989, 1994, 1999, 2004, 2009, 2014, 2019, and 2024. A final future block (winters 2025–2026, 41 pairs) is reserved for out-of-sample prediction and is disjoint from every historical test winter. Because training-block sizes are nearly but not exactly equal, the transmitted state uses a task-balanced cumulative objective: each task contributes its empirical average loss.

G.2 Evaluation and tuning

We compare Current-only Lasso, Offline pooled Lasso, Direct center, and SP-CL under a common penalty schedule $\lambda_t = \lambda_1/\sqrt{t}$, with $\lambda_1 = 0.20$ selected by rolling-origin validation on the training blocks only (grid $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.80\}$; all held-out test winters and the future block are excluded from tuning), together with a zero-anomaly and a persistence (current-week temperature) baseline. For task-specific test block j, let $A_{t,j}$ denote the test mean squared forecast error (MSFE) of the state deployed after task t, evaluated on block $j \le t$; define the historical forgetting $\mathrm{MaxF}_t = \max_{j<t}[A_{t,j} - A_{j,j}]_+$ and the historical regret relative to Offline pooled Lasso $\mathrm{MReg}_t = \max_{j<t}[A_{t,j} - A^{\mathrm{off}}_{t,j}]_+$, summarized by $\mathrm{MReg}^{\mathrm{hist}}_T = \max_{2\le t\le T} \mathrm{MReg}_t$ (and similarly $\mathrm{MaxF}^{\mathrm{hist}}_T$). Historical-test uncertainty uses a paired percentile bootstrap (2000 replications) over the eight test winters; the short future block instead uses a winter-stratified moving-block bootstrap (four-week blocks).

G.3 Prediction and historical risk

Table 3: Climate-stream results with historical test blocks. "Future MSFE" uses the disjoint future block (winters 2025–2026). "Hist. MSFE" is the final-state average over the eight held-out historical test winters. $\mathrm{MaxF}^{\mathrm{hist}}_T$ and $\mathrm{MReg}^{\mathrm{hist}}_T$ are the maxima over time of historical forgetting and of historical regret relative to Offline pooled Lasso, respectively.

Method	Future MSFE	R^2_{OOS}	Hist. MSFE	$\mathrm{MaxF}^{\mathrm{hist}}_T$	$\mathrm{MReg}^{\mathrm{hist}}_T$	Final support
Zero anomaly	5.256	0.000	–	–	–	–
Persistence	5.902	-0.123	–	–	–	–
Current-only Lasso	4.944	0.059	3.709	2.475	1.911	13
Offline pooled Lasso	3.914	0.255	3.551	0.891	0.000	20
Direct center	4.524	0.139	3.980	1.639	1.353	10
SP-CL	3.914	0.255	3.551	0.891	0.000	20

Table 3 shows the sparse SLP models improving on the zero-anomaly and persistence baselines on the future block. SP-CL coincides with Offline pooled Lasso to numerical tolerance, as Theorem 1 predicts: the empirical score-transmission gap is 3.6×10^{-11} for SP-CL against 0.391 for Direct center at the final task, mirroring the mechanism in Figure 1. The future block contains only 41 weekly pairs from two winters, so these comparisons are descriptive: winter-specific future MSFEs are 1.555 (2025) and 6.391 (2026) for Offline pooled Lasso/SP-CL against 1.480 and 7.720 for Direct center, and a moving-block bootstrap gives a mean paired MSFE difference (Direct center minus Offline pooled Lasso) of 0.610 on the combined future block, 95% interval $[-0.164, 1.103]$; Offline pooled

Lasso against zero-anomaly and persistence gives mean differences of -1.342 ($[-2.106, 0.039]$) and -1.988 ($[-5.046, -0.628]$), respectively.

The historical test blocks give a more direct forgetting check. A paired percentile bootstrap over the eight held-out winters gives an average final historical-test excess MSFE for Direct center relative to Offline pooled Lasso of 0.430, 95% interval $[0.059, 0.861]$; the corresponding Current-only excess is 0.158, interval $[-0.226, 0.553]$ (includes zero). Direct center selects fewer grid cells (10, against 20 for Offline pooled Lasso/SP-CL) – consistent with the accumulated-shrinkage mechanism of Proposition 1 rather than superior variable identification: the Offline pooled Lasso/SP-CL supports coincide exactly, while Direct center's support has no exact overlap with them, even though 70% of its selected cells lie within 5° and 80% within 10° of a shared-support cell, and the two fitted-value series remain highly correlated (0.952 historical, 0.967 future).

G.4 Spatial diagnostics and limitations

Figure 8 plots the future-block MSFE, forgetting and regret curves, and a batchwise covariance-distance diagnostic $\|\widehat{\Sigma}_j - \widehat{\Sigma}_k\|_F / (\|\widehat{\Sigma}_j\|_F + \|\widehat{\Sigma}_k\|_F)$; its between-task mean (0.268) and within-task split-half mean (0.295) are close, and a sample-size-matched calibration in the leading ten empirical PC directions (0.306 versus 0.313) remains close, so this diagnostic is descriptive rather than confirmatory evidence of covariate shift. Figure 9 compares final supports spatially, and Figure 10 additionally shows coefficient signs. These are exploratory summaries of correlated pressure-grid proxies, not a causal attribution, and this application illustrates score transmission on one real chronological stream; it neither verifies the common-covariance approximation of Section 4.1 nor rules out climate-regime drift more generally.

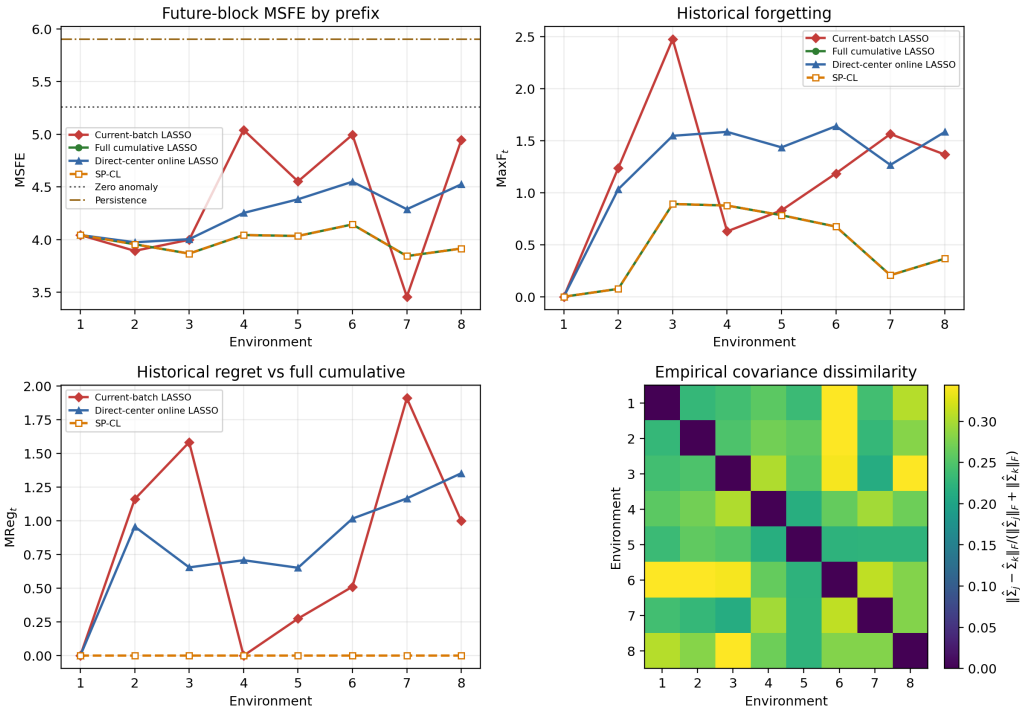

Figure 8: Climate-stream historical test-block analysis. Top left: raw future-block MSFE for each method, with zero-anomaly and persistence baselines. Top right: historical forgetting MaxF_t. Bottom left: historical regret MReg_t relative to Offline pooled Lasso. Bottom right: pairwise empirical covariance dissimilarities between chronological training tasks. The common penalty is $\lambda_t = 0.20/\sqrt{t}$.

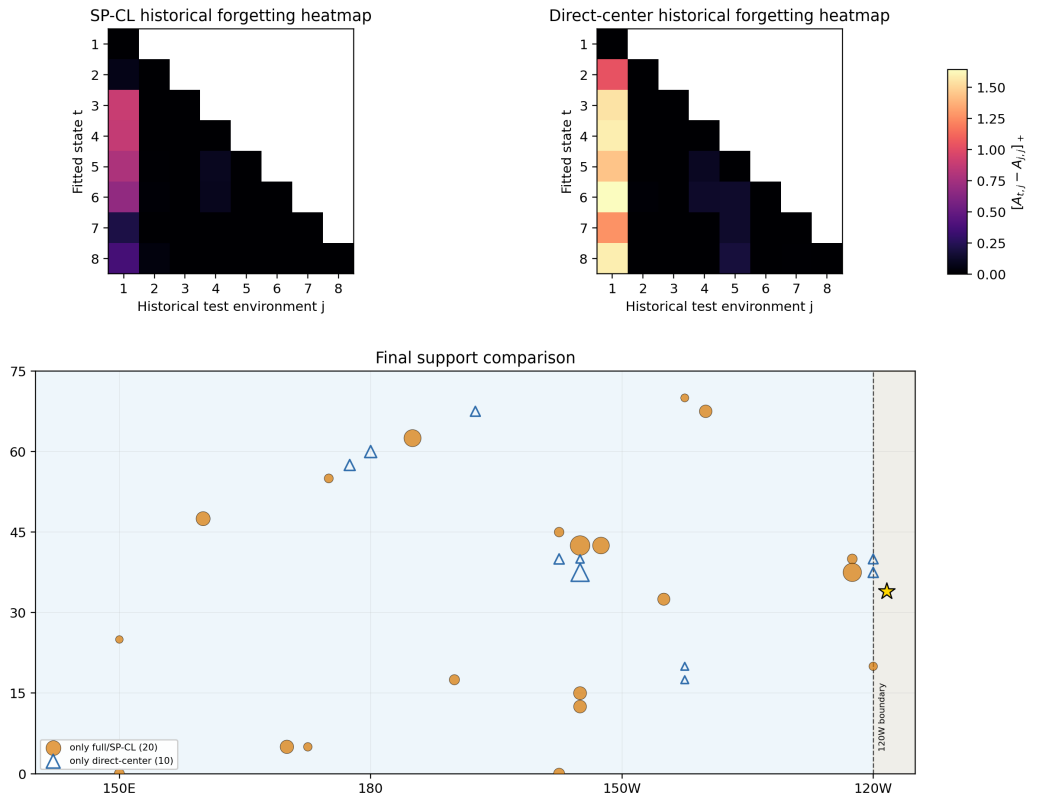

Figure 9: Historical forgetting heatmaps and final support comparison. The heatmaps show $[A_{t,j} - A_{j,j}]_+$ for SP-CL and Direct center on a common color scale. The map compares final supports: orange points are selected only by Offline pooled Lasso/SP-CL, blue triangles only by Direct center; point size uses a common coefficient-magnitude scale, the dashed line marks the eastern edge of the SLP predictor window at 120°W, and the star marks Los Angeles immediately east of that window.

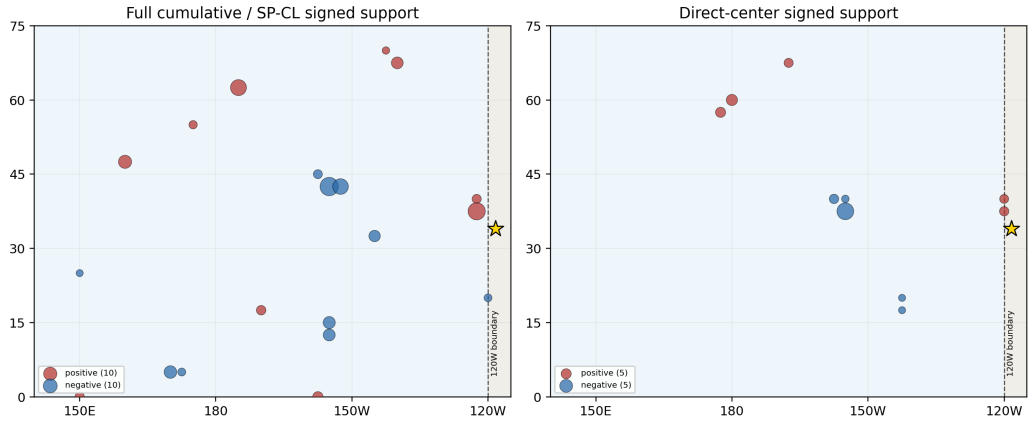

Figure 10: Signed final support maps for the climate-stream application. Red points denote positive coefficients, blue points denote negative coefficients, with point size on a common absolute-magnitude scale across panels. These maps separate coefficient sign from the method-overlap comparison in Figure 9.