Online Directional Regression for Streaming Sufficient Dimension Reduction

Dohyun Kim*
Department of Statistics and Data Science
Yonsei University, Seoul, Republic of Korea
anna01new@yonsei.ac.kr

Hyungryul Park*
Department of Statistics and Data Science
Yonsei University, Seoul, Republic of Korea
parkhryul@yonsei.ac.kr

Kyongwon Kim†
Department of Applied Statistics
Department of Statistics and Data Science
Yonsei University, Seoul, Republic of Korea
kimk@yonsei.ac.kr

Abstract

We develop an online directional regression for sufficient dimension reduction with streaming data. Unlike first-moment methods, directional regression exploits both inverse conditional means and inverse conditional variances, and can therefore recover central subspace directions generated by linear structures, symmetric dependencies, and interaction driven relationships. Extending directional regression to the online setting is nontrivial because its kernel depends on slice-wise inverse moments and on standardization by an unknown covariance matrix. We address these challenges by utilizing a stable denominator-free kernel, a ridge-stabilized recursive least-squares estimator of the standardization operator, metric-aware subspace tracking under a vanishing-ridge covariance metric, and a weighted-bootstrap ladle criterion for online structural dimension selection. Under standard regularity conditions, the online kernel estimator is root-t consistent, the estimated generalized eigenspace consistently recovers the central subspace, and the proposed dimension selector consistently estimates the structural dimension. Simulation studies and real-data analyses demonstrate that the proposed method serves as a robust online alternative when the dominant data structure is unknown, adapting to both first-moment and second-moment-driven dependencies while yielding substantial computational savings relative to repeated batch directional regression.

1 Introduction

Preliminaries. Modern data streams in biomedical monitoring, finance, sensor networks, recommender systems, and imaging often contain many predictors but must be processed one observation at a time. In such problems, directly estimating the conditional distribution of a response Y given a high-dimensional predictor $X \in \mathbb{R}^p$ is statistically fragile and computationally expensive. Sufficient dimension reduction (SDR) addresses this difficulty by seeking a low-dimensional subspace that preserves the full conditional distribution,

$$Y \perp\!\!\!\perp X \mid B^\top X,$$

where $B \in \mathbb{R}^{p\times d}$ and $d \ll p$ [11, 18]. Since B is identifiable only through its column space, the central object is the central subspace $\mathcal{S}_{Y|X}$ [9, 10, 30].

*Dohyun Kim and Hyungryul Park contributed equally to this work and are co-first authors.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Related work. Inverse-moment SDR methods provide some of the most widely used estimators of $\mathcal{S}_{Y|X}$. Sliced inverse regression (SIR) uses the inverse mean $E(X \mid Y)$ [20], while sliced average variance estimation (SAVE) uses inverse conditional variances and can recover symmetric dependencies that SIR misses [12]. Directional regression (DR) combines first and second-order inverse moments in a single kernel and typically spans a richer identification set than either first-moment methods or SAVE alone [19]. Existing online dimension reduction methods, however, are mostly based on SIR or related first-moment constructions [28, 7, 4, 6, 2]. Thus they may fail on quadratic, absolute-value, and interaction structures that are common in streaming applications.

Challenges. Taking DR online is not a simple extension of the previous method. The DR kernel is a quartic function of slice-wise first and second inverse moments, and naïvely updating all pieces can destroy the root-t accuracy of the batch estimator. The batch kernel is also defined on the standardized Z-scale, with $Z = \Sigma^{-1/2}(X - \mu)$ where $\mu = E(X)$ $\Sigma = \mathrm{cov}(X)$, and therefore depends on $\Sigma^{-1/2}$, which is unstable early in the stream. Finally, the method must track a generalized eigenspace under the time-varying covariance metric and preferably select the structural dimension online. Batch order-determination tools such as sequential tests, BIC-type criteria, bootstrap methods, and the ladle estimator [5, 33, 22, 31, 23] do not directly provide such a streaming solution.

Our contribution. We propose online directional regression (ODR), a streaming framework for DR with online subspace tracking and order determination. First, we introduce a p_h-free DR kernel $K^\star$ whose standardized range agrees with that of the population DR target and avoids unstable division by empirical slice probabilities. Second, we compute the Σ^{-1}-standardized slice blocks through a ridge-stabilized recursive least-squares (RLS) update, so the standardized-block construction avoids directly inverting $\widehat{\Sigma}_t$. Third, under the vanishing ridge metric $S_t = \widehat{\Sigma}_t + \rho_t I_p$, we provide two streaming subspace estimators: a generalized perturbation update for the exact instance-average generalized eigenspace and a generalized-gradient update for memory-efficient stochastic approximation. Fourth, we adapt the ladle criterion to the streaming data using a Poisson-multiplier weighted bootstrap evaluated on the plug-in standardized kernel matrix.

Main Theoretical Results. The main theoretical guarantees are provided in Section 3. In particular, the online kernel and averaged online kernel are root-t consistent, the exact generalized eigenspace converges at the root-t rate, the perturbation update is first-order valid with a localized $O_p(t^{-2})$ remainder, the generalized-gradient update converges almost surely under a local stability condition, and the weighted-bootstrap ladle selector is consistent for the true structural dimension. Empirically, ODR matches or improves over online SIR and online SAVE on nine simulated regression structures and two benchmark datasets while providing a substantial runtime advantage over repeatedly recomputing batch DR.

2 Online Directional Regression and Structural Dimension Selection

2.1 Online Estimation of the Directional Regression Kernel

Directional regression preliminaries. Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ with $\mu = E[X]$ and $\Sigma = \mathrm{Cov}(X) \succ 0$, and write $\delta_{t,h} = \mathbf{1}\{Y_t \in I_h\}$. In its batch form [19], DR operates on the standardized predictor $Z = \Sigma^{-1/2}(X - \mu)$ and combines inverse moments of Z given Y through the kernel

$$\Lambda_{\mathrm{DR}} = 2E\big[M(Y)^2\big] + 2\,E\big[m(Y)m(Y)^\top\big]^2 + 2\,\mathrm{tr}\big(E[m(Y)m(Y)^\top]\big)\,E[m(Y)m(Y)^\top] - 2I_p,$$

where $m(Y) = E[Z \mid Y]$ and $M(Y) = E[ZZ^\top \mid Y]$. Because the target is the central subspace $\mathcal{S}_{Y|X}$, which lives on the X-scale, we work with the conjugated kernel $K = \Sigma^{1/2}\Lambda_{\mathrm{DR}}\Sigma^{1/2}$. After slicing Y into intervals $I_h = (q_{h-1}, q_h]$ for $h = 1, \ldots, H$ with $p_h = \mathbb{P}(Y \in I_h) > 0$, and writing

$$U_h = E[X - \mu \mid Y \in I_h], \qquad V_h = E[(X-\mu)(X-\mu)^\top \mid Y \in I_h] - \Sigma,$$

the X-scale DR kernel is

$$K = 2\sum_{h=1}^{H} p_h V_h \Sigma^{-1} V_h + 2\Big(\sum_{h=1}^{H} p_h U_h U_h^\top\Big)\Sigma^{-1}\Big(\sum_{h=1}^{H} p_h U_h U_h^\top\Big) + 2\,\mathrm{tr}\Big\{\Sigma^{-1}\sum_{h=1}^{H} p_h U_h U_h^\top\Big\}\Big(\sum_{h=1}^{H} p_h U_h U_h^\top\Big).$$

Stable kernel construction. Each p_h in this expression has to be estimated from the streaming data, and repeatedly dividing by an empirical $\widehat{p}_{t,h}$ is a source of numerical instability early in the stream and for rare slices. We therefore absorb p_h into the slice blocks themselves. Define the unnormalized slice blocks

$$A_h = E[(X-\mu)\mathbf{1}\{Y \in I_h\}] = p_h U_h,$$
$$B_h = E[(X-\mu)(X-\mu)^\top \mathbf{1}\{Y \in I_h\}] - p_h\Sigma = p_h V_h,$$

and the p_h-free kernel

$$K^\star = 2\sum_{h=1}^{H} B_h\Sigma^{-1}B_h + 2\Big(\sum_{h=1}^{H} A_hA_h^\top\Big)\Sigma^{-1}\Big(\sum_{h=1}^{H} A_hA_h^\top\Big) + 2\,\mathrm{tr}\Big\{\Sigma^{-1}\sum_{h=1}^{H} A_hA_h^\top\Big\}\Big(\sum_{h=1}^{H} A_hA_h^\top\Big).$$

Although $K^\star \neq K$ in general, their standardized forms have the same range under the mild slice condition $p_h > 0$ for all h. Indeed, with

$$M_K = \Sigma^{-1/2}K\Sigma^{-1/2}, \qquad M^\star = \Sigma^{-1/2}K^\star\Sigma^{-1/2},$$

Lemma 5 in Appendix A gives

$$\mathrm{range}(M_K) = \mathrm{range}(M^\star).$$

Consequently, under the usual DR exhaustiveness condition, if $v_1, \ldots, v_d$ are the leading ordinary eigenvectors of $M^\star$, then

$$\mathcal{S}_{Y|X} = \Sigma^{-1/2}\,\mathrm{span}(v_1, \ldots, v_d).$$

Working with $K^\star$ is what makes the online procedure numerically robust, because every quantity that would otherwise be divided by $\widehat{p}_{t,h}$ is replaced by an expectation over the full stream.

Notation and online moments. From $\{(X_i, Y_i)\}_{i=1}^t$, for each slice $h = 1, \ldots, H$, write $\delta_{i,h} = \mathbf{1}\{Y_i \in I_h\}$. For a generic pair (X, Y), let $\delta_h = \mathbf{1}\{Y \in I_h\}$, and define $u_h = E(X\delta_h)$ and $S_h^{(2)} = E(XX^\top\delta_h)$. The raw online estimators are $\widehat{p}_{t,h} = \frac{1}{t}\sum_{i=1}^t \delta_{i,h}, \widehat{u}_{t,h} = \frac{1}{t}\sum_{i=1}^t X_i\delta_{i,h}, \widehat{S}_{t,h}^{(2)} = \frac{1}{t}\sum_{i=1}^t X_iX_i^\top\delta_{i,h}$, with $\widehat{\mu}_t = \frac{1}{t}\sum_{i=1}^t X_i, \widehat{\Sigma}_t = \frac{1}{t}\sum_{i=1}^t (X_i - \widehat{\mu}_t)(X_i - \widehat{\mu}_t)^\top$. The raw online moment estimators are

$$\widehat{\mu}_{t+1} = \widehat{\mu}_t + \frac{X_{t+1} - \widehat{\mu}_t}{t+1},$$

$$\widehat{\Sigma}_{t+1} = \frac{t}{t+1}\widehat{\Sigma}_t + \frac{t}{(t+1)^2}(X_{t+1} - \widehat{\mu}_t)(X_{t+1} - \widehat{\mu}_t)^\top,$$

$$\begin{pmatrix} \widehat{p}_{t+1,h} \\ \widehat{u}_{t+1,h} \\ \widehat{S}_{t+1,h}^{(2)} \end{pmatrix} = \frac{t}{t+1}\begin{pmatrix} \widehat{p}_{t,h} \\ \widehat{u}_{t,h} \\ \widehat{S}_{t,h}^{(2)} \end{pmatrix} + \frac{1}{t+1}\begin{pmatrix} \delta_{t+1,h} \\ X_{t+1}\,\delta_{t+1,h} \\ X_{t+1}X_{t+1}^\top\delta_{t+1,h} \end{pmatrix}.$$

Each update is $O(p^2)$ in time and $O(Hp^2)$ in memory, independent of t.

Recursive standardization. The p_h-free kernel $K^\star$ couples the unnormalized blocks A_h and B_h to the Σ^{-1}-standardized blocks $\widetilde{A}_h = \Sigma^{-1}A_h$ and $\widetilde{B}_h = \Sigma^{-1}B_h$. A direct plug-in $\widetilde{A}_{t,h} = \widehat{\Sigma}_t^{-1}\widehat{A}_{t,h}$ is dangerous for small t, because $\widehat{\Sigma}_t^{-1}$ is ill-conditioned before enough observations have arrived. We bypass this by computing the standardized blocks through a ridge-stabilized RLS problem with extended regressor $x_t^{\mathrm{e}} = (1, X_t^\top)^\top$ and slope selector $\widetilde{I} = (0_{p\times 1}, I_p)$. Let

$$G_t = \lambda I_{p+1} + \sum_{i=1}^{t} x_i^{\mathrm{e}}x_i^{\mathrm{e}\top}, \qquad \lambda > 0.$$

The ridge constant λ is used to stabilize the early RLS inverse. It is kept fixed as $t \to \infty$, so the normalized ridge λ/t is $O(t^{-1})$. With standardized predictors, we use $\lambda = 10^{-6}$ in numerical experiments. By the Sherman-Morrison identity,

$$G_{t+1}^{-1} = G_t^{-1} - \frac{G_t^{-1}x_{t+1}^{\mathrm{e}}x_{t+1}^{\mathrm{e}\top}G_t^{-1}}{1 + x_{t+1}^{\mathrm{e}\top}G_t^{-1}x_{t+1}^{\mathrm{e}}},$$

so G_t^{-1} is updated at $O(p^2)$ cost. With the slice-wise regression sums

$$s_{t,h} = \sum_{i=1}^{t} x_i^{\mathrm{e}} \delta_{i,h}, \qquad R_{t,h} = \sum_{i=1}^{t} x_i^{\mathrm{e}} X_i^\top \delta_{i,h},$$

which are also rank one updatable, the RLS coefficients

$$\widehat{\vartheta}_{t,h} = G_t^{-1} s_{t,h}, \qquad \widehat{\Theta}_{t,h} = G_t^{-1} R_{t,h}$$

have slope blocks

$$\widetilde{A}_{t,h} = \widetilde{I}\,\widehat{\vartheta}_{t,h}, \qquad \widehat{Q}_{t,h} = \widetilde{I}\,\widehat{\Theta}_{t,h}$$

that converge almost surely and at rate root-t to the population slopes $\Sigma^{-1}A_h$ and $\Sigma^{-1}S_h^{(2)} - (\Sigma^{-1}\mu)u_h^\top$, respectively, as established in Lemma 7. The symbols $\widehat{\vartheta}_{t,h}$ and $\widehat{\Theta}_{t,h}$ are reserved for these RLS coefficients throughout, while eigenvalues of the ladle matrix introduced later are denoted separately by $\widehat{\nu}_{t,j}$.

Online kernel estimator. The unstandardized sample blocks are the natural plug-in expressions

$$\widehat{A}_{t,h} = \widehat{u}_{t,h} - \widehat{p}_{t,h}\widehat{\mu}_t, \qquad \widehat{\mathcal{B}}_{t,h} = \widehat{S}_{t,h}^{(2)} - \widehat{\mu}_t\widehat{u}_{t,h}^\top - \widehat{u}_{t,h}\widehat{\mu}_t^\top + \widehat{p}_{t,h}\widehat{\mu}_t\widehat{\mu}_t^\top - \widehat{p}_{t,h}\widehat{\Sigma}_t,$$

where the $\widehat{\mathcal{B}}_{t,h}$ is used to avoid collision with the subspace basis $\widehat{B}_t$ introduced in Section 2.2. The standardized sample blocks are built from the RLS slopes through

$$\widetilde{\mathcal{B}}_{t,h} = \widehat{Q}_{t,h} - \widetilde{A}_{t,h}\widehat{\mu}_t^\top - \widehat{p}_{t,h} I_p,$$

which is the empirical counterpart of $\widetilde{B}_h = \Sigma^{(-1)} B_h = (\Sigma^{-1}S_h^{(2)} - (\Sigma^{-1}\mu)u_h^\top) - \widetilde{A}_h\mu^\top - p_h I_p$. In particular, $\widetilde{A}_{t,h}$ and $\widetilde{\mathcal{B}}_{t,h}$ are obtained without explicitly inverting $\widehat{\Sigma}_t$ in the standardized slice-block construction. Finally, setting

$$T_t^{(U)} = \sum_{h=1}^{H} \widehat{A}_{t,h}\widehat{A}_{t,h}^\top, \qquad \widetilde{T}_t^{(U)} = \sum_{h=1}^{H} \widetilde{A}_{t,h}\widehat{A}_{t,h}^\top, \qquad \widetilde{T}_t^{(V)} = \sum_{h=1}^{H} \widehat{\mathcal{B}}_{t,h}\widetilde{\mathcal{B}}_{t,h}, \qquad \widetilde{s}_t = \mathrm{tr}\big(\widetilde{T}_t^{(U)}\big),$$

the ODR kernel estimator is

$$\widehat{K}_t^\star = 2\,\mathrm{Sym}\big(\widetilde{T}_t^{(V)}\big) + 2\,\mathrm{Sym}\big(T_t^{(U)}\widetilde{T}_t^{(U)}\big) + 2\,\widetilde{s}_t\,T_t^{(U)},$$

where $\mathrm{Sym}(M) = (M + M^\top)/2$. The raw moment and RLS summaries are updated in $O(Hp^2)$ time per observation. If the full $p \times p$ objective matrix is explicitly formed, the second-order product terms require $O(Hp^3)$ naïvely.

2.2 Streaming Estimation of the Central Subspace

Subspace-tracking preliminaries. The online kernel $\widehat{K}_t^\star$ alone does not deliver an estimate of the central subspace. For that we need a running d-dimensional basis aligned with the leading generalized eigenvectors of $(K^\star, \Sigma)$. Two obstacles have to be addressed simultaneously. First, the metric is a moving target, since $\widehat{\Sigma}_t$ drifts across t and its inverse is unreliable at small t. Second, the streaming environment precludes a full eigendecomposition at every step, so the basis must be tracked by cheap local updates. We address the first issue through vanishing-ridge stabilization and the second through two complementary subspace-tracking updates: a generalized perturbation step, which closely tracks the instance-average generalized eigenspace, and a generalized-gradient step, which provides a memory-efficient stochastic-approximation update. Both are computed from the same online moment and RLS summaries.

Metric stabilization. Let

$$S_t = \widehat{\Sigma}_t + \rho_t I_p, \qquad \rho_t = \frac{\rho_0}{t}, \qquad \rho_0 > 0.$$

The constant ρ_0 controls early metric stabilization. More generally, ρ_0 can be selected to ensure that the initial covariance estimate S_{t_0} is sufficiently well conditioned. In our numerical experiments with standardized predictors, we set $\rho_0 = 10^{-4}$. Because $\rho_t = \rho_0/t = O(t^{-1})$, this ridge stabilizes early generalized eigenproblems without affecting the root-t limit or the limiting generalized eigenspace (Lemma 10). Adding the $O(t^{-1})$ ridge guarantees $S_t \succ 0$ for every t while preserving the root-t rate and the limiting generalized eigenspace (Lemma 10). Two companion matrices play complementary roles. The averaged online kernel

$$D_t = \frac{1}{t}\sum_{i=1}^{t} \widehat{K}_i^{\star}, \qquad W_t = S_t^{-1/2} D_t S_t^{-1/2},$$

is updated recursively through $D_{t+1} = D_t - (t+1)^{-1}(D_t - \widehat{K}_{t+1}^{\star})$. This averaging smooths the sample path and remains root-t consistent, but we do not use or need a general claim that it has smaller variance than $\widehat{K}_t^{\star}$ at the same time point. It is the natural object for the perturbation-based analysis. The current-instance standardization

$$\widetilde{W}_t = S_t^{-1/2} \widehat{K}_t^{\star} S_t^{-1/2},$$

on the other hand, is what drives the gradient-based update. For order determination in Section 2.3 we use exactly the same matrix and rename it $\mathcal{M}_t \equiv \widetilde{W}_t$ to emphasize its role. The instance-average standardized matrix

$$W_t = S_t^{-1/2} D_t S_t^{-1/2}$$

is used for the perturbation method and for the exact generalized-eigenspace analysis. The current-kernel standardized matrix

$$\widetilde{W}_t = S_t^{-1/2} \widehat{K}_t^{\star} S_t^{-1/2}$$

is used for the gradient approach. For structural dimension selection, we use the same current-kernel matrix and write

$$\mathcal{M}_t \equiv \widetilde{W}_t = S_t^{-1/2} \widehat{K}_t^{\star} S_t^{-1/2}.$$

Thus W_t and $\mathcal{M}_t$ are different finite-time matrices, whereas $\widetilde{W}_t$ and $\mathcal{M}_t$ are identical by notation.

For positive definite S and full-column-rank $B \in \mathbb{R}^{p\times d}$, define the S-metric orthonormalization operator

$$\mathcal{P}_{\mathrm{orth}}^{(S)}(B) = S^{-1/2}\mathcal{P}_{\mathrm{orth}}(S^{1/2}B), \qquad \mathcal{P}_{\mathrm{orth}}(Z) = Z(Z^\top Z)^{-1/2},$$

so that $\{\mathcal{P}_{\mathrm{orth}}^{(S)}(B)\}^\top S\, \mathcal{P}_{\mathrm{orth}}^{(S)}(B) = I_d$.

Generalized perturbation update. The first update approach is based on a first-order perturbation analysis of the generalized eigenproblem (D_t, S_t). Let $(\widehat{\lambda}_{t,j}, \widehat{\beta}_{t,j})$, $j = 1, \ldots, d$, be the leading generalized eigenpairs of (D_t, S_t) normalized by $\widehat{\beta}_{t,j}^\top S_t \widehat{\beta}_{t,j} = 1$. Equivalently, if $(\widehat{\lambda}_{t,j}, \widehat{v}_{t,j}^{(W)})$ is the jth ordinary eigenpair of W_t, then $\widehat{\beta}_{t,j} = S_t^{-1/2}\widehat{v}_{t,j}^{(W)}$. The superscript (W) distinguishes these eigenvectors from the ladle eigenvectors of $\mathcal{M}_t$ introduced in Section 2.3. Write

$$\Delta D_t = D_{t+1} - D_t, \qquad \Delta S_t = S_{t+1} - S_t,$$

and define

$$\delta_{t,j} = \widehat{\beta}_{t,j}^\top(\Delta D_t - \widehat{\lambda}_{t,j}\Delta S_t)\widehat{\beta}_{t,j}, \qquad r_{t,j} = (\Delta D_t - \widehat{\lambda}_{t,j}\Delta S_t)\widehat{\beta}_{t,j} - \delta_{t,j} S_t \widehat{\beta}_{t,j}.$$

The first-order generalized perturbation step is

$$\widehat{\lambda}_{t+1,j} = \widehat{\lambda}_{t,j} + \delta_{t,j}, \qquad \widehat{\beta}^{-}_{t+1,j} = \widehat{\beta}_{t,j} + (\widehat{\lambda}_{t,j} S_t - D_t)^{+} r_{t,j},$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse, followed by S_{t+1}-metric orthonormalization

$$\widehat{B}_{t+1} = \mathcal{P}_{\mathrm{orth}}^{(S_{t+1})}\big(\widehat{\beta}^{-}_{t+1,1}, \ldots, \widehat{\beta}^{-}_{t+1,d}\big).$$

Here $\widehat{B}_t$ denotes the basis estimate produced by this method, and Proposition 15 shows that the update coincides with the exact generalized eigenspace of (D_{t+1}, S_{t+1}) up to a localized $O_p(t^{-2})$

remainder. The generalized perturbation step can also be compared with the ordinary symmetric perturbation step for

$$W_t = S_t^{-1/2} D_t S_t^{-1/2}.$$

Because the metric S_t changes with t, the two formulations should be viewed as first-order equivalent representations of the same local generalized eigenspace, rather than as literally identical updates.

Generalized-gradient update. The second update approach takes an optimization view of the same problem. At the population level, the leading d-dimensional generalized eigenspace of $(K^\star, \Sigma)$ solves the constrained trace maximization

$$\max_{B \in \mathbb{R}^{p \times d}} \operatorname{tr}(B^\top K^\star B) \qquad \text{subject to } B^\top \Sigma B = I_d,$$

which is a generalized Brockett-type problem on the Σ-Stiefel manifold. After replacing $(K^\star, \Sigma)$ with their online counterparts $(\widehat{K}^\star_{t+1}, S_{t+1})$, we use the step-size sequence $\gamma_t = c/t$. Throughout the numerical experiments, we set $c = 0.1$. The corresponding online solution is the generalized-gradient update

$$B_{t+1} = \mathcal{P}_{\mathrm{orth}}^{(S_{t+1})}\Big(B_t + \gamma_{t+1} S_{t+1}^{-1} \widehat{K}^\star_{t+1} B_t\Big).$$

Here B_t denotes the current gradient iterate. Since the generalized-gradient update uses the kernel only through the product $\widehat{K}^\star_{t+1} B_t$, the full matrix $\widehat{K}^\star_{t+1}$ need not be explicitly materialized. This product can be computed directly from the stored slice blocks, avoiding unnecessary full $p \times p$ matrix products. In this framework, we reserve $\widehat{B}_t$ for the exact generalized eigenspace of (D_t, S_t) and use B_t for the gradient iterate. The generalized-gradient update is a distinctive strength of our framework for several reasons. First, it uses only the current-time kernel $\widehat{K}^\star_{t+1}$ rather than the accumulator D_t, so it requires no running average matrix and its per-iteration memory footprint is the smallest possible for a recursive subspace estimator of this form. Second, it avoids any solve with $(\widehat{\lambda}_{t,j} S_t - D_t)^+$ and replaces it with a single matrix-vector product against $S_{t+1}^{-1} \widehat{K}^\star_{t+1}$, which in practice gives the procedure the lowest per-step cost in our simulations (Table 12). Third, because it is a constrained stochastic approximation, its convergence analysis is a natural instance of the Kushner-Clark theory [16], and Theorem 22 shows that the iterate converges almost surely to the central subspace under a local basin-containment condition. Fourth, the almost-sure convergence theorem in Section 3 is stated for the projected generalized-gradient update, where S_{t+1}-metric orthonormalization is performed at each iteration. In large-scale implementations, less frequent orthonormalization may be used as a numerical heuristic.

Implementation. The two update methods above are computed from the same online moment and RLS summaries and target the same population generalized eigenspace, but they emphasize different computational tradeoffs. The perturbation method is preferable when the eigengap is well separated, p is moderate, and a first-order approximation close to the exact generalized eigensystem is desired after every update. Its exact dense implementation requires d pseudoinverse or constrained linear-solve operations and can cost $O(dp^3)$. The gradient approach is preferable when d and p are moderate to large and minimum per-step cost is the dominant consideration. A complete pseudocode implementation is deferred to Appendix E. The formal guarantees for both subspace estimators are stated in Section 3.

2.3 A Weighted-Bootstrap Ladle Criterion for Order Determination

Ladle setup. Selecting the structural dimension d in a streaming setting is qualitatively harder than selecting it in batch. Every monitoring time t requires a separate stability assessment, and re-running a full bootstrap from scratch would defeat the purpose of an online procedure. We address this through two ingredients. First, we evaluate the criterion on the plug-in standardized kernel matrix $\mathcal{M}_t = \widetilde{W}_t$, which is a smooth function of the time-t online moment and RLS summaries and therefore inherits root-t accuracy (Proposition 23). Second, we derive the ladle variability term [23] by a Poisson-multiplier weighted bootstrap whose weighted moment and RLS summaries are updated sequentially, so the cost per replicate is the same as the cost of the unweighted stream. Using the direct kernel rather than the instance average is a deliberate choice. It gives the order-determination statistic a clean influence-function expansion, which in turn lets us prove a consistent bootstrap representation of the ladle criterion under exactly the moment conditions already imposed for subspace estimation.

Write the eigendecomposition of $\mathcal{M}_t$ as

$$\mathcal{M}_t = \widehat{U}_t^{(\mathcal{M})} \widehat{\Xi}_t^{(\mathcal{M})} \widehat{U}_t^{(\mathcal{M})\top},$$
$$\widehat{\Xi}_t^{(\mathcal{M})} = \operatorname{diag}(\widehat{\nu}_{t,1}, \dots, \widehat{\nu}_{t,p}), \qquad \widehat{\nu}_{t,1} \geq \cdots \geq \widehat{\nu}_{t,p},$$

where $\widehat{U}_t^{(\mathcal{M})} = (\widehat{u}_{t,1}^{(\mathcal{M})}, \dots, \widehat{u}_{t,p}^{(\mathcal{M})})$ collects the orthonormal eigenvectors of $\mathcal{M}_t$. The symbols $\widehat{\nu}_{t,j}$ and $\widehat{\Xi}_t^{(\mathcal{M})}$ are used instead of $\widehat{\vartheta}_{t,j}$ and $\widehat{\Theta}_t$ to avoid collision with the RLS coefficients $\widehat{\vartheta}_{t,h}$ and $\widehat{\Theta}_{t,h}$. For $k \geq 1$, let

$$\widehat{U}_{t,k}^{(\mathcal{M})} = (\widehat{u}_{t,1}^{(\mathcal{M})}, \dots, \widehat{u}_{t,k}^{(\mathcal{M})}) \in \mathbb{R}^{p \times k}, \qquad \widehat{U}_{t,0}^{(\mathcal{M})} \text{ empty},$$

and for column-orthonormal $U, V \in \mathbb{R}^{p \times k}$ define the subspace discrepancy

$$\operatorname{dist}(U, V) = 1 - \left|\det(U^\top V)\right| \in [0, 1].$$

Let $\widehat{U}_{t,k}^{(\mathcal{M})*(b)}$ denote the leading k orthonormal eigenvectors of the bootstrap matrix $\mathcal{M}_t^{*(b)}$ constructed below. The raw variability term is

$$\widehat{f}_{t,0}(0) = 0, \qquad \widehat{f}_{t,0}(k) = \frac{1}{N_t} \sum_{b=1}^{N_t} \operatorname{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(b)}\big), \qquad k = 1, \dots, q,$$

where $q \leq p - 1$ is a fixed search cap and N_t denotes the number of bootstrap replicates used at time t, which is not to be confused with the cumulative Poisson weight $N_t^{*(b)} = \sum_{i=1}^t w_{b,i}$ of replicate b introduced below. Normalizing by

$$\widehat{f}_t(k) = \frac{\widehat{f}_{t,0}(k)}{1 + \sum_{j=0}^q \widehat{f}_{t,0}(j)}, \qquad k = 0, \dots, q,$$

and combining with the eigenvalue term

$$\widehat{\phi}_t(k) = \frac{\widehat{\nu}_{t,k+1}^+}{1 + \sum_{j=1}^{q+1} \widehat{\nu}_{t,j}^+}, \qquad \widehat{\nu}_{t,j}^+ = \max(\widehat{\nu}_{t,j}, 0), \qquad k = 0, \dots, q,$$

yields the ladle objective

$$\widehat{g}_t(k) = \widehat{f}_t(k) + \widehat{\phi}_t(k)$$

and the structural dimension estimator

$$\widehat{d}_t^{\,\mathrm{ladle}} = \arg\min_{0 \leq k \leq q} \widehat{g}_t(k).$$

If an original-scale basis is needed after order selection, we use $S_t^{-1/2} \widehat{U}_{t,\widehat{d}_t^{\,\mathrm{ladle}}}^{(\mathcal{M})}$.

Weighted-bootstrap update. To evaluate the variability term in the stream, we use a weighted bootstrap with Poisson multipliers. For each bootstrap replicate $b = 1, \dots, N$ and time t, draw $w_{b,t} \sim \mathrm{Poisson}(1)$ independently across b and t. For replicate b, define $N_t^{*(b)} = \sum_{i=1}^t w_{b,i}$. Here $N_t^{*(b)}$ is the cumulative bootstrap weight, or equivalently the effective weighted sample size, of the bth bootstrap stream up to time t. It is used as the normalizing denominator in the weighted raw moments and weighted RLS summaries for that replicate. The weighted raw moments, weighted RLS blocks, bootstrap kernel, and full streaming implementation of this ladle method are given in Appendix F. The consistency statement for $\widehat{d}_t^{\,\mathrm{ladle}}$ is included with the other main asymptotic results in Section 3 and Appendix A. The corresponding full streaming implementation is presented in Algorithm 2.

3 Large-Sample Theory for Online Directional Regression

Overview. We now state the main asymptotic theory for the online kernel, the two subspace-tracking updates, and the weighted-bootstrap ladle selector. The complete lemmas, perturbation expansions, stochastic-approximation argument, and bootstrap proof are collected in Appendix A.

Notation and assumptions. Recall that

$$M^\star = \Sigma^{-1/2} K^\star \Sigma^{-1/2}, \qquad S_t = \widehat{\Sigma}_t + \rho_t I_p, \qquad \rho_t = \rho_0/t,$$

and define the instance-average and current-kernel standardized matrices

$$D_t = \frac{1}{t}\sum_{i=1}^{t} \widehat{K}_i^\star, \qquad W_t = S_t^{-1/2} D_t S_t^{-1/2}, \qquad \mathcal{M}_t = S_t^{-1/2} \widehat{K}_t^\star S_t^{-1/2}.$$

Let the positive eigenvalues of $M^\star$ satisfy

$$\lambda_1 > \cdots > \lambda_d > \lambda_{d+1} = \cdots = \lambda_p = 0,$$

and let $V_\star = (v_1, \ldots, v_d)$ be the associated orthonormal eigenvectors. The target DR subspace is

$$\mathcal{S}_{Y|X} = \operatorname{span}\{\Sigma^{-1/2} v_1, \ldots, \Sigma^{-1/2} v_d\}.$$

Here, p and H are fixed, the observations are i.i.d., $E\|X\|^{64} < \infty$, $\Sigma \succ 0$, all slices satisfy $p_h > 0$, the usual DR exhaustiveness conditions hold, and the signal eigenvalues above are distinct. In the generalized-gradient update, we use the step-size sequence $\gamma_t = c/t$ with fixed $c > 0$. These assumptions correspond to (A0)-(A6) in Appendix A. We refer the reader to that appendix for the detailed technical arguments.

Kernel and exact eigenspace.

Theorem 1 (Root-t kernel and exact eigenspace guarantees)**.** *Under the assumptions above,*

$$\|\widehat{K}_t^\star - K^\star\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad \widehat{K}_t^\star \to K^\star \quad \textit{almost surely},$$

and the instance average satisfies

$$\|D_t - K^\star\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad D_t \to K^\star \quad \textit{almost surely}.$$

Moreover, if $\widehat{V}_t^{(W)}$ contains the leading d eigenvectors of W_t and

$$\widehat{B}_t^{\mathrm{ex}} = S_t^{-1/2} \widehat{V}_t^{(W)}, \qquad B_\star = \Sigma^{-1/2} V_\star,$$

then, with $P(B) = B(B^\top B)^{-1} B^\top$,

$$\|P(\widehat{B}_t^{\mathrm{ex}}) - P(B_\star)\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad \operatorname{span}(\widehat{B}_t^{\mathrm{ex}}) \to \mathcal{S}_{Y|X} \quad \textit{almost surely}.$$

Equivalently, the projection matrix distance between the exact online generalized eigenspace and the central subspace is $O_p(t^{-1/2})$. This theorem follows by combining Theorems 9 and 13. See Appendix A for the detailed arguments.

Perturbation update.

Proposition 2 (First-order validity of the generalized perturbation step)**.** *Let $(\widehat{\lambda}_{t,j}, \widehat{\beta}_{t,j})$ be a simple signal generalized eigenpair of (D_t, S_t) with $\widehat{\beta}_{t,j}^\top S_t \widehat{\beta}_{t,j} = 1$, and define ΔD_t, ΔS_t, $\delta_{t,j}$, and $r_{t,j}$ as in Section 2.2. If $(\widehat{\lambda}_{t+1,j}^{\mathrm{ex}}, \widehat{\beta}_{t+1,j}^{\mathrm{ex}})$ denotes the corresponding exact local generalized eigenpair of (D_{t+1}, S_{t+1}), then*

$$\begin{aligned}
\widehat{\lambda}_{t+1,j}^{\mathrm{ex}} &= \widehat{\lambda}_{t,j} + \delta_{t,j} + O_p(t^{-2}), \\
\widehat{\beta}_{t+1,j}^{\mathrm{ex}} &= \widehat{\beta}_{t,j} + (\widehat{\lambda}_{t,j} S_t - D_t)^+ r_{t,j} + O_p(t^{-2}),
\end{aligned}$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse, and the expansion is taken under the same local normalization used in Appendix A. Thus the generalized perturbation update differs from recomputing the exact generalized eigenpair only by a localized quadratic one-step remainder. The detailed argument is in Proposition 15 in Appendix A.

Gradient update.

Theorem 3 (Almost sure convergence of the generalized-gradient update)**.** *Consider the projected generalized-gradient update with S_{t+1}-metric orthonormalization at every iteration,*

$$B_{t+1} = \mathcal{P}_{\mathrm{orth}}^{(S_{t+1})}\Big(B_t + \gamma_{t+1} S_{t+1}^{-1} \widehat{K}_{t+1}^{\star} B_t\Big), \qquad \gamma_t = c/t.$$

After the change of variables $V_t = S_t^{1/2} B_t$, the generalized-gradient update can be viewed as a stochastic approximation to the population constrained optimization problem that maximizes $\mathrm{tr}(B^\top K^\star B)$ over the Σ-Stiefel manifold. If the transformed iterates eventually remain in a compact subset of the basin of attraction of the leading eigenspace of $M^\star$, then

$$\mathrm{span}(B_t) \to \mathcal{S}_{Y|X} \qquad \textit{almost surely}.$$

The proof accounts for both the moving-metric error and the stochastic-approximation error in the every-step projected update. See Theorems 20 and 22 in Appendix A.

Dimension selection.

Theorem 4 (Consistency of the online weighted-bootstrap ladle selector)**.** *Fix a search grid $q \in \{d, \ldots, p-1\}$, and assume $N_t \to \infty$ bootstrap replicates, each built from the full weighted history up to time t. Under the assumptions above, and under the nondegeneracy condition for null-space bootstrap variability stated in Appendix A, the ladle objective*

$$\widehat{g}_t(k) = \widehat{f}_t(k) + \widehat{\phi}_t(k), \qquad 0 \le k \le q,$$

selects the true structural dimension with probability tending to one. That is,

$$\mathbb{P}\{\widehat{d}_t^{\,\mathrm{ladle}} = d\} \longrightarrow 1.$$

The criterion follows the same logic as the batch ladle method, with the bootstrap component adapted to the streaming setting. The eigenvalue term prevents selecting $k < d$, whereas the bootstrap subspace-variability term prevents selecting $k > d$. The full justification is given by the influence-function expansion and multiplier-bootstrap argument in Theorem 25.

4 Numerical Results

Simulation setup. We include a compact finite-sample summary here to connect the theory with the empirical behavior. The full model definitions, complete simulation settings, results tables, runtime study, and ladle study are deferred to Appendix B, with real-data experiments and results in Appendix C. The simulations use nine regression structures, $p \in \{20, 50, 100\}$, $n \in \{1000, 5000, 10000\}$, and 20 replications. The simulation experiments used $H = 5$ and initialized each online procedure with the first 40% of the stream. For the online ladle procedure, the search cap was chosen as described in Appendix B.

Table 1: Representative simulation results. OSIR, OSAVE, and ODR report subspace distances, with smaller values indicating better performance. Runtime notes compare the proposed generalized-gradient ODR estimator (GD) with repeatedly recomputed batch DR for each model at $p = 100, n = 10000$.

Setting	Dependence type	OSIR	OSAVE	ODR	Runtime note
Model 1, $p = 100, n = 10000$	first moment	0.01	0.04	0.01	GD: 64.61s vs batch DR: 781.29s
Model 5, $p = 100, n = 10000$	symmetric	0.99	0.04	0.02	GD: 62.01s vs batch DR: 828.39s
Model 9, $p = 100, n = 10000$	symmetric + correlated	0.99	0.19	0.04	GD: 68.21s vs batch DR: 841.43s

Simulation findings. The first row of Table 1 shows that ODR remains competitive with OSIR in an inverse mean-driven setting. The next two rows show the main advantage of DR. First-moment online SDR fails on symmetric structures, while ODR continues to recover the central subspace under both independent and correlated predictor settings. The online ladle method exhibits the same large-sample pattern reported in Table 13 of Appendix B. On representative Models 1, 5, and 7, it selects the true structural dimension with probability one for every tested dimension. The experiments show that the proposed method remains competitive with batch DR while improving runtime.

5 Discussion

Conclusion and future work. We propose a streaming framework for DR that makes the identifying power of its batch counterpart available online, providing a robust alternative when the dominant data structure is unknown because it adapts to first-moment, symmetric, and some interaction-driven structures. The perturbation and gradient updates offer complementary computational profiles, closely tracking the exact generalized eigenspace or operating in a memory-efficient stochastic-approximation mode for long streams. Theory establishes root-t consistency of the online kernel and generalized eigenspace together with consistent structural dimension selection, and the numerical results confirm these guarantees across first-moment, symmetric, and computationally demanding settings. Future work includes developing growing-p theory, allowing dependent or nonstationary streams, designing adaptive choices of slicing, ridge, and step-size parameters, and studying global stability properties of the recursive subspace updates. Another promising direction is to combine the proposed ODR statistics with sparse or regularized subspace estimation [17], which would further improve interpretability in high-dimensional streaming problems.

Acknowledgments and Disclosure of Funding

We sincerely thank the anonymous reviewers and area chairs for their careful reading and constructive feedback, which greatly helped us improve this work Kyongwon Kim is supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. RS-2026-25469371).

Competing interests

The authors declare that they have no competing financial or non-financial interests relevant to this work.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-PCA: A global, gap-free, and near-optimal rate. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 487–492, 2017.

[2] Andreas Artemiou, Yuexiao Dong, and Seung Jun Shin. Real-time sufficient dimension reduction through principal least squares support vector machines. *Pattern Recognition*, 112:107768, 2021.

[3] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[4] Bernard Bercu, Thi Mong Ngoc Nguyen, and Jérôme Saracco. On the asymptotic behaviour of the recursive nadaraya–watson estimator associated with the recursive sliced inverse regression method. *Statistics*, 49:660 – 679, 2015.

[5] Efstathia Bura and R. Dennis Cook. Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association*, 96(455):996–1003, 2001.

[6] Zhanrui Cai, Runze Li, and Liping Zhu. Online sufficient dimension reduction through sliced inverse regression. *Journal of Machine Learning Research*, 21(10):1–25, 2020.

[7] Marie Chavent, Stéphane Girard, Vanessa Kuentz-Simonet, Benoit Liquet, Thi Mong Ngoc Nguyen, and Jérôme Saracco. A sliced inverse regression approach for data stream. *Computational Statistics*, 29:1129–1152, 2014.

[8] Xin Chen, Changliang Zou, and R. Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723, 2010.

[9] R. Dennis Cook. On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189, 1994.

[10] R. Dennis Cook. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992, 1996.

[11] R. Dennis Cook. *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York, 1998.

[12] R. Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[13] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[14] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer, 1966.

[15] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

[16] Harold Joseph Kushner and Dean S Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Springer Science & Business Media, 2012.

[17] Gayun Kwon, Gijeong Noh, and Kyongwon Kim. Sparse sufficient dimension reduction for directional regression. *Journal of Big Data*, 12(1):163, 2025.

[18] Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman and Hall/CRC, Boca Raton, FL, 2018.

[19] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[20] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[21] Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

[22] Benoît Liquet and Jérôme Saracco. Application of the bootstrap approach to the choice of dimension and the α parameter in the SIRα method. *Communications in Statistics—Simulation and Computation*, 37(6):1198–1218, 2008.

[23] Wei Luo and Bing Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887, 2016.

[24] Erkki Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992.

[25] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106 (1):69–84, 1985.

[26] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 2000.

[27] Aad W. van der Vaart and Jon A. Wellner. Weak convergence and empirical processes. In *Springer Series in Statistics*. Springer, New York, 1996.

[28] Manfred K Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(10):2287–2320, 2008.

[29] Jianjun Xu, Wenquan Cui, and Haoyang Cheng. Online sparse sliced inverse regression for high-dimensional streaming data. *International Journal of Wavelets, Multiresolution and Information Processing*, 21(2):2250055, 2023.

[30] Xiangrong Yin, Bing Li, and R Dennis Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8): 1733–1757, 2008.

[31] Jae Keun Yoo. Unified predictor hypothesis tests in sufficient dimension reduction: A bootstrap approach. *Journal of the Korean Statistical Society*, 40(2):217–225, 2011.

[32] Xue-Qiang Zeng and Guo-Zheng Li. Incremental partial least squares analysis of big streaming data. *Pattern Recognition*, 47(11):3726–3735, 2014.

[33] Liping Zhu, Tao Wang, Lixing Zhu, and Louis Ferré. Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 97(2):295–304, 2010.

Table 2: Summary of notation used in the online directional regression framework.

Group	Symbol	Role
Population	K, $K^\star$	DR kernel and denominator-free target kernel
	A_h, B_h	Slice-wise first- and second-moment blocks
	$M^\star = \Sigma^{-1/2} K^\star \Sigma^{-1/2}$	Standardized population target
Online estimators	$\widehat{\mu}_t$, $\widehat{\Sigma}_t$, $\widehat{p}_{t,h}$, $\widehat{u}_{t,h}$, $\widehat{S}^{(2)}_{t,h}$	Running moment summaries
	$\widehat{\vartheta}_{t,h}$, $\widehat{\Theta}_{t,h}$	RLS coefficients
	$\widehat{A}_{t,h}$, $\widehat{\mathcal{B}}_{t,h}$, $\widetilde{A}_{t,h}$, $\widetilde{\mathcal{B}}_{t,h}$	Online slice blocks
	$\widehat{K}^\star_t$, S_t	Current kernel and stabilized metric
Tracking	$D_t = t^{-1}\sum_{i=1}^t \widehat{K}^\star_i$	Averaged online kernel
	$W_t = S_t^{-1/2} D_t S_t^{-1/2}$	Matrix for exact and perturbation tracking
	$\widetilde{W}_t = S_t^{-1/2} \widehat{K}^\star_t S_t^{-1/2}$	Matrix for the generalized-gradient update
Ladle	$\mathcal{M}_t \equiv \widetilde{W}_t$	Matrix used for order determination
	$\widehat{\nu}_{t,j}$, $\widehat{U}^{(\mathcal{M})}_{t,k}$, $\widehat{d}^{\,\mathrm{ladle}}_t$	Eigensystem and selected dimension
	$\mathcal{M}^{*(b)}_t$, N_{boot}	Bootstrap matrices and replicate count

A Detailed Asymptotic Theory and Proofs

Recall that the p_h-free directional-regression kernel is

$$K^\star = 2\sum_{h=1}^H B_h \Sigma^{-1} B_h + 2\Big(\sum_{h=1}^H A_h A_h^\top\Big)\Sigma^{-1}\Big(\sum_{h=1}^H A_h A_h^\top\Big) + 2\,\mathrm{tr}\Big\{\Sigma^{-1}\sum_{h=1}^H A_h A_h^\top\Big\}\Big(\sum_{h=1}^H A_h A_h^\top\Big),$$

where

$$A_h = E[(X-\mu)\mathbf{1}\{Y \in I_h\}], \qquad B_h = E\big[(X-\mu)(X-\mu)^\top \mathbf{1}\{Y \in I_h\}\big] - p_h\Sigma.$$

Let

$$M^\star = \Sigma^{-1/2} K^\star \Sigma^{-1/2}.$$

Lemma 5. *Let*

$$M_K = \Sigma^{-1/2} K \Sigma^{-1/2}, \qquad M^\star = \Sigma^{-1/2} K^\star \Sigma^{-1/2}.$$

If $p_h > 0$ for every $h = 1, \dots, H$, then

$$\mathrm{range}(M_K) = \mathrm{range}(M^\star).$$

Proof. Write

$$a_h = \Sigma^{-1/2} U_h, \qquad C_h = \Sigma^{-1/2} V_h \Sigma^{-1/2}.$$

On the standardized scale,

$$M_K = 2\sum_{h=1}^H p_h C_h^2 + 2T_1^2 + 2\,\mathrm{tr}(T_1)T_1, \qquad T_1 = \sum_{h=1}^H p_h a_h a_h^\top,$$
$$M^\star = 2\sum_{h=1}^H p_h^2 C_h^2 + 2T_2^2 + 2\,\mathrm{tr}(T_2)T_2, \qquad T_2 = \sum_{h=1}^H p_h^2 a_h a_h^\top.$$

For any symmetric matrix C, $\mathrm{range}(C^2) = \mathrm{range}(C)$. Also, for positive semidefinite matrices R_ℓ and positive constants c_ℓ,

$$\mathrm{range}\Big(\sum_\ell c_\ell R_\ell\Big) = \sum_\ell \mathrm{range}(R_\ell),$$

because the null space of the sum is the intersection of the null spaces. Since p_h and p_h^2 are both positive, replacing p_h by p_h^2 changes weights but not the span of the slice mean blocks or the slice variance blocks. Hence both M_K and $M^\star$ have range

$$\operatorname{span}\Big(\{a_h : h = 1, \dots, H\} \cup \{\operatorname{range}(C_h) : h = 1, \dots, H\}\Big),$$

which proves the claim. □

By Lemma 5, the p_h-free kernel has the same standardized range as the usual sliced DR kernel. The generalized eigenspace of $(K^\star, \Sigma)$ is equivalently the ordinary eigenspace of $M^\star$. Write the positive eigenvalues of $M^\star$ as

$$\lambda_1 > \cdots > \lambda_d > \lambda_{d+1} = \cdots = \lambda_p = 0, \tag{1}$$

and let $v_1, \dots, v_d$ be the associated orthonormal eigenvectors. The target directional-regression subspace is therefore

$$\mathcal{S}_{Y|X} = \operatorname{span}\big(\Sigma^{-1/2} v_1, \dots, \Sigma^{-1/2} v_d\big).$$

At the sample level we analyze the stabilized metric

$$S_t = \widehat{\Sigma}_t + \rho_t I_p, \qquad \rho_t = \frac{\rho_0}{t}, \qquad \rho_0 > 0 \text{ fixed}, \tag{2}$$

and the averaged online kernel

$$D_t = \widehat{\Gamma}_t^{\mathrm{DR}\star} = \frac{1}{t} \sum_{i=1}^{t} \widehat{K}_i^\star.$$

The vanishing ridge in (2) guarantees $S_t \succ 0$ for every t and is of order $O(t^{-1})$, so it does not change the limiting generalized eigenspace or any root-t rate. Define the matrices

$$\begin{aligned} W_t &= S_t^{-1/2} D_t S_t^{-1/2}, \\ \widetilde{W}_t &= S_t^{-1/2} \widehat{K}_t^\star S_t^{-1/2}, \\ \mathcal{M}_t &\equiv \widetilde{W}_t. \end{aligned}$$

The matrix W_t is the standardized instance average used in the exact generalized-eigenspace analysis. By contrast, $\widetilde{W}_t$ denotes the standardized current kernel used in the generalized-gradient update, while the same matrix is denoted by $\mathcal{M}_t$ when used for order determination.

A.1 Assumptions

Scope of the theory. All guarantees in this appendix, and the main-text statements in Section 3 that summarize them, hold for fixed p and H, i.i.d. streams, slice boundaries fixed in advance or frozen after a pilot sample, conditions (A0)-(A6) below, and DR exhaustiveness. Theorem 22 (Theorem 3 in the main text) is a local stochastic-approximation result, conditional on the transformed iterates remaining in a compact subset of the attraction basin, and no global convergence of the generalized-gradient update is established. Dependent or nonstationary streams, concept drift, adaptive slicing, and growing dimension lie outside what is proved here.

Throughout this section we impose the following conditions.

- (A0) The dimensions p and H are fixed and do not depend on t.
- (A1) $\Sigma \succ 0$.
- (A2) The observations $\{(X_t, Y_t),\ t \geq 1\}$ are independent and identically distributed, and $E\|X\|^{64} < \infty$.
- (A3) The slices $I_1, \dots, I_H$ are fixed in advance, or are chosen from an independent pilot sample and then held fixed; in the pilot-frozen case all statements are conditional on the pilot sample. They satisfy $p_h = \mathbb{P}(Y \in I_h) > 0$ for every $h = 1, \dots, H$.

(A4) The usual directional-regression exhaustiveness conditions hold, so that the nonzero generalized eigenspace of $(K^\star, \Sigma)$ equals the central subspace $\mathcal{S}_{Y|X}$ (see [19]).

(A5) The positive eigenvalues of $M^\star$ are distinct, namely (1) holds.

(A6) For the generalized-gradient update, the step-size sequence is $\gamma_t = c/t$ for some fixed constant $c > 0$.

To avoid overloading the superscript e, we use $x_i^{\mathrm{e}} = (1, X_i^\top)^\top$ for the extended regressor in the RLS step, and we use $\widetilde{A}_h = \Sigma^{-1} A_h$ and $\widetilde{B}_h = \Sigma^{-1} B_h$ for the Σ^{-1}-standardized slice blocks.

The strengthened moment condition in (A2) is used to control exceptional inverse events in the RLS step and to justify the uniform fourth-moment and one-step bounds needed by the multilinear kernel expansion. Under (A2), every entry of $XX^\top$ and $x^{\mathrm{e}}x^{\mathrm{e}\top}$ has finite thirty-second moment, which yields $O(t^{-16})$ probabilities for the bad inverse events by Markov's inequality.

A.2 Root-t consistency of the online blocks and kernels

For each slice $h = 1, \ldots, H$, write

$$\delta_{i,h} = \mathbf{1}\{Y_i \in I_h\}, \qquad u_h = E(X\delta_{1,h}), \qquad S_h^{(2)} = E(XX^\top \delta_{1,h}).$$

The raw online estimators are

$$\widehat{p}_{t,h} = \frac{1}{t}\sum_{i=1}^{t} \delta_{i,h}, \qquad \widehat{u}_{t,h} = \frac{1}{t}\sum_{i=1}^{t} X_i \delta_{i,h}, \qquad \widehat{S}_{t,h}^{(2)} = \frac{1}{t}\sum_{i=1}^{t} X_i X_i^\top \delta_{i,h},$$

with

$$\widehat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t} X_i, \qquad \widehat{\Sigma}_t = \frac{1}{t}\sum_{i=1}^{t} (X_i - \widehat{\mu}_t)(X_i - \widehat{\mu}_t)^\top.$$

The unstandardized slice blocks are

$$\begin{aligned} \widehat{A}_{t,h} &= \widehat{u}_{t,h} - \widehat{p}_{t,h}\widehat{\mu}_t, \\ \widehat{\mathcal{B}}_{t,h} &= \widehat{S}_{t,h}^{(2)} - \widehat{\mu}_t \widehat{u}_{t,h}^\top - \widehat{u}_{t,h}\widehat{\mu}_t^\top + \widehat{p}_{t,h}\widehat{\mu}_t\widehat{\mu}_t^\top - \widehat{p}_{t,h}\widehat{\Sigma}_t. \end{aligned} \tag{3}$$

The $\widehat{\mathcal{B}}_{t,h}$ is used to avoid confusion with the subspace bases B_t and $\widehat{B}_t$ that appear later in the generalized-eigenspace analysis.

For the standardized blocks, let $\widetilde{I} = (0_{p\times 1}, I_p)$ and define the ridge-stabilized normal matrix

$$G_t = \lambda I_{p+1} + \sum_{i=1}^{t} x_i^{\mathrm{e}} x_i^{\mathrm{e}\top}, \qquad \lambda > 0 \quad \text{fixed},$$

together with its average-scale version

$$\bar{G}_t = t^{-1} G_t, \qquad G = E\big(x_1^{\mathrm{e}} x_1^{\mathrm{e}\top}\big).$$

For each slice h, set

$$s_{t,h} = \sum_{i=1}^{t} x_i^{\mathrm{e}} \delta_{i,h}, \qquad R_{t,h} = \sum_{i=1}^{t} x_i^{\mathrm{e}} X_i^\top \delta_{i,h}, \qquad \bar{s}_{t,h} = t^{-1} s_{t,h}, \qquad \bar{R}_{t,h} = t^{-1} R_{t,h},$$

and define the RLS coefficients

$$\widehat{\vartheta}_{t,h} = G_t^{-1} s_{t,h} = \bar{G}_t^{-1} \bar{s}_{t,h}, \qquad \widehat{\Theta}_{t,h} = G_t^{-1} R_{t,h} = \bar{G}_t^{-1} \bar{R}_{t,h}.$$

Here $\widehat{\vartheta}_{t,h}$ and $\widehat{\Theta}_{t,h}$ are reserved for the RLS coefficients. The eigenvalues of $\mathcal{M}_t$ are denoted later by $\widehat{\nu}_{t,j}$ to avoid overloading this notation. The sample standardized blocks are

$$\widetilde{A}_{t,h} = \widetilde{I}\,\widehat{\vartheta}_{t,h}, \qquad \widehat{Q}_{t,h} = \widetilde{I}\,\widehat{\Theta}_{t,h}, \qquad \widetilde{\mathcal{B}}_{t,h} = \widehat{Q}_{t,h} - \widetilde{A}_{t,h}\widehat{\mu}_t^\top - \widehat{p}_{t,h} I_p. \tag{4}$$

To define the sample kernel, let

$$T_t^{(U)} = \sum_{h=1}^{H} \widehat{A}_{t,h}\widehat{A}_{t,h}^\top, \qquad T^{(U)} = \sum_{h=1}^{H} A_h A_h^\top, \tag{5}$$

$$\widetilde{T}_t^{(U)} = \sum_{h=1}^{H} \widetilde{A}_{t,h}\widehat{A}_{t,h}^\top, \qquad \widetilde{T}^{(U)} = \sum_{h=1}^{H} \widetilde{A}_h A_h^\top, \tag{6}$$

$$\widetilde{T}_t^{(V)} = \sum_{h=1}^{H} \widehat{\mathcal{B}}_{t,h}\widetilde{\mathcal{B}}_{t,h}, \qquad \widetilde{T}^{(V)} = \sum_{h=1}^{H} B_h \widetilde{B}_h,$$

$$\widetilde{s}_t = \operatorname{tr}\big(\widetilde{T}_t^{(U)}\big), \qquad \widetilde{s} = \operatorname{tr}\big(\widetilde{T}^{(U)}\big).$$

Writing $\operatorname{Sym}(M) = (M + M^\top)/2$, define

$$\widehat{K}_t^\star = 2\,\operatorname{Sym}\big(\widetilde{T}_t^{(V)}\big) + 2\,\operatorname{Sym}\big(T_t^{(U)}\widetilde{T}_t^{(U)}\big) + 2\widetilde{s}_t\, T_t^{(U)}.$$

Lemma 6. *Assume (A1)-(A3). Then, for each fixed $h = 1, \dots, H$, there exists a constant $C > 0$ such that*

$$E\big|\widehat{p}_{t,h} - p_h\big|^4 + E\big\|\widehat{u}_{t,h} - u_h\big\|^4 + E\big\|\widehat{S}_{t,h}^{(2)} - S_h^{(2)}\big\|_{\mathrm{F}}^4 + E\big\|\widehat{\mu}_t - \mu\big\|^4 + E\big\|\widehat{\Sigma}_t - \Sigma\big\|_{\mathrm{F}}^4 \le Ct^{-2}, \tag{7}$$

$$E\big\|\bar{G}_t - G\big\|_{\mathrm{F}}^8 + E\big\|\widehat{\Sigma}_t - \Sigma\big\|_{\mathrm{F}}^8 \le Ct^{-4}, \tag{8}$$

$$E\big|\widehat{p}_{t+1,h} - \widehat{p}_{t,h}\big|^2 + E\big\|\widehat{u}_{t+1,h} - \widehat{u}_{t,h}\big\|^2 + E\big\|\widehat{S}_{t+1,h}^{(2)} - \widehat{S}_{t,h}^{(2)}\big\|_{\mathrm{F}}^2 + E\big\|\widehat{\mu}_{t+1} - \widehat{\mu}_t\big\|^2$$
$$+ E\big\|\widehat{\Sigma}_{t+1} - \widehat{\Sigma}_t\big\|_{\mathrm{F}}^2 \le Ct^{-2}. \tag{9}$$

Moreover, all raw estimators converge almost surely to their population counterparts.

Proof. Under (A2), every entry of X, $XX^\top\delta_{1,h}$, $(X-\mu)(X-\mu)^\top$, and $x_1^{\mathrm{e}}x_1^{\mathrm{e}\top}$ has finite eighth moment. Standard moment inequalities for sample averages of i.i.d. random vectors therefore give the fourth-moment bounds for $\widehat{p}_{t,h} - p_h$, $\widehat{u}_{t,h} - u_h$, $\widehat{S}_{t,h}^{(2)} - S_h^{(2)}$, and $\widehat{\mu}_t - \mu$. Since

$$\widehat{\Sigma}_t = \frac{1}{t}\sum_{i=1}^{t} X_i X_i^\top - \widehat{\mu}_t\widehat{\mu}_t^\top, \qquad \bar{G}_t = \frac{\lambda}{t} I_{p+1} + \frac{1}{t}\sum_{i=1}^{t} x_i^{\mathrm{e}} x_i^{\mathrm{e}\top},$$

the same argument yields (7) and (8). The almost sure convergence is immediate from the strong law of large numbers.

Each raw estimator satisfies an update of the form

$$\widehat{\theta}_{t+1} - \widehat{\theta}_t = \frac{1}{t+1}\big\{\psi_{t+1} - \widehat{\theta}_t\big\},$$

where ψ_{t+1} has finite second moment. Hence $E\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|^2 = O(t^{-2})$, which proves (9). □

Lemma 7. *Assume (A1)-(A3). For each fixed $h = 1, \dots, H$, let*

$$g_h = E(x_1^{\mathrm{e}}\delta_{1,h}), \qquad M_h = E\big(x_1^{\mathrm{e}}X_1^\top\delta_{1,h}\big),$$

and define the population least-squares coefficients $\vartheta_h \in \mathbb{R}^{p+1}$ and $\Theta_h \in \mathbb{R}^{(p+1)\times p}$ by

$$G\vartheta_h = g_h, \qquad G\Theta_h = M_h.$$

Then the slope block of ϑ_h equals

$$\widetilde{A}_h = \Sigma^{-1}A_h,$$

and the slope block of Θ_h equals

$$Q_h = \Sigma^{-1}S_h^{(2)} - (\Sigma^{-1}\mu)u_h^\top.$$

Moreover, there exists a constant $C > 0$ such that

$$E\big\|\widehat{\vartheta}_{t,h} - \vartheta_h\big\|^2 + E\big\|\widehat{\Theta}_{t,h} - \Theta_h\big\|_{\mathrm{F}}^2 \le Ct^{-1}, \tag{10}$$

$$\sup_{t\ge1} E\big\|\widehat{\vartheta}_{t,h}\big\|^4 + \sup_{t\ge1} E\big\|\widehat{\Theta}_{t,h}\big\|_{\mathrm{F}}^4 \le C, \tag{11}$$

$$E\big\|\widehat{\vartheta}_{t+1,h} - \widehat{\vartheta}_{t,h}\big\|^2 + E\big\|\widehat{\Theta}_{t+1,h} - \widehat{\Theta}_{t,h}\big\|_{\mathrm{F}}^2 \le Ct^{-2}, \tag{12}$$

and

$$\widehat{\vartheta}_{t,h} \to \vartheta_h, \qquad \widehat{\Theta}_{t,h} \to \Theta_h \quad \textit{almost surely.} \tag{13}$$

Consequently,

$$E\big\|\widetilde{A}_{t,h} - \widetilde{A}_h\big\|^2 + E\big\|\widehat{Q}_{t,h} - Q_h\big\|_{\mathrm{F}}^2 + E\big\|\widetilde{\mathcal{B}}_{t,h} - \widetilde{B}_h\big\|_{\mathrm{F}}^2 \le Ct^{-1}, \tag{14}$$

$$\sup_{t\ge1} E\big\|\widetilde{A}_{t,h}\big\|^4 + \sup_{t\ge1} E\big\|\widehat{Q}_{t,h}\big\|_{\mathrm{F}}^4 + \sup_{t\ge1} E\big\|\widetilde{\mathcal{B}}_{t,h}\big\|_{\mathrm{F}}^4 \le C, \tag{15}$$

$$E\big\|\widetilde{A}_{t+1,h} - \widetilde{A}_{t,h}\big\|^2 + E\big\|\widehat{Q}_{t+1,h} - \widehat{Q}_{t,h}\big\|_{\mathrm{F}}^2 + E\big\|\widetilde{\mathcal{B}}_{t+1,h} - \widetilde{\mathcal{B}}_{t,h}\big\|_{\mathrm{F}}^2 \le Ct^{-2}. \tag{16}$$

Moreover, $\widetilde{A}_{t,h} \to \widetilde{A}_h$ and $\widetilde{\mathcal{B}}_{t,h} \to \widetilde{B}_h$ almost surely.

Proof. A standard regression-with-intercept calculation shows that the slope block of ϑ_h equals $\Sigma^{-1}\operatorname{Cov}(X, \delta_{1,h}) = \Sigma^{-1}(u_h - p_h\mu) = \widetilde{A}_h$, and the slope block of Θ_h equals $\Sigma^{-1}\operatorname{Cov}(X, X\delta_{1,h}) = \Sigma^{-1}S_h^{(2)} - (\Sigma^{-1}\mu)u_h^\top = Q_h$.

For the scalar-response regression, define the centered residuals

$$\zeta_{i,h} = x_i^{\mathrm{e}}\big(\delta_{i,h} - x_i^{\mathrm{e}\top}\vartheta_h\big), \qquad \bar{\zeta}_{t,h} = t^{-1}\sum_{i=1}^{t}\zeta_{i,h}.$$

Then

$$\widehat{\vartheta}_{t,h} - \vartheta_h = \bar{G}_t^{-1}\Big(\bar{\zeta}_{t,h} - \frac{\lambda}{t}\vartheta_h\Big). \tag{17}$$

Similarly, with the matrix-valued residuals

$$\Xi_{i,h} = x_i^{\mathrm{e}}X_i^\top\delta_{i,h} - x_i^{\mathrm{e}}x_i^{\mathrm{e}\top}\Theta_h, \qquad \bar{\Xi}_{t,h} = t^{-1}\sum_{i=1}^{t}\Xi_{i,h},$$

we have

$$\widehat{\Theta}_{t,h} - \Theta_h = \bar{G}_t^{-1}\Big(\bar{\Xi}_{t,h} - \frac{\lambda}{t}\Theta_h\Big). \tag{18}$$

Because $\Sigma \succ 0$, the matrix G is positive definite. Let

$$E_t^G = \Big\{\lambda_{\min}(\bar{G}_t) \ge \tfrac{1}{2}\lambda_{\min}(G)\Big\}.$$

By (8) and Markov's inequality,

$$\mathbb{P}\big\{(E_t^G)^c\big\} = O(t^{-4}). \tag{19}$$

On E_t^G, $\|\bar{G}_t^{-1}\|_{\mathrm{op}} \le 2/\lambda_{\min}(G)$. Since $\zeta_{i,h}$ and $\Xi_{i,h}$ are i.i.d. mean-zero vectors or matrices with finite eighth moments under (A2), (17) and (18) imply

$$E\big[\|\widehat{\vartheta}_{t,h} - \vartheta_h\|^2\mathbf{1}\{E_t^G\}\big] + E\big[\|\widehat{\Theta}_{t,h} - \Theta_h\|_{\mathrm{F}}^2\mathbf{1}\{E_t^G\}\big] = O(t^{-1}),$$

and

$$E\big[\|\widehat{\vartheta}_{t,h} - \vartheta_h\|^4\mathbf{1}\{E_t^G\}\big] + E\big[\|\widehat{\Theta}_{t,h} - \Theta_h\|_{\mathrm{F}}^4\mathbf{1}\{E_t^G\}\big] = O(t^{-2}).$$

On $(E_t^G)^c$, the ridge bound $\bar{G}_t \succeq (\lambda/t)I_{p+1}$ implies $\|\bar{G}_t^{-1}\|_{\mathrm{op}} \le t/\lambda$. Hence, by Cauchy-Schwarz,

$$E\big[\|\widehat{\vartheta}_{t,h} - \vartheta_h\|^2\mathbf{1}\{(E_t^G)^c\}\big] \le C\Big\{E\Big\|\sum_{i=1}^{t}\zeta_{i,h} - \lambda\vartheta_h\Big\|^4\Big\}^{1/2}\mathbb{P}\{(E_t^G)^c\}^{1/2} = O(t^{-1}),$$

and likewise

$$E\big[\|\widehat{\Theta}_{t,h} - \Theta_h\|_{\mathrm{F}}^2 \mathbf{1}\{(E_t^G)^c\}\big] = O(t^{-1}).$$

The same argument with eighth moments gives

$$E\big[\|\widehat{\vartheta}_{t,h} - \vartheta_h\|^4 \mathbf{1}\{(E_t^G)^c\}\big] + E\big[\|\widehat{\Theta}_{t,h} - \Theta_h\|_{\mathrm{F}}^4 \mathbf{1}\{(E_t^G)^c\}\big] = O(1).$$

Combining the good and bad events proves (10) and (11). Since $\bar{G}_t \to G$, $\bar{s}_{t,h} \to g_h$, and $\bar{R}_{t,h} \to M_h$ almost surely by the strong law of large numbers, the continuous mapping theorem gives (13).

It remains to prove the one-step bounds. The Sherman-Morrison identity yields

$$G_{t+1}^{-1} = G_t^{-1} - \frac{G_t^{-1} x_{t+1}^{\mathrm{e}} x_{t+1}^{\mathrm{e}\top} G_t^{-1}}{1 + x_{t+1}^{\mathrm{e}\top} G_t^{-1} x_{t+1}^{\mathrm{e}}},$$

and therefore

$$G_t^{-1} - G_{t+1}^{-1} = G_{t+1}^{-1} x_{t+1}^{\mathrm{e}} x_{t+1}^{\mathrm{e}\top} G_t^{-1}.$$

Using $s_{t+1,h} = s_{t,h} + x_{t+1}^{\mathrm{e}} \delta_{t+1,h}$ and $R_{t+1,h} = R_{t,h} + x_{t+1}^{\mathrm{e}} X_{t+1}^{\top} \delta_{t+1,h}$, we obtain the exact rank one update formulas

$$\begin{aligned}
\widehat{\vartheta}_{t+1,h} - \widehat{\vartheta}_{t,h} &= G_{t+1}^{-1} x_{t+1}^{\mathrm{e}} \Big(\delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\vartheta}_{t,h}\Big), \\
\widehat{\Theta}_{t+1,h} - \widehat{\Theta}_{t,h} &= G_{t+1}^{-1} x_{t+1}^{\mathrm{e}} \Big(X_{t+1}^{\top} \delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\Theta}_{t,h}\Big).
\end{aligned}$$

The key deterministic bound is

$$\big\|G_{t+1}^{-1} x_{t+1}^{\mathrm{e}}\big\|^2 = x_{t+1}^{\mathrm{e}\top} G_{t+1}^{-2} x_{t+1}^{\mathrm{e}} \le \|G_{t+1}^{-1}\|_{\mathrm{op}}\, x_{t+1}^{\mathrm{e}\top} G_{t+1}^{-1} x_{t+1}^{\mathrm{e}} \le \lambda^{-1}, \tag{20}$$

because $G_{t+1} \succeq \lambda I_{p+1}$ and $G_{t+1} \succeq x_{t+1}^{\mathrm{e}} x_{t+1}^{\mathrm{e}\top}$, so $x_{t+1}^{\mathrm{e}\top} G_{t+1}^{-1} x_{t+1}^{\mathrm{e}} \le 1$.

On the good event E_{t+1}^G,

$$\|G_{t+1}^{-1} x_{t+1}^{\mathrm{e}}\| \le \|\bar{G}_{t+1}^{-1}\|_{\mathrm{op}} \frac{\|x_{t+1}^{\mathrm{e}}\|}{t+1} \le C(t+1)^{-1} \|x_{t+1}^{\mathrm{e}}\|.$$

Since (X_{t+1}, Y_{t+1}) is independent of $\mathcal{F}_t$, and $\sup_t E\|\widehat{\vartheta}_{t,h}\|^4 + \sup_t E\|\widehat{\Theta}_{t,h}\|_{\mathrm{F}}^4 < \infty$ by (11),

$$\begin{aligned}
E\Big[\|\widehat{\vartheta}_{t+1,h} - \widehat{\vartheta}_{t,h}\|^2 \mathbf{1}\{E_{t+1}^G\}\Big] &\le Ct^{-2} E\Big[\|x_{t+1}^{\mathrm{e}}\|^2 \big|\delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\vartheta}_{t,h}\big|^2\Big] \le Ct^{-2}, \\
E\Big[\|\widehat{\Theta}_{t+1,h} - \widehat{\Theta}_{t,h}\|_{\mathrm{F}}^2 \mathbf{1}\{E_{t+1}^G\}\Big] &\le Ct^{-2} E\Big[\|x_{t+1}^{\mathrm{e}}\|^2 \big\|X_{t+1}^{\top} \delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\Theta}_{t,h}\big\|^2\Big] \le Ct^{-2}.
\end{aligned}$$

On the bad event $(E_{t+1}^G)^c$, (20) implies

$$\|\widehat{\vartheta}_{t+1,h} - \widehat{\vartheta}_{t,h}\|^4 \le \lambda^{-2} \big|\delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\vartheta}_{t,h}\big|^4,$$

and similarly

$$\|\widehat{\Theta}_{t+1,h} - \widehat{\Theta}_{t,h}\|_{\mathrm{F}}^4 \le \lambda^{-2} \big\|X_{t+1}^{\top} \delta_{t+1,h} - x_{t+1}^{\mathrm{e}\top} \widehat{\Theta}_{t,h}\big\|^4.$$

The right-hand sides have uniformly bounded expectations under (A2) and (11). Therefore, by Cauchy-Schwarz and (19),

$$E\Big[\|\widehat{\vartheta}_{t+1,h} - \widehat{\vartheta}_{t,h}\|^2 \mathbf{1}\{(E_{t+1}^G)^c\}\Big] + E\Big[\|\widehat{\Theta}_{t+1,h} - \widehat{\Theta}_{t,h}\|_{\mathrm{F}}^2 \mathbf{1}\{(E_{t+1}^G)^c\}\Big] \le Ct^{-2}.$$

This proves (12).

Finally, (14)-(16) follow from (4), Lemma 6, and the bounds already proved for $\widehat{\vartheta}_{t,h}$ and $\widehat{\Theta}_{t,h}$. □

Lemma 8. *Assume (A1)-(A3). Let Z_t denote the finite-dimensional vector obtained by stacking*

$$\{\widehat{A}_{t,h}, \widehat{\mathcal{B}}_{t,h}, \widetilde{A}_{t,h}, \widetilde{\mathcal{B}}_{t,h} : h = 1, \dots, H\},$$

and let Z be the corresponding population vector obtained by replacing these blocks by $\{A_h, B_h, \widetilde{A}_h, \widetilde{B}_h\}$. Then there exists $C < \infty$ such that

$$E\|Z_t - Z\|^2 \le Ct^{-1}, \qquad E\|Z_t - Z\|^4 \le Ct^{-2}, \tag{21}$$

$$\sup_{t\ge 1} E\|Z_t\|^{16} \le C, \tag{22}$$

$$E\|Z_{t+1} - Z_t\|^2 \le Ct^{-2}, \qquad E\|Z_{t+1} - Z_t\|^4 \le Ct^{-4}. \tag{23}$$

Moreover, $Z_t \to Z$ almost surely.

Proof. The raw blocks are polynomial functions of the empirical averages

$$\widehat{p}_{t,h}, \quad \widehat{u}_{t,h}, \quad \widehat{S}^{(2)}_{t,h}, \quad \widehat{\mu}_t, \quad \widehat{\Sigma}_t.$$

Under (A2), the standard moment inequalities for averages of i.i.d. random vectors give the corresponding fourth and sixteenth moment rates, as well as the second and fourth moment one-step bounds.

For the RLS blocks, use the expansions

$$\widehat{\vartheta}_{t,h} - \vartheta_h = \bar{G}_t^{-1}\Big(\bar{\zeta}_{t,h} - \frac{\lambda}{t}\vartheta_h\Big), \qquad \widehat{\Theta}_{t,h} - \Theta_h = \bar{G}_t^{-1}\Big(\bar{\Xi}_{t,h} - \frac{\lambda}{t}\Theta_h\Big).$$

Since,

$$E_t^G = \{\lambda_{\min}(\bar{G}_t) \ge \tfrac{1}{2}\lambda_{\min}(G)\},$$

under (A2), the coordinates of $x_i^{\mathrm{e}}x_i^{\mathrm{e}\top}$, $x_i^{\mathrm{e}}\delta_{i,h}$, and $x_i^{\mathrm{e}}X_i^\top\delta_{i,h}$ have sufficiently high moments. Hence the standard Rosenthal inequality for fixed-dimensional sample averages gives, in particular,

$$E\|\bar{G}_t - G\|_{\mathrm{F}}^{32} + E\|\bar{s}_{t,h} - g_h\|^{32} + E\|\bar{R}_{t,h} - M_h\|_{\mathrm{F}}^{32} \le Ct^{-16}.$$

Since $E\|\bar{G}_t - G\|_{\mathrm{F}}^{32} = O(t^{-16})$, Markov's inequality gives $\mathbb{P}\{(E_t^G)^c\} = O(t^{-16})$. On E_t^G, $\bar{G}_t^{-1}$ is uniformly bounded and the desired moment rates follow from the moment bounds for the sample averages $\bar{\zeta}_{t,h}$ and $\bar{\Xi}_{t,h}$. For the RLS blocks, use the expansions

$$\widehat{\vartheta}_{t,h} - \vartheta_h = \bar{G}_t^{-1}\left(\bar{\zeta}_{t,h} - \frac{\lambda}{t}\vartheta_h\right),$$

$$\widehat{\Theta}_{t,h} - \Theta_h = \bar{G}_t^{-1}\left(\bar{\Xi}_{t,h} - \frac{\lambda}{t}\Theta_h\right).$$

Let

$$E_t^G = \left\{\lambda_{\min}(\bar{G}_t) \ge \frac{1}{2}\lambda_{\min}(G)\right\}.$$

On E_t^G, $\bar{G}_t^{-1}$ is uniformly bounded, and the required moment bounds follow from the moment bounds for the fixed-dimensional sample averages $\bar{\zeta}_{t,h}$ and $\bar{\Xi}_{t,h}$.

On $(E_t^G)^c$, we avoid using only the crude bound $\|\bar{G}_t^{-1}\|_{\mathrm{op}} \le t/\lambda$. Instead we use the deterministic structure of the ridge least-squares estimators. The coefficient $\widehat{\vartheta}_{t,h}$ is the minimizer of

$$L_t(\vartheta) = \sum_{i=1}^{t}\left(\delta_{i,h} - x_i^{\mathrm{e}\top}\vartheta\right)^2 + \lambda\|\vartheta\|^2.$$

Since $L_t(\widehat{\vartheta}_{t,h}) \le L_t(0)$,

$$\lambda\|\widehat{\vartheta}_{t,h}\|^2 \le \sum_{i=1}^{t}\delta_{i,h} \le t.$$

Similarly, $\widehat{\Theta}_{t,h}$ minimizes

$$L_t(\Theta) = \sum_{i=1}^{t}\left\|X_i^\top\delta_{i,h} - x_i^{\mathrm{e}\top}\Theta\right\|^2 + \lambda\|\Theta\|_{\mathrm{F}}^2,$$

and therefore

$$\lambda\|\widehat{\Theta}_{t,h}\|_{\mathrm{F}}^2 \le \sum_{i=1}^{t} \delta_{i,h}\|X_i\|^2.$$

By Markov's inequality applied to the thirty-second moment of $\bar{G}_t - G$,

$$\mathbb{P}\{(E_t^G)^c\} = O(t^{-16}).$$

Consequently,

$$E\left[\|\widehat{\vartheta}_{t,h}\|^{16}\mathbf{1}\{(E_t^G)^c\}\right] \le \lambda^{-8}t^8\mathbb{P}\{(E_t^G)^c\} = O(t^{-8}),$$

and, by Cauchy–Schwarz and $E\|X\|^{64} < \infty$,

$$\begin{aligned} E\left[\|\widehat{\Theta}_{t,h}\|_{\mathrm{F}}^{16}\mathbf{1}\{(E_t^G)^c\}\right] &\le \lambda^{-8}E\left[\left(\sum_{i=1}^{t}\delta_{i,h}\|X_i\|^2\right)^8 \mathbf{1}\{(E_t^G)^c\}\right] \\ &\le \lambda^{-8}\left\{E\left[\left(\sum_{i=1}^{t}\delta_{i,h}\|X_i\|^2\right)^{16}\right]\right\}^{1/2} \mathbb{P}\{(E_t^G)^c\}^{1/2} \\ &\le Ct^8 \cdot t^{-8} \le C. \end{aligned}$$

Thus the RLS blocks have uniformly bounded sixteenth moments on both E_t^G and $(E_t^G)^c$. The same deterministic bounds, together with the rank one update formulas from Lemma 7, give the one-step bounds by the same good-event and bad event decomposition. The rank one Sherman-Morrison update formulas in Lemma 7 give the stated one-step bounds by the same good-event or bad-event decomposition.

Finally, $\widetilde{A}_{t,h}$, $\widehat{Q}_{t,h}$, and $\widetilde{\mathcal{B}}_{t,h}$ are polynomial functions of the raw and RLS quantities. Since H is fixed, the displayed bounds follow. Almost sure convergence follows from the strong law and the continuous mapping theorem. □

Theorem 9. *Assume (A1)-(A3). Then, for each fixed $h = 1, \dots, H$, there exists a constant $C > 0$ such that*

$$E\big\|\widehat{A}_{t,h} - A_h\big\|^4 + E\big\|\widehat{\mathcal{B}}_{t,h} - B_h\big\|_{\mathrm{F}}^4 \le Ct^{-2}, \tag{24}$$

$$E\big\|\widetilde{A}_{t,h} - \widetilde{A}_h\big\|^2 + E\big\|\widetilde{\mathcal{B}}_{t,h} - \widetilde{B}_h\big\|_{\mathrm{F}}^2 \le Ct^{-1},$$

$$E\big\|\widehat{K}_t^\star - K^\star\big\|_{\mathrm{op}}^2 \le Ct^{-1}, \tag{25}$$

$$\sup_{t\ge 1} E\big\|\widehat{K}_t^\star\big\|_{\mathrm{op}}^4 \le C, \tag{26}$$

$$E\big\|\widehat{K}_{t+1}^\star - \widehat{K}_t^\star\big\|_{\mathrm{op}}^2 \le Ct^{-2}. \tag{27}$$

Moreover,

$$\big\|\widehat{K}_t^\star - K^\star\big\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad \widehat{K}_t^\star \to K^\star \quad \textit{almost surely.} \tag{28}$$

The same is true for the instance average

$$E\big\|D_t - K^\star\big\|_{\mathrm{op}}^2 \le Ct^{-1}, \qquad \big\|D_t - K^\star\big\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad D_t \to K^\star \quad \textit{almost surely.} \tag{29}$$

In addition,

$$E\big\|D_{t+1} - D_t\big\|_{\mathrm{op}}^2 \le Ct^{-2}. \tag{30}$$

Proof. The raw block fourth-moment bound follows from Lemma 8. The second-moment bound for the standardized blocks is also contained in (21).

Let $\mathcal{P}$ be the polynomial map that sends the stacked block vector Z_t to $\widehat{K}_t^\star$. The degree of $\mathcal{P}$ is at most four. Hence, on a finite-dimensional space, there exists $C < \infty$ such that

$$\begin{aligned}\|\mathcal{P}(z) - \mathcal{P}(z_0)\| &\le C(1 + \|z\|^3 + \|z_0\|^3)\|z - z_0\|,\\ \|\mathcal{P}(z)\|^4 &\le C(1 + \|z\|^{16}),\\ \|\mathcal{P}(z_1) - \mathcal{P}(z_0)\|^2 &\le C(1 + \|z_1\|^6 + \|z_0\|^6)\|z_1 - z_0\|^2.\end{aligned}$$

Applying these inequalities with $z = Z_t$, $z_0 = Z$, and using Lemma 8 gives

$$E\|\widehat{K}_t^\star - K^\star\|_{\mathrm{op}}^2 \le Ct^{-1}, \qquad \sup_{t\ge 1} E\|\widehat{K}_t^\star\|_{\mathrm{op}}^4 \le C.$$

Similarly, applying the third inequality with $z_1 = Z_{t+1}$ and $z_0 = Z_t$, and using (22)-(23), yields

$$E\|\widehat{K}_{t+1}^\star - \widehat{K}_t^\star\|_{\mathrm{op}}^2 \le Ct^{-2}.$$

The stochastic root-t rate follows by Markov's inequality, and the almost sure convergence follows from $Z_t \to Z$ and continuity of $\mathcal{P}$.

For the instance average, the almost sure convergence of D_t follows by Cesàro averaging. Moreover,

$$\Big\{E\|D_t - K^\star\|_{\mathrm{op}}^2\Big\}^{1/2} \le \frac{1}{t}\sum_{i=1}^{t}\Big\{E\|\widehat{K}_i^\star - K^\star\|_{\mathrm{op}}^2\Big\}^{1/2} \le \frac{\sqrt{C}}{t}\sum_{i=1}^{t} i^{-1/2} = O(t^{-1/2}).$$

Finally,

$$D_{t+1} - D_t = \frac{1}{t+1}(\widehat{K}_{t+1}^\star - D_t),$$

and the bounded second moments of $\widehat{K}_{t+1}^\star$ and D_t imply (30). □

Lemma 10. *Assume (A1)-(A3). Then there exists a constant $C > 0$ such that*

$$E\big\|S_t^{-1/2} - \Sigma^{-1/2}\big\|_{\mathrm{op}}^2 \le Ct^{-1}, \tag{31}$$

$$E\big\|\widetilde{W}_t - M^\star\big\|_{\mathrm{op}} + E\big\|W_t - M^\star\big\|_{\mathrm{op}} \le Ct^{-1/2}, \tag{32}$$

$$\sup_{t\ge 1} E\big\|\widetilde{W}_t\big\|_{\mathrm{op}}^2 + \sup_{t\ge 1} E\big\|W_t\big\|_{\mathrm{op}}^2 \le C, \tag{33}$$

$$E\big\|\Delta S_t\big\|_{\mathrm{F}}^2 \le Ct^{-2}, \qquad \Delta S_t = S_{t+1} - S_t, \tag{34}$$

$$E\big\|\widetilde{W}_{t+1} - \widetilde{W}_t\big\|_{\mathrm{op}} \le Ct^{-1}. \tag{35}$$

Moreover,

$$\big\|W_t - M^\star\big\|_{\mathrm{op}} = O_p(t^{-1/2}), \qquad W_t \to M^\star \quad \textit{almost surely.} \tag{36}$$

Proof. Because $\rho_t = \rho_0/t$, Theorem 9 and (8) imply

$$E\big\|S_t - \Sigma\big\|_{\mathrm{F}}^8 \le Ct^{-4}.$$

Let

$$F_t = \Big\{\|S_t - \Sigma\|_{\mathrm{op}} \le \tfrac{1}{2}\lambda_{\min}(\Sigma)\Big\}.$$

Markov's inequality gives $\mathbb{P}(F_t^c) = O(t^{-4})$. On F_t, the inverse-square-root map is Lipschitz on a compact subset of the positive-definite cone, so

$$\big\|S_t^{-1/2} - \Sigma^{-1/2}\big\|_{\mathrm{op}} \le C\|S_t - \Sigma\|_{\mathrm{op}}.$$

On F_t^c, the ridge stabilization gives $S_t \succeq (\rho_0/t)I_p$, hence

$$\big\|S_t^{-1/2} - \Sigma^{-1/2}\big\|_{\mathrm{op}}^2 \le C\Big(1 + \frac{t}{\rho_0}\Big).$$

Therefore,

$$E\big\|S_t^{-1/2} - \Sigma^{-1/2}\big\|_{\mathrm{op}}^2 \le CE\|S_t - \Sigma\|_{\mathrm{op}}^2 + C\Big(1 + \frac{t}{\rho_0}\Big)\mathbb{P}(F_t^c) = O(t^{-1}),$$

which proves (31). Almost sure convergence follows from the continuous mapping theorem and $S_t \to \Sigma$ almost surely.

On F_t, the matrices $S_t^{-1/2}$ and $\Sigma^{-1/2}$ are uniformly bounded, so

$$\begin{aligned}\|\widetilde{W}_t - M^\star\|_{\mathrm{op}}\mathbf{1}\{F_t\} \le& C\|S_t^{-1/2} - \Sigma^{-1/2}\|_{\mathrm{op}}\big(\|\widehat{K}_t^\star\|_{\mathrm{op}} + \|K^\star\|_{\mathrm{op}}\big)\\ &+ C\|\widehat{K}_t^\star - K^\star\|_{\mathrm{op}},\\ \|W_t - M^\star\|_{\mathrm{op}}\mathbf{1}\{F_t\} \le& C\|S_t^{-1/2} - \Sigma^{-1/2}\|_{\mathrm{op}}\big(\|D_t\|_{\mathrm{op}} + \|K^\star\|_{\mathrm{op}}\big)\\ &+ C\|D_t - K^\star\|_{\mathrm{op}}.\end{aligned}$$

By Cauchy-Schwarz, (31), Theorem 9, and the bounded second moments of $\widehat{K}_t^\star$ and D_t, the expectations of the right-hand sides are $O(t^{-1/2})$.

On F_t^c, the ridge bound gives

$$\|S_t^{-1/2}\|_{\mathrm{op}} \le Ct^{1/2},$$

so

$$\|\widetilde{W}_t - M^\star\|_{\mathrm{op}} + \|W_t - M^\star\|_{\mathrm{op}} \le Ct\big(\|\widehat{K}_t^\star\|_{\mathrm{op}} + \|D_t\|_{\mathrm{op}} + 1\big).$$

Hence, again by Cauchy-Schwarz,

$$E\Big[\big(\|\widetilde{W}_t - M^\star\|_{\mathrm{op}} + \|W_t - M^\star\|_{\mathrm{op}}\big)\mathbf{1}\{F_t^c\}\Big] \le Ct\mathbb{P}(F_t^c)^{1/2} = O(t^{-1}).$$

Combining the good and bad events proves (32). By Jensen's inequality and (26),

$$\sup_{t\ge1} E\|D_t\|_{\mathrm{op}}^4 \le \sup_{t\ge1}\frac{1}{t}\sum_{i=1}^t E\|\widehat{K}_i^\star\|_{\mathrm{op}}^4 < \infty.$$

For the uniform second moments, on F_t

$$\|\widetilde{W}_t\|_{\mathrm{op}}^2 + \|W_t\|_{\mathrm{op}}^2 \le C\big(\|\widehat{K}_t^\star\|_{\mathrm{op}}^2 + \|D_t\|_{\mathrm{op}}^2\big),$$

whereas on F_t^c,

$$\|\widetilde{W}_t\|_{\mathrm{op}}^2 + \|W_t\|_{\mathrm{op}}^2 \le Ct^2\big(\|\widehat{K}_t^\star\|_{\mathrm{op}}^2 + \|D_t\|_{\mathrm{op}}^2\big).$$

Therefore,

$$\begin{aligned}E\big(\|\widetilde{W}_t\|_{\mathrm{op}}^2 + \|W_t\|_{\mathrm{op}}^2\big) \le& CE\big(\|\widehat{K}_t^\star\|_{\mathrm{op}}^2 + \|D_t\|_{\mathrm{op}}^2\big) + Ct^2\Big\{E\big(\|\widehat{K}_t^\star\|_{\mathrm{op}}^4 + \|D_t\|_{\mathrm{op}}^4\big)\Big\}^{1/2}\mathbb{P}(F_t^c)^{1/2}\\ \le& C,\end{aligned}$$

which proves (33).

The covariance update and the ridge correction give

$$\Delta S_t = -\frac{1}{t+1}\widehat{\Sigma}_t + \frac{t}{(t+1)^2}(X_{t+1} - \widehat{\mu}_t)(X_{t+1} - \widehat{\mu}_t)^\top - \frac{\rho_0}{t(t+1)}I_p. \tag{37}$$

Since $E\|X_{t+1} - \widehat{\mu}_t\|^4 < \infty$ uniformly in t and $\widehat{\Sigma}_t$ has bounded second moment, (34) follows.

To prove (35), write

$$\begin{aligned}\widetilde{W}_{t+1} - \widetilde{W}_t =& (S_{t+1}^{-1/2} - S_t^{-1/2})\widehat{K}_{t+1}^\star S_{t+1}^{-1/2} + S_t^{-1/2}(\widehat{K}_{t+1}^\star - \widehat{K}_t^\star)S_{t+1}^{-1/2}\\ &+ S_t^{-1/2}\widehat{K}_t^\star(S_{t+1}^{-1/2} - S_t^{-1/2}).\end{aligned}$$

On the good event $F_t \cap F_{t+1}$, the inverse-square-root map is Lipschitz, so $\|S_{t+1}^{-1/2} - S_t^{-1/2}\|_{\mathrm{op}} \le C\|\Delta S_t\|_{\mathrm{op}}$. Hence

$$\|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} \mathbf{1}\{F_t \cap F_{t+1}\} \le C\|\Delta S_t\|_{\mathrm{op}}\big(\|\widehat{K}_t^\star\|_{\mathrm{op}} + \|\widehat{K}_{t+1}^\star\|_{\mathrm{op}}\big) + C\|\widehat{K}_{t+1}^\star - \widehat{K}_t^\star\|_{\mathrm{op}}.$$

By Cauchy-Schwarz, (34), (26), and (27),

$$E\Big[\|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} \mathbf{1}\{F_t \cap F_{t+1}\}\Big] \le Ct^{-1}.$$

On $(F_t \cap F_{t+1})^c$, the ridge gives $\|S_t^{-1/2}\|_{\mathrm{op}} + \|S_{t+1}^{-1/2}\|_{\mathrm{op}} \le Ct^{1/2}$, and hence

$$\|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} \le Ct\big(\|\widehat{K}_t^\star\|_{\mathrm{op}} + \|\widehat{K}_{t+1}^\star\|_{\mathrm{op}}\big).$$

Since $\mathbb{P}((F_t \cap F_{t+1})^c) = O(t^{-4})$ and $\sup_t E\|\widehat{K}_t^\star\|_{\mathrm{op}}^2 < \infty$,

$$E\Big[\|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} \mathbf{1}\{(F_t \cap F_{t+1})^c\}\Big] \le Ct\mathbb{P}((F_t \cap F_{t+1})^c)^{1/2} \le Ct^{-1}.$$

Combining the two parts proves (35).

Moreover,

$$\sup_{t\ge 1} \mathbb{P}\big(t^{1/2}\|W_t - M^\star\|_{\mathrm{op}} > M\big) \le M^{-1} \sup_{t\ge 1} t^{1/2} E\|W_t - M^\star\|_{\mathrm{op}} \le CM^{-1},$$

so (36) follows by Markov's inequality. The almost sure convergence is obtained from $D_t \to K^\star$ and $S_t \to \Sigma$ almost surely, together with the continuity of the standardization map. □

Lemma 11. *Let*

$$\mathbb{S}^p = \{A \in \mathbb{R}^{p\times p} : A^\top = A\}$$

be the space of real symmetric $p \times p$ matrices, and let

$$\mathbb{S}_{++}^p = \{A \in \mathbb{S}^p : x^\top Ax > 0 \text{ for all } x \in \mathbb{R}^p \setminus \{0\}\}$$

be the cone of symmetric positive definite matrices. For $A, B \in \mathbb{S}^p$, write $A \preceq B$ if $B - A$ is positive semidefinite, and let I_p denote the $p \times p$ identity matrix.

Fix $0 < m < M < \infty$ and define

$$\mathcal{K}_{m,M} = \{S \in \mathbb{S}_{++}^p : \ mI_p \preceq S \preceq MI_p\}.$$

For $S \in \mathbb{S}_{++}^p$, let $S^{1/2}$ denote the principal positive definite square root of S, and set $S^{-1/2} = (S^{1/2})^{-1}$. Then the maps

$$S \longmapsto S^{1/2}, \qquad S \longmapsto S^{-1/2}, \qquad (D, S) \longmapsto S^{-1/2} D S^{-1/2}$$

are restrictions of C^∞ maps on $\mathbb{S}_{++}^p$ and $\mathbb{S}^p \times \mathbb{S}_{++}^p$, respectively. In particular, their first Fréchet derivatives are locally Lipschitz on compact subsets. Hence, this property holds on $\mathcal{K}_{m,M}$ and on $\mathbb{S}^p \times \mathcal{K}_{m,M}$.

Proof. For $S \in \mathcal{K}_{m,M}$, the standard resolvent representations are

$$S^{-1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}(S + tI_p)^{-1}\, dt, \qquad S^{1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2} S(S + tI_p)^{-1}\, dt.$$

For the inverse square root, the first two Fréchet derivatives of $(S + tI_p)^{-1}$ are bounded by

$$C(m + t)^{-2}\|H\|, \qquad C(m + t)^{-3}\|H_1\|\|H_2\|,$$

and hence the corresponding differentiated integrands are dominated by

$$Ct^{-1/2}(m + t)^{-2}, \qquad Ct^{-1/2}(m + t)^{-3},$$

which are integrable over $(0, \infty)$.

For the square root, use the identity

$$S(S+tI_p)^{-1} = I_p - t(S+tI_p)^{-1}.$$

Thus the first and second differentiated integrands are dominated by

$$Ct^{1/2}(m+t)^{-2}, \qquad Ct^{1/2}(m+t)^{-3},$$

respectively, which are also integrable over $(0,\infty)$. Differentiation under the integral sign is therefore justified for both maps. Iterating the same argument gives C^∞ smoothness on $\mathcal{K}_{m,M}$.

The map $(D,S) \mapsto S^{-1/2}DS^{-1/2}$ is a finite composition of bilinear maps with $S \mapsto S^{-1/2}$, so it is C^∞ on $\mathbb{S}^p \times \mathcal{K}_{m,M}$. The local Lipschitz property of the first derivatives on compact subsets follows from boundedness of the second derivatives. □

A.3 Consistency of the exact generalized eigenspace

Let $\widehat{\lambda}_{t,1} \geq \cdots \geq \widehat{\lambda}_{t,p}$ denote the eigenvalues of W_t, and let $\widehat{v}_{t,1}^{(W)}, \ldots, \widehat{v}_{t,p}^{(W)}$ be the associated orthonormal eigenvectors. Set

$$V_\star = (v_1, \ldots, v_d), \qquad \widehat{V}_t^{(W)} = (\widehat{v}_{t,1}^{(W)}, \ldots, \widehat{v}_{t,d}^{(W)}), \qquad \widehat{B}_t = S_t^{-1/2}\widehat{V}_t^{(W)}, \qquad B_\star = \Sigma^{-1/2}V_\star.$$

Write $P(B) = B(B^\top B)^{-1}B^\top$ for the orthogonal projector onto $\mathrm{span}(B)$ whenever B has full column rank.

Lemma 12. *Let $B_0 \in \mathbb{R}^{p\times d}$ have full column rank. Let*

$$\mathbb{O}(d) = \{O \in \mathbb{R}^{d\times d} : O^\top O = I_d\}$$

denote the orthogonal group, where I_d is the $d \times d$ identity matrix. Then the map $B \mapsto P(B)$ is locally Lipschitz at B_0. There exist $\varepsilon > 0$ and $C > 0$ such that

$$\big\|P(B) - P(B_0)\big\|_{\mathrm{op}} \leq C \inf_{O\in\mathbb{O}(d)} \big\|B - B_0O\big\|_{\mathrm{op}} \tag{38}$$

for every B with $\inf_{O\in\mathbb{O}(d)} \|B - B_0O\|_{\mathrm{op}} < \varepsilon$.

Proof. In a neighborhood of B_0, the matrix $B^\top B$ remains invertible. The map $B \mapsto B(B^\top B)^{-1}B^\top$ is continuously Fréchet differentiable on the open set of full-column-rank matrices, hence locally Lipschitz there. Right multiplication by an orthogonal matrix does not change the projector, which gives (38). □

Theorem 13. *Assume (A1)-(A5). Then*

$$\max_{1\leq j\leq p} \big|\widehat{\lambda}_{t,j} - \lambda_j\big| = O_p(t^{-1/2}),$$

$$\widehat{\lambda}_{t,j} \to \lambda_j \quad \textit{almost surely for all } j, \tag{39}$$

$$\inf_{O\in\mathbb{O}(d)} \big\|\widehat{V}_t^{(W)} - V_\star O\big\|_{\mathrm{op}} = O_p(t^{-1/2}), \tag{40}$$

$$\big\|P(\widehat{B}_t) - P(B_\star)\big\|_{\mathrm{op}} = O_p(t^{-1/2}),$$

and $\mathrm{span}(\widehat{B}_t) \to \mathcal{S}_{Y|X}$ almost surely in projector distance. Equivalently,

$$\big\|\sin\Theta\big(\mathrm{span}(\widehat{B}_t),\, \mathcal{S}_{Y|X}\big)\big\|_{\mathrm{op}} = O_p(t^{-1/2}). \tag{41}$$

Proof. Weyl's inequality [13] gives

$$\max_{1\leq j\leq p} \big|\widehat{\lambda}_{t,j} - \lambda_j\big| \leq \|W_t - M^\star\|_{\mathrm{op}},$$

so (39) follows from Lemma 10. Since the signal-space eigengap equals $\lambda_d - \lambda_{d+1} = \lambda_d > 0$, standard symmetric subspace perturbation bounds imply

$$\inf_{O\in\mathbb{O}(d)} \|\widehat{V}_t^{(W)} - V_\star O\|_{\mathrm{op}} \leq C\|W_t - M^\star\|_{\mathrm{op}}$$

for all sufficiently large t. See, for example, [14]. This proves the first part of (40).

Next, for any $O \in \mathbb{O}(d)$,

$$\|\widehat{B}_t - B_\star O\|_{\mathrm{op}} \le \|S_t^{-1/2} - \Sigma^{-1/2}\|_{\mathrm{op}} \|\widehat{V}_t^{(W)}\|_{\mathrm{op}} + \|\Sigma^{-1/2}\|_{\mathrm{op}} \|\widehat{V}_t^{(W)} - V_\star O\|_{\mathrm{op}} = O_p(t^{-1/2}),$$

because $\|\widehat{V}_t^{(W)}\|_{\mathrm{op}} = 1$. Since $B_\star$ has full column rank, Lemma 12 yields

$$\|P(\widehat{B}_t) - P(B_\star)\|_{\mathrm{op}} = O_p(t^{-1/2}).$$

For equal-dimensional subspaces, the operator norm of the sine-theta matrix equals the projector distance, so (41) follows.

Almost sure convergence is obtained in the same way by combining $W_t \to M^\star$ and $S_t^{-1/2} \to \Sigma^{-1/2}$ almost surely with continuity of spectral projections. □

A.4 First-order perturbation for generalized eigenpairs

The next lemma gives a first-order expansion for a simple generalized eigenpair of the symmetric matrix pair (D, S). Since a generalized eigenvector is defined only up to scale, we fix a local representative by imposing a Euclidean identifiability constraint. The result is the generalized-eigenvalue counterpart of the standard first-order perturbation formula for a simple eigenpair, see [14].

Lemma 14. *Let*

$$D(\varepsilon) = D + \varepsilon G + O(\varepsilon^2), \qquad S(\varepsilon) = S + \varepsilon H + O(\varepsilon^2),$$

where D, S, G, and H are symmetric and $S \succ 0$. Let (λ, β) be a simple generalized eigenpair of (D, S) normalized by $\beta^\top S \beta = 1$. Then there exists a differentiable branch $(\lambda(\varepsilon), \beta(\varepsilon))$ of generalized eigenpairs of $(D(\varepsilon), S(\varepsilon))$ such that $\lambda(0) = \lambda$, $\beta(0) = \beta$, and the Euclidean orthogonality constraint

$$\beta^\top \big\{\beta(\varepsilon) - \beta\big\} = 0 \tag{42}$$

holds for all sufficiently small ε. Under this constraint,

$$\lambda(\varepsilon) = \lambda + \varepsilon\, \beta^\top (G - \lambda H)\beta + O(\varepsilon^2), \tag{43}$$

$$\beta(\varepsilon) = \beta + \varepsilon (\lambda S - D)^+ r + O(\varepsilon^2), \tag{44}$$

where

$$\begin{aligned} r &= (G - \lambda H)\beta - \big\{\beta^\top (G - \lambda H)\beta\big\} S\beta \\ &= \big(I_p - S\beta\beta^\top\big)(G - \lambda H)\beta. \end{aligned} \tag{45}$$

Proof. Define

$$\mathcal{F}(u, \ell, \varepsilon) = \begin{pmatrix} D(\varepsilon)u - \ell S(\varepsilon)u \\ \beta^\top (u - \beta) \end{pmatrix}.$$

At $(u, \ell, \varepsilon) = (\beta, \lambda, 0)$, the derivative of $\mathcal{F}$ with respect to (u, ℓ) is

$$\mathrm{D}_{(u,\ell)} \mathcal{F}(\beta, \lambda, 0)[h, a] = \begin{pmatrix} (D - \lambda S)h - aS\beta \\ \beta^\top h \end{pmatrix}.$$

Because λ is simple, $\ker(D - \lambda S) = \mathrm{span}(\beta)$. Hence the displayed linear map is invertible. If

$$(D - \lambda S)h - aS\beta = 0, \qquad \beta^\top h = 0,$$

then left-multiplying the first equation by $\beta^\top$ and using $D\beta = \lambda S\beta$ and $\beta^\top S\beta = 1$ gives $a = 0$, and therefore $(D - \lambda S)h = 0$, so $h \in \mathrm{span}(\beta)$, combined with $\beta^\top h = 0$, this yields $h = 0$. The implicit function theorem therefore yields a differentiable local branch $(\beta(\varepsilon), \lambda(\varepsilon))$ such that

$$D(\varepsilon)\beta(\varepsilon) = \lambda(\varepsilon) S(\varepsilon)\beta(\varepsilon), \qquad \beta^\top \big(\beta(\varepsilon) - \beta\big) = 0.$$

Now write

$$\lambda(\varepsilon) = \lambda + \varepsilon\dot{\lambda} + O(\varepsilon^2), \qquad \beta(\varepsilon) = \beta + \varepsilon\dot{\beta} + O(\varepsilon^2).$$

Substituting into $D(\varepsilon)\beta(\varepsilon) = \lambda(\varepsilon)S(\varepsilon)\beta(\varepsilon)$ and collecting first-order terms gives

$$(\lambda S - D)\dot{\beta} = (G - \lambda H)\beta - \dot{\lambda}\, S\beta. \tag{46}$$

Left-multiplying by $\beta^\top$ and using $D\beta = \lambda S\beta$ together with $\beta^\top S\beta = 1$ yields

$$\dot{\lambda} = \beta^\top (G - \lambda H)\beta,$$

which proves (43). Substituting this identity back into (46) gives

$$(\lambda S - D)\dot{\beta} = r, \tag{47}$$

with r defined in (45).

Let $A = \lambda S - D$. Because A is symmetric and λ is simple, $\ker(A) = \operatorname{span}(\beta)$. Hence

$$\operatorname{range}(A) = \ker(A)^\perp = \{x \in \mathbb{R}^p : \ \beta^\top x = 0\}$$

in the Euclidean inner product. Moreover,

$$\beta^\top r = \beta^\top (G - \lambda H)\beta - \dot{\lambda}\, \beta^\top S\beta = 0,$$

so $r \in \operatorname{range}(A)$ and $A^+ r$ is a well-defined particular solution of (47). Since the image of the Moore-Penrose pseudoinverse is Euclidean-orthogonal to the null space, $\beta^\top A^+ r = 0$. The general solution of (47) is $\dot{\beta} = A^+ r + \eta\beta$ for some scalar η, and the Euclidean orthogonality constraint (42) is equivalent to $\beta^\top \dot{\beta} = 0$ at first order. Therefore $\eta = 0$, which proves (44). □

The Euclidean orthogonality constraint in Lemma 14 is essential. In general, $(\lambda S - D)^+ r$ is Euclidean-orthogonal to β but need not be S-orthogonal to β. If one imposed the S-metric constraint $\beta^\top S\dot{\beta} = 0$ at the perturbation stage, an additional multiple of β would be required. Our algorithm does not need that correction because the raw perturbation direction is followed by an explicit $\mathcal{P}^{(S)}_{\mathrm{orth}}$ normalization step.

Proposition 15. *Assume (A1)-(A5). Fix $j \in \{1, \dots, d\}$ and let $(\widehat{\lambda}_{t,j}, \widehat{\beta}_{t,j})$ be the jth exact generalized eigenpair of (D_t, S_t), normalized by $\widehat{\beta}_{t,j}^\top S_t \widehat{\beta}_{t,j} = 1$. Define*

$$\Delta D_t = D_{t+1} - D_t, \qquad \Delta S_t = S_{t+1} - S_t,$$
$$\delta_{t,j} = \widehat{\beta}_{t,j}^\top (\Delta D_t - \widehat{\lambda}_{t,j}\Delta S_t)\widehat{\beta}_{t,j}, \qquad r_{t,j} = (\Delta D_t - \widehat{\lambda}_{t,j}\Delta S_t)\widehat{\beta}_{t,j} - \delta_{t,j} S_t \widehat{\beta}_{t,j}.$$

Then there exist deterministic constants $\varepsilon_j > 0$ and $C_j < \infty$, an event Ω_j with $\mathbb{P}(\Omega_j) = 1$, and an a.s. finite random time T_j such that the following holds. For every $\omega \in \Omega_j$ and every $t \geq T_j(\omega)$ satisfying

$$\|\Delta D_t(\omega)\|_{\mathrm{op}} + \|\Delta S_t(\omega)\|_{\mathrm{op}} \leq \varepsilon_j, \tag{48}$$

one can choose a representative $\widehat{\beta}^{\mathrm{ex}}_{t+1,j}(\omega)$ of the exact jth generalized eigendirection of (D_{t+1}, S_{t+1}) such that

$$\widehat{\beta}_{t,j}^\top \big(\widehat{\beta}^{\mathrm{ex}}_{t+1,j} - \widehat{\beta}_{t,j}\big) = 0 \tag{49}$$

and

$$\widehat{\lambda}^{\mathrm{ex}}_{t+1,j} = \widehat{\lambda}_{t,j} + \delta_{t,j} + R_{\lambda,t,j}, \qquad |R_{\lambda,t,j}| \leq C_j\big(\|\Delta D_t\|_{\mathrm{op}} + \|\Delta S_t\|_{\mathrm{op}}\big)^2, \tag{50}$$
$$\widehat{\beta}^{\mathrm{ex}}_{t+1,j} = \widehat{\beta}_{t,j} + (\widehat{\lambda}_{t,j} S_t - D_t)^+ r_{t,j} + R_{\beta,t,j}, \quad \|R_{\beta,t,j}\| \leq C_j\big(\|\Delta D_t\|_{\mathrm{op}} + \|\Delta S_t\|_{\mathrm{op}}\big)^2. \tag{51}$$

In particular,

$$\widehat{\lambda}^{\mathrm{ex}}_{t+1,j} = \widehat{\lambda}_{t,j} + \delta_{t,j} + O_p(t^{-2}), \qquad \widehat{\beta}^{\mathrm{ex}}_{t+1,j} = \widehat{\beta}_{t,j} + (\widehat{\lambda}_{t,j} S_t - D_t)^+ r_{t,j} + O_p(t^{-2}). \tag{52}$$

Proof. Fix $j \in \{1, \dots, d\}$. By (A5) and continuity of the generalized spectrum, there exist an open neighborhood $\mathcal{U}_j$ of $(K^\star, \Sigma)$ in $\mathbb{S}^p \times \mathbb{S}^p_{++}$ and a compact neighborhood $\mathcal{N}_j \Subset \mathcal{U}_j$, meaning that $\mathcal{N}_j$ is compact and contained in the interior of $\mathcal{U}_j$, such that, for every $(D, S) \in \mathcal{U}_j$, the jth generalized eigenvalue of (D, S) is simple and remains uniformly separated from the rest of the generalized spectrum.

For each base point $(D, S) \in \mathcal{N}_j$, the simplicity of the jth generalized eigenvalue and Lemma 14 imply, via the implicit function theorem, that there exists a local C^2 map

$$(\Delta D, \Delta S) \longmapsto \big(\lambda_j(D + \Delta D, S + \Delta S), \beta_j(D + \Delta D, S + \Delta S)\big)$$

defined for $\|\Delta D\|_{\mathrm{op}} + \|\Delta S\|_{\mathrm{op}}$ sufficiently small, where $\beta_j(D + \Delta D, S + \Delta S)$ is a representative of the perturbed generalized eigendirection satisfying

$$\beta_j(D, S)^\top \Big\{ \beta_j(D + \Delta D, S + \Delta S) - \beta_j(D, S) \Big\} = 0.$$

Because the generalized eigengap is uniformly positive on the compact set $\mathcal{N}_j$, the radii of these neighborhoods and the bounds on the second Fréchet derivatives can be chosen uniformly over $(D, S) \in \mathcal{N}_j$. Hence there exist deterministic constants $\varepsilon_j > 0$ and $C_j < \infty$ such that, whenever $(D, S) \in \mathcal{N}_j$ and $(\Delta D, \Delta S)$ satisfies $\|\Delta D\|_{\mathrm{op}} + \|\Delta S\|_{\mathrm{op}} \le \varepsilon_j$ with $(D + \Delta D, S + \Delta S) \in \mathcal{U}_j$, one has

$$\begin{aligned}
\lambda_j(D + \Delta D, S + \Delta S) &= \lambda_j(D, S) + \mathrm{D}\lambda_j(D, S)[\Delta D, \Delta S] + R_\lambda, \\
\beta_j(D + \Delta D, S + \Delta S) &= \beta_j(D, S) + \mathrm{D}\beta_j(D, S)[\Delta D, \Delta S] + R_\beta,
\end{aligned}$$

with

$$|R_\lambda| + \|R_\beta\| \le C_j \big(\|\Delta D\|_{\mathrm{op}} + \|\Delta S\|_{\mathrm{op}}\big)^2.$$

The first Fréchet derivative is exactly the one given by Lemma 14, which yields (50)-(51).

By Theorem 9 and Lemma 10, $D_t \to K^\star$ and $S_t \to \Sigma$ almost surely. Therefore

$$\Omega_j = \Big\{ \exists\, T_j < \infty \text{ such that } (D_t, S_t) \in \mathcal{N}_j \text{ for all } t \ge T_j \Big\}$$

has probability one. On Ω_j, the above Taylor expansion is valid for every $t \ge T_j(\omega)$ satisfying (48). Finally, (30) and (34) imply $\|\Delta D_t\|_{\mathrm{op}} + \|\Delta S_t\|_{\mathrm{op}} = O_p(t^{-1})$, so the quadratic remainder is $O_p(t^{-2})$. This proves (52). □

A.5 Generalized perturbation: exact eigenspace and first-order update

For the perturbation approach, it is convenient to work with the ordinary symmetric eigenproblem induced by standardization. Let

$$\Delta W_t = W_{t+1} - W_t.$$

For a positive definite matrix S, define the S-orthonormalization operator by

$$\mathcal{P}^{(S)}_{\mathrm{orth}}(B) = S^{-1/2} \mathcal{P}_{\mathrm{orth}}(S^{1/2} B), \qquad \mathcal{P}_{\mathrm{orth}}(Z) = Z(Z^\top Z)^{-1/2},$$

whenever $Z^\top Z$ is nonsingular. Suppose that the leading d generalized directions are updated numerically through the standardized ordinary perturbation step

$$\begin{aligned}
\widetilde{\delta}_{t,j} &= \widetilde{v}^{(W)\top}_{t,j} \Delta W_t \, \widetilde{v}^{(W)}_{t,j}, \\
\widetilde{q}_{t,j} &= \Delta W_t \, \widetilde{v}^{(W)}_{t,j} - \widetilde{\delta}_{t,j} \widetilde{v}^{(W)}_{t,j}, \\
\widetilde{\lambda}_{t+1,j} &= \widetilde{\lambda}_{t,j} + \widetilde{\delta}_{t,j}, \qquad \widetilde{v}^{-}_{t+1,j} = \widetilde{v}^{(W)}_{t,j} + (\widetilde{\lambda}_{t,j} I_p - W_t)^{+} \widetilde{q}_{t,j},
\end{aligned} \tag{53}$$

for $j = 1, \dots, d$, followed by Euclidean orthonormalization

$$\widetilde{V}^{(W)}_{t+1} = \mathcal{P}_{\mathrm{orth}}\big(\widetilde{v}^{-}_{t+1,1}, \dots, \widetilde{v}^{-}_{t+1,d}\big), \qquad \widetilde{B}^{(W)}_{t+1} = S^{-1/2}_{t+1} \widetilde{V}^{(W)}_{t+1}. \tag{54}$$

Theorem 13 already identifies the exact leading generalized signal subspace of (D_t, S_t) and proves that it converges almost surely to $\mathcal{S}_{Y|X}$. Lemma 16 gives the one-step expansion underlying the standardized update (53)-(54). Proposition 15 gives the corresponding original-scale generalized-eigenpair expansion, showing that whenever the current iterate equals the exact eigensystem of (D_t, S_t), the perturbation step reproduces the exact eigensystem of (D_{t+1}, S_{t+1}) up to a quadratic remainder of order $O\big((\|\Delta D_t\|_{\mathrm{op}} + \|\Delta S_t\|_{\mathrm{op}})^2\big) = O_p(t^{-2})$.

Lemma 16. *Assume (A1)-(A5). Let*

$$\Delta W_t = W_{t+1} - W_t.$$

Then

$$\|\Delta W_t\|_{\mathrm{op}} = O_p(t^{-1}).$$

Moreover, for each $j = 1, \dots, d$, if $(\widehat{\lambda}_{t,j}, \widehat{v}_{t,j}^{(W)})$ is the exact jth eigenpair of W_t with $\|\widehat{v}_{t,j}^{(W)}\| = 1$, then one can choose a representative $\widehat{v}_{t+1,j}^{(W),\mathrm{ex}}$ of the exact jth eigendirection of W_{t+1} such that

$$\widehat{v}_{t,j}^{(W)\top}\big(\widehat{v}_{t+1,j}^{(W),\mathrm{ex}} - \widehat{v}_{t,j}^{(W)}\big) = 0$$

and

$$\begin{aligned}
\widehat{\lambda}_{t+1,j}^{\mathrm{ex}} &= \widehat{\lambda}_{t,j} + \widehat{v}_{t,j}^{(W)\top}\Delta W_t \widehat{v}_{t,j}^{(W)} + O_p(t^{-2}),\\
\widehat{v}_{t+1,j}^{(W),\mathrm{ex}} &= \widehat{v}_{t,j}^{(W)} + (\widehat{\lambda}_{t,j} I_p - W_t)^{+}\Big\{\Delta W_t \widehat{v}_{t,j}^{(W)} - (\widehat{v}_{t,j}^{(W)\top}\Delta W_t \widehat{v}_{t,j}^{(W)})\widehat{v}_{t,j}^{(W)}\Big\} + O_p(t^{-2}).
\end{aligned}$$

Proof. The bound $\|\Delta D_t\|_{\mathrm{op}} = O_p(t^{-1})$ follows from (30), and $\|\Delta S_t\|_{\mathrm{op}} = O_p(t^{-1})$ follows from (34). On the event where S_t and S_{t+1} lie in a fixed compact subset of $\mathbb{S}_{++}^p$, Lemma 11 gives

$$\|S_{t+1}^{-1/2} - S_t^{-1/2}\|_{\mathrm{op}} \le C\|\Delta S_t\|_{\mathrm{op}}.$$

Combining this with the uniform boundedness in probability of D_t gives

$$\|\Delta W_t\|_{\mathrm{op}} = \|S_{t+1}^{-1/2} D_{t+1} S_{t+1}^{-1/2} - S_t^{-1/2} D_t S_t^{-1/2}\|_{\mathrm{op}} = O_p(t^{-1}).$$

Since the jth positive eigenvalue of $M^\star$ is simple and the eigengap is positive, the exact eigenpair map for the symmetric matrix W is C^2 in a neighborhood of $M^\star$. The ordinary symmetric first-order perturbation expansion applied to $W_{t+1} = W_t + \Delta W_t$ therefore yields the two displayed expansions, with a remainder bounded by $C\|\Delta W_t\|_{\mathrm{op}}^2 = O_p(t^{-2})$. □

The asymptotic theory established here for the perturbation method is therefore the following. First, the exact generalized eigenspace of (D_t, S_t) is root-t consistent and converges almost surely by Theorem 13. Second, the sample generalized perturbation update is mathematically correct at first order, with a localized quadratic remainder, by Proposition 15. Third, the standardized perturbation step (53)-(54) is therefore justified as a one-step approximation to the exact next eigensystem. What is not asserted here is a separate almost-sure convergence theorem for the standalone iterated first-order perturbation update rule (53)-(54). Such a theorem would require an additional stability analysis of the repeated approximate-eigenpair map, which is distinct from the localized one-step perturbation argument proved above. For the asymptotic target of the ODR procedure, Theorem 13 already provides the exact almost sure limit.

The perturbation step for the generalized eigenproblem (D_t, S_t) can be viewed as the natural counterpart, on the original scale, of the usual first-order perturbation step for the S_t-standardized matrix $W_t = S_t^{-1/2} D_t S_t^{-1/2}$. This representation is convenient because it separates the effect of the varying metric S_t from the spectral perturbation itself, so the first-order analysis can be carried out using standard perturbation results for symmetric eigenproblems.

A.6 Online subspace estimation via stochastic approximation

The every-step projected generalized-gradient update analyzed in the theorem is

$$B_{t+1} = \mathcal{P}_{\mathrm{orth}}^{(S_{t+1})}\Big(B_t + \gamma_{t+1} S_{t+1}^{-1}\widehat{K}_{t+1}^{\star} B_t\Big), \qquad \gamma_{t+1} = \frac{c}{t+1}. \tag{55}$$

For the convergence analysis, it is convenient to work in standardized coordinates. We therefore consider the idealized one-step-lag update

$$\bar{V}_{t+1} = \mathcal{P}_{\mathrm{orth}}\Big(\bar{V}_t + \gamma_{t+1}\widetilde{W}_t \bar{V}_t\Big), \qquad \bar{V}_t^{\top}\bar{V}_t = I_d. \tag{56}$$

The associated estimator on the original scale is then defined by

$$\bar{B}_t = S_t^{-1/2}\bar{V}_t. \tag{57}$$

Write

$$h(V) = M^\star V - V(V^\top M^\star V), \qquad \mathcal{E}_\star = \{V_\star O : \, O \in \mathbb{O}(d)\},$$

where $\mathbb{O}(d) = \{O \in \mathbb{R}^{d\times d} : O^\top O = I_d\}$ is the orthogonal group. For $V \in \mathbb{R}^{p\times d}$, write

$$\operatorname{dist}(V, \mathcal{E}_\star) = \inf_{O\in\mathbb{O}(d)} \|V - V_\star O\|_{\mathrm{F}}.$$

Lemma 17. *Let $V \in \mathbb{R}^{p\times d}$ satisfy $V^\top V = I_d$, and define $\mathcal{P}_{\mathrm{orth}}(Z) = Z(Z^\top Z)^{-1/2}$ in a neighborhood of the Stiefel manifold. Then, for every perturbation $E \in \mathbb{R}^{p\times d}$ with $\|E\|_{\mathrm{F}}$ sufficiently small,*

$$\mathcal{P}_{\mathrm{orth}}(V + E) = V + E - \frac{1}{2}V\big(V^\top E + E^\top V\big) + R(V, E), \tag{58}$$

where $\|R(V, E)\|_{\mathrm{F}} \le C\|E\|_{\mathrm{F}}^2$ uniformly for V on the Stiefel manifold. In particular, if A is symmetric and $E = \gamma AV$, then

$$\mathcal{P}_{\mathrm{orth}}(V + \gamma AV) = V + \gamma\{AV - V(V^\top AV)\} + R_\gamma(V, A), \tag{59}$$

$$\|R_\gamma(V, A)\|_{\mathrm{F}} \le C\gamma^2\|A\|_{\mathrm{op}}^2. \tag{60}$$

Proof. Let $Z = V + E$. Since $V^\top V = I_d$,

$$Z^\top Z = I_d + H, \qquad H = V^\top E + E^\top V + E^\top E.$$

For small $\|E\|_{\mathrm{F}}$, the Taylor expansion of the inverse square root gives

$$(I_d + H)^{-1/2} = I_d - \frac{1}{2}H + O(\|H\|_{\mathrm{F}}^2) = I_d - \frac{1}{2}\big(V^\top E + E^\top V\big) + O(\|E\|_{\mathrm{F}}^2).$$

Multiplying by $Z = V + E$ yields (58). If $E = \gamma AV$ and A is symmetric, then $V^\top E + E^\top V = 2\gamma V^\top AV$, which proves (59). □

Lemma 18. *Let $\{\xi_t\}$ be an adapted matrix-valued sequence such that*

$$E\|\xi_t\|_{\mathrm{F}} \le Ct^{-1/2}.$$

If $\gamma_t = c/t$, then

$$\sum_{t=1}^{\infty} \gamma_t\|\xi_t\|_{\mathrm{F}} < \infty \qquad \textit{almost surely.} \tag{61}$$

Consequently,

$$\sup_{m\ge k}\Big\|\sum_{i=k}^{m}\gamma_i\xi_i\Big\|_{\mathrm{F}} \longrightarrow 0 \qquad \textit{almost surely as } k \to \infty. \tag{62}$$

Proof. By Tonelli's theorem,

$$E\sum_{t=1}^{\infty}\gamma_t\|\xi_t\|_{\mathrm{F}} = \sum_{t=1}^{\infty}\gamma_t E\|\xi_t\|_{\mathrm{F}} \le C\sum_{t=1}^{\infty} t^{-3/2} < \infty.$$

Therefore (61) holds almost surely. The tail bound (62) follows from

$$\sup_{m\ge k}\Big\|\sum_{i=k}^{m}\gamma_i\xi_i\Big\|_{\mathrm{F}} \le \sum_{i=k}^{\infty}\gamma_i\|\xi_i\|_{\mathrm{F}}.$$

□

Lemma 19. *Assume $\lambda_d > \lambda_{d+1}$. Let*

$$\begin{aligned}\mathcal{V}_{p,d} &= \{V \in \mathbb{R}^{p\times d} : V^\top V = I_d\},\\ L(V) &= \operatorname{tr}(V^\top M^\star V),\\ \mathcal{E}_\star &= \{V_\star O :\ O \in \mathbb{O}(d)\}.\end{aligned}$$

Then the vector field

$$h(V) = M^\star V - V(V^\top M^\star V)$$

is the Riemannian gradient of L on $\mathcal{V}_{p,d}$. Along every solution of

$$\dot V = h(V), \qquad V(0) \in \mathcal{V}_{p,d},$$

we have

$$\frac{d}{dt}L(V(t)) = 2\big\|(I_p - V(t)V(t)^\top)M^\star V(t)\big\|_{\mathrm{F}}^2 \geq 0.$$

Moreover, $L(VO) = L(V)$ for every $O \in \mathbb{O}(d)$, so L descends to a smooth function on the Grassmann manifold $\mathcal{G}_{p,d}$. By the Ky Fan maximum principle, the descended function has a unique maximizer, namely the leading d-dimensional eigenspace of $M^\star$, and this maximizer is isolated because $\lambda_d > \lambda_{d+1}$. Consequently, $\mathcal{E}_\star$ is a compact locally asymptotically stable invariant set of the ODE.

Proof. For $V \in \mathcal{V}_{p,d}$,

$$\begin{aligned}h(V) &= M^\star V - V(V^\top M^\star V)\\ &= (I_p - VV^\top)M^\star V,\end{aligned}$$

which is the orthogonal projection of the Euclidean gradient $2M^\star V$ onto the tangent space of $\mathcal{V}_{p,d}$ at V. Hence h is the Riemannian gradient of L up to the irrelevant factor 2. Therefore,

$$\begin{aligned}\frac{d}{dt}L(V(t)) &= 2\operatorname{tr}\big(V(t)^\top M^\star \dot V(t)\big)\\ &= 2\operatorname{tr}\Big(V(t)^\top M^\star (I_p - V(t)V(t)^\top)M^\star V(t)\Big)\\ &= 2\big\|(I_p - V(t)V(t)^\top)M^\star V(t)\big\|_{\mathrm{F}}^2.\end{aligned}$$

The final claim now follows from the gradient-flow structure on the Grassmann manifold, the Ky Fan characterization of the leading eigenspace, and the eigengap $\lambda_d > \lambda_{d+1}$, which makes the maximizer isolated and strict on $\mathcal{G}_{p,d}$. □

Theorem 20. *Assume (A1)-(A6). Let $\bar V_t$ be generated by* (56)*, and define $\bar B_t$ by* (57)*. Consider the ODE*

$$\dot V = h(V), \qquad V^\top V = I_d. \tag{63}$$

Let $\mathcal{B}_\star$ denote the basin of attraction of $\mathcal{E}_\star$ for this ODE. Assume that, with probability one, the iterates eventually remain in a compact subset of this basin. That is,

$$\mathbb{P}\Big\{\exists\, T < \infty,\ \exists\, \eta > 0 \text{ such that } \operatorname{dist}(\bar V_t, \mathcal{B}_\star^c) \geq \eta \text{ for all } t \geq T\Big\} = 1.$$

Then

$$\operatorname{dist}\big(\bar V_t, \mathcal{E}_\star\big) \longrightarrow 0 \qquad \text{almost surely,} \tag{64}$$

and hence

$$\operatorname{span}(\bar B_t) \longrightarrow \mathcal{S}_{Y|X} \qquad \text{almost surely.} \tag{65}$$

Proof. Set $E_t = \gamma_{t+1}\widetilde{W}_t\bar{V}_t$. By (33), for every $\varepsilon > 0$,

$$\sum_{t=1}^{\infty}\mathbb{P}\big(\|E_t\|_{\mathrm{F}} > \varepsilon\big) \le \frac{1}{\varepsilon^2}\sum_{t=1}^{\infty}\gamma_{t+1}^2 E\|\widetilde{W}_t\bar{V}_t\|_{\mathrm{F}}^2$$
$$\le \frac{d}{\varepsilon^2}\sum_{t=1}^{\infty}\gamma_{t+1}^2 E\|\widetilde{W}_t\|_{\mathrm{op}}^2 < \infty.$$

Hence $\|E_t\|_{\mathrm{F}} \to 0$ almost surely by the Borel-Cantelli lemma, and Lemma 17 applies for all sufficiently large t on an event of probability one. Using (59), we obtain

$$\bar{V}_{t+1} = \bar{V}_t + \gamma_{t+1}\Big\{\widetilde{W}_t\bar{V}_t - \bar{V}_t\big(\bar{V}_t^{\top}\widetilde{W}_t\bar{V}_t\big)\Big\} + r_t, \tag{66}$$

with

$$\|r_t\|_{\mathrm{F}} \le C\gamma_{t+1}^2\|\widetilde{W}_t\|_{\mathrm{op}}^2. \tag{67}$$

Since

$$E\sum_{t=1}^{\infty}\|r_t\|_{\mathrm{F}} \le C\sum_{t=1}^{\infty}\gamma_{t+1}^2 E\|\widetilde{W}_t\|_{\mathrm{op}}^2 < \infty$$

by (33), we have $\sum_t \|r_t\|_{\mathrm{F}} < \infty$ almost surely.

Next define

$$\xi_t = (\widetilde{W}_t - M^\star)\bar{V}_t - \bar{V}_t\Big\{\bar{V}_t^{\top}(\widetilde{W}_t - M^\star)\bar{V}_t\Big\}.$$

Because $\bar{V}_t \in \mathcal{V}_{p,d} = \{V : V^{\top}V = I_d\}$,

$$\|\xi_t\|_{\mathrm{F}} \le 2\sqrt{d}\,\|\widetilde{W}_t - M^\star\|_{\mathrm{op}}.$$

Hence (32) implies $E\|\xi_t\|_{\mathrm{F}} \le Ct^{-1/2}$. Lemma 18 therefore yields

$$\sum_{t=1}^{\infty}\gamma_{t+1}\|\xi_t\|_{\mathrm{F}} < \infty \qquad \text{almost surely.} \tag{68}$$

Combining (66), (67), and (68), we see that $\{\bar{V}_t\}$ is an asymptotic pseudo-trajectory of the ODE (63).

By Lemma 19, the ODE (63) has a compact locally asymptotically stable invariant set $\mathcal{E}_\star$. Since $\{\bar{V}_t\}$ is an asymptotic pseudo-trajectory and is eventually contained in a compact subset of the basin of attraction of $\mathcal{E}_\star$, the stochastic-approximation theorem of Kushner and Clark [16] implies that

$$\operatorname{dist}(\bar{V}_t, \mathcal{E}_\star) \longrightarrow 0 \qquad \text{almost surely.}$$

Finally, $\bar{B}_t = S_t^{-1/2}\bar{V}_t$ and $S_t^{-1/2} \to \Sigma^{-1/2}$ almost surely, so (65) follows. □

Proposition 21. *Assume (A1)-(A6). Let B_t be generated by the implemented update* (55)*, and define $V_t = S_t^{1/2}B_t$. Then*

$$V_{t+1} = \mathcal{P}_{\mathrm{orth}}\Big(V_t + \gamma_{t+1}\widetilde{W}_t V_t + e_t\Big),$$

where the perturbation sequence admits a decomposition

$$e_t = a_t + m_t + b_t \tag{69}$$

such that

$$\sum_{t=1}^{\infty}\|a_t\|_{\mathrm{F}} < \infty \text{ a.s.}, \qquad \sum_{t=1}^{\infty} m_t \text{ converges a.s.}, \qquad \sum_{t=1}^{\infty}\|b_t\|_{\mathrm{F}} < \infty \text{ a.s.} \tag{70}$$

Moreover,

$$\sum_{t=1}^{\infty}\|e_t\|_{\mathrm{F}}^2 < \infty \qquad \text{a.s.} \tag{71}$$

Consequently, the partial sums of e_t converge almost surely. In particular, $e_t \to 0$ almost surely and

$$S_{t+1}^{1/2}S_t^{-1/2} - I_p = O_p(t^{-1}). \tag{72}$$

Proof. Using $\mathcal{P}_{\mathrm{orth}}^{(S)}(B) = S^{-1/2}\mathcal{P}_{\mathrm{orth}}(S^{1/2}B)$ and $B_t = S_t^{-1/2}V_t$, we obtain the exact identity

$$V_{t+1} = \mathcal{P}_{\mathrm{orth}}\Big(R_tV_t + \gamma_{t+1}\widetilde{W}_{t+1}R_tV_t\Big), \qquad R_t = S_{t+1}^{1/2}S_t^{-1/2}.$$

Hence

$$e_t = (R_t - I_p)V_t + \gamma_{t+1}(\widetilde{W}_{t+1} - \widetilde{W}_t)V_t + \gamma_{t+1}\widetilde{W}_{t+1}(R_t - I_p)V_t.$$

From (37),

$$E\|\Delta S_t\|_{\mathrm{F}}^2 \le Ct^{-2}.$$

Moreover,

$$\begin{aligned} E(\Delta S_t \mid \mathcal{F}_t) &= -\frac{1}{t+1}\widehat{\Sigma}_t + \frac{t}{(t+1)^2}\Sigma + \frac{t}{(t+1)^2}(\widehat{\mu}_t - \mu)(\widehat{\mu}_t - \mu)^\top - \frac{\rho_0}{t(t+1)}I_p \\ &= \frac{1}{t+1}(\Sigma - \widehat{\Sigma}_t) - \frac{1}{(t+1)^2}\Sigma + \frac{t}{(t+1)^2}(\widehat{\mu}_t - \mu)(\widehat{\mu}_t - \mu)^\top - \frac{\rho_0}{t(t+1)}I_p. \end{aligned}$$

Therefore,

$$E\big\|E(\Delta S_t \mid \mathcal{F}_t)\big\|_{\mathrm{op}} \le Ct^{-3/2}, \tag{73}$$

because $E\|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} = O(t^{-1/2})$ and $E\|\widehat{\mu}_t - \mu\|^2 = O(t^{-1})$.

Fix the compact neighborhood

$$\mathcal{K} = \Big\{S \succ 0 :\ \|S - \Sigma\|_{\mathrm{op}} \le \tfrac{1}{2}\lambda_{\min}(\Sigma)\Big\},$$

and let $\mathbf{1}_t = \mathbf{1}\{S_t \in \mathcal{K}\}$. Since $S_t \to \Sigma$ almost surely and $\Delta S_t \to 0$ almost surely, we have $\mathbf{1}_t = 1$ eventually almost surely and the second-order Taylor expansion of the square-root map is valid for all sufficiently large t. After absorbing the finitely many exceptional indices into u_t below, we may therefore assume that there exist $\mathcal{F}_t$-measurable linear operators $\mathcal{L}_t$ and remainders H_t such that

$$\mathbf{1}_t(R_t - I_p) = \mathbf{1}_t\mathcal{L}_t(\Delta S_t) + H_t, \qquad \|\mathcal{L}_t\| \le C, \qquad \|H_t\|_{\mathrm{op}} \le C\mathbf{1}_t\|\Delta S_t\|_{\mathrm{op}}^2. \tag{74}$$

Now set

$$\begin{aligned} u_t &= (1 - \mathbf{1}_t)(R_t - I_p)V_t, \\ a_t &= \mathbf{1}_t\mathcal{L}_t\big(E(\Delta S_t \mid \mathcal{F}_t)\big)V_t, \\ m_t &= \mathbf{1}_t\mathcal{L}_t\Big(\Delta S_t - E(\Delta S_t \mid \mathcal{F}_t)\Big)V_t, \\ b_t &= H_tV_t + u_t + \gamma_{t+1}(\widetilde{W}_{t+1} - \widetilde{W}_t)V_t + \gamma_{t+1}\widetilde{W}_{t+1}(R_t - I_p)V_t. \end{aligned}$$

Then (69) holds with $e_t = a_t + m_t + b_t$.

We now verify (70). Since $\|V_t\|_{\mathrm{op}} = 1$ and $\|\mathcal{L}_t\| \le C$ on $\mathcal{K}$,

$$\sum_{t=1}^{\infty}\|a_t\|_{\mathrm{F}} \le C\sum_{t=1}^{\infty}\mathbf{1}_t\big\|E(\Delta S_t \mid \mathcal{F}_t)\big\|_{\mathrm{op}}.$$

By Tonelli's theorem and (73), the right-hand side is almost surely finite. Hence $\sum_t \|a_t\|_{\mathrm{F}} < \infty$ almost surely.

Next, m_t is a martingale difference sequence because $\mathbf{1}_t$, $\mathcal{L}_t$, and V_t are $\mathcal{F}_t$-measurable. Moreover,

$$E\big(\|m_t\|_{\mathrm{F}}^2 \mid \mathcal{F}_t\big) \le C\mathbf{1}_tE\Big(\big\|\Delta S_t - E(\Delta S_t \mid \mathcal{F}_t)\big\|_{\mathrm{F}}^2 \mid \mathcal{F}_t\Big).$$

Taking expectations and using (34) yields

$$E\sum_{t=1}^{\infty}E\big(\|m_t\|_{\mathrm{F}}^2 \mid \mathcal{F}_t\big) \le C\sum_{t=1}^{\infty}E\|\Delta S_t\|_{\mathrm{F}}^2 < \infty.$$

Hence $\sum_t E(\|m_t\|_{\mathrm{F}}^2 \mid \mathcal{F}_t) < \infty$ almost surely, and the martingale convergence theorem implies that $\sum_t m_t$ converges almost surely.

For b_t, the term u_t is nonzero only finitely many times almost surely, because $\mathbf{1}_t = 1$ eventually. Also,

$$\sum_{t=1}^{\infty} \|H_t V_t\|_{\mathrm{F}} \le C \sum_{t=1}^{\infty} \|\Delta S_t\|_{\mathrm{F}}^2,$$

which is almost surely finite by Tonelli and (34). Furthermore,

$$E \sum_{t=1}^{\infty} \gamma_{t+1} \|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} \le \sum_{t=1}^{\infty} \gamma_{t+1} E\|\widetilde{W}_{t+1} - \widetilde{W}_t\|_{\mathrm{op}} < \infty$$

by (35), so this series is absolutely summable almost surely by Tonelli's theorem.

Finally, from the explicit representation (37) and (A2),

$$E\|\Delta S_t\|_{\mathrm{F}}^4 \le C t^{-4}.$$

On the event $\{\mathbf{1}_t = 1\}$,

$$\|R_t - I_p\|_{\mathrm{op}} \le C\|\Delta S_t\|_{\mathrm{op}} + C\|\Delta S_t\|_{\mathrm{op}}^2.$$

The contribution from $\{\mathbf{1}_t = 0\}$ is finite almost surely, because $\mathbf{1}_t = 1$ eventually almost surely. Hence, by Tonelli's theorem, Cauchy-Schwarz, (33), (34), and the fourth-moment bound above,

$$\begin{aligned} E \sum_{t=1}^{\infty} \gamma_{t+1} \|\widetilde{W}_{t+1}(R_t - I_p) V_t\|_{\mathrm{F}} \mathbf{1}_t &\le C \sum_{t=1}^{\infty} \gamma_{t+1} \big\{E\|\widetilde{W}_{t+1}\|_{\mathrm{op}}^2\big\}^{1/2} \big\{E\|\Delta S_t\|_{\mathrm{F}}^2\big\}^{1/2} \\ &\quad + C \sum_{t=1}^{\infty} \gamma_{t+1} \big\{E\|\widetilde{W}_{t+1}\|_{\mathrm{op}}^2\big\}^{1/2} \big\{E\|\Delta S_t\|_{\mathrm{F}}^4\big\}^{1/2} < \infty. \end{aligned}$$

Thus $\sum_t \|b_t\|_{\mathrm{F}} < \infty$ almost surely. This proves (70). The proof above also gives

$$E \sum_{t=1}^{\infty} \|m_t\|_{\mathrm{F}}^2 < \infty.$$

Hence $\sum_t \|m_t\|_{\mathrm{F}}^2 < \infty$ almost surely. Since $\sum_t \|a_t\|_{\mathrm{F}} < \infty$ and $\sum_t \|b_t\|_{\mathrm{F}} < \infty$ almost surely, we obtain (71). Since $\Delta S_t = O_p(t^{-1})$ by (34), (72) follows from (74). □

Theorem 22 (Local almost-sure convergence under basin containment)**.** *Assume (A1)-(A6). Let B_t be generated by* (55)*, and set $V_t = S_t^{1/2} B_t$. Let $\mathcal{B}_\star$ be the basin of attraction of $\mathcal{E}_\star$ for the ODE* (63)*. Assume that, with probability one, the transformed iterates eventually remain in a compact subset of this basin. That is,*

$$\mathbb{P}\Big\{\exists\, T < \infty,\ \exists\, \eta > 0 \text{ such that } \operatorname{dist}(V_t, \mathcal{B}_\star^c) \ge \eta \text{ for all } t \ge T\Big\} = 1.$$

Then

$$\operatorname{span}(B_t) \longrightarrow \mathcal{S}_{Y|X} \qquad \text{almost surely.} \tag{75}$$

Proof. By Proposition 21,

$$V_{t+1} = \mathcal{P}_{\mathrm{orth}}\Big(V_t + \gamma_{t+1}\widetilde{W}_t V_t + e_t\Big),$$

where $e_t \to 0$ almost surely and $\sum_t \|e_t\|_{\mathrm{F}}^2 < \infty$ almost surely. Also,

$$\sum_{t=1}^{\infty} \gamma_{t+1}^2 \|\widetilde{W}_t\|_{\mathrm{op}}^2 < \infty \qquad \text{almost surely}$$

by (33) and Tonelli's theorem. Hence $\|\gamma_{t+1}\widetilde{W}_t V_t + e_t\|_{\mathrm{F}} \to 0$ almost surely, and Lemma 17 applies for all sufficiently large t.

Let

$$\mathrm{St}(p,d) = \{V \in \mathbb{R}^{p\times d} : V^\top V = I_d\}.$$

Let $V_t \in \mathrm{St}(p,d)$ denote the orthonormalized iterate at time t. For $V \in \mathrm{St}(p,d)$ and $E \in \mathbb{R}^{p\times d}$, define

$$\Pi_V(E) = E - \frac{1}{2}V\big(V^\top E + E^\top V\big).$$

Equivalently, $\Pi_V(E)$ is the orthogonal projection of E onto the tangent space $T_V\mathrm{St}(p,d)$.

Using Lemma 17 with

$$E_t = \gamma_{t+1}\widetilde{W}_t V_t + e_t,$$

we obtain

$$V_{t+1} = V_t + \gamma_{t+1}\Big\{\widetilde{W}_t V_t - V_t(V_t^\top \widetilde{W}_t V_t)\Big\} + \Pi_{V_t}(e_t) + r_t^{\mathrm{impl}},$$

where r_t^{impl} is the remainder induced by the orthonormalization step, defined by

$$r_t^{\mathrm{impl}} = V_{t+1} - V_t - \gamma_{t+1}\Big\{\widetilde{W}_t V_t - V_t(V_t^\top \widetilde{W}_t V_t)\Big\} - \Pi_{V_t}(e_t).$$

Moreover,

$$\|r_t^{\mathrm{impl}}\|_{\mathrm{F}} \le C\Big(\gamma_{t+1}^2\|\widetilde{W}_t\|_{\mathrm{op}}^2 + \|e_t\|_{\mathrm{F}}^2 + \gamma_{t+1}\|\widetilde{W}_t\|_{\mathrm{op}}\|e_t\|_{\mathrm{F}}\Big).$$

By Cauchy-Schwarz, the three terms on the right are summable almost surely. Therefore

$$\sum_{t=1}^{\infty}\|r_t^{\mathrm{impl}}\|_{\mathrm{F}} < \infty \qquad \text{almost surely.}$$

Next decompose $e_t = a_t + m_t + b_t$ as in Proposition 21. The terms $\sum_t \Pi_{V_t}(a_t)$ and $\sum_t \Pi_{V_t}(b_t)$ converge absolutely almost surely. Moreover, Π_{V_t} is $\mathcal{F}_t$-measurable and uniformly bounded, so $\{\Pi_{V_t}(m_t)\}$ is a martingale difference sequence with square-summable conditional second moments. Hence

$$\sum_{t=1}^{\infty}\Pi_{V_t}(m_t)$$

converges almost surely. Thus the partial sums of $\Pi_{V_t}(e_t)$ converge almost surely.

Finally, write

$$\xi_t = (\widetilde{W}_t - M^\star)V_t - V_t\{V_t^\top(\widetilde{W}_t - M^\star)V_t\}.$$

As in the proof of Theorem 20, (32) and Lemma 18 imply

$$\sum_{t=1}^{\infty}\gamma_{t+1}\|\xi_t\|_{\mathrm{F}} < \infty \qquad \text{almost surely.}$$

Therefore $\{V_t\}$ is an asymptotic pseudo-trajectory of the ODE (63). By the eventual compact-containment assumption inside $\mathcal{B}_\star$, the same stochastic-approximation argument used in Theorem 20 gives $\mathrm{dist}(V_t, \mathcal{E}_\star) \to 0$ almost surely. Since $B_t = S_t^{-1/2}V_t$ and $S_t^{-1/2} \to \Sigma^{-1/2}$ almost surely, (75) follows. □

A.7 Structural dimension selection by a weighted-bootstrap ladle method

For order determination it is convenient to work with the direct-kernel standardized matrix

$$\mathcal{M}_t = \widetilde{W}_t = S_t^{-1/2}\widehat{K}_t^\star S_t^{-1/2}.$$

This keeps the order-determination statistic as a smooth function of the time-t online moment and RLS summaries. As noted above, the notation $\mathcal{M}_t$ is identical to $\widetilde{W}_t$. Here we use $\mathcal{M}_t$ only to emphasize the order-selection role of the direct-kernel matrix. Any alternative order-determination matrix may be used in place of $\widehat{K}_t^\star$ once it is shown to admit the same first-order expansion stated below.

Write the eigendecomposition of $\mathcal{M}_t$ as

$$\mathcal{M}_t = \widehat{U}_t^{(\mathcal{M})}\widehat{\Xi}_t^{(\mathcal{M})}\widehat{U}_t^{(\mathcal{M})\top}, \qquad \widehat{\Xi}_t^{(\mathcal{M})} = \mathrm{diag}(\widehat{\nu}_{t,1},\dots,\widehat{\nu}_{t,p}), \qquad \widehat{\nu}_{t,1} \geq \cdots \geq \widehat{\nu}_{t,p},$$

where $\widehat{U}_t^{(\mathcal{M})} = (\widehat{u}_{t,1}^{(\mathcal{M})},\dots,\widehat{u}_{t,p}^{(\mathcal{M})})$ is orthogonal. The symbols $\widehat{\nu}_{t,j}$ and $\widehat{\Xi}_t^{(\mathcal{M})}$ are used here instead of $\widehat{\vartheta}_{t,j}$ and $\widehat{\Theta}_t$ to avoid confusion with the RLS coefficients $\widehat{\vartheta}_{t,h}$ and $\widehat{\Theta}_{t,h}$. For $k \geq 1$, let

$$\widehat{U}_{t,k}^{(\mathcal{M})} = (\widehat{u}_{t,1}^{(\mathcal{M})},\dots,\widehat{u}_{t,k}^{(\mathcal{M})}) \in \mathbb{R}^{p\times k}, \qquad \widehat{U}_{t,0}^{(\mathcal{M})} \text{ empty.}$$

For each bootstrap replicate b, let

$$\mathcal{M}_t^{*(b)} = (S_t^{*(b)})^{-1/2}\widehat{K}_t^{\star *(b)}(S_t^{*(b)})^{-1/2}. \tag{76}$$

Here the weighted-bootstrap quantities are defined as follows. For replicate b, let $w_{b,1},\dots,w_{b,t}$ be independent Poisson(1) weights, independent of the data, and set

$$N_t^{*(b)} = \sum_{i=1}^t w_{b,i}.$$

On the event $N_t^{*(b)} > 0$, define

$$\widehat{p}_{t,h}^{*(b)} = (N_t^{*(b)})^{-1}\sum_{i=1}^t w_{b,i}\delta_{i,h}, \quad \widehat{u}_{t,h}^{*(b)} = (N_t^{*(b)})^{-1}\sum_{i=1}^t w_{b,i}X_i\delta_{i,h},$$

$$\widehat{S}_{t,h}^{(2)*(b)} = (N_t^{*(b)})^{-1}\sum_{i=1}^t w_{b,i}X_iX_i^\top\delta_{i,h}, \quad \widehat{\mu}_t^{*(b)} = (N_t^{*(b)})^{-1}\sum_{i=1}^t w_{b,i}X_i,$$

$$\widehat{\Sigma}_t^{*(b)} = (N_t^{*(b)})^{-1}\sum_{i=1}^t w_{b,i}(X_i-\widehat{\mu}_t^{*(b)})(X_i-\widehat{\mu}_t^{*(b)})^\top.$$

The weighted RLS quantities are

$$G_t^{*(b)} = \lambda I_{p+1} + \sum_{i=1}^t w_{b,i}x_i^{\mathrm{e}}x_i^{\mathrm{e}\top}, \quad s_{t,h}^{*(b)} = \sum_{i=1}^t w_{b,i}x_i^{\mathrm{e}}\delta_{i,h}, \quad R_{t,h}^{*(b)} = \sum_{i=1}^t w_{b,i}x_i^{\mathrm{e}}X_i^\top\delta_{i,h},$$

$$\widehat{\vartheta}_{t,h}^{*(b)} = (G_t^{*(b)})^{-1}s_{t,h}^{*(b)}, \qquad \widehat{\Theta}_{t,h}^{*(b)} = (G_t^{*(b)})^{-1}R_{t,h}^{*(b)}.$$

Then $\widehat{A}_{t,h}^{*(b)}$, $\widehat{\mathcal{B}}_{t,h}^{*(b)}$, $\widetilde{A}_{t,h}^{*(b)}$, $\widetilde{\mathcal{B}}_{t,h}^{*(b)}$, and $\widehat{K}_t^{\star *(b)}$ are obtained from the same formulas as (3)-(4) and the displayed definition of $\widehat{K}_t^\star$, replacing each unweighted object by its weighted analogue. Finally,

$$S_t^{*(b)} = \widehat{\Sigma}_t^{*(b)} + \rho_t I_p.$$

On the event $N_t^{*(b)} = 0$, define the bootstrap statistic arbitrarily. This event has probability e^{-t} and is asymptotically negligible. Let $\widehat{U}_{t,k}^{(\mathcal{M})*(b)}$ denote the leading k orthonormal eigenvectors of $\mathcal{M}_t^{*(b)}$.

For column-orthonormal $U, V \in \mathbb{R}^{p\times k}$ define the subspace discrepancy

$$\mathrm{dist}(U,V) = 1 - \left|\det(U^\top V)\right| \in [0,1].$$

Let the ideal and Monte Carlo raw bootstrap variability terms be

$$f_{t,0}^{\mathrm{id}}(0) = 0, \qquad \widehat{f}_{t,0}(0) = 0,$$

$$f_{t,0}^{\mathrm{id}}(k) = E^*\left\{\mathrm{dist}(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(1)})\right\}, \qquad \widehat{f}_{t,0}(k) = \frac{1}{N_t}\sum_{b=1}^{N_t}\mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(b)}\big),$$

where $k = 1, \dots, q$ and $q \le p-1$ is a fixed search cap. N_t denotes the number of bootstrap replicates used at time t. Normalize them by

$$f_t^{\mathrm{id}}(k) = \frac{f_{t,0}^{\mathrm{id}}(k)}{1 + \sum_{j=0}^{q} f_{t,0}^{\mathrm{id}}(j)}, \qquad \widehat{f}_t(k) = \frac{\widehat{f}_{t,0}(k)}{1 + \sum_{j=0}^{q} \widehat{f}_{t,0}(j)}. \tag{77}$$

Define the eigenvalue term

$$\widehat{\phi}_t(k) = \frac{\widehat{\nu}_{t,k+1}^{+}}{1 + \sum_{j=1}^{q+1} \widehat{\nu}_{t,j}^{+}}, \qquad \widehat{\nu}_{t,j}^{+} = \max(\widehat{\nu}_{t,j}, 0), \qquad k = 0, \dots, q.$$

Proposition 23. *Assume (A1)-(A3). There exists a mean-zero symmetric matrix-valued random variable $\Psi(Z_1)$, with $Z_1 = (X_1, Y_1)$ and $E\|\Psi(Z_1)\|_{\mathrm{F}}^4 < \infty$, such that*

$$\mathcal{M}_t - M^\star = \frac{1}{t} \sum_{i=1}^{t} \Psi(Z_i) + R_t^{(M)}, \qquad \|R_t^{(M)}\|_{\mathrm{op}} = O_p(t^{-1}). \tag{78}$$

In particular,

$$\sqrt{t}\ \mathrm{vec}(\mathcal{M}_t - M^\star) \rightsquigarrow N\big(0, \Omega_M\big) \tag{79}$$

for some covariance matrix Ω_M. Here, $\rightsquigarrow$ denotes convergence in distribution as $t \to \infty$.

Proof. Let $\zeta(Z_i)$ be the finite-dimensional vector obtained by stacking the coordinates

$$X_i, \quad \mathrm{vec}(X_i X_i^\top), \quad \{\delta_{i,h}\}_{h=1}^{H}, \quad \{X_i \delta_{i,h}\}_{h=1}^{H}, \quad \{\mathrm{vec}(X_i X_i^\top \delta_{i,h})\}_{h=1}^{H},$$

$$\mathrm{vec}(x_i^{\mathrm{e}} x_i^{\mathrm{e}\top}), \quad \{x_i^{\mathrm{e}} \delta_{i,h}\}_{h=1}^{H}, \quad \{\mathrm{vec}(x_i^{\mathrm{e}} X_i^\top \delta_{i,h})\}_{h=1}^{H},$$

where $x_i^{\mathrm{e}} = (1, X_i^\top)^\top$. Write

$$\bar{\zeta}_t = \frac{1}{t} \sum_{i=1}^{t} \zeta(Z_i), \qquad \zeta_0 = E\{\zeta(Z_1)\}.$$

Under (A2), the vector $\zeta(Z_1)$ has finite fourth moment, so

$$\bar{\zeta}_t - \zeta_0 = \frac{1}{t} \sum_{i=1}^{t} \big\{\zeta(Z_i) - \zeta_0\big\}, \qquad \|\bar{\zeta}_t - \zeta_0\| = O_p(t^{-1/2}). \tag{80}$$

All raw online moment summaries are either linear functions of $\bar{\zeta}_t$ or smooth functions of those linear functions. In particular,

$$\widehat{\mu}_t,\ \widehat{\Sigma}_t,\ \{\widehat{p}_{t,h}, \widehat{u}_{t,h}, \widehat{S}_{t,h}^{(2)}\}_{h=1}^{H},\ \bar{G}_t,\ \{\bar{s}_{t,h}, \bar{R}_{t,h}\}_{h=1}^{H}$$

are smooth functions of $\bar{\zeta}_t$ and t^{-1}. Since $\Sigma \succ 0$ and $G = E(x_1^{\mathrm{e}} x_1^{\mathrm{e}\top}) \succ 0$, Lemma 11 and the smooth inverse-map calculus imply that there exists a neighborhood $\mathcal{N}$ of $(\zeta_0, 0)$ and a C^2 map

$$F : \mathcal{N} \subset \mathbb{R}^m \times \mathbb{R} \longrightarrow \mathbb{S}^p$$

for some finite m, such that

$$\mathcal{M}_t = F(\bar{\zeta}_t, t^{-1}), \qquad M^\star = F(\zeta_0, 0). \tag{81}$$

The second-order Taylor expansion of F at $(\zeta_0, 0)$ gives

$$\mathcal{M}_t - M^\star = \mathrm{D}_1 F(\zeta_0, 0)[\bar{\zeta}_t - \zeta_0] + t^{-1} \partial_2 F(\zeta_0, 0) + R_t^{(M)}, \tag{82}$$

$$\|R_t^{(M)}\|_{\mathrm{op}} \le C\big(\|\bar{\zeta}_t - \zeta_0\|^2 + t^{-2}\big) = O_p(t^{-1}), \tag{83}$$

where (83) follows from (80). Define

$$\Psi(Z_i) = \mathrm{D}_1 F(\zeta_0, 0)\big[\zeta(Z_i) - \zeta_0\big].$$

Then $E\{\Psi(Z_1)\} = 0$, $\Psi(Z_1)$ is symmetric because F takes values in $\mathbb{S}^p$, and $E\|\Psi(Z_1)\|_{\mathrm{F}}^4 < \infty$ because $\mathrm{D}_1 F(\zeta_0, 0)$ is a fixed linear map and $E\|\zeta(Z_1)\|^4 < \infty$. Absorbing the deterministic bias term $t^{-1}\partial_2 F(\zeta_0, 0)$ into the remainder yields (78). Finally, (79) follows from the multivariate central limit theorem applied to $t^{-1/2} \sum_{i=1}^{t} \mathrm{vec}\{\Psi(Z_i)\}$ together with (83) and Slutsky's theorem. □

Proposition 24. *Assume (A1)-(A3). For each bootstrap replicate b, let $\mathcal{M}_t^{*(b)}$ be defined by* (76)*. Then, conditionally on the data stream and in outer probability,*

$$\mathcal{M}_t^{*(b)} - \mathcal{M}_t = \frac{1}{t}\sum_{i=1}^{t}\big(w_{b,i}-1\big)\Psi(Z_i) + R_t^{*(b)}, \qquad \|R_t^{*(b)}\|_{\mathrm{op}} = o_{P^*}(t^{-1/2}), \tag{84}$$

where Ψ is the influence function from Proposition 23. Consequently, conditionally on the data,

$$\sqrt{t}\ \mathrm{vec}\big(\mathcal{M}_t^{*(b)} - \mathcal{M}_t\big) \rightsquigarrow^* N\big(0, \Omega_M\big) \tag{85}$$

in probability. Here, $\rightsquigarrow^*$ denotes weak convergence of the conditional bootstrap law given the data, in probability.

Proof. Fix one bootstrap replicate and suppress the replicate index b. Let

$$N_t^* = \sum_{i=1}^{t} w_i, \qquad \bar{\zeta}_t^* = (N_t^*)^{-1}\sum_{i=1}^{t} w_i\zeta(Z_i),$$

where $\zeta(Z_i)$ is the finite-dimensional vector introduced in the proof of Proposition 23. The weighted raw moments, weighted RLS objects, weighted kernel, and weighted stabilized metric are all smooth functions of $\bar{\zeta}_t^*$ and t^{-1}. Hence, with the same C^2 map F from (81), there is a C^2 map $\widetilde{F}$ such that, on the event $N_t^* > 0$,

$$\mathcal{M}_t^* = \widetilde{F}\big(\bar{\zeta}_t^*, (N_t^*)^{-1}, t^{-1}\big),$$
$$\mathcal{M}_t = \widetilde{F}\big(\bar{\zeta}_t, t^{-1}, t^{-1}\big).$$

Here the second scalar argument accounts for the RLS ridge term λ/N_t^*, while the third scalar argument accounts for the metric ridge term ρ_0/t. Since

$$(N_t^*)^{-1} - t^{-1} = O_{P^*}(t^{-3/2}),$$

the contribution of the random denominator in the RLS ridge is $o_{P^*}(t^{-1/2})$ and is absorbed into the bootstrap remainder. Since the weights are Poisson(1), conditionally on the data stream we have

$$N_t^* - t = \sum_{i=1}^{t}(w_i - 1) = O_{P^*}(t^{1/2}) \tag{86}$$

in probability.

Using

$$\bar{\zeta}_t^* - \bar{\zeta}_t = \frac{1}{N_t^*}\sum_{i=1}^{t}(w_i-1)\{\zeta(Z_i) - \bar{\zeta}_t\},$$

together with $N_t^* - t = O_{P^*}(t^{1/2})$, we obtain

$$\bar{\zeta}_t^* - \bar{\zeta}_t = \frac{1}{t}\sum_{i=1}^{t}(w_i-1)\{\zeta(Z_i) - \bar{\zeta}_t\} + r_t^*,$$
$$\|r_t^*\| = O_{P^*}(t^{-1})$$

in outer probability. In particular,

$$\|\bar{\zeta}_t^* - \bar{\zeta}_t\| = O_{P^*}(t^{-1/2}). \tag{87}$$

Applying the conditional second-order Taylor expansion of $\widetilde{F}$ at

$$(\bar{\zeta}_t, t^{-1}, t^{-1})$$

and using (87) together with

$$(N_t^*)^{-1} - t^{-1} = O_{P^*}(t^{-3/2}),$$

we obtain

$$\begin{aligned}\mathcal{M}_t^* - \mathcal{M}_t =& \mathrm{D}_1\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1})[\bar{\zeta}_t^* - \bar{\zeta}_t] \\ &+ \partial_2\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1})\big\{(N_t^*)^{-1} - t^{-1}\big\} + \widetilde{R}_t^*, \\ \|\widetilde{R}_t^*\|_{\mathrm{op}} \le C\Big(&\|\bar{\zeta}_t^* - \bar{\zeta}_t\|^2 + \big|(N_t^*)^{-1} - t^{-1}\big|^2\Big) = O_{P^*}(t^{-1}) = o_{P^*}(t^{-1/2}).\end{aligned}$$

Since

$$\partial_2\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1})\big\{(N_t^*)^{-1} - t^{-1}\big\} = O_{P^*}(t^{-3/2}),$$

this term is also absorbed into the bootstrap remainder. Therefore

$$\begin{aligned}\mathcal{M}_t^* - \mathcal{M}_t &= \mathrm{D}_1\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1})[\bar{\zeta}_t^* - \bar{\zeta}_t] + \widetilde{R}_t^*, \\ \|\widetilde{R}_t^*\|_{\mathrm{op}} &= o_{P^*}(t^{-1/2}).\end{aligned}$$

Because $\mathrm{D}_1\widetilde{F}$ is locally Lipschitz and $\bar{\zeta}_t \to \zeta_0$ in probability, we have

$$\big\|\mathrm{D}_1\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1}) - \mathrm{D}_1\widetilde{F}(\zeta_0, 0, 0)\big\| = O_p(t^{-1/2}).$$

Combining this with (87) shows that replacing

$$\mathrm{D}_1\widetilde{F}(\bar{\zeta}_t, t^{-1}, t^{-1})$$

by

$$\mathrm{D}_1\widetilde{F}(\zeta_0, 0, 0)$$

incurs only an $O_{P^*}(t^{-1})$ error. Using the definition $\Psi(Z_i) = \mathrm{D}_1F(\zeta_0, 0)[\zeta(Z_i) - \zeta_0]$ and the identity

$$\frac{1}{t}\sum_{i=1}^{t}(w_i - 1)\big\{\zeta(Z_i) - \bar{\zeta}_t\big\} = \frac{1}{t}\sum_{i=1}^{t}(w_i - 1)\big\{\zeta(Z_i) - \zeta_0\big\} - \Big(\frac{1}{t}\sum_{i=1}^{t}(w_i - 1)\Big)(\bar{\zeta}_t - \zeta_0),$$

where the last term is $O_{P^*}(t^{-1})$ by (86) and (80), we conclude that (84) holds.

Finally, conditionally on the data stream, the centered finite-dimensional weighted sum $t^{-1/2}\sum_{i=1}^{t}(w_i - 1)\,\mathrm{vec}\{\Psi(Z_i)\}$ satisfies the Poisson-multiplier central limit theorem with covariance Ω_M. The sample covariance converges in probability to Ω_M by the law of large numbers and $E\|\Psi(Z_1)\|_{\mathrm{F}}^4 < \infty$. Therefore (85) follows from (84) and Slutsky's theorem. □

Theorem 25. *Assume (A1)-(A5) and fix $q \in \{d, \dots, p-1\}$. Suppose the number of bootstrap replicates N_t (the replicate count, which is distinct from the per-replicate cumulative Poisson weight $N_t^{*(b)}$) satisfies*

$$N_t \longrightarrow \infty,$$

and that every active bootstrap replicate is constructed with a complete weighted history through time t. Assume further that, for each $k = d+1, \dots, q$, the ideal raw bootstrap variability satisfies

$$f_{t,0}^{\mathrm{id}}(k) \xrightarrow{P} c_k, \qquad c_k > 0. \tag{88}$$

Then

$$f_{t,0}^{\mathrm{id}}(k) \xrightarrow{P} 0, \qquad k = 0, 1, \dots, d, \tag{89}$$

$$\widehat{\phi}_t(k) \xrightarrow{P} \frac{\lambda_{k+1}}{1 + \sum_{j=1}^{d}\lambda_j}, \qquad k = 0, 1, \dots, d-1, \tag{90}$$

$$\widehat{\phi}_t(k) \xrightarrow{P} 0, \qquad k = d, d+1, \dots, q. \tag{91}$$

Consequently the ideal ladle criterion

$$g_t^{\mathrm{id}}(k) = f_t^{\mathrm{id}}(k) + \widehat{\phi}_t(k) \tag{92}$$

is uniquely minimized at $k = d$ with probability tending to one. Moreover,

$$\sup_{0\le k\le q} \left|\widehat{f}_t(k) - f_t^{\mathrm{id}}(k)\right| \xrightarrow{P} 0, \tag{93}$$

and hence the Monte Carlo ladle selector

$$\widehat{d}_t^{\mathrm{ladle}} = \arg\min_{0\le k\le q} \left\{\widehat{f}_t(k) + \widehat{\phi}_t(k)\right\}$$

is consistent. That is

$$\mathbb{P}\big(\widehat{d}_t^{\mathrm{ladle}} = d\big) \longrightarrow 1. \tag{94}$$

Proof. For $k \le d$, the leading k-dimensional eigenspace of $M^\star$ is isolated by a strict eigengap because $\lambda_k > \lambda_{k+1}$ for $k < d$ and $\lambda_d > \lambda_{d+1} = 0$ for $k = d$. Proposition 23 and Proposition 24, together with Davis-Kahan perturbation theory, imply that both $\widehat{U}_{t,k}^{(\mathcal{M})}$ and $\widehat{U}_{t,k}^{(\mathcal{M})*(1)}$ are consistent for the same population eigenspace.

Therefore

$$\mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(1)}\big) \xrightarrow{P^*} 0 \qquad \text{in probability.}$$

Since

$$0 \le \mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(1)}\big) \le 1,$$

for every fixed $\varepsilon > 0$ we have

$$\begin{aligned} f_{t,0}^{\mathrm{id}}(k) &= E^*\Big\{\mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(1)}\big)\Big\} \\ &\le \varepsilon + P^*\Big(\mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(1)}\big) > \varepsilon\Big). \end{aligned}$$

The second term converges to 0 in probability. Since $\varepsilon > 0$ is arbitrary, this proves (89).

Next, Proposition 23 shows that $\widehat{\nu}_{t,j}^{+} \to \lambda_j$ in probability for $j \le d$ and $\widehat{\nu}_{t,j}^{+} \to 0$ in probability for $j \ge d+1$. This proves (90) and (91). Hence, for $k < d$, the limiting eigenvalue term is strictly positive. For $k = d$, both the variability term and the eigenvalue term vanish. And for $k > d$, the ideal variability term is asymptotically strictly positive by (88) while the eigenvalue term vanishes. Therefore the ideal criterion (92) is uniquely minimized at $k = d$ with probability tending to one.

It remains to control the Monte Carlo approximation. Conditionally on the data stream, the variables $\mathrm{dist}(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(b)})$, $b = 1, \dots, N_t$, are i.i.d. and bounded in $[0,1]$. Because q is fixed and $N_t \to \infty$, Hoeffding's inequality and a union bound yield

$$\sup_{0\le k\le q} \left|\widehat{f}_{t,0}(k) - f_{t,0}^{\mathrm{id}}(k)\right| \xrightarrow{P} 0.$$

The normalization map in (77) is Lipschitz on $[0,1]^{q+1}$, so (93) follows. Since the ideal criterion is asymptotically uniquely minimized at d, the same is true for the Monte Carlo criterion, which proves (94). □

Theorem 25 corresponds to a growing-replicate construction in which the number of bootstrap replicates used for order selection is N_t and $N_t \to \infty$, with complete weighted histories for all active replicates. If the ladle criterion is evaluated only on a monitoring grid $\{t_m\}$ rather than at every time, it is enough to require $N_{t_m} \to \infty$ along that grid. A strictly one-pass implementation with a fixed replicate count B remains useful as a Monte Carlo approximation to the ideal bootstrap criterion, but the finite-B Monte Carlo error is not covered by (93) without an additional argument.

Condition (88) is the precise mathematical form of the ladle heuristic beyond the structural dimension. Once the bootstrap enters the zero-eigenvalue block, the additional directions are not identifiable, so the bootstrap subspace is expected to wander inside the null space rather than collapse onto one deterministic limit. In other words, the bootstrap basis undergoes a random rotation in the null block. Whenever this null-block limit is genuinely nondegenerate and is not supported on one fixed $(k-d)$-dimensional subspace, the constants c_k are automatically strictly positive.

The theorem above is stated for the direct-kernel statistic $\mathcal{M}_t$ because that is the method for which the first-order expansion closes cleanly under the moment and smoothness conditions already used in this section. If one prefers to use the instance-average order-determination matrix $S_t^{-1/2} D_t S_t^{-1/2}$ instead, the same ladle proof goes through once the corresponding influence-function expansion and weighted-bootstrap representation are established.

Theorem 9 establishes root-t accuracy of the online $K^\star$-kernel, Theorem 13 identifies the exact generalized signal subspace of (D_t, S_t), and Proposition 15 proves that the numerical generalized perturbation step is the correct first-order update under the Euclidean orthogonality constraint, with a localized quadratic remainder. Theorems 20 and 22 establish almost-sure convergence of the idealized and implemented generalized-gradient updates in the appropriate local basin, while Propositions 23 and 24, together with Theorem 25, justify the weighted-bootstrap ladle selector for the plug-in standardized kernel matrix. Thus the issues are addressed explicitly. The RLS one-step bounds use the exact Sherman–Morrison rank one identity and the deterministic bound (20) to avoid the rare ill-conditioned events in the L^2 bound. The moving metric $S_{t+1}^{1/2} S_t^{-1/2}$ is treated through a summable perturbation decomposition rather than suppressed. The gradient results are stated as local almost sure convergence theorems. And the structural-dimension method is backed by an explicit first-order expansion and conditional multiplier-bootstrap argument. We do not claim a separate almost-sure convergence theorem for the standalone iteration of the first-order perturbation update itself. That would require an additional stability lemma for the repeated approximate-eigenpair map beyond the one-step perturbation theory proved in Proposition 15.

B Simulation Studies

This section examines the finite-sample behavior of the proposed ODR procedure on nine simulated regression structures. The design covers a broad range of conditional dependencies, including (i) models in which the regression is driven by the first inverse conditional moment and therefore favors SIR, (ii) models in which the dependence is symmetric and therefore favors SAVE, and (iii) models in which symmetric dependencies are combined with a correlated covariance structure. Across all nine settings we compare the online estimators to their batch counterparts in terms of subspace recovery, computational cost, and structural dimension selection.

Throughout the simulation experiments, the response was divided into $H = 5$ slices. Each online method was initialized using the first 40% of the data stream. Thus, for a stream of length n, the initialization size was $t_0 = \lfloor 0.40n \rfloor$. For the online weighted-bootstrap ladle selector, the maximum search dimension was set to

$$q = \begin{cases} p - 2, & p < 10, \\ \lfloor p/\log(p) \rfloor, & p \geq 10, \end{cases}$$

where $\log$ denotes the natural logarithm. The number of bootstrap replicates in the ladle computation was fixed at $N_{\mathrm{boot}} = 10$. This fixed replicate count was used for the numerical implementation of Algorithm 2.

To measure subspace recovery on the original predictor scale we follow [6] and adapt their accuracy measure to the Σ-metric. For the true central-subspace basis B and its estimate $\widehat{B}$, we report

$$d(B, \widehat{B}) = 1 - \left| \det\!\big(B^\top \Sigma\, \widehat{B}\big) \right|.$$

This quantity takes values in $[0, 1]$, equals zero when $\widehat{B}$ spans the same Σ-orthonormal subspace as B, and approaches one when the two subspaces are nearly orthogonal in the Σ-metric. All reported values are averaged over 20 replications, with the corresponding standard deviations shown in parentheses.

We include six online competitors, arising from three moment-based SDR methods and two modes of standardization. The three methods are online SIR (OSIR,[6]), online sliced average variance estimation (OSAVE, [6]), and the proposed ODR. For each method we consider two versions, an ordinary version that operates on the empirically standardized predictor and a generalized version that operates on the original predictor under the Σ-metric. For each of the six method–standardization combinations, we run two subspace update procedures, a gradient descent (GD) method and a perturbation-based (PT) method. The six methods are labeled as follows.

- M1: Ordinary OSIR.

- M2: Generalized OSIR.
- M3: Ordinary OSAVE.
- M4: Generalized OSAVE.
- M5: Ordinary ODR.
- M6: Generalized ODR.

Throughout, the predictor X is drawn from a multivariate normal distribution and the error ε from a standard normal distribution. For each model we vary the dimension $p \in \{20, 50, 100\}$ and the sample size $n \in \{1000, 5000, 10000\}$. The number of slices is fixed throughout the experiments and chosen in line with standard SDR practice.

B.1 Models with a First-Moment Dependence

The first group of designs places the regression signal in the first inverse conditional moment, so that methods based on $E(X \mid Y)$ such as SIR are expected to perform strongly. Here d denotes the structural dimension, that is, the dimension of the central subspace under the corresponding data generating model. We consider three such models,

$$\begin{aligned}
&\text{Model 1 } (d=1): \quad Y = \beta_1^\top X + 0.2(\beta_1^\top X)^3 + \varepsilon,\\
&\text{Model 2 } (d=2): \quad Y = (\beta_2^\top X)^3 + 1.5(\beta_3^\top X) + 0.2\varepsilon,\\
&\text{Model 3 } (d=2): \quad Y = \frac{1}{\exp\{0.2\big((\beta_2^\top X)^3 + (\beta_3^\top X)\big)\}} + 0.2\varepsilon,
\end{aligned}$$

with $\beta_1 = \frac{1}{\sqrt{8}}(1,2,-1,1,-1,0,\ldots,0)^\top$, $\beta_2 = \frac{1}{\sqrt{10}}(1,-2,2,-1,0,\ldots,0)^\top$, and $\beta_3 = \frac{1}{\sqrt{10}}(2,1,-1,-2,0,\ldots,0)^\top$.

Tables 3-5 report the subspace distances for Models 1 through 3. Two qualitative trends are immediate. First, the subspace distance decreases monotonically with the sample size n, in line with the root-t consistency established in Section A. Second, at fixed n, the estimation problem becomes harder as the dimension p grows, which is expected given that the online kernel couples Hp^2 slice-wise summaries. SIR-based methods perform strongly in these first-moment models, as anticipated. Note that, the two ODR variants M5 and M6 are fully competitive with OSIR across all three models and attain comparable accuracy at $n = 10000$ even in the hardest case $p = 100$. In several cells of the tables, the generalized ODR variant M6 achieves the smallest subspace distance among all six methods, indicating that the additional second-moment information carried by ODR does not degrade performance in settings where only the first inverse moment is informative.

B.2 Models with a Symmetric Dependence

The second group of designs places the regression signal in the inverse conditional variance, so that the conditional mean structure carries little information about the central subspace. Such symmetric dependencies are a classical failure mode of first-moment SDR, and are precisely the setting in which DR is expected to provide a clear benefit. We consider

$$\begin{aligned}
&\text{Model 4 } (d=2): \quad Y = (\beta_1^\top X)^2 + 0.5(\beta_2^\top X)^2 + 0.3(\beta_1^\top X) + 0.2\varepsilon,\\
&\text{Model 5 } (d=2): \quad Y = (\beta_3^\top X)^2 + 0.7(\beta_4^\top X)^2 + 0.3(\beta_3^\top X)(\beta_4^\top X) + 0.2\varepsilon,\\
&\text{Model 6 } (d=2): \quad Y = \log\big(1 + (\beta_5^\top X)^2\big) + \frac{0.7}{1 + (\beta_6^\top X)^2} + 0.2\varepsilon,
\end{aligned}$$

with $\beta_1 = \frac{1}{\sqrt{8}}(1,2,-1,1,-1,0,\ldots,0)^\top$, $\beta_2 = \frac{1}{\sqrt{8}}(-1,2,1,-1,1,0,\ldots,0)^\top$, $\beta_3 = \frac{1}{\sqrt{10}}(1,-2,2,-1,0,\ldots,0)^\top$, $\beta_4 = \frac{1}{\sqrt{10}}(2,1,-1,-2,0,\ldots,0)^\top$, $\beta_5 = \frac{1}{\sqrt{2}}(1,1,0,\ldots,0)^\top$, and $\beta_6 = \frac{1}{\sqrt{2}}(1,-1,0,\ldots,0)^\top$. The predictor and noise distributions are as in the preceding subsection. Here again, d denotes the structural dimension.

The results for Models 4 through 6 are reported in Tables 6-8 and display the pattern predicted by SDR theory. Both OSIR variants, M1 and M2, fail to recover the central subspace in essentially every cell,

Table 3: Subspace distance for Model 1. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	**0.02** (0.01)	**0.02** (0.02)	**0.00** (0.00)	**0.00** (0.01)	**0.00** (0.00)	0.01 (0.05)
	M2	0.05 (0.01)	0.03 (0.01)	0.01 (0.00)	**0.00** (0.00)	**0.00** (0.00)	**0.00** (0.00)
	M3	0.19 (0.09)	0.19 (0.10)	0.01 (0.00)	0.01 (0.01)	0.01 (0.00)	0.01 (0.00)
	M4	0.44 (0.19)	0.19 (0.25)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	**0.00** (0.00)
	M5	0.04 (0.01)	0.08 (0.16)	0.01 (0.00)	0.01 (0.00)	**0.00** (0.00)	**0.00** (0.00)
	M6	0.04 (0.01)	0.03 (0.01)	0.01 (0.00)	**0.00** (0.00)	**0.00** (0.00)	**0.00** (0.00)
50	M1	**0.05** (0.01)	**0.04** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.00** (0.00)	**0.00** (0.00)
	M2	0.13 (0.02)	0.06 (0.01)	0.02 (0.00)	**0.01** (0.00)	0.01 (0.00)	0.01 (0.00)
	M3	0.83 (0.12)	0.83 (0.13)	0.04 (0.01)	0.04 (0.01)	0.02 (0.00)	0.02 (0.00)
	M4	0.86 (0.12)	0.82 (0.12)	0.08 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.00)
	M5	0.17 (0.05)	0.18 (0.05)	0.02 (0.00)	0.02 (0.00)	0.01 (0.00)	0.03 (0.08)
	M6	0.17 (0.05)	0.18 (0.26)	0.02 (0.00)	**0.01** (0.00)	0.01 (0.00)	0.01 (0.00)
100	M1	**0.11** (0.02)	**0.12** (0.04)	**0.02** (0.00)	**0.02** (0.00)	**0.01** (0.00)	**0.01** (0.00)
	M2	0.35 (0.16)	0.21 (0.26)	0.05 (0.01)	**0.02** (0.00)	0.02 (0.00)	**0.01** (0.00)
	M3	0.91 (0.07)	0.91 (0.07)	0.16 (0.04)	0.17 (0.04)	0.04 (0.01)	0.04 (0.01)
	M4	0.91 (0.07)	0.88 (0.09)	0.55 (0.20)	0.33 (0.38)	0.09 (0.01)	0.04 (0.01)
	M5	0.68 (0.20)	0.72 (0.19)	0.04 (0.01)	0.06 (0.06)	0.02 (0.00)	0.04 (0.06)
	M6	0.70 (0.19)	0.95 (0.05)	0.04 (0.01)	**0.02** (0.00)	**0.01** (0.00)	**0.01** (0.00)

Table 4: Subspace distance for Model 2. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.44 (0.17)	0.44 (0.17)	0.15 (0.06)	0.17 (0.09)	0.07 (0.02)	0.17 (0.26)
	M2	0.50 (0.20)	0.37 (0.19)	0.17 (0.06)	0.11 (0.04)	0.07 (0.02)	0.05 (0.01)
	M3	0.48 (0.21)	0.48 (0.22)	**0.03** (0.01)	0.03 (0.01)	0.02 (0.00)	0.02 (0.01)
	M4	0.66 (0.19)	0.48 (0.26)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	**0.01** (0.00)
	M5	**0.35** (0.22)	0.36 (0.24)	**0.03** (0.01)	0.03 (0.01)	**0.01** (0.00)	0.02 (0.01)
	M6	**0.35** (0.22)	**0.23** (0.30)	**0.03** (0.01)	**0.02** (0.00)	**0.01** (0.00)	**0.01** (0.00)
50	M1	**0.68** (0.13)	0.69 (0.13)	0.27 (0.06)	0.28 (0.09)	0.15 (0.03)	0.18 (0.16)
	M2	0.75 (0.13)	**0.65** (0.15)	0.31 (0.07)	0.21 (0.05)	0.16 (0.03)	0.11 (0.02)
	M3	0.93 (0.07)	0.93 (0.07)	0.13 (0.03)	0.13 (0.03)	0.05 (0.01)	0.09 (0.19)
	M4	0.96 (0.04)	0.95 (0.04)	0.23 (0.06)	0.10 (0.02)	0.07 (0.02)	0.04 (0.01)
	M5	0.69 (0.20)	0.69 (0.20)	**0.10** (0.02)	0.12 (0.07)	**0.04** (0.01)	0.04 (0.01)
	M6	0.69 (0.20)	0.84 (0.13)	**0.10** (0.02)	**0.05** (0.01)	**0.04** (0.01)	**0.02** (0.00)
100	M1	**0.84** (0.08)	0.84 (0.08)	0.40 (0.08)	0.42 (0.15)	0.28 (0.03)	0.30 (0.17)
	M2	0.90 (0.06)	**0.83** (0.11)	0.45 (0.08)	0.34 (0.06)	0.30 (0.03)	0.21 (0.02)
	M3	0.99 (0.01)	0.99 (0.01)	0.52 (0.18)	0.53 (0.18)	0.13 (0.02)	0.14 (0.03)
	M4	0.99 (0.01)	0.99 (0.01)	0.79 (0.14)	0.61 (0.29)	0.23 (0.03)	0.10 (0.01)
	M5	0.92 (0.06)	0.92 (0.06)	**0.30** (0.09)	0.35 (0.15)	**0.11** (0.02)	0.25 (0.30)
	M6	0.92 (0.05)	0.99 (0.01)	**0.30** (0.09)	**0.12** (0.02)	**0.11** (0.02)	**0.05** (0.01)

Table 5: Subspace distance for Model 3. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	**0.57** (0.18)	0.57 (0.18)	0.19 (0.09)	0.17 (0.08)	0.10 (0.03)	0.14 (0.18)
	M2	0.61 (0.21)	**0.49** (0.17)	0.20 (0.09)	0.13 (0.06)	0.10 (0.03)	0.07 (0.02)
	M3	0.77 (0.15)	0.76 (0.15)	0.14 (0.08)	0.14 (0.08)	0.05 (0.02)	0.05 (0.02)
	M4	0.84 (0.11)	0.77 (0.14)	0.18 (0.09)	0.09 (0.04)	0.06 (0.02)	0.04 (0.01)
	M5	0.73 (0.21)	0.72 (0.21)	**0.10** (0.05)	0.10 (0.05)	**0.04** (0.02)	0.04 (0.02)
	M6	0.73 (0.20)	0.62 (0.23)	**0.10** (0.05)	**0.04** (0.01)	**0.04** (0.02)	**0.02** (0.01)
50	M1	**0.78** (0.15)	0.78 (0.15)	0.40 (0.13)	0.46 (0.20)	0.22 (0.05)	0.26 (0.22)
	M2	0.85 (0.11)	**0.73** (0.17)	0.43 (0.14)	0.30 (0.06)	0.23 (0.06)	0.16 (0.03)
	M3	0.96 (0.03)	0.96 (0.03)	0.48 (0.19)	0.49 (0.20)	0.18 (0.05)	0.24 (0.19)
	M4	0.97 (0.02)	0.96 (0.03)	0.65 (0.17)	0.44 (0.27)	0.24 (0.09)	0.12 (0.03)
	M5	0.88 (0.11)	0.88 (0.10)	**0.38** (0.20)	0.39 (0.21)	**0.14** (0.04)	0.18 (0.20)
	M6	0.88 (0.10)	0.89 (0.08)	**0.38** (0.20)	**0.19** (0.17)	**0.14** (0.04)	**0.06** (0.01)
100	M1	**0.86** (0.07)	**0.86** (0.07)	**0.55** (0.13)	0.59 (0.17)	**0.35** (0.07)	0.39 (0.17)
	M2	0.92 (0.04)	0.89 (0.06)	0.59 (0.13)	**0.44** (0.07)	0.39 (0.08)	0.28 (0.06)
	M3	0.99 (0.01)	0.99 (0.01)	0.86 (0.09)	0.86 (0.09)	0.69 (0.16)	0.75 (0.19)
	M4	0.99 (0.01)	0.99 (0.01)	0.91 (0.06)	0.85 (0.12)	0.85 (0.12)	0.57 (0.22)
	M5	0.95 (0.06)	0.96 (0.05)	0.85 (0.11)	0.86 (0.11)	0.50 (0.18)	0.55 (0.21)
	M6	0.95 (0.05)	0.99 (0.01)	0.85 (0.11)	0.86 (0.09)	0.50 (0.18)	**0.23** (0.21)

with subspace distances hovering near 0.9 and showing no noticeable improvement as n increases. Because the conditional mean of X given Y is nearly non-informative in these symmetric designs, no amount of additional first-moment data can correct this limitation. The two OSAVE variants, M3 and M4, perform better than OSIR but remain visibly inferior to the ODR variants, reflecting the fact that OSAVE captures only the second-moment component of the directional-regression kernel. In contrast, the ODR variants M5 and M6 achieve small subspace distances across all three models and their accuracy improves steadily with n. Among all six methods the generalized ODR variant M6 attains the smallest subspace distance in a majority of cells, particularly in the larger-sample settings $n \in \{5000, 10000\}$ and at larger p, providing empirical support for the theoretical claim that the generalized metric produces a more stable recursive subspace estimator.

B.3 Models with Symmetric Dependence and a Correlated Covariance

The third group of designs combines the symmetric regression structure of the second group with a non-diagonal predictor covariance whose off-diagonal entries decay with the distance between variables. This setting is intended to stress-test online SDR procedures, in that the moving sample covariance $\widehat{\Sigma}_t$ now carries substantial off-diagonal signal that must be tracked accurately for the standardized slice blocks to be estimated consistently. We consider

$$\text{Model 7 } (d=2): \quad Y = (\beta_1^\top X)^2 + 0.7(\beta_2^\top X)^2 + 0.3(\beta_1^\top X)(\beta_2^\top X) + 0.2\varepsilon,$$

$$\text{Model 8 } (d=2): \quad Y = \log\big(1 + (\beta_3^\top X)^2 + 0.7|\beta_4^\top X|\big) + 0.2\varepsilon,$$

$$\text{Model 9 } (d=2): \quad Y = \frac{\exp\big((\beta_3^\top X)^2 + (\beta_4^\top X)^2\big)}{1 + (\beta_4^\top X)^2} + 0.2\varepsilon,$$

with $\beta_1 = \frac{1}{\sqrt{10}}(1, -2, 2, -1, 0, \ldots, 0)^\top$, $\beta_2 = \frac{1}{\sqrt{10}}(2, 1, -1, -2, 0, \ldots, 0)^\top$, $\beta_3 = \frac{1}{\sqrt{2}}(1, 1, 0, \ldots, 0)^\top$, and $\beta_4 = \frac{1}{\sqrt{2}}(1, -1, 0, \ldots, 0)^\top$. The predictor and noise distributions are as before. Specifically, $X \sim N_p(0, \Sigma)$, where $\Sigma_{ij} = \exp(-|i-j|)$ for $1 \le i, j \le p$. Here again, d denotes the structural dimension.

Table 6: Subspace distance for Model 4. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.91 (0.08)	0.91 (0.08)	0.87 (0.09)	0.88 (0.09)	0.85 (0.09)	0.83 (0.07)
	M2	0.91 (0.08)	0.92 (0.08)	0.86 (0.10)	0.85 (0.09)	0.86 (0.09)	0.86 (0.10)
	M3	0.11 (0.03)	0.11 (0.03)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	0.02 (0.04)
	M4	0.27 (0.08)	0.10 (0.03)	0.03 (0.01)	**0.01** (0.00)	**0.01** (0.00)	0.01 (0.00)
	M5	**0.09** (0.02)	0.09 (0.02)	**0.01** (0.00)	**0.01** (0.01)	**0.01** (0.00)	0.01 (0.00)
	M6	**0.09** (0.02)	**0.05** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	**0.00** (0.00)
50	M1	0.96 (0.03)	0.96 (0.03)	0.91 (0.07)	0.91 (0.07)	0.92 (0.06)	0.95 (0.04)
	M2	0.97 (0.04)	0.97 (0.02)	0.92 (0.06)	0.91 (0.07)	0.93 (0.06)	0.94 (0.05)
	M3	0.77 (0.15)	0.78 (0.15)	**0.04** (0.01)	0.05 (0.02)	**0.02** (0.00)	0.02 (0.01)
	M4	0.93 (0.06)	0.90 (0.14)	0.09 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.00)
	M5	**0.47** (0.18)	**0.48** (0.16)	**0.04** (0.01)	0.04 (0.01)	**0.02** (0.00)	0.02 (0.01)
	M6	0.48 (0.18)	0.61 (0.36)	**0.04** (0.01)	**0.02** (0.00)	**0.02** (0.00)	**0.01** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.96 (0.03)	0.96 (0.03)	0.95 (0.03)	0.96 (0.04)
	M2	0.99 (0.01)	0.98 (0.01)	0.96 (0.03)	0.95 (0.04)	0.96 (0.03)	0.94 (0.04)
	M3	0.98 (0.02)	0.98 (0.02)	0.12 (0.01)	0.13 (0.01)	0.05 (0.01)	0.05 (0.01)
	M4	0.99 (0.01)	0.98 (0.02)	0.30 (0.06)	0.12 (0.01)	0.09 (0.01)	0.04 (0.00)
	M5	**0.83** (0.11)	**0.86** (0.10)	**0.10** (0.01)	0.11 (0.03)	**0.04** (0.01)	0.11 (0.22)
	M6	0.85 (0.10)	0.99 (0.03)	**0.10** (0.01)	**0.06** (0.01)	**0.04** (0.01)	**0.02** (0.00)

Table 7: Subspace distance for Model 5. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.95 (0.05)	0.95 (0.05)	0.94 (0.06)	0.95 (0.06)	0.98 (0.03)	0.97 (0.04)
	M2	0.95 (0.04)	0.95 (0.05)	0.94 (0.07)	0.94 (0.07)	0.98 (0.02)	0.97 (0.03)
	M3	0.10 (0.02)	0.10 (0.02)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	0.02 (0.03)
	M4	0.25 (0.09)	0.09 (0.03)	0.02 (0.01)	**0.01** (0.00)	**0.01** (0.00)	0.01 (0.00)
	M5	**0.08** (0.02)	0.09 (0.02)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	0.01 (0.00)
	M6	**0.08** (0.02)	**0.04** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	**0.00** (0.00)
50	M1	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.01)	0.98 (0.02)	0.98 (0.02)
	M2	0.98 (0.01)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)
	M3	0.69 (0.18)	0.69 (0.17)	**0.04** (0.01)	0.04 (0.01)	**0.02** (0.00)	0.02 (0.00)
	M4	0.93 (0.04)	0.92 (0.10)	0.08 (0.01)	0.04 (0.00)	0.03 (0.00)	0.02 (0.00)
	M5	**0.37** (0.12)	**0.40** (0.12)	**0.04** (0.01)	0.09 (0.21)	**0.02** (0.00)	0.02 (0.01)
	M6	**0.37** (0.12)	0.43 (0.37)	**0.04** (0.01)	**0.02** (0.00)	**0.02** (0.00)	**0.01** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M2	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M3	0.97 (0.02)	0.97 (0.02)	0.10 (0.01)	0.11 (0.03)	**0.04** (0.00)	0.05 (0.02)
	M4	0.99 (0.01)	0.98 (0.02)	0.24 (0.03)	0.10 (0.01)	0.08 (0.01)	0.04 (0.01)
	M5	**0.76** (0.12)	**0.78** (0.11)	0.09 (0.01)	0.09 (0.01)	**0.04** (0.00)	0.04 (0.00)
	M6	0.77 (0.12)	0.99 (0.01)	**0.08** (0.01)	**0.05** (0.00)	**0.04** (0.00)	**0.02** (0.00)

Table 8: Subspace distance for Model 6. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.93 (0.07)	0.93 (0.07)	0.94 (0.05)	0.94 (0.05)	0.93 (0.08)	0.95 (0.04)
	M2	0.94 (0.07)	0.93 (0.08)	0.95 (0.04)	0.95 (0.06)	0.93 (0.08)	0.93 (0.06)
	M3	0.38 (0.21)	0.38 (0.21)	**0.04** (0.01)	0.04 (0.01)	**0.02** (0.01)	0.10 (0.25)
	M4	0.62 (0.23)	0.36 (0.30)	0.06 (0.01)	0.03 (0.01)	**0.02** (0.01)	**0.01** (0.00)
	M5	**0.24** (0.12)	0.26 (0.13)	**0.04** (0.01)	0.04 (0.01)	**0.02** (0.01)	0.02 (0.01)
	M6	**0.24** (0.13)	**0.15** (0.17)	**0.04** (0.01)	**0.02** (0.01)	**0.02** (0.01)	**0.01** (0.00)
50	M1	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)
	M2	0.99 (0.02)	0.98 (0.02)	0.98 (0.01)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)
	M3	0.83 (0.13)	0.83 (0.13)	0.12 (0.03)	0.12 (0.03)	**0.05** (0.01)	0.10 (0.16)
	M4	0.90 (0.07)	0.88 (0.12)	0.18 (0.04)	0.09 (0.02)	0.07 (0.01)	0.04 (0.01)
	M5	**0.71** (0.22)	**0.71** (0.21)	**0.11** (0.02)	0.17 (0.19)	**0.05** (0.01)	0.06 (0.02)
	M6	**0.71** (0.22)	0.90 (0.11)	**0.11** (0.02)	**0.06** (0.01)	**0.05** (0.01)	**0.03** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)
	M2	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M3	0.97 (0.03)	0.97 (0.03)	0.33 (0.13)	0.35 (0.15)	0.12 (0.02)	0.13 (0.03)
	M4	0.99 (0.01)	0.98 (0.02)	0.62 (0.12)	0.45 (0.29)	0.19 (0.02)	0.09 (0.01)
	M5	**0.91** (0.05)	**0.92** (0.05)	**0.24** (0.07)	0.26 (0.07)	**0.11** (0.02)	0.11 (0.03)
	M6	**0.91** (0.05)	0.99 (0.01)	**0.24** (0.07)	**0.12** (0.02)	**0.11** (0.02)	**0.06** (0.01)

Tables 9-11 show that the overall ordering among the six methods is preserved under the more challenging covariance structure, but that the performance gaps are amplified by the correlated design. The OSIR variants remain incapable of recovering the symmetric signal. The OSAVE variants improve substantially in the lowest-dimension setting as n grows, but degrade markedly at $p = 100$, where OSAVE requires accurate inversion of the full $p \times p$ sample covariance and is therefore penalized by the correlated design. The ODR variants remain comparatively robust, with M6 again attaining the smallest subspace distance in the majority of cells, especially when $n \geq 5000$. These results support two conclusions. First, the additional second-moment information encoded by DR is decisive under symmetric dependence, and this advantage is preserved under correlated predictor designs that degrade the simpler OSAVE kernel. Second, the generalized (M6) variant is more stable than the ordinary (M5) variant as p grows, consistent with the vanishing-ridge metric analysis of Section A.

B.4 Computational Efficiency

To quantify the computational advantage of the proposed streaming procedure, we compare the runtime of the two ODR variants against the batch DR baseline under Model 9. In the online setting, observations arrive sequentially and the estimator is updated recursively at each step. In the batch setting, DR is re-run from scratch on all observations available up to step t. All experiments are conducted on a server with an Intel Xeon Gold 6526Y CPU with 4 cores and 251GiB of RAM. Runtimes are reported in seconds and averaged over 20 replications.

Table 12 shows that the online procedure is faster than the batch alternative across every combination of p and n, and that the computational gap widens substantially as either dimension grows. At $p = 100$ and $n = 10000$, for example, batch DR takes on the order of fourteen minutes, whereas the generalized ODR variant with gradient descent completes in just over a minute. Between the two subspace update methods, gradient descent is consistently faster than perturbation, which is consistent with the analysis in Section 2.2 where gradient descent avoids the generalized eigendecomposition at every step and relies instead on a single matrix–vector product. The generalized variant M6 is

Table 9: Subspace distance for Model 7. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.95 (0.05)	0.94 (0.05)	0.95 (0.05)	0.96 (0.04)	0.95 (0.04)	0.94 (0.05)
	M2	0.96 (0.03)	0.95 (0.04)	0.96 (0.04)	0.95 (0.05)	0.95 (0.05)	0.96 (0.04)
	M3	0.16 (0.06)	0.17 (0.07)	0.02 (0.00)	0.02 (0.00)	**0.01** (0.00)	0.01 (0.00)
	M4	0.58 (0.19)	0.33 (0.13)	0.07 (0.01)	0.06 (0.01)	0.05 (0.01)	0.05 (0.01)
	M5	0.10 (0.02)	0.09 (0.02)	0.02 (0.00)	0.02 (0.02)	**0.01** (0.00)	0.01 (0.00)
	M6	**0.09** (0.02)	**0.05** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)	**0.00** (0.00)
50	M1	0.99 (0.01)	0.99 (0.01)	0.98 (0.01)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)
	M2	0.99 (0.01)	0.98 (0.01)	0.98 (0.01)	0.99 (0.01)	0.98 (0.02)	0.98 (0.02)
	M3	0.86 (0.12)	0.87 (0.12)	0.06 (0.01)	0.10 (0.21)	**0.02** (0.00)	0.03 (0.06)
	M4	0.96 (0.03)	0.95 (0.03)	0.22 (0.06)	0.12 (0.02)	0.09 (0.01)	0.07 (0.01)
	M5	0.39 (0.16)	**0.39** (0.19)	0.05 (0.01)	0.07 (0.10)	**0.02** (0.00)	0.09 (0.23)
	M6	**0.35** (0.17)	0.41 (0.36)	**0.04** (0.00)	**0.02** (0.00)	**0.02** (0.00)	**0.01** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M2	0.99 (0.01)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M3	0.98 (0.03)	0.98 (0.03)	0.21 (0.05)	0.21 (0.05)	0.06 (0.01)	0.09 (0.08)
	M4	0.98 (0.02)	0.99 (0.01)	0.87 (0.11)	0.62 (0.23)	0.22 (0.03)	0.12 (0.01)
	M5	0.87 (0.11)	**0.86** (0.12)	0.10 (0.01)	0.15 (0.20)	0.05 (0.01)	0.11 (0.18)
	M6	**0.85** (0.13)	1.00 (0.00)	**0.09** (0.01)	**0.05** (0.01)	**0.04** (0.00)	**0.02** (0.00)

marginally faster than the ordinary variant M5 across all settings, reflecting the numerical savings of avoiding direct inversion of $\widehat{\Sigma}_t$ in favor of the vanishing-ridge metric.

B.5 Structural Dimension Selection

We next evaluate the online ladle method introduced in Section 2.3 for estimating the structural dimension d. To cover a representative range of data-generating structures, we apply the method to Model 1 (first-moment dependence), Model 5 (symmetric dependence), and Model 7 (symmetric dependence with a correlated covariance), with 20 independent replications per configuration. For each replication, we record whether the ladle estimator $\widehat{d}_t^{\,\mathrm{ladle}}$ equals the true structural dimension at the final time point, and report the empirical selection accuracy in Table 13.

The results indicate that the ladle method is a reliable order-selection device provided that the sample size is adequate for the dimension. For the first-moment Model 1 the selector achieves perfect accuracy across all combinations except in the low-sample, high-dimensional corner cases, where the weighted bootstrap cannot yet distinguish the leading eigenvalue from the bulk. For Model 5 and the more challenging Model 7, the selector achieves perfect accuracy at $n = 10000$ for every dimension, and achieves perfect or near-perfect accuracy at $n = 5000$ when $p \leq 50$. The poor performance at $n = 1000$ in higher dimensions is expected, because the weighted-bootstrap eigenvectors require enough effective sample size within the null block to produce a stable null variability term, and this sample-size requirement grows with p. Overall, the table supports the consistency result in Theorem 25 and provides practical guidance that the proposed online ladle method becomes reliable once the sample size is moderate relative to the dimension.

B.6 Sensitivity Analysis

We conducted a compact sensitivity study for the main tuning choices. Table 14 varies the number of slices, the slice boundaries, both ridge constants, the step-size constant, and the initialization fraction. Table 15 varies the ladle search cap and the number of bootstrap replicates.

Table 10: Subspace distance for Model 8. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.97 (0.03)	0.96 (0.04)	0.93 (0.07)	0.93 (0.07)	0.95 (0.04)	0.95 (0.04)
	M2	0.97 (0.03)	0.98 (0.03)	0.96 (0.05)	0.96 (0.04)	0.97 (0.03)	0.97 (0.03)
	M3	0.61 (0.23)	0.61 (0.22)	0.04 (0.01)	0.04 (0.01)	**0.01** (0.00)	0.02 (0.01)
	M4	0.83 (0.11)	0.75 (0.19)	0.17 (0.05)	0.12 (0.02)	0.10 (0.01)	0.09 (0.01)
	M5	**0.34** (0.19)	0.34 (0.19)	0.03 (0.01)	0.02 (0.01)	0.02 (0.00)	0.05 (0.18)
	M6	**0.34** (0.19)	**0.21** (0.28)	**0.02** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)
50	M1	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)
	M2	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)
	M3	0.92 (0.08)	0.92 (0.07)	0.39 (0.22)	0.38 (0.24)	0.07 (0.02)	0.07 (0.02)
	M4	0.98 (0.02)	0.98 (0.02)	0.72 (0.20)	0.46 (0.21)	0.23 (0.06)	0.14 (0.02)
	M5	**0.82** (0.15)	**0.82** (0.15)	**0.10** (0.02)	0.15 (0.20)	0.04 (0.01)	0.10 (0.19)
	M6	**0.82** (0.15)	0.86 (0.09)	**0.10** (0.02)	**0.04** (0.01)	**0.03** (0.01)	**0.02** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)
	M2	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M3	0.98 (0.02)	0.98 (0.02)	0.84 (0.13)	0.88 (0.10)	0.41 (0.17)	0.45 (0.22)
	M4	0.99 (0.01)	0.99 (0.01)	0.95 (0.03)	0.90 (0.08)	0.76 (0.13)	0.46 (0.18)
	M5	0.93 (0.06)	**0.93** (0.06)	0.39 (0.15)	0.41 (0.16)	0.11 (0.01)	0.15 (0.20)
	M6	**0.92** (0.06)	0.99 (0.01)	**0.38** (0.16)	**0.13** (0.07)	**0.10** (0.01)	**0.04** (0.00)

Every mean lies between 0.00 and 0.09, the largest spans are 0.03, 0.06, and 0.07, and no single perturbation caused severe deterioration, so the defaults are not narrowly optimized. Table 15 is a ladle study, separate from Table 13. The replicate rows fix the default cap $q = \lfloor p/\log p \rfloor$.

The 30 rates lie between 0.85 and 1.00, with 24 of them at least 0.95. Raising the replicate count never lowers a rate, and every model and sample size reaches 1.00 at 50 replicates. Raising n improves 11 of 15 comparisons, and every $n = 10000$ rate is at least 0.95. The default cap stays within 0.05 of the best cap.

C Real Data Analysis

We apply the proposed ODR procedure to two benchmark datasets from the UCI Machine Learning Repository, one for regression and one for classification. The regression task uses the concrete compressive strength data ($n = 1030$, $p = 8$), available at

`https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength`

where the goal is to predict the compressive strength of concrete from its mixture composition and production conditions. The classification task uses the electrical grid stability data ($n = 10000$, $p = 12$), available at

`https://archive.ics.uci.edu/dataset/471/electrical+grid+stability+simulated+data`

where the goal is to classify the stability of a decentralized grid system as stable or unstable from measured system parameters.

For the real-data experiments, the number of response slices was chosen according to the response type. For the concrete compressive strength regression data, we used $H = 10$. For the electrical grid stability classification data, the binary response labels were used as two slices, so $H = 2$. For

Table 11: Subspace distance for Model 9. Entries report mean (standard deviation) over 20 replications.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M1	0.93 (0.07)	0.93 (0.07)	0.95 (0.05)	0.95 (0.05)	0.95 (0.04)	0.96 (0.05)
	M2	0.97 (0.04)	0.97 (0.02)	0.97 (0.03)	0.97 (0.03)	0.97 (0.03)	0.97 (0.03)
	M3	0.59 (0.25)	0.60 (0.25)	0.03 (0.01)	0.03 (0.01)	**0.01** (0.00)	**0.01** (0.00)
	M4	0.79 (0.12)	0.65 (0.18)	0.13 (0.02)	0.11 (0.01)	0.10 (0.01)	0.09 (0.01)
	M5	0.28 (0.16)	0.28 (0.15)	0.03 (0.01)	0.02 (0.01)	0.02 (0.00)	**0.01** (0.00)
	M6	**0.27** (0.16)	**0.13** (0.18)	**0.02** (0.01)	**0.01** (0.00)	**0.01** (0.00)	**0.01** (0.00)
50	M1	0.98 (0.02)	0.98 (0.02)	0.99 (0.02)	0.99 (0.02)	0.98 (0.03)	0.99 (0.01)
	M2	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M3	0.90 (0.07)	0.90 (0.06)	0.21 (0.09)	0.25 (0.19)	0.06 (0.02)	0.09 (0.17)
	M4	0.98 (0.01)	0.97 (0.02)	0.58 (0.21)	0.31 (0.14)	0.19 (0.03)	0.13 (0.02)
	M5	**0.83** (0.14)	**0.84** (0.14)	0.08 (0.02)	0.13 (0.19)	0.04 (0.01)	0.08 (0.21)
	M6	**0.83** (0.14)	**0.84** (0.08)	**0.07** (0.01)	**0.04** (0.01)	**0.03** (0.01)	**0.02** (0.00)
100	M1	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	M2	0.99 (0.01)	1.00 (0.00)	0.99 (0.01)	0.99 (0.01)	0.99 (0.00)	0.99 (0.01)
	M3	0.98 (0.01)	0.98 (0.01)	0.78 (0.14)	0.79 (0.14)	0.19 (0.06)	0.20 (0.07)
	M4	0.99 (0.01)	0.99 (0.00)	0.94 (0.05)	0.85 (0.10)	0.67 (0.19)	0.34 (0.20)
	M5	**0.92** (0.05)	**0.92** (0.06)	0.27 (0.09)	0.30 (0.15)	0.09 (0.01)	0.14 (0.20)
	M6	**0.92** (0.06)	0.98 (0.03)	**0.25** (0.09)	**0.09** (0.01)	**0.08** (0.01)	**0.04** (0.00)

Table 12: Runtime comparison under Model 9. Entries report mean (standard deviation) runtime in seconds over 20 replications. M5 and M6 denote the ordinary and generalized ODR variants, respectively.

p	Method	$n = 1000$		$n = 5000$		$n = 10000$	
		GD	PT	GD	PT	GD	PT
20	M5	0.25 (0.04)	0.39 (0.03)	1.17 (0.01)	1.92 (0.02)	2.26 (0.02)	3.73 (0.04)
	M6	**0.24 (0.04)**	0.40 (0.03)	**1.15 (0.01)**	1.90 (0.01)	**2.22 (0.02)**	3.70 (0.04)
	Batch DR	1.27 (0.02)		16.19 (0.13)		56.86 (0.23)	
50	M5	1.13 (0.01)	2.05 (0.02)	5.60 (0.02)	10.18 (0.05)	11.42 (0.60)	20.67 (0.95)
	M6	**1.09 (0.01)**	1.91 (0.01)	**5.40 (0.02)**	9.52 (0.02)	**10.96 (0.48)**	19.29 (0.92)
	Batch DR	3.62 (0.01)		57.57 (0.09)		230.28 (11.12)	
100	M5	7.32 (0.47)	12.70 (0.81)	37.79 (2.13)	64.87 (3.14)	74.20 (4.22)	127.39 (7.03)
	M6	**6.90 (0.49)**	11.46 (0.76)	**35.62 (2.01)**	59.19 (3.21)	**68.21 (5.01)**	113.19 (7.24)
	Batch DR	13.96 (0.78)		225.82 (10.36)		841.43 (23.09)	

each randomly ordered stream, the online algorithms were initialized using the first 40% of the observations available to the online SDR update. For both regression and classification data $q = 5$ was used as the maximum search dimension. The number of weighted-bootstrap replicates used in the real-data ladle computation was fixed at $N_{\text{boot}} = 10$.

For each dataset, we evaluate the proposed procedure along four axes. Table 16 reports the subspace distance between the batch and online estimators, averaged over 50 randomly permuted replications, quantifying how faithfully the online procedure tracks the batch target. Table 17 reports prediction error, with mean squared error for regression and misclassification rate for classification, under three settings, namely the original predictors, the batch SDR estimator, and the online SDR estimator.

Table 13: Empirical accuracy of the online ladle selector under Models 1, 5, and 7. Entries report the proportion of 20 replications in which the selector returns the true structural dimension.

Model	d	p	$n = 1000$	$n = 5000$	$n = 10000$
1	1	20	1.00	1.00	1.00
		50	0.00	1.00	1.00
		100	0.00	0.75	1.00
5	2	20	0.85	1.00	1.00
		50	0.00	1.00	1.00
		100	0.00	0.60	1.00
7	2	20	0.75	1.00	1.00
		50	0.00	1.00	1.00
		100	0.00	0.65	1.00

Table 14: Sensitivity of generalized ODR, 20 paired streams, $p = 50$, $n = 5000$. Entries are min–max mean subspace distances.

Varied item	Grid	Model 1	Model 5	Model 9
Slices H	$3, 5, 10$	0.01–0.02	0.02–0.07	0.01–0.07
RLS ridge λ	$10^{-8}, 10^{-6}, 10^{-4}$	0.01–0.02	0.01–0.07	0.02–0.07
Metric ridge ρ_0	$10^{-6}, 10^{-4}, 10^{-2}$	0.00–0.02	0.01–0.06	0.01–0.08
Step-size constant c	$0.05, 0.10, 0.20$	0.00–0.02	0.03–0.05	0.03–0.09
Initialization fraction t_0/n	$0.20, 0.40, 0.60$	0.00–0.02	0.02–0.06	0.03–0.07
Slice boundaries	source and pilot size	0.00–0.03	0.01–0.06	0.01–0.07

Table 20 compares the runtime of the ODR variants with that of the batch DR baseline. Finally, Table 21 reports the output of the online ladle method for structural dimension selection. Throughout, the six online competitors are M1 through M6 as defined in Section B.

ODR should not be presented as a uniformly better downstream predictor (Tables 18 and 19); the contribution is an efficient online approximation to batch DR with clear advantages under second-moment structure.

C.1 Estimation of the Central Subspace

Table 16 reveals a clear separation among the six online methods across both datasets. For the regression data, the ODR-based methods (M5 and M6) yield moderately small subspace distances from the batch estimator, performing comparably to ordinary OSIR (M1). By contrast, the OSAVE-based methods (M3 and M4) yield substantially larger distances, both exceeding 0.75, indicating that they fail to faithfully track their batch counterparts on this dataset. This pattern persists on the classification data, where ordinary OSIR (M1) again achieves the smallest subspace distance and the OSAVE variants (M3, M4) demonstrate notably poor tracking performance. The ODR-based methods (M5, M6) remain competitive across both datasets, with the generalized variant M6 demonstrating notably stable performance. This robustness suggests that ODR can accommodate diverse response types without requiring prior knowledge of the dominant moment structure.

Table 15: Online ladle accuracy at $p = 50$. Entries are correct-d proportions at $n = 5000$ / $n = 10000$ for Models 1, 5, and 7 with $d = 1, 2, 2$, respectively.

Ladle setting	Model 1	Model 5	Model 7
$N_{\text{boot}} = 10$, default cap	$0.95/1.00$	$0.90/0.95$	$0.90/0.95$
$N_{\text{boot}} = 30$, default cap	$1.00/1.00$	$0.95/1.00$	$0.95/1.00$
$N_{\text{boot}} = 50$, default cap	$1.00/1.00$	$1.00/1.00$	$1.00/1.00$
$q = 5$	$0.90/0.95$	$0.85/0.95$	$0.85/0.95$
$q = 10$	$0.95/1.00$	$0.90/1.00$	$0.95/1.00$

Table 16: Average subspace distance $d(\widehat{B}_{\mathrm{batch}}, \widehat{B}_{\mathrm{online}})$ between the batch and online estimators, over 50 replications, for the concrete compressive strength (regression) and electrical grid stability (classification) datasets.

Method	Regression		Classification	
	GD	PT	GD	PT
M1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
M2	0.72 (0.07)	0.68 (0.06)	0.24 (0.20)	0.20 (0.19)
M3	0.76 (0.16)	0.76 (0.16)	0.75 (0.14)	0.76 (0.14)
M4	0.94 (0.02)	0.94 (0.01)	0.97 (0.01)	0.97 (0.01)
M5	0.02 (0.02)	0.02 (0.03)	0.11 (0.12)	0.13 (0.16)
M6	0.01 (0.01)	0.01 (0.01)	0.11 (0.12)	0.03 (0.03)

C.2 Predictive Performance

Table 17 shows how well each estimator preserves the predictive information in the recovered subspace. For this evaluation, we repeated 50 random train/test splits, using 70% of the observations for training and the remaining 30% for testing. Prediction was performed on the reduced predictors using support vector machine. For the regression task, the ODR-based methods M5 and M6 produce mean squared errors of roughly 78 to 85, close to the batch DR baseline of 67.12 and to the raw-predictor benchmark of 71.47. The ordinary OSIR variants give comparable performance, whereas OSAVE and the generalized OSIR variant suffer much larger errors. Notably, the ODR estimators deliver comparable predictive accuracy as their batch counterpart at a fraction of the computational cost, which is the main practical motivation for the online procedure. For the classification task, the raw-predictor benchmark is already very small at 0.07, reflecting the fact that the grid stability response is nearly linearly separable in its original 12-dimensional feature space, so dimension reduction cannot improve on the full predictor set in terms of predictive error. Among the online SDR procedures, however, the generalized ODR variant M6 achieves a misclassification rate of 0.20 under perturbation, which is reasonably close to the batch DR baseline of 0.18. The ODR-based methods are also competitive with the best online alternative, ordinary OSIR at 0.19, and dominate the OSAVE variants, whose error rates remain near 0.36.

Table 17: Average prediction error on the two datasets, with mean squared error reported for the regression task and misclassification rate reported for the classification task. Three reference rows are included for comparison, namely the batch SDR estimators (Batch SIR, Batch SAVE, Batch DR) and the full original-predictor baseline.

Method	Regression		Classification	
	GD	PT	GD	PT
M1	70.87 (9.61)	71.02 (9.74)	0.19 (0.00)	0.19 (0.00)
M2	269.95 (19.41)	265.24 (16.06)	0.27 (0.00)	0.27 (0.00)
Batch SIR	65.34(6.07)		0.18(0.00)	
M3	260.79 (27.26)	261.67 (27.45)	0.34 (0.00)	0.34 (0.00)
M4	282.41 (19.74)	282.32 (19.51)	0.36 (0.00)	0.36 (0.00)
Batch SAVE	74.49 (8.65)		0.18 (0.00)	
M5	85.14 (11.27)	81.84 (10.43)	0.23 (0.04)	0.24 (0.05)
M6	81.49 (10.18)	77.81 (9.55)	0.23 (0.04)	0.20 (0.01)
Batch DR	67.12 (5.94)		0.18 (0.00)	
Original Predictors	71.47 (7.91)		0.07 (0.00)	

Held-out prediction error offers a direct real-data assessment of downstream utility, and simulations designed around second-moment structure should not be used to claim general superiority. We therefore reran all methods under matched downstream-prediction setups on seven real datasets

and three representative simulation models. We further compare the performance of ODR with IPLS1, incremental PLS [32], which tracks a prediction subspace that need not be central, and OSSIR-CCIPCA, a sparse online SIR with CCIPCA updates [29]. For real data, we used 50 identical random 70/30 splits for every method, with the same support-vector learner and tuning grid within each dataset. All preprocessing, representation fitting, and tuning used training data only. Table 18 reports means with standard deviations, and lower is better.

Table 18: Matched held-out error. The first three datasets report classification error and the last four report MSE. Each dataset name is followed by its number of observations and predictors.

Dataset (n, p)	Tracker	ODR	IPLS1	OSSIR-CCIPCA	OSIR	OSAVE
Ionosphere	GD	0.17 (0.05)	0.18 (0.02)	0.20 (0.03)	0.23 (0.03)	0.24 (0.04)
(351, 34)	PT	0.23 (0.05)			0.24 (0.03)	0.23 (0.04)
Grid Stability	GD	0.23 (0.04)	0.23 (0.02)	0.27 (0.02)	0.27 (0.00)	0.36 (0.00)
(10,000, 12)	PT	0.20 (0.01)			0.27 (0.00)	0.36 (0.00)
MAGIC	GD	0.23 (0.00)	0.25 (0.00)	0.28 (0.01)	0.27 (0.01)	0.32 (0.01)
(19,020, 10)	PT	0.23 (0.00)			0.27 (0.01)	0.31 (0.01)
Energy	GD	67.51 (29.83)	55.01 (9.03)	98.08 (19.07)	91.52 (15.10)	100.53 (13.42)
(768, 8)	PT	73.80 (21.21)			93.40 (15.23)	100.31 (11.74)
Yacht	GD	39.73 (18.75)	95.01 (23.90)	242.07 (58.03)	266.09 (59.18)	267.04 (60.52)
(308, 6)	PT	35.25 (16.84)			269.05 (60.85)	269.02 (61.38)
Concrete	GD	85.14 (11.27)	91.04 (15.02)	81.01 (12.12)	70.87 (9.61)	260.79 (27.26)
(1,030, 8)	PT	81.84 (10.43)			71.02 (9.74)	261.67 (27.45)
Parkinsons	GD	114.12 (3.53)	116.05 (3.65)	117.23 (7.55)	128.41 (5.32)	137.51 (5.89)
(5,875, 18)	PT	115.19 (3.91)			120.22 (4.45)	137.42 (4.82)

Among the compared streaming methods, at least one of the two separately reported ODR trackers has the numerically lowest mean error on five of seven datasets. IPLS1 is lowest on Energy, and OSIR is lowest on Concrete. ODR is lowest on all three datasets with more than five thousand observations. The clearest ODR gain occurs on Yacht, where ODR-PT gives 35.25 compared with 95.01 for the lowest non-ODR mean. The differences on Ionosphere and Parkinsons are modest. Thus, the real-data evidence supports competitiveness, not uniform superiority, and no single tracker is best in every dataset.

We also ran a matched simulation study (Table 19) using 20 paired runs, 50 predictors, 1,000 training observations, 5,000 test observations, and the same RBF-SVR procedure for all methods. Standardization and tuning used training data only.

Table 19: Matched held-out simulation MSE.

Model	Tracker	ODR	IPLS1	OSSIR-CCIPCA	OSIR	OSAVE
Model 1	GD	1.11 (0.01)	1.16 (0.04)	1.29 (0.03)	1.16 (0.03)	1.59 (0.01)
	PT	1.10 (0.02)			1.10 (0.02)	1.29 (0.01)
Model 5	GD	0.32 (0.06)	3.17 (0.02)	3.17 (0.01)	3.20 (0.03)	0.55 (0.03)
	PT	0.19 (0.01)			3.20 (0.03)	0.32 (0.03)
Model 7	GD	0.19 (0.06)	1.72 (0.07)	1.75 (0.04)	1.76 (0.02)	0.72 (0.06)
	PT	0.13 (0.03)			1.76 (0.02)	0.45 (0.05)

Model 1 is a first-moment setting. ODR-PT and OSIR-PT have the same rounded mean MSE of 1.10, while IPLS1 gives 1.16, showing that ODR remains competitive in this setting. Models 5 and 7 favor methods that use second-moment information. We therefore interpret them only as evidence that this additional information can translate into downstream prediction gains. They are not evidence of universal superiority.

Practical utility. These results do not justify ODR as an automatic default. When second-moment dependence is scientifically plausible, ODR offers a way to retain that information without committing to a first-moment procedure alone. When the signal structure is uncertain, ODR draws on both moments rather than relying on one of them. Energy and Concrete show that other procedures can be preferable on particular datasets. Across both tables, however, no ODR branch exceeds 1.7 times the lowest error in its row, while every competitor exceeds 7.5 times in at least one setting.

C.3 Computational Efficiency

Table 20 compares the runtime of the two ODR variants with the batch DR baseline on both datasets. On the small regression problem, the online procedure runs in roughly 0.2 seconds while batch DR requires 0.84 seconds, yielding a speedup of roughly four times. On the larger classification problem the gap grows substantially, with the online procedure completing in about 1 to 2 seconds compared to 27.93 seconds for batch DR, a speedup of more than fifteen times. The gradient descent method is consistently faster than the perturbation method, which is consistent with the analysis in Section A. Among the ODR-based methods, the M6 with gradient descent gives the fastest runtime on both datasets. Taken together with the preceding two subsections, these runtimes demonstrate that the proposed procedure reproduces the batch estimator to high fidelity while scaling far more favorably with the sample size.

Table 20: Runtime comparison between the ODR variants (M5, M6) and the batch DR baseline on the two datasets. Entries report the mean runtime in seconds over 20 replications, with standard deviations in parentheses.

Method	Regression		Classification	
	GD	PT	GD	PT
M5	0.17 (0.04)	0.20 (0.02)	1.05 (0.07)	1.62 (0.04)
M6	0.16 (0.02)	0.22 (0.01)	1.03 (0.02)	1.80 (0.04)
Batch DR	0.84 (0.04)		27.93 (0.54)	

C.4 Online Order Determination

Finally, Table 21 reports the output of the online ladle method for the two ODR variants. On both datasets, the average selected dimension $\widehat{d}_{\mathrm{ladle}}$ is very close to one, which is consistent with the interpretation that a single linear direction captures most of the conditional dependence in each problem. The subspace distance between the batch estimator and the online ladle-based estimator is very small on the regression data (0.01 and 0.02) for the two variants, and moderate on the classification data (0.11 and 0.06). Most importantly, the prediction errors of the online ladle-based estimator closely track those of batch DR on both datasets, with regression MSE values of 77.76 and 76.63 against a batch baseline of 67.10 and 67.18, and classification error rates of 0.19 and 0.18 against 0.19 and 0.19. These results confirm that the proposed online ladle method produces a dimension estimate consistent with the batch analysis while preserving predictive accuracy.

Table 21: Online ladle results for the two ODR variants, averaged over 50 replications. The GD-based and PT-based rows correspond to the gradient-descent-based and perturbation-based online updates, respectively. The column labeled $\mathrm{dist}(\widehat{B}_{\mathrm{batch}}, \widehat{B}_{\mathrm{ladle}})$ reports the subspace distance between the batch DR estimator and the online ladle-based estimator. The Online and Batch columns report mean squared error for the regression task and misclassification rate for the classification task. The final column reports the average ladle-selected dimension $\widehat{d}_{\mathrm{ladle}}$.

Type	Method	$\mathrm{dist}(\widehat{B}_{\mathrm{batch}}, \widehat{B}_{\mathrm{ladle}})$	Online	Batch	Avg. $\widehat{d}_{\mathrm{ladle}}$
Regression	GD-based	0.01 (0.03)	77.76 (9.08)	67.10 (6.03)	1.04
	PT-based	0.02 (0.12)	76.63 (8.77)	67.18 (6.17)	1.06
Classification	GD-based	0.11 (0.14)	0.19 (0.01)	0.19 (0.04)	1.00
	PT-based	0.06 (0.17)	0.18 (0.01)	0.19 (0.04)	1.06

D Discussion

This paper introduces ODR, a streaming framework for SDR that combines a p_h-free recursive kernel, a ridge-stabilized RLS standardization, two complementary subspace update methods, and a weighted-bootstrap ladle method for structural dimension selection. The framework fills a gap

in the online SDR literature. No prior procedure has, to our knowledge, taken DR online together with a provably consistent streaming approach for the structural dimension. The resulting method inherits the broader identification set of DR [19, 18] at essentially the per-observation cost of online SIR, providing a robust alternative when the dominant data structure is unknown across first-moment, symmetric, and certain interaction driven settings. The asymptotic theory we develop shows that this gain does not come at the expense of statistical guarantees.

Three aspects of the framework deserve emphasis. First, the p_h free kernel $\widehat{K}_t^\star$ and the ridge-stabilized RLS standardization jointly eliminate the two most fragile operations in a naïve approach, namely division by rare slice probabilities and inversion of an ill-conditioned early-stage $\widehat{\Sigma}_t$. The shared RLS inverse G_t^{-1} is maintained by the Sherman–Morrison rank one update in $O(p^2)$ time, because $G_t = \lambda I_{p+1} + \sum_{i=1}^t x_i^{\mathrm{e}} x_i^{\mathrm{e}\top}$ is common to all slices. The full per-observation update is larger. Updating the raw slice moments and the RLS slice summaries over H slices costs $O(Hp^2)$ time and $O(Hp^2)$ memory. If the full kernel matrix $\widehat{K}_t^\star$ is explicitly materialized, the second-order matrix products can cost $O(Hp^3)$ naïvely. In the gradient implementation, however, $\widehat{K}_t^\star B_t$ is applied directly from the stored slice blocks, so the full per-observation gradient update costs $O(Hp^2d + p^2d + pd^2 + d^3)$, which is $O(Hp^2)$ for fixed d and H. And the vanishing-ridge metric $S_t = \widehat{\Sigma}_t + \rho_t I_p$ preserves the limiting generalized eigenspace while guaranteeing positive definiteness at every time. Second, the two subspace update methods occupy genuinely distinct regions of the cost-accuracy trade-off. The generalized perturbation method tracks the exact generalized eigenspace of (D_t, S_t) with a first-order update whose remainder is $O_p(t^{-2})$, and the generalized gradient method casts the same problem as a constrained stochastic approximation on the Σ-Stiefel manifold, extending the classical Oja-type analysis [25, 24, 3, 1] to a moving metric. The almost sure convergence result in Theorem 22 appears to be the first such guarantee for online SDR under a time-varying standardization, and the simulations in Section B confirm that gradient descent is the fastest option in large p settings without sacrificing accuracy. Third, the online ladle method provides, to our knowledge, the first consistent streaming order-determination procedure for moment-based SDR. The influence-function expansion of the plug-in standardized kernel matrix $\mathcal{M}_t$, combined with the Poisson-multiplier weighted bootstrap [15, 26, 27], yields a criterion that retains the main advantages of the batch ladle [23] while using only weighted moment and RLS summaries updated sequentially along the stream.

The empirical findings are consistent with these theoretical properties. Across nine simulated models spanning SIR-favorable, SAVE-favorable, and correlated-covariance settings, the generalized ODR variant matches or outperforms online SIR and online SAVE, with the largest gains in the symmetric designs where first-moment methods are provably blind. The runtime comparison indicates a speedup of more than an order of magnitude over the batch alternative at $p = 100$ and $n = 10^4$, and the two UCI benchmarks show that the online procedure recovers a subspace close to the batch DR target while delivering predictive accuracy that is competitive with, and in some cases indistinguishable from, the batch baseline at a small fraction of its cost.

Several directions remain open. The present analysis fixes p and H, and a natural extension is to allow p to grow with t, potentially with sparsity regularization in the spirit of sparse SDR [21, 8]. Relaxing the i.i.d. assumption to stationary or weakly dependent streams would broaden the applicability of the method to time-series and sensor data, and would require adapting the moment and martingale arguments in Section A accordingly. Data-driven choices of the slice boundaries, the ridge constants λ and ρ_0, and the step-size constant c would also be desirable, and a principled online calibration of these tuning parameters is a nontrivial problem that we leave for future work. Finally, although the generalized-gradient update converges almost surely within any local basin of attraction of the underlying ODE, we do not establish global convergence for the generalized-gradient update. We also do not establish global convergence for the standalone iterated perturbation update; doing so would require a stability analysis of the repeated approximate-eigenpair map.

This work is methodological in nature and is intended to improve online SDR for streaming high-dimensional data. A potential positive impact is that the proposed procedure can reduce storage and computational costs while producing low-dimensional representations that may be easier to interpret and use in statistical learning tasks. These benefits may be useful in scientific and engineering applications where observations arrive sequentially and repeated batch recomputation is impractical. At the same time, dimension reduction can also propagate or obscure biases already present in the data. If a learned reduced representation is used in high-stakes downstream settings, such as biomedical monitoring, finance, infrastructure monitoring, or other decision-support systems, systematic bias in

the original data or distributional drift in the data stream may be carried into the reduced features and subsequently affect predictions or decisions. Because our method estimates a low-dimensional subspace rather than making decisions directly, these risks arise primarily through downstream use. Practical deployment should therefore include careful validation of the estimated subspace and the resulting downstream model across relevant subgroups and time periods, monitoring for distributional drift, sensitivity analysis with respect to the selected structural dimension and tuning parameters, and domain-expert review. The real-data examples in this paper use public benchmark datasets and are intended only to illustrate empirical performance, not to support deployment in high-stakes applications.

E Full Online Update Algorithm

Algorithm 1 collects the complete ODR procedure, consolidating the kernel construction, ridge-stabilized standardization, and subspace tracking steps described in the main text into a single reference.

Algorithmic roadmap. One full pass of ODR consists of four steps.

1. **Standardize.** Update the running moments, the RLS inverse G_t^{-1} that whitens the predictors, and the slicewise RLS sums that carry the response information.
2. **Build the kernel and metric.** Combine the standardized first- and second-moment blocks into the current kernel $\widehat{K}_t^\star$, and set the metric S_t to the running covariance plus a shrinking ridge.
3. **Track the subspace with one tracker.** GD steps directly on $(\widehat{K}_t^\star, S_t)$ and never forms an average. PT and the exact reference instead maintain the averaged kernel D_t and work with (D_t, S_t).
4. **Select the dimension when asked.** At monitoring times only, form $\mathcal{M}_t$ and its bootstrap copies and evaluate the ladle, leaving the per-observation cost untouched.

F Continuation of the Weighted-Bootstrap Ladle Implementation

On the event $N_t^{*(b)} > 0$, define the weighted raw moments

$$\widehat{p}_{t,h}^{*(b)} = (N_t^{*(b)})^{-1} \sum_{i=1}^{t} w_{b,i}\delta_{i,h}, \qquad \widehat{u}_{t,h}^{*(b)} = (N_t^{*(b)})^{-1} \sum_{i=1}^{t} w_{b,i}X_i\delta_{i,h},$$

$$\widehat{S}_{t,h}^{(2)*(b)} = (N_t^{*(b)})^{-1} \sum_{i=1}^{t} w_{b,i}X_iX_i^\top\delta_{i,h},$$

$$\widehat{\mu}_t^{*(b)} = (N_t^{*(b)})^{-1} \sum_{i=1}^{t} w_{b,i}X_i, \qquad \widehat{\Sigma}_t^{*(b)} = (N_t^{*(b)})^{-1} \sum_{i=1}^{t} w_{b,i}(X_i - \widehat{\mu}_t^{*(b)})(X_i - \widehat{\mu}_t^{*(b)})^\top.$$

The weighted RLS quantities are

$$G_t^{*(b)} = \lambda I_{p+1} + \sum_{i=1}^{t} w_{b,i}x_i^{\mathrm{e}}x_i^{\mathrm{e}\top},$$

$$s_{t,h}^{*(b)} = \sum_{i=1}^{t} w_{b,i}x_i^{\mathrm{e}}\delta_{i,h}, \qquad R_{t,h}^{*(b)} = \sum_{i=1}^{t} w_{b,i}x_i^{\mathrm{e}}X_i^\top\delta_{i,h},$$

$$\widehat{\vartheta}_{t,h}^{*(b)} = \big(G_t^{*(b)}\big)^{-1}s_{t,h}^{*(b)}, \qquad \widehat{\Theta}_{t,h}^{*(b)} = \big(G_t^{*(b)}\big)^{-1}R_{t,h}^{*(b)}.$$

The weighted block estimators are

$$\begin{aligned}
\widehat{A}_{t,h}^{*(b)} &= \widehat{u}_{t,h}^{*(b)} - \widehat{p}_{t,h}^{*(b)}\widehat{\mu}_t^{*(b)}, \qquad \widetilde{A}_{t,h}^{*(b)} = \widetilde{I}\,\widehat{\vartheta}_{t,h}^{*(b)},\\
\widehat{\mathcal{B}}_{t,h}^{*(b)} &= \widehat{S}_{t,h}^{(2)*(b)} - \widehat{\mu}_t^{*(b)}\widehat{u}_{t,h}^{*(b)\top} - \widehat{u}_{t,h}^{*(b)}\widehat{\mu}_t^{*(b)\top}\\
&\quad + \widehat{p}_{t,h}^{*(b)}\widehat{\mu}_t^{*(b)}\widehat{\mu}_t^{*(b)\top} - \widehat{p}_{t,h}^{*(b)}\widehat{\Sigma}_t^{*(b)},\\
\widehat{Q}_{t,h}^{*(b)} &= \widetilde{I}\,\widehat{\Theta}_{t,h}^{*(b)}, \qquad \widetilde{\mathcal{B}}_{t,h}^{*(b)} = \widehat{Q}_{t,h}^{*(b)} - \widetilde{A}_{t,h}^{*(b)}\widehat{\mu}_t^{*(b)\top} - \widehat{p}_{t,h}^{*(b)}I_p.
\end{aligned}$$

Aggregating as in the unweighted case,

$$T_t^{(U)*(b)} = \sum_{h=1}^{H} \widehat{A}_{t,h}^{*(b)}\widehat{A}_{t,h}^{*(b)\top}, \qquad \widetilde{T}_t^{(U)*(b)} = \sum_{h=1}^{H} \widetilde{A}_{t,h}^{*(b)}\widehat{A}_{t,h}^{*(b)\top},$$

$$\widetilde{T}_t^{(V)*(b)} = \sum_{h=1}^{H} \widehat{\mathcal{B}}_{t,h}^{*(b)}\widetilde{\mathcal{B}}_{t,h}^{*(b)}, \qquad \widetilde{s}_t^{*(b)} = \mathrm{tr}\big(\widetilde{T}_t^{(U)*(b)}\big),$$

we obtain the bootstrap kernel

$$\widehat{K}_t^{\star *(b)} = 2\ \mathrm{Sym}\big(\widetilde{T}_t^{(V)*(b)}\big) + 2\ \mathrm{Sym}\big(T_t^{(U)*(b)}\widetilde{T}_t^{(U)*(b)}\big) + 2\,\widetilde{s}_t^{*(b)}\,T_t^{(U)*(b)},$$

and the bootstrap standardized matrix

$$\begin{aligned} S_t^{*(b)} &= \widehat{\Sigma}_t^{*(b)} + \rho_t I_p, \\ \mathcal{M}_t^{*(b)} &= \big(S_t^{*(b)}\big)^{-1/2}\,\widehat{K}_t^{\star *(b)}\,\big(S_t^{*(b)}\big)^{-1/2}. \end{aligned}$$

On the event $N_t^{*(b)} = 0$, which has probability e^{-t} and is asymptotically negligible, the bootstrap statistic may be defined arbitrarily. Algorithm 2 presents the full ladle-based selection procedure.

Algorithm 2 is a strict single-pass implementation with a fixed number N of bootstrap replicates. Each replicate is initialized once, receives one new Poisson multiplier per incoming observation, and never revisits discarded historical data. Its per-time cost is

$$O(Hp^2 + Np^3),$$

where the p^3 term arises from the eigendecomposition of $\mathcal{M}_t$ and its N bootstrap replicates. The asymptotic consistency theorem with $N_t \to \infty$ is a separate theoretical triangular-array construction with complete multiplier histories, and is not an operational step in Algorithm 2.

Algorithm 1 Online Directional Regression

Require: Stream $\{(X_t, Y_t)\}_{t=1}^{n_{\mathrm{stream}}}$; slices $I_1, \dots, I_H$; target rank d; initialization size t_0; update method $\texttt{method} \in \{\texttt{perturb}, \texttt{gd}\}$; ridge parameters λ, ρ_0; step constant c

Initialize raw moments and RLS summaries

$$\mathcal{S}_{t_0} = \big\{\widehat{\mu}, \widehat{\Sigma}, \widehat{p}_h, \widehat{u}_h, \widehat{S}_h^{(2)}, G^{-1}, s_h, R_h\big\}_{h=1}^{H}.$$

Construct $\widehat{K}_{t_0}^{\star}$, set

$$S_{t_0} = \widehat{\Sigma}_{t_0} + \rho_0 I_p, \qquad D_{t_0} = \widehat{K}_{t_0}^{\star},$$

and initialize $\widehat{B}_{t_0}$ by the leading d generalized eigenvectors of (D_{t_0}, S_{t_0}).

if $\texttt{method} = \texttt{gd}$ **then**

  Set $B_{t_0} = \widehat{B}_{t_0}$.

end if

for $t = t_0 + 1, t_0 + 2, \dots$ **do**

  Observe (X_t, Y_t) and update all raw moments and RLS summaries in $\mathcal{S}_t$ by rank one updates.

  For $h = 1, \dots, H$, construct the unstandardized and RLS-standardized slice blocks as defined in Section 2

$$\widehat{A}_{t,h} \qquad \widehat{\mathcal{B}}_{t,h} \qquad \widetilde{A}_{t,h} \qquad \widetilde{\mathcal{B}}_{t,h}.$$

  Construct the p_h-free ODR kernel and stabilized metric as defined in Section 2

$$T_t \qquad \widetilde{T}_t \qquad V_t \qquad \widetilde{s}_t \qquad \widehat{K}_t^{\star} \qquad S_t = \widehat{\Sigma}_t + \frac{\rho_0}{t} I_p.$$

  if $\texttt{method} = \texttt{perturb}$ **then**

    For $j = 1, \dots, d$, apply the labeled generalized perturbation update as defined in Section 2

$$\begin{aligned}
\delta_{t-1,j} &= \widehat{\beta}_{t-1,j}^{\top}\big(\Delta D_{t-1} - \widehat{\lambda}_{t-1,j}\Delta S_{t-1}\big)\widehat{\beta}_{t-1,j},\\
r_{t-1,j} &= \big(\Delta D_{t-1} - \widehat{\lambda}_{t-1,j}\Delta S_{t-1}\big)\widehat{\beta}_{t-1,j} - \delta_{t-1,j} S_{t-1}\widehat{\beta}_{t-1,j},\\
\widehat{\lambda}_{t,j} &= \widehat{\lambda}_{t-1,j} + \delta_{t-1,j},\\
\widehat{\beta}_{t,j}^{-} &= \widehat{\beta}_{t-1,j} + (\widehat{\lambda}_{t-1,j} S_{t-1} - D_{t-1})^{+} r_{t-1,j},\\
\widehat{\beta}_{t,j} &= \frac{\widehat{\beta}_{t,j}^{-}}{\{(\widehat{\beta}_{t,j}^{-})^{\top} S_t \widehat{\beta}_{t,j}^{-}\}^{1/2}}.
\end{aligned}$$

    If necessary, flip the sign of $\widehat{\beta}_{t,j}$ so that

$$\widehat{\beta}_{t,j}^{\top} S_{t-1} \widehat{\beta}_{t-1,j} \geq 0.$$

    Then set the reported subspace basis as

$$\widehat{B}_t = \mathrm{P}_{\mathrm{orth}}^{(S_t)}\Big((\widehat{\beta}_{t,1}, \dots, \widehat{\beta}_{t,d})\Big).$$

  else

    Take the generalized-gradient step

$$B_t = \mathrm{P}_{\mathrm{orth}}^{(S_t)}\Big(B_{t-1} + \frac{c}{t} S_t^{-1} \widehat{K}_t^{\star} B_{t-1}\Big).$$

  end if

end for

Output: $\widehat{B}_t$ for perturbation; $\mathrm{span}(B_t)$ for gradient descent.

Algorithm 2 Online Weighted-Bootstrap Ladle Selector for Online Directional Regression

Require: Stream $\{(X_t, Y_t)\}_{t=1}^{n_{\mathrm{stream}}}$; ODR summaries from Algorithm 1; fixed slices $I_1, \ldots, I_H$; t_0; ridge constants $\lambda, \rho_0 > 0$; bootstrap replicate count N_{boot}

Output: Online structural dimension $\widehat{d}_t^{\mathrm{ladle}}$ and basis $\widehat{B}_t^{\mathrm{ladle}}$

Set the ladle search cap by

$$q = \begin{cases} p-2, & p < 10, \\ \lfloor p/\log(p) \rfloor, & p \geq 10. \end{cases}$$

Initialize the unweighted ODR summaries $\mathcal{S}_{t_0}$ and, for $b = 1, \ldots, N_{\mathrm{boot}}$, initialize weighted summaries $\mathcal{S}_{t_0}^{*(b)}$ using independent weights $w_{b,i} \sim \mathrm{Poisson}(1)$ for $i = 1, \ldots, t_0$.

for $t = t_0 + 1, t_0 + 2, \ldots, n_{\mathrm{stream}}$ **do**

Update the unweighted summaries as in Algorithm 1 and compute

$$S_t = \widehat{\Sigma}_t + \frac{\rho_0}{t} I_p, \qquad \mathcal{M}_t = S_t^{-1/2} \widehat{K}_t^{\star} S_t^{-1/2}.$$

Compute the eigendecomposition

$$\mathcal{M}_t = \widehat{U}_t^{(\mathcal{M})} \widehat{\Xi}_t^{(\mathcal{M})} \widehat{U}_t^{(\mathcal{M})\top}, \qquad \widehat{\Xi}_t^{(\mathcal{M})} = \mathrm{diag}\big(\widehat{\nu}_{t,1}, \ldots, \widehat{\nu}_{t,p}\big), \qquad \widehat{\nu}_{t,1} \geq \cdots \geq \widehat{\nu}_{t,p}.$$

for $b = 1, \ldots, N_{\mathrm{boot}}$ **do**

Draw $w_{b,t} \sim \mathrm{Poisson}(1)$ independently and update the weighted summaries $\mathcal{S}_t^{*(b)}$ by the weighted analogue of Algorithm 1; equivalently, each incoming contribution at time t is multiplied by $w_{b,t}$.

From $\mathcal{S}_t^{*(b)}$, compute

$$\widehat{K}_t^{\star *(b)}, \qquad S_t^{*(b)} = \widehat{\Sigma}_t^{*(b)} + \frac{\rho_0}{t} I_p, \qquad \mathcal{M}_t^{*(b)} = \big(S_t^{*(b)}\big)^{-1/2} \widehat{K}_t^{\star *(b)} \big(S_t^{*(b)}\big)^{-1/2}.$$

Let $\widehat{U}_{t,k}^{(\mathcal{M})*(b)}$ be the leading k eigenvectors of $\mathcal{M}_t^{*(b)}$ for $k = 1, \ldots, q$.

end for

For $k = 0, \ldots, q$, compute

$$\widehat{f}_{t,0}(0) = 0, \qquad \widehat{f}_{t,0}(k) = \frac{1}{N_{\mathrm{boot}}} \sum_{b=1}^{N_{\mathrm{boot}}} \mathrm{dist}\big(\widehat{U}_{t,k}^{(\mathcal{M})}, \widehat{U}_{t,k}^{(\mathcal{M})*(b)}\big), \qquad k = 1, \ldots, q,$$

$$\widehat{f}_t(k) = \frac{\widehat{f}_{t,0}(k)}{1 + \sum_{\ell=0}^{q} \widehat{f}_{t,0}(\ell)}, \qquad \widehat{\phi}_t(k) = \frac{\widehat{\nu}_{t,k+1}^{+}}{1 + \sum_{\ell=1}^{q+1} \widehat{\nu}_{t,\ell}^{+}}, \qquad \widehat{\nu}_{t,\ell}^{+} = \max(\widehat{\nu}_{t,\ell}, 0),$$

$$\widehat{g}_t(k) = \widehat{f}_t(k) + \widehat{\phi}_t(k).$$

Select

$$\widehat{d}_t^{\mathrm{ladle}} = \operatorname*{arg\,min}_{0 \leq k \leq q} \widehat{g}_t(k), \qquad \widehat{B}_t^{\mathrm{ladle}} = S_t^{-1/2} \widehat{U}_{t,\widehat{d}_t^{\mathrm{ladle}}}^{(\mathcal{M})}.$$

end for

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the proposed online directional regression kernel, metric-aware subspace tracking, weighted-bootstrap ladle selector, and the scope of the theoretical and empirical claims. These claims are supported by the method in Section 2, the main theorems in Section 3, and the simulation and real-data results in Section 4, Appendices B and C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 and Appendix D discuss limitations including fixed dimension and fixed number of slices, the i.i.d. data-stream assumption, tuning choices for ridge and step-size parameters. The theoretical statements also explicitly state local-basin and moment assumptions in Section 3 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main assumptions are summarized in Section 3 and stated formally as (A0)-(A6) in Appendix A. The corresponding lemmas, propositions, theorems, and proofs for the kernel, generalized eigenspace, perturbation update, gradient recursion, and ladle selector are provided and cross-referenced in Section 3 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the proposed algorithms and update formulas in Sections 2 and 3, and gives detailed experimental settings in Section 4, Appendices B, and C, including simulation models, dimensions, tuning parameters, sample sizes, repetition counts, evaluation metrics, benchmark datasets, and runtime environment. In addition, the implementation code is openly available in the public GitHub repository `https://github.com/zxsderty-sys/ODR`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses public benchmark datasets for the real-data experiments and the implementation code is openly available in the public GitHub repository `https://github.com/zxsderty-sys/ODR`. The repository provides the main algorithm code, simulation scripts, and real-data analysis scripts. The accompanying experimental descriptions in Section 4 and Appendices B and C provide the additional implementation details needed to reproduce the main empirical results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The manuscript specifies the simulation models, sample sizes, predictor dimensions, repetition counts, evaluation metrics, online update schemes, benchmark methods, and runtime setup in Section 4, Appendices B, and C. It also describes the real-data train/test protocol and the prediction procedure, which together provide the essential experimental details needed to understand the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Section 4 and Appendices B and C, the simulation and runtime tables report means with standard deviations over repeated replications, and the text states the number of replications used for simulations, runtime studies, ladle studies, and real-data experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The manuscript reports the computing environment used for the runtime experiments, including the CPU model, total memory, and runtime measurements for the main simulation study in Section B. It also reports real-data runtimes in Table 20, thereby providing the essential compute-resource information needed to interpret and reproduce the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work is methodological and uses simulated data together with public benchmark datasets. It does not involve human-subject experiments, private data collection, deception, surveillance, or release of high-risk models.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses broader impacts in Appendix D. The proposed method is methodological and may improve computational efficiency and interpretability for streaming high-dimensional data. The discussion also notes potential negative impacts, including the possibility that biased or drifting data streams may lead to biased reduced representations in high-stakes downstream settings such as biomedical monitoring, finance, or infrastructure monitoring.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk pretrained models, image generators, scraped datasets, or other assets with an evident high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section C identifies the UCI Machine Learning Repository datasets used in the real-data experiments. No external code or pretrained model asset is used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new algorithmic method, and the corresponding code is publicly available at `https://github.com/zxsderty-sys/ODR`. The repository provides the accompanying asset for implementation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or other research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology, theoretical analysis, and experiments do not use LLMs as an important, original, or non-standard research component. Any LLM use limited to writing, editing, or formatting would not affect the scientific methodology or originality and does not require declaration under this checklist item.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.