STATE-DIMENSION LAWS FOR PREDICTION HANKEL GEOMETRY OF FINITE-MEMORY PREDICTORS

Abstract. Modern sequence models succeed by compressing long histories into finite internal states, but there is still no general theory that measures how far a learned state is from the canonical predictor generated by the full past. We propose a prediction-geometric framework for causal sequence modeling. The central benchmark is the full-past Bayes predictor

$$m_t^\star := \mathbb{E}[x_{t+1} \mid \mathcal{F}_t],$$

which provides the canonical one-step predictor for any square-integrable stationary process. For a finite-context predictor and for a learned state z_t, we define context defect and state defect as excess mean-square risk over the full-past benchmark. These defects admit exact geometric identities as squared L^2-distances between conditional expectations. In the Gaussian linear regime this viewpoint recovers the classical Wiener–Kolmogorov framework. This gives the first rigorous bridge from prediction geometry to a quantitative theory of context length and state compression. We then import the explicit phase-series representation of the PACF and derive phase-based and AR/MA-based context laws, including exponential and polynomial short-memory regimes and an L^{-1} context law for FARIMA processes with long memory. Finally, on solvable structured state classes we prove matched two-sided state-dimension laws and explicit memory-allocation principles across heterogeneous channels, and we show that these laws persist under constant channel mixing up to sharp condition-number factors. We also prove that the same context and state exponents survive under invertible frequency-dependent finite-memory channel mixing, with only finite window shifts and additive state overheads.

1. Introduction

A central problem in sequence modeling is to represent the cumulative past by a finite internal state that can be updated online and used for prediction. Several influential modern architectures already move in this direction. HiPPO formulates online memory as optimal projection of the history onto polynomial bases; S4 turns structured continuous-time state-space models into efficient long-sequence architectures; Mamba introduces selective state spaces that adapt their dynamics to the input; and KalmanNet shows that classical state-estimation structure can be combined with learned modules in partially known dynamical systems. These developments strongly suggest that memory, state, and causality should be treated as first-class mathematical objects rather than as incidental features of a particular architecture.

At the same time, current theory still lacks a canonical language for asking whether a learned hidden state is *good* as a summary of the past. Task loss alone does not answer this question. A representation may achieve low empirical error while still mixing predictable structure with innovation, or while failing to approximate the predictor induced by the whole past. Classical prediction theory offers exactly the missing benchmark: the predictor obtained from the full information generated by the past, together with its finite-past analogues, past–future intersection properties, and minimal-observation-length questions.

The aim of this paper is to translate these ideas into a language suitable for deep sequence modeling. Our proposal is simple: a causal sequence model should be viewed as a finite-dimensional compression of the past, and its quality should be measured relative to the full-past Bayes predictor. This leads to three quantitative objects:

(i) context defect, (ii) state defect, (iii) innovation defect.

The first measures the loss from restricting attention to a finite context window; the second measures the loss from encoding the past in a finite state; and the third measures the failure to separate predictable structure from innovation. The long-term objective is a prediction-theoretic foundation for sequence models.

The present draft has two goals. First, it lays out the general geometric framework valid for arbitrary square-integrable stationary processes. Second, it develops the scalar Gaussian/Hardy-space reduction in enough detail that the principal analytic objects are completely explicit: outer factor, innovation variance, finite predictor coefficients, PACF, and the exact relation between context defect and finite-prediction error. This second part is the first genuinely technical step toward the broader program.

2. General Setting and Main Definitions

Let $(x_t)_{t\in\mathbb{Z}}$ be a centered, square-integrable, stationary process with values in $\mathbb{R}^m$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For each time t, let

$$\mathcal{F}_t := \sigma(x_s : s \le t)$$

denote the sigma-algebra generated by the full past.

Definition 2.1 (Full-past Bayes predictor)**.** The *full-past Bayes predictor* is

$$m_t^\star := \mathbb{E}[x_{t+1} \mid \mathcal{F}_t].$$

Definition 2.2 (Finite-context information set)**.** For $L \in \mathbb{N}$, define

$$\mathcal{F}_t^{(L)} := \sigma(x_{t-L+1}, \dots, x_t).$$

Definition 2.3 (Context defect)**.** The *context defect* at window length L is

$$\mathcal{E}_{\mathrm{ctx}}(L) := \mathbb{E}\Big[\|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{F}_t^{(L)}]\|^2\Big] - \mathbb{E}\Big[\|x_{t+1} - m_t^\star\|^2\Big].$$

Now consider a causal representation $z_t \in \mathbb{R}^d$ generated by a measurable update rule

$$z_{t+1} = F_\theta(z_t, x_t).$$

Let

$$\mathcal{G}_t^\theta := \sigma(z_t)$$

be the information carried by the state at time t.

Assumption 2.4 (Causality)**.** For every t, the learned state z_t is $\mathcal{F}_t$-measurable. Equivalently,

$$\mathcal{G}_t^\theta \subseteq \mathcal{F}_t.$$

Definition 2.5 (State defect)**.** The *state defect* of a causal representation z_t is

$$\mathcal{E}_{\mathrm{state}}(\theta) := \mathbb{E}\Big[\|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{G}_t^\theta]\|^2\Big] - \mathbb{E}\Big[\|x_{t+1} - m_t^\star\|^2\Big].$$

Definition 2.6 (Best d-state defect)**.** Let $\mathfrak{F}_d$ be a class of causal d-dimensional state models. Define

$$\mathcal{E}_d^\star := \inf_{\theta\in\mathfrak{F}_d} \mathcal{E}_{\mathrm{state}}(\theta).$$

Definition 2.7 (Innovation)**.** The *innovation* relative to the full past is

$$\eta_{t+1} := x_{t+1} - m_t^\star.$$

By construction,

$$\mathbb{E}[\eta_{t+1} \mid \mathcal{F}_t] = 0.$$

3. The First Exact Geometric Identity

Proposition 3.1 (Prediction-geometric identity)**.** *Let $\mathcal{H} \subseteq \mathcal{F}_t$ be any sub-sigma-algebra. Then*

$$\mathbb{E}\Big[\|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{H}]\|^2\Big] - \mathbb{E}\Big[\|x_{t+1} - m_t^\star\|^2\Big] = \mathbb{E}\Big[\|m_t^\star - \mathbb{E}[x_{t+1} \mid \mathcal{H}]\|^2\Big].$$

In particular,

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathbb{E}\Big[\|m_t^\star - \mathbb{E}[x_{t+1} \mid \mathcal{F}_t^{(L)}]\|^2\Big],$$

and

$$\mathcal{E}_{\mathrm{state}}(\theta) = \mathbb{E}\Big[\|m_t^\star - \mathbb{E}[x_{t+1} \mid \mathcal{G}_t^\theta]\|^2\Big].$$

Proof. Since $\mathcal{H} \subseteq \mathcal{F}_t$, the tower property gives

$$\mathbb{E}[x_{t+1} \mid \mathcal{H}] = \mathbb{E}[m_t^\star \mid \mathcal{H}].$$

Hence

$$x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{H}] = (x_{t+1} - m_t^\star) + (m_t^\star - \mathbb{E}[x_{t+1} \mid \mathcal{H}]).$$

The first term has zero conditional expectation given $\mathcal{F}_t$, hence also given $\mathcal{H}$, while the second term is $\mathcal{H}$-measurable. Therefore the two terms are orthogonal in L^2. Taking squared norms and expectations yields the claim. □

Remark 3.2. Proposition 3.1 says that excess prediction risk is exactly a squared L^2-distance between the full-past predictor and the predictor recoverable from the reduced information set. This is the basic geometric fact behind the entire paper.

4. Gaussian Linear Reduction and a First Context-Length Law

In this section we specialize to the scalar Gaussian linear regime. Then the full-past Bayes predictor becomes the classical Wiener–Kolmogorov predictor, and the finite-context defect becomes the gap between finite-past and whole-past linear prediction errors.

Let $(x_t)_{t\in\mathbb{Z}}$ be a centered scalar stationary Gaussian process. For $L \geq 1$, define

$$H_{t,L}^- := \mathrm{span}\{x_{t-L+1}, \ldots, x_t\} \subset L^2(\Omega)$$

and

$$H_t^- := \overline{\mathrm{span}}\{x_s : s \leq t\}.$$

Let

$$m_t^{(L)} := P_{H_{t,L}^-} x_{t+1}, \qquad m_t^\star := P_{H_t^-} x_{t+1},$$

and define the corresponding prediction-error variances

$$v_L := \mathbb{E}|x_{t+1} - m_t^{(L)}|^2, \qquad v_\infty := \mathbb{E}|x_{t+1} - m_t^\star|^2.$$

Proposition 4.1 (Context defect = finite-past prediction gap)**.** *In the scalar Gaussian linear setting,*

$$\mathcal{E}_{\mathrm{ctx}}(L) = v_L - v_\infty = \mathbb{E}|m_t^\star - m_t^{(L)}|^2.$$

Proof. For Gaussian processes, conditional expectation onto $\mathcal{F}_t^{(L)}$ is the orthogonal projection onto $H_{t,L}^-$, while conditional expectation onto $\mathcal{F}_t$ is the orthogonal projection onto H_t^-. Thus

$$\mathbb{E}[x_{t+1} \mid \mathcal{F}_t^{(L)}] = m_t^{(L)}, \qquad \mathbb{E}[x_{t+1} \mid \mathcal{F}_t] = m_t^\star.$$

The conclusion follows from Proposition 3.1. □

Theorem 4.2 (Exact context-length law for AR(p))**.** *Let (x_t) be a scalar stationary Gaussian autoregressive process of order p,*

$$x_{t+1} = a_1 x_t + \cdots + a_p x_{t-p+1} + \varepsilon_{t+1}, \qquad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma_\varepsilon^2),$$

with stable AR polynomial. Then:

(i) *the full-past predictor is*

$$m_t^\star = a_1 x_t + \cdots + a_p x_{t-p+1};$$

(ii) *for every $L \geq p$,*

$$\mathcal{E}_{\mathrm{ctx}}(L) = 0;$$

(iii) *for $1 \leq L < p$,*

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathbb{E}\left| \sum_{j=L+1}^{p} a_j \Big(x_{t+1-j} - \mathbb{E}[x_{t+1-j} \mid \mathcal{F}_t^{(L)}] \Big) \right|^2 .$$

Proof. Since ε_{t+1} is independent of $\mathcal{F}_t$ and has zero mean,

$$\mathbb{E}[x_{t+1} \mid \mathcal{F}_t] = a_1 x_t + \cdots + a_p x_{t-p+1}.$$

This proves (i).

If $L \geq p$, then $x_t, \ldots, x_{t-p+1} \in \mathcal{F}_t^{(L)}$, hence

$$\mathbb{E}[x_{t+1} \mid \mathcal{F}_t^{(L)}] = a_1 x_t + \cdots + a_p x_{t-p+1} = m_t^\star,$$

so $\mathcal{E}_{\mathrm{ctx}}(L) = 0$, proving (ii).

For $L < p$, write

$$m_t^\star = \sum_{j=1}^{L} a_j x_{t+1-j} + \sum_{j=L+1}^{p} a_j x_{t+1-j}.$$

The first sum is $\mathcal{F}_t^{(L)}$-measurable, so

$$\mathbb{E}[m_t^\star \mid \mathcal{F}_t^{(L)}] = \sum_{j=1}^{L} a_j x_{t+1-j} + \sum_{j=L+1}^{p} a_j \mathbb{E}[x_{t+1-j} \mid \mathcal{F}_t^{(L)}].$$

Since $\mathbb{E}[x_{t+1} \mid \mathcal{F}_t^{(L)}] = \mathbb{E}[m_t^\star \mid \mathcal{F}_t^{(L)}]$, Proposition 3.1 yields the claimed formula. □

Corollary 4.3 (Effective context length for AR(p))**.** *For an AR(p) process, the effective context length at tolerance $\delta = 0$ is exactly p:*

$$L_0 := \inf\{L \geq 1 : \mathcal{E}_{\mathrm{ctx}}(L) = 0\} = p.$$

Proof. This is immediate from Theorem 4.2. □

Theorem 4.4 (Imported classical baseline: exponential finite-past convergence)**.** *Let (x_t) be a q-variate stationary process whose spectral density matrix is sufficiently smooth in the sense of the classical finite-prediction theory literature. Then the finite linear least-squares predictor converges to the Kolmogorov–Wiener predictor at an exponential rate. Consequently, after translating predictor convergence into excess risk, there exist constants $C, c > 0$ such that*

$$0 \leq \mathcal{E}_{\mathrm{ctx}}(L) \leq C e^{-cL}, \qquad L \geq 1.$$

Remark 4.5. Theorem 4.4 is imported as a classical baseline. In the present paper it serves only as motivation: it says that spectrally smooth infinite-memory processes already exhibit a quantitative context-length law.

Corollary 4.6 (Logarithmic effective context length in smooth classes)**.** *Under the assumptions of Theorem 4.4,*

$$L_\delta := \inf\{L \geq 1 : \mathcal{E}_{\mathrm{ctx}}(L) \leq \delta\} \leq \frac{1}{c} \log \frac{C}{\delta}.$$

Proof. If $Ce^{-cL} \leq \delta$, then $\mathcal{E}_{\mathrm{ctx}}(L) \leq \delta$. Solving for L gives the claim. □

5. Scalar Gaussian / Hardy-Space Reduction

This section develops the scalar Gaussian reduction in a form suitable for later quantitative work. We separate the discussion into three layers:

(i) the spectral/Hardy-space layer: outer factor, innovation variance, and phase function;
(ii) the finite-predictor layer: Yule–Walker equations and error variances;
(iii) the recursion layer: Durbin–Levinson, partial autocorrelation, and exact PACF formulas for the context defect.

The main point is that, in the scalar Gaussian regime, the context defect can be written exactly in terms of finite predictor error variances and, therefore, exactly in terms of PACF tails.

5.1. **Spectral factorization, outer factor, and phase function.** Throughout this section, let $(x_t)_{t\in\mathbb{Z}}$ be a centered scalar stationary Gaussian process with autocovariance function

$$\gamma(k) := \mathbb{E}[x_0 x_k], \qquad k \in \mathbb{Z},$$

and spectral density $w \in L^1(\mathbb{T})$, so that

$$\gamma(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\theta} w(e^{i\theta})\, d\theta.$$

We impose the standard Szegö non-determinism condition.

Assumption 5.1 (Szegö condition)**.** The spectral density satisfies

$$w(e^{i\theta}) > 0 \quad \text{for a.e. } \theta, \qquad \log w \in L^1(\mathbb{T}).$$

Under Assumption 5.1, set

$$\sigma^2 := \exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \log w(e^{i\theta})\, d\theta\right) > 0.$$

Then there exists a unique outer function $D \in H^2$ such that

$$D(0) = 1, \qquad w(e^{i\theta}) = \sigma^2 |D(e^{i\theta})|^2 \quad \text{for a.e. } \theta.$$

Write

$$D(z) = \sum_{k=0}^{\infty} c_k z^k, \qquad c_0 = 1.$$

The coefficients (c_k) are the moving-average coefficients associated with the outer factor.

Theorem 5.2 (Imported classical baseline: Wold–Szegö representation)**.** *Under Assumption 5.1, there exists an innovation sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ with*

$$\mathbb{E}[\varepsilon_t] = 0, \qquad \mathbb{E}[\varepsilon_t \varepsilon_s] = 0 \ (t \neq s), \qquad \mathbb{E}[\varepsilon_t^2] = \sigma^2,$$

such that

$$x_t = \sum_{k=0}^{\infty} c_k\, \varepsilon_{t-k} \quad \text{in } L^2.$$

Moreover,

$$m_t^\star = \sum_{k=1}^{\infty} c_k\, \varepsilon_{t+1-k}, \qquad v_\infty = \sigma^2.$$

Remark 5.3. The representation theorem itself is classical and is imported here. What we shall use explicitly are the displayed formulas for the predictor and its error variance.

Proof of the predictor and error formulas from the representation. From the Wold representation,

$$x_{t+1} = c_0\varepsilon_{t+1} + \sum_{k=1}^{\infty} c_k\varepsilon_{t+1-k} = \varepsilon_{t+1} + \sum_{k=1}^{\infty} c_k\varepsilon_{t+1-k},$$

since $c_0 = 1$. The second sum is measurable with respect to the closed span of the past, while ε_{t+1} is orthogonal to that span. Therefore

$$m_t^\star = \sum_{k=1}^{\infty} c_k\varepsilon_{t+1-k}, \qquad x_{t+1} - m_t^\star = \varepsilon_{t+1},$$

whence

$$v_\infty = \mathbb{E}|x_{t+1} - m_t^\star|^2 = \mathbb{E}[\varepsilon_{t+1}^2] = \sigma^2.$$

□

To introduce the phase function, we strengthen the assumptions slightly whenever inverse-factor manipulations are needed.

Assumption 5.4 (Inverse-square-integrable regime)**.** In addition to Assumption 5.1, assume

$$\frac{1}{w} \in L^1(\mathbb{T}).$$

Equivalently, $D^{-1} \in H^2$.

Under Assumption 5.4, write

$$A(z) := \frac{1}{D(z)} = \sum_{k=0}^{\infty} a_k z^k, \qquad a_0 = 1.$$

The coefficients (a_k) play the role of one-sided autoregressive coefficients.

Definition 5.5 (Phase function and phase coefficients)**.** Under Assumption 5.4, define the *phase function*

$$q(e^{i\theta}) := \frac{\overline{D(e^{i\theta})}}{D(e^{i\theta})} = \overline{D(e^{i\theta})}\, A(e^{i\theta}), \qquad |q(e^{i\theta})| = 1 \text{ a.e.}$$

For $n \in \mathbb{Z}$, define its Fourier coefficients by

$$\beta_n := \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\theta} q(e^{i\theta})\, d\theta.$$

Lemma 5.6 (Phase coefficients as outer/inverse-factor convolutions)**.** *Assume Assumption 5.4 and, in addition, that (a_k) and (c_k) belong to ℓ^1. Then for every $n \geq 0$,*

$$\beta_n = \sum_{v=0}^{\infty} c_v a_{n+v}.$$

Proof. By the ℓ^1 assumptions,

$$D(e^{i\theta}) = \sum_{k=0}^{\infty} c_k e^{ik\theta}, \qquad A(e^{i\theta}) = \sum_{k=0}^{\infty} a_k e^{ik\theta}$$

converge absolutely and uniformly on the unit circle. Hence

$$\overline{D(e^{i\theta})} = \sum_{v=0}^{\infty} c_v e^{-iv\theta}$$

(the coefficients are real because the process is real-valued), and therefore

$$q(e^{i\theta}) = \left(\sum_{v=0}^{\infty} c_v e^{-iv\theta}\right)\left(\sum_{k=0}^{\infty} a_k e^{ik\theta}\right).$$

Absolute convergence allows termwise multiplication and rearrangement:

$$q(e^{i\theta}) = \sum_{n\in\mathbb{Z}} \left(\sum_{v-k=-n} c_v a_k\right) e^{in\theta}.$$

The coefficient of $e^{in\theta}$ for $n \geq 0$ is obtained by setting $k = n + v$, namely

$$\beta_n = \sum_{v=0}^{\infty} c_v a_{n+v}.$$

□

Remark 5.7. Lemma 5.6 is the cleanest coefficient-level identity connecting the outer factor, its inverse, and the phase function. In more general regimes, one should interpret the same identity in weaker function-space senses. For the present draft, the ℓ^1 assumption is imposed only to keep every manipulation elementary and fully rigorous.

5.2. **Finite predictor coefficients and Yule–Walker equations.** For $L \geq 1$, let

$$\Gamma_L := \big(\gamma(i-j)\big)_{1\leq i,j\leq L}$$

be the $L \times L$ Toeplitz covariance matrix, and let

$$\gamma_L := (\gamma(1), \ldots, \gamma(L))^\top.$$

Since $w > 0$ a.e., the matrix Γ_L is strictly positive definite for each L, hence invertible.

Proposition 5.8 (Finite predictor coefficients and Yule–Walker equations)**.** *For every $L \geq 1$, there exists a unique vector*

$$\phi_L = (\phi_{L,1}, \ldots, \phi_{L,L})^\top \in \mathbb{R}^L$$

such that

$$m_t^{(L)} = \sum_{j=1}^{L} \phi_{L,j} x_{t+1-j}.$$

Moreover, the coefficient vector is characterized by the Yule–Walker system

$$\Gamma_L \phi_L = \gamma_L,$$

and the finite-predictor error variance is

$$v_L = \gamma(0) - \gamma_L^\top \Gamma_L^{-1} \gamma_L.$$

Consequently,

$$\mathcal{E}_{\mathrm{ctx}}(L) = \gamma(0) - \gamma_L^\top \Gamma_L^{-1} \gamma_L - \sigma^2.$$

Proof. Because $H_{t,L}^- = \operatorname{span}\{x_t, \dots, x_{t-L+1}\}$ is finite-dimensional, the orthogonal projection $m_t^{(L)} = P_{H_{t,L}^-} x_{t+1}$ admits a unique representation

$$m_t^{(L)} = \sum_{j=1}^{L} \phi_{L,j} x_{t+1-j}$$

for a unique coefficient vector ϕ_L.

By the orthogonality characterization of projection,

$$\mathbb{E}\Big[\Big(x_{t+1} - \sum_{j=1}^{L} \phi_{L,j} x_{t+1-j}\Big) x_{t+1-i}\Big] = 0, \qquad i = 1, \dots, L.$$

Using stationarity, this becomes

$$\gamma(i) - \sum_{j=1}^{L} \phi_{L,j} \gamma(i-j) = 0, \qquad i = 1, \dots, L,$$

which is exactly the matrix equation $\Gamma_L \phi_L = \gamma_L$.

For the error variance,

$$\begin{aligned} v_L &= \mathbb{E}\left| x_{t+1} - \sum_{j=1}^{L} \phi_{L,j} x_{t+1-j} \right|^2 \\ &= \gamma(0) - 2\phi_L^\top \gamma_L + \phi_L^\top \Gamma_L \phi_L. \end{aligned}$$

Since $\Gamma_L \phi_L = \gamma_L$, this simplifies to

$$v_L = \gamma(0) - \phi_L^\top \gamma_L = \gamma(0) - \gamma_L^\top \Gamma_L^{-1} \gamma_L.$$

Finally, Theorem 5.2 gives $v_\infty = \sigma^2$, so Proposition 4.1 yields

$$\mathcal{E}_{\mathrm{ctx}}(L) = v_L - v_\infty = \gamma(0) - \gamma_L^\top \Gamma_L^{-1} \gamma_L - \sigma^2.$$

□

Remark 5.9. Proposition 5.8 makes the first exact finite-dimensional reduction of context defect completely explicit: in the scalar Gaussian regime, $\mathcal{E}_{\mathrm{ctx}}(L)$ is exactly a Toeplitz inverse formula minus the innovation variance.

5.3. **Durbin–Levinson recursion and PACF formulas.**

Definition 5.10 (Partial autocorrelation coefficients)**.** For $L \geq 1$, define the partial autocorrelation coefficient

$$\alpha_L := \phi_{L,L}.$$

Let J_{L-1} denote the $(L-1) \times (L-1)$ reversal matrix,

$$J_{L-1}(u_1, \dots, u_{L-1})^\top = (u_{L-1}, \dots, u_1)^\top.$$

Theorem 5.11 (Durbin–Levinson recursion)**.** *For each $L \geq 2$, the finite predictor coefficients satisfy*

$$\begin{aligned} \alpha_L &= \frac{\gamma(L) - \sum_{j=1}^{L-1} \phi_{L-1,j} \gamma(L-j)}{v_{L-1}}, \\ \phi_{L,1:L-1} &= \phi_{L-1} - \alpha_L J_{L-1} \phi_{L-1}, \\ v_L &= v_{L-1}(1 - \alpha_L^2), \end{aligned}$$

where $\phi_{L,1:L-1} := (\phi_{L,1}, \dots, \phi_{L,L-1})^\top$ *and* $\phi_{L-1} := (\phi_{L-1,1}, \dots, \phi_{L-1,L-1})^\top$.

Proof. Write the Yule–Walker system for order L in block form:

$$\begin{pmatrix} \Gamma_{L-1} & J_{L-1}\gamma_{L-1} \\ (J_{L-1}\gamma_{L-1})^\top & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_{L,1:L-1} \\ \alpha_L \end{pmatrix} = \begin{pmatrix} \gamma_{L-1} \\ \gamma(L) \end{pmatrix}.$$

The top block equation reads

$$\Gamma_{L-1}\phi_{L,1:L-1} + (J_{L-1}\gamma_{L-1})\alpha_L = \gamma_{L-1}.$$

Since $\Gamma_{L-1}\phi_{L-1} = \gamma_{L-1}$ and Γ_{L-1} commutes with J_{L-1} (because Γ_{L-1} is symmetric Toeplitz), we have

$$\Gamma_{L-1}(J_{L-1}\phi_{L-1}) = J_{L-1}\Gamma_{L-1}\phi_{L-1} = J_{L-1}\gamma_{L-1}.$$

Thus

$$\Gamma_{L-1}\big(\phi_{L,1:L-1} - \phi_{L-1} + \alpha_L J_{L-1}\phi_{L-1}\big) = 0.$$

Since Γ_{L-1} is invertible,

$$\phi_{L,1:L-1} = \phi_{L-1} - \alpha_L J_{L-1}\phi_{L-1}.$$

This proves the second identity.

Now examine the bottom block equation:

$$(J_{L-1}\gamma_{L-1})^\top \phi_{L,1:L-1} + \gamma(0)\alpha_L = \gamma(L).$$

Substituting the coefficient recursion gives

$$(J_{L-1}\gamma_{L-1})^\top \phi_{L-1} - \alpha_L (J_{L-1}\gamma_{L-1})^\top J_{L-1}\phi_{L-1} + \gamma(0)\alpha_L = \gamma(L).$$

Because J_{L-1} is orthogonal,

$$(J_{L-1}\gamma_{L-1})^\top J_{L-1}\phi_{L-1} = \gamma_{L-1}^\top \phi_{L-1}.$$

Using

$$v_{L-1} = \gamma(0) - \gamma_{L-1}^\top \phi_{L-1},$$

we obtain

$$\alpha_L = \frac{\gamma(L) - (J_{L-1}\gamma_{L-1})^\top \phi_{L-1}}{v_{L-1}} = \frac{\gamma(L) - \sum_{j=1}^{L-1} \phi_{L-1,j}\gamma(L-j)}{v_{L-1}},$$

which proves the first identity.

Finally,

$$\begin{aligned} v_L &= \gamma(0) - \gamma_L^\top \phi_L \\ &= \gamma(0) - \gamma_{L-1}^\top \phi_{L,1:L-1} - \gamma(L)\alpha_L \\ &= \gamma(0) - \gamma_{L-1}^\top (\phi_{L-1} - \alpha_L J_{L-1}\phi_{L-1}) - \gamma(L)\alpha_L \\ &= v_{L-1} + \alpha_L \gamma_{L-1}^\top J_{L-1}\phi_{L-1} - \gamma(L)\alpha_L. \end{aligned}$$

Since

$$\gamma_{L-1}^\top J_{L-1}\phi_{L-1} = (J_{L-1}\gamma_{L-1})^\top \phi_{L-1},$$

and

$$\gamma(L) - (J_{L-1}\gamma_{L-1})^\top \phi_{L-1} = \alpha_L v_{L-1},$$

we conclude that

$$v_L = v_{L-1} - \alpha_L^2 v_{L-1} = v_{L-1}(1 - \alpha_L^2).$$

This proves the third identity. □

Corollary 5.12 (Finite-product formula for prediction error variances)**.** *For every $L \geq 1$,*

$$v_L = \gamma(0) \prod_{j=1}^{L} (1 - \alpha_j^2).$$

Moreover,

$$v_\infty = \gamma(0) \prod_{j=1}^{\infty} (1 - \alpha_j^2).$$

Proof. The finite product formula follows by iterating Theorem 5.11, starting from $v_0 = \gamma(0)$. Since $v_L \downarrow v_\infty$ by monotone convergence of orthogonal projections, taking $L \to \infty$ yields the infinite product identity. □

Theorem 5.13 (Exact PACF representation of the context defect)**.** *For every $L \geq 0$,*

$$\mathcal{E}_{\mathrm{ctx}}(L) = v_L - v_\infty = \sum_{n=L+1}^{\infty} (v_{n-1} - v_n) = \sum_{n=L+1}^{\infty} \alpha_n^2 v_{n-1}.$$

Consequently,

$$v_\infty \sum_{n=L+1}^{\infty} \alpha_n^2 \leq \mathcal{E}_{\mathrm{ctx}}(L) \leq \gamma(0) \sum_{n=L+1}^{\infty} \alpha_n^2.$$

Proof. By Proposition 4.1,

$$\mathcal{E}_{\mathrm{ctx}}(L) = v_L - v_\infty.$$

Since $v_n \downarrow v_\infty$, telescoping gives

$$v_L - v_\infty = \sum_{n=L+1}^{\infty} (v_{n-1} - v_n).$$

By Theorem 5.11,

$$v_{n-1} - v_n = v_{n-1} - v_{n-1}(1 - \alpha_n^2) = \alpha_n^2 v_{n-1}.$$

Thus

$$\mathcal{E}_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} \alpha_n^2 v_{n-1}.$$

Because

$$v_\infty \leq v_{n-1} \leq v_0 = \gamma(0)$$

for all n, the two-sided PACF bound follows immediately. □

Corollary 5.14 (Exponential and polynomial context laws from PACF decay)**.** *The following implications hold.*

(i) *If there exist constants $C > 0$ and $\rho \in (0, 1)$ such that*

$$|\alpha_n| \leq C\rho^n \qquad (n \geq 1),$$

then

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq \gamma(0) C^2 \sum_{n=L+1}^{\infty} \rho^{2n} = \frac{\gamma(0) C^2}{1 - \rho^2} \rho^{2L+2}.$$

(ii) *If there exist constants $C > 0$ and $s > \frac{1}{2}$ such that*

$$|\alpha_n| \leq C n^{-s} \qquad (n \geq 1),$$

then there exists a constant $C_s > 0$ such that

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq C_s L^{1-2s} \qquad (L \geq 1).$$

Proof. Part (i) follows from the upper bound in Theorem 5.13 and the geometric-series formula.

For part (ii), Theorem 5.13 yields

$$\mathcal{E}_{\mathrm{ctx}}(L) \le \gamma(0)C^2 \sum_{n=L+1}^{\infty} n^{-2s}.$$

Since $2s > 1$, the tail of the p-series satisfies

$$\sum_{n=L+1}^{\infty} n^{-2s} \le \int_L^{\infty} x^{-2s}\,dx = \frac{1}{2s-1} L^{1-2s}.$$

Thus one may take $C_s = \gamma(0)C^2/(2s-1)$. □

Remark 5.15. Corollary 5.14 is one of the first genuinely useful quantitative consequences of the present framework. It says that any class of processes for which one can control PACF decay automatically yields a context-length law.

5.4. **How phase information enters the context law.** We now summarize the role of the phase function. The outer factor D determines the Wold coefficients, its inverse A determines the one-sided inverse filter whenever Assumption 5.4 holds, and the phase function

$$q = \overline{D}\, A$$

packages the relative position of past and future in a single unimodular function. In structured classes, one expects to estimate finite predictor coefficients and PACF tails through Fourier information carried by q.

The exact quantitative mechanism is class-dependent. In later stages of the project, this will be connected to alternating projections, forward/backward innovations, and phase-based formulas in finite-prediction theory. For the purposes of the present draft, the key logical point is simpler:

> *Once phase information yields usable estimates on α_n, Theorem 5.13 immediately turns those estimates into a context-length law for $\mathcal{E}_{\mathrm{ctx}}(L)$.*

Thus the flow of ideas is

$$\text{outer factor } D \longrightarrow \text{phase function } q \longrightarrow \text{PACF estimates } \alpha_n \longrightarrow \mathcal{E}_{\mathrm{ctx}}(L).$$

5.5. **Summary of the scalar Gaussian reduction.**

Proposition 5.16 (Summary formulas)**.** *Under Assumption 5.1, the scalar Gaussian context defect satisfies*

$$\mathcal{E}_{\mathrm{ctx}}(L) = v_L - \sigma^2,$$

where v_L is the finite-prediction error variance. More explicitly,

$$v_L = \gamma(0) - \gamma_L^{\top}\Gamma_L^{-1}\gamma_L, \qquad \mathcal{E}_{\mathrm{ctx}}(L) = \gamma(0) - \gamma_L^{\top}\Gamma_L^{-1}\gamma_L - \sigma^2.$$

If $\alpha_L = \phi_{L,L}$ denotes the PACF, then

$$v_L = \gamma(0)\prod_{j=1}^{L}(1-\alpha_j^2), \qquad \mathcal{E}_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} \alpha_n^2 v_{n-1}.$$

Therefore,

$$v_\infty \sum_{n=L+1}^{\infty} \alpha_n^2 \le \mathcal{E}_{\mathrm{ctx}}(L) \le \gamma(0) \sum_{n=L+1}^{\infty} \alpha_n^2.$$

Proof. The first two identities are Proposition 5.8. The product formula is Corollary 5.12. The PACF-tail formula and the two-sided bounds are Theorem 5.13. □

Remark 5.17. Proposition 5.16 is the main technical endpoint of the present progress draft. It reduces the scalar Gaussian context-defect problem to completely classical objects: Toeplitz covariance matrices, finite predictor coefficients, error variances, and PACF tails. This is the precise sense in which the prediction-geometric framework has now been made concrete.

6. Phase-Series Representation and Phase-Based Context Laws

Proposition 5.16 reduces the scalar Gaussian context defect to the PACF. The next step is therefore to estimate the PACF from analytic data attached to the spectral density. The most direct such data are the Fourier coefficients of the phase function. In this section we import the explicit phase-series representation of the PACF due to Bingham, Inoue, and Kasahara and combine it with the exact defect identity from Theorem 5.13. This yields a general mechanism:

$$\text{phase coefficients } \beta_n \Longrightarrow \text{PACF bounds } \alpha_n \Longrightarrow \text{context laws for } \mathcal{E}_{\mathrm{ctx}}(L).$$

The first half of the section works directly with the phase coefficients. The second half shows how to translate short-memory bounds on the AR coefficients into PACF and context-defect bounds. We then record an imported long-memory PACF asymptotic for FARIMA and deduce the resulting context law.

6.1. Short-memory hypothesis and phase-series representation.

Assumption 6.1 (Short-memory regime)**.** The process is purely nondeterministic and both coefficient sequences $(a_n)_{n\ge 0}$ and $(c_n)_{n\ge 0}$ belong to ℓ^1.

Under Assumption 6.1, the phase coefficients are absolutely summable.

Lemma 6.2 (Summability of the phase coefficients)**.** *Under Assumption 6.1, the sequence $(\beta_n)_{n\ge 0}$ belongs to ℓ^1. More precisely,*

$$\sum_{n=0}^{\infty} |\beta_n| \le \Big(\sum_{v=0}^{\infty} |c_v|\Big)\Big(\sum_{u=0}^{\infty} |a_u|\Big).$$

Consequently, the tails

$$B(n) := \sum_{u=0}^{\infty} |\beta_{n+u}|, \qquad n \ge 0,$$

are finite and satisfy $B(n) \downarrow 0$ as $n \to \infty$.

Proof. Using the coefficient identity from Lemma 5.6,

$$|\beta_n| \le \sum_{v=0}^{\infty} |c_v|\,|a_{n+v}|.$$

Summing in n and applying Tonelli's theorem gives

$$\sum_{n=0}^{\infty} |\beta_n| \le \sum_{n=0}^{\infty}\sum_{v=0}^{\infty} |c_v|\,|a_{n+v}| = \sum_{v=0}^{\infty} |c_v| \sum_{m=v}^{\infty} |a_m| \le \Big(\sum_{v=0}^{\infty} |c_v|\Big)\Big(\sum_{m=0}^{\infty} |a_m|\Big).$$

Hence $(\beta_n) \in \ell^1$, and therefore $B(n) \downarrow 0$. □

To write the phase-series representation, we introduce the auxiliary kernels

$$\begin{aligned} \delta_0(n,j) &:= \mathbb{1}_{\{j=0\}}, \\ \delta_1(n,j) &:= \beta_{n+j}, \\ \delta_{k+1}(n,j) &:= \sum_{v=0}^{\infty} \beta_{n+j+v}\,\delta_k(n,v), \qquad k \ge 1, \end{aligned}$$

for $n, j \in \mathbb{N} \cup \{0\}$. We also define the odd phase iterates by

$$\vartheta_1(n) := \beta_n, \qquad \vartheta_{2k+1}(n) := \sum_{v=0}^{\infty} \beta_{n+v}\, \delta_{2k}(n+1, v), \qquad k \geq 1.$$

Expanding the recursion shows that

$$\vartheta_{2k+1}(n) = \sum_{v_1,\dots,v_{2k} \geq 0} \beta_{n+v_1} \beta_{n+1+v_1+v_2} \cdots \beta_{n+1+v_{2k-1}+v_{2k}} \beta_{n+1+v_{2k}},$$

so $\vartheta_{2k+1}(n)$ is the $(2k+1)$-fold odd phase convolution at lag n.

Theorem 6.3 (Imported phase-series representation of the PACF)**.** *Assume Assumption 6.1. Then for every $n \geq 2$,*

$$\alpha_n = -\sum_{k=0}^{\infty} \vartheta_{2k+1}(n),$$

where the series is absolutely convergent for all sufficiently large n. In the notation of Bingham–Inoue–Kasahara, this is the explicit representation of the Verblunsky coefficients in terms of the Fourier coefficients of the phase function, translated into the sign convention of the present draft.[11, 12]

Remark 6.4. Theorem 6.3 is the precise bridge between Hardy-space data and the PACF. Section 5 identified β_n as the nth Fourier coefficient of the phase function $\overline{D}/D$ in our sign convention; the imported theorem then expresses the PACF as the negative odd phase series built from these coefficients. Since all quantitative bounds below are taken in absolute value, the sign plays no further role in the context-law estimates.[11]

6.2. **A direct phase-tail bound for the PACF.**

Lemma 6.5 (Tail-kernel estimate)**.** *Assume Assumption 6.1. Then for every $k \geq 1$ and every $n \geq 0$,*

$$\sum_{j=0}^{\infty} |\delta_k(n, j)| \leq B(n)^k.$$

Proof. We argue by induction on k. For $k = 1$,

$$\sum_{j=0}^{\infty} |\delta_1(n, j)| = \sum_{j=0}^{\infty} |\beta_{n+j}| = B(n).$$

Now assume the claim holds for some $k \geq 1$. Then, by the defining recursion for δ_{k+1} and Tonelli's theorem,

$$\begin{aligned}
\sum_{j=0}^{\infty} |\delta_{k+1}(n, j)| &\leq \sum_{j=0}^{\infty} \sum_{v=0}^{\infty} |\beta_{n+j+v}|\, |\delta_k(n, v)| \\
&= \sum_{v=0}^{\infty} |\delta_k(n, v)| \sum_{j=0}^{\infty} |\beta_{n+v+j}| \\
&\leq B(n) \sum_{v=0}^{\infty} |\delta_k(n, v)| \\
&\leq B(n)^{k+1}.
\end{aligned}$$

This closes the induction. □

Theorem 6.6 (Phase-tail PACF bound)**.** *Assume Assumption 6.1. If $B(N+1) < 1$ for some $N \in \mathbb{N}$, then for every $n \geq N$,*

$$|\alpha_n| \leq \frac{B(n)}{1 - B(n+1)^2}.$$

In particular, $\alpha_n = O(B(n))$ as $n \to \infty$.

Proof. Fix $n \geq N$. By Theorem 6.3,

$$\alpha_n = \sum_{k=0}^{\infty} \vartheta_{2k+1}(n).$$

For the first term,

$$|\vartheta_1(n)| = |\beta_n| \leq B(n).$$

For $k \geq 1$, Lemma 6.5 gives

$$\begin{aligned} |\vartheta_{2k+1}(n)| &\leq \sum_{v=0}^{\infty} |\beta_{n+v}|\, |\delta_{2k}(n+1, v)| \\ &\leq \Big(\sup_{v \geq 0} |\beta_{n+v}|\Big) \sum_{v=0}^{\infty} |\delta_{2k}(n+1, v)| \\ &\leq B(n)\, B(n+1)^{2k}. \end{aligned}$$

Therefore,

$$|\alpha_n| \leq \sum_{k=0}^{\infty} |\vartheta_{2k+1}(n)| \leq B(n) \sum_{k=0}^{\infty} B(n+1)^{2k} = \frac{B(n)}{1 - B(n+1)^2},$$

as claimed. □

Corollary 6.7 (Exponential phase decay implies exponential context law)**.** *Assume Assumption 6.1, and suppose there exist constants $C_\beta > 0$ and $\rho \in (0, 1)$ such that*

$$|\beta_n| \leq C_\beta \rho^n, \qquad n \geq 0.$$

Then there exist constants $C_\alpha, C_{\mathrm{ctx}} > 0$ such that

$$|\alpha_n| \leq C_\alpha \rho^n, \qquad n \geq 1,$$

and

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq C_{\mathrm{ctx}} \rho^{2L}, \qquad L \geq 1.$$

Proof. The geometric bound on β_n implies

$$B(n) = \sum_{u=0}^{\infty} |\beta_{n+u}| \leq \frac{C_\beta}{1 - \rho} \rho^n.$$

Hence $B(n+1) < 1/2$ for all sufficiently large n. By Theorem 6.6, there exist constants N and C_α such that

$$|\alpha_n| \leq C_\alpha \rho^n, \qquad n \geq N.$$

Enlarging C_α if necessary to absorb the finitely many small values of n gives the first bound for all $n \geq 1$.

Now apply Theorem 5.13:

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq \gamma(0) \sum_{n=L+1}^{\infty} \alpha_n^2 \leq \gamma(0) C_\alpha^2 \sum_{n=L+1}^{\infty} \rho^{2n} = \frac{\gamma(0) C_\alpha^2 \rho^2}{1 - \rho^2} \rho^{2L}.$$

This proves the claim. □

Corollary 6.8 (A coarse polynomial phase law)**.** *Assume Assumption 6.1, and suppose that for some $p > 1$ and some $C_\beta > 0$,*

$$|\beta_n| \le C_\beta n^{-p}, \qquad n \ge 1.$$

Then there exists a constant $C_\alpha > 0$ such that

$$|\alpha_n| \le C_\alpha n^{1-p}, \qquad n \ge 1.$$

If in addition $p > \frac{3}{2}$, then there exists $C_{\mathrm{ctx}} > 0$ such that

$$\mathcal{E}_{\mathrm{ctx}}(L) \le C_{\mathrm{ctx}} L^{3-2p}, \qquad L \ge 1.$$

Proof. Since $p > 1$, the tail estimate for a p-series gives

$$B(n) \le C_\beta \sum_{m=n}^{\infty} m^{-p} \le \frac{C_\beta}{p-1}(n-1)^{1-p}, \qquad n \ge 2.$$

Hence $B(n+1) < 1/2$ for all sufficiently large n, so Theorem 6.6 yields

$$|\alpha_n| \le 2B(n) \le C_\alpha n^{1-p}$$

for all large n, and therefore for all $n \ge 1$ after enlarging C_α.

If $p > \frac{3}{2}$, then $2p - 2 > 1$, so Theorem 5.13 implies

$$\mathcal{E}_{\mathrm{ctx}}(L) \le \gamma(0) C_\alpha^2 \sum_{n=L+1}^{\infty} n^{-2p+2}.$$

Applying the integral test,

$$\sum_{n=L+1}^{\infty} n^{-2p+2} \le \int_L^{\infty} x^{-2p+2}\, dx = \frac{1}{2p-3} L^{3-2p},$$

which proves the context bound. □

Remark 6.9. Corollary 6.8 is intentionally stated as a coarse direct consequence of the phase-series representation. It loses one power because $B(n)$ is a tail sum of the phase coefficients. The next subsection shows how, in short-memory classes, one can recover sharper PACF bounds from the AR coefficients and thereby improve the context law.

6.3. **Sharpening via the AR coefficients.** Define

$$\|c\|_{\ell^1} := \sum_{v=0}^{\infty} |c_v|, \qquad F(n) := \|c\|_{\ell^1} \sum_{u=n}^{\infty} |a_u|, \qquad n \ge 0.$$

Lemma 6.10 (Phase tails controlled by AR tails)**.** *Under Assumption 6.1, for every $n \ge 0$,*

$$B(n) \le F(n), \qquad \sup_{v \ge 0} |\beta_{n+v}| \le \|c\|_{\ell^1} \sup_{j \ge n} |a_j|.$$

Proof. Using Lemma 5.6,

$$|\beta_{n+u}| \le \sum_{v=0}^{\infty} |c_v|\, |a_{n+u+v}|.$$

Summing in u yields

$$\begin{aligned} B(n) &= \sum_{u=0}^{\infty} |\beta_{n+u}| \\ &\le \sum_{u=0}^{\infty} \sum_{v=0}^{\infty} |c_v|\,|a_{n+u+v}| \\ &= \sum_{v=0}^{\infty} |c_v| \sum_{m=n+v}^{\infty} |a_m| \\ &\le \|c\|_{\ell^1} \sum_{m=n}^{\infty} |a_m| = F(n). \end{aligned}$$

Likewise,

$$|\beta_{n+u}| \le \sum_{v=0}^{\infty} |c_v| \sup_{j\ge n} |a_j| = \|c\|_{\ell^1} \sup_{j\ge n} |a_j|,$$

which proves the second estimate after taking the supremum over $u \ge 0$. □

Theorem 6.11 (Sharper short-memory PACF bound)**.** *Assume Assumption 6.1. If $F(N+1) < 1$ for some $N \in \mathbb{N}$, then for every $n \ge N$,*

$$|\alpha_n| \le \frac{\|c\|_{\ell^1}}{1 - F(n+1)^2} \sup_{j\ge n} |a_j|.$$

In particular, $\alpha_n = O\big(\sup_{j\ge n} |a_j|\big)$ as $n \to \infty$.

Proof. By Lemma 6.10,

$$|\vartheta_1(n)| = |\beta_n| \le \|c\|_{\ell^1} \sup_{j\ge n} |a_j|.$$

For $k \ge 1$, Lemmas 6.5 and 6.10 give

$$\begin{aligned} |\vartheta_{2k+1}(n)| &\le \Big(\sup_{v\ge 0} |\beta_{n+v}|\Big) \sum_{v=0}^{\infty} |\delta_{2k}(n+1, v)| \\ &\le \|c\|_{\ell^1} \sup_{j\ge n} |a_j|\, B(n+1)^{2k} \\ &\le \|c\|_{\ell^1} \sup_{j\ge n} |a_j|\, F(n+1)^{2k}. \end{aligned}$$

Therefore, using Theorem 6.3,

$$|\alpha_n| \le \|c\|_{\ell^1} \sup_{j\ge n} |a_j| \sum_{k=0}^{\infty} F(n+1)^{2k} = \frac{\|c\|_{\ell^1}}{1 - F(n+1)^2} \sup_{j\ge n} |a_j|,$$

which proves the claim. □

Corollary 6.12 (Exponential short-memory law from AR tails)**.** *Assume Assumption 6.1, and suppose there exist constants $A > 0$ and $\rho \in (0,1)$ such that*

$$|a_n| \le A\rho^n, \qquad n \ge 0.$$

Then there exist constants $C_\alpha, C_{\mathrm{ctx}} > 0$ such that

$$|\alpha_n| \le C_\alpha \rho^n, \qquad n \ge 1,$$

and

$$\mathcal{E}_{\mathrm{ctx}}(L) \le C_{\mathrm{ctx}} \rho^{2L}, \qquad L \ge 1.$$

Proof. Because

$$F(n) = \|c\|_{\ell^1} \sum_{u=n}^{\infty} |a_u| \le \frac{\|c\|_{\ell^1} A}{1-\rho} \rho^n,$$

we have $F(n+1) < 1/2$ for all sufficiently large n. Theorem 6.11 then yields

$$|\alpha_n| \le 2\|c\|_{\ell^1} A \rho^n$$

for all large n, and hence for all $n \ge 1$ after adjusting the constant. The context-defect bound then follows exactly as in the proof of Corollary 6.7. □

Corollary 6.13 (Polynomial short-memory law from AR tails)**.** *Assume Assumption 6.1, and suppose there exist constants $A > 0$ and $p > 1$ such that*

$$|a_n| \le A n^{-p}, \qquad n \ge 1.$$

Then there exist constants $C_\alpha, C_{\mathrm{ctx}} > 0$ such that

$$|\alpha_n| \le C_\alpha n^{-p}, \qquad n \ge 1,$$

and

$$\mathcal{E}_{\mathrm{ctx}}(L) \le C_{\mathrm{ctx}} L^{1-2p}, \qquad L \ge 1.$$

Proof. Because $p > 1$,

$$F(n) = \|c\|_{\ell^1} \sum_{u=n}^{\infty} |a_u| \le A\|c\|_{\ell^1} \sum_{u=n}^{\infty} u^{-p} \le \frac{A\|c\|_{\ell^1}}{p-1} n^{1-p},$$

so $F(n+1) < 1/2$ for all sufficiently large n. Theorem 6.11 therefore yields

$$|\alpha_n| \le 2A\|c\|_{\ell^1} n^{-p}$$

for all large n, and hence for all $n \ge 1$ after enlarging the constant.

Applying Theorem 5.13,

$$\mathcal{E}_{\mathrm{ctx}}(L) \le \gamma(0) C_\alpha^2 \sum_{n=L+1}^{\infty} n^{-2p}.$$

Since $2p > 1$, the integral test gives

$$\sum_{n=L+1}^{\infty} n^{-2p} \le \int_L^{\infty} x^{-2p}\, dx = \frac{1}{2p-1} L^{1-2p},$$

which proves the claim. □

Remark 6.14. Corollary 6.13 is substantially sharper than the coarse phase-tail bound of Corollary 6.8: the exponent p is preserved. This is exactly why the AR coefficients remain useful even after the phase-series representation is available.

6.4. **Imported long-memory baseline: FARIMA.** The explicit phase-series representation is also useful in long-memory classes. The cleanest baseline is the FARIMA case. Bingham, Inoue, and Kasahara prove that for a causal invertible FARIMA(p, d, q) process with $0 < d < \frac{1}{2}$, the PACF has the sharp asymptotic expansion

$$\alpha_n = \frac{d}{n} + O(n^{-1-d}).$$

This is more precise than merely saying that α_n is of order n^{-1}.[11, 12]

Theorem 6.15 (Imported PACF asymptotic for FARIMA)**.** *Let (x_t) be a causal invertible FARIMA(p,d,q) process with $0<d<\frac{1}{2}$. Then*

$$\alpha_n = \frac{d}{n} + O(n^{-1-d}), \qquad n \to \infty.$$

Corollary 6.16 (Long-memory context law for FARIMA)**.** *Under the assumptions of Theorem 6.15,*

$$\mathcal{E}_{\mathrm{ctx}}(L) \sim \sigma^2 d^2 L^{-1}, \qquad L \to \infty.$$

Proof. By Theorem 6.15,

$$\alpha_n^2 = \frac{d^2}{n^2} + O(n^{-2-d}).$$

Hence

$$\sum_{n=L+1}^{\infty} \alpha_n^2 = d^2 \sum_{n=L+1}^{\infty} n^{-2} + O\left(\sum_{n=L+1}^{\infty} n^{-2-d} \right) = \frac{d^2}{L} + O(L^{-1-d}).$$

Now write

$$\mathcal{E}_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} \alpha_n^2 v_{n-1} = \sigma^2 \sum_{n=L+1}^{\infty} \alpha_n^2 + \sum_{n=L+1}^{\infty} \alpha_n^2 (v_{n-1} - \sigma^2),$$

using Theorem 5.13. Since $v_{n-1} \downarrow \sigma^2$, given $\varepsilon > 0$ there exists N such that

$$0 \le v_{n-1} - \sigma^2 \le \varepsilon, \qquad n \ge N.$$

For $L \ge N$ this yields

$$0 \le \sum_{n=L+1}^{\infty} \alpha_n^2 (v_{n-1} - \sigma^2) \le \varepsilon \sum_{n=L+1}^{\infty} \alpha_n^2.$$

Therefore

$$\mathcal{E}_{\mathrm{ctx}}(L) = \sigma^2 \sum_{n=L+1}^{\infty} \alpha_n^2 + o\left(\sum_{n=L+1}^{\infty} \alpha_n^2 \right) \sim \sigma^2 \frac{d^2}{L},$$

which proves the result. □

7. Linear State Compression and Weighted Rational Approximation

The context-defect theory developed above studies one very special family of causal compressions: the raw shift-register state $(x_t, \dots, x_{t-L+1})$. To move toward a genuine theory of hidden-state dimension, one needs a richer class of compressions. The natural next class is the family of stable finite-dimensional linear time-invariant state representations.

The key point of this section is that, in the scalar Gaussian regime, the best achievable excess risk in this class is an *exact weighted rational approximation problem*. In innovation coordinates this becomes an exact Hardy-space approximation problem.

7.1. Innovation-space coordinates and the prediction symbol.

Under Assumption 5.1, Theorem 5.2 gives the Wold–Szegö representation

$$x_t = \sum_{k=0}^{\infty} c_k \varepsilon_{t-k}, \qquad c_0 = 1, \qquad \mathbb{E}[\varepsilon_t^2] = \sigma^2.$$

It is convenient to encode the predictable part of x_{t+1} directly in innovation coordinates.

Definition 7.1 (Prediction symbol)**.** Define the *prediction symbol*

$$F(z) := \sum_{j=0}^{\infty} c_{j+1} z^j = \frac{D(z) - 1}{z}, \qquad z \in \mathbb{D}.$$

Since $D \in H^2$ and $D(0) = 1$, we have $F \in H^2$.

Lemma 7.2 (Innovation-space isometry)**.** *Define $U_t : H_t^- \to H^2$ on finite linear combinations of past innovations by*

$$U_t\Big(\sum_{j=0}^{N} b_j \varepsilon_{t-j}\Big) := \frac{1}{\sigma} \sum_{j=0}^{N} b_j z^j.$$

Then U_t extends uniquely to a unitary map from the closed span of $(\varepsilon_t, \varepsilon_{t-1}, \dots)$ onto H^2.

Proof. Because the innovations are orthogonal and all have variance σ^2,

$$\Big\| \sum_{j=0}^{N} b_j \varepsilon_{t-j} \Big\|_{L^2(\Omega)}^2 = \sigma^2 \sum_{j=0}^{N} |b_j|^2 = \sigma^2 \Big\| \sum_{j=0}^{N} b_j z^j \Big\|_{H^2}^2.$$

Thus U_t is an isometry on a dense subspace of the innovation-past space, and it extends uniquely to a unitary map onto H^2. □

Proposition 7.3 (The full-past predictor as a Hardy-space vector)**.** *Under the unitary map U_t,*

$$U_t m_t^\star = F.$$

Equivalently,

$$m_t^\star = \sum_{j=0}^{\infty} c_{j+1} \varepsilon_{t-j}.$$

Proof. The second identity is the predictor formula already obtained from the Wold representation in Theorem 5.2. Applying Lemma 7.2 to

$$m_t^\star = \sum_{j=0}^{\infty} c_{j+1} \varepsilon_{t-j}$$

gives

$$U_t m_t^\star = \sum_{j=0}^{\infty} c_{j+1} z^j = F.$$

□

7.2. **General causal linear predictors.** We first analyze arbitrary one-step linear predictors before restricting to finite-dimensional state-space realizations.

Definition 7.4 (Causal linear predictor kernel)**.** Let $r = (r_j)_{j \ge 1}$ be a real sequence such that $\sum_{j \ge 1} |r_j| < \infty$. Its *predictor symbol* is

$$R(z) := \sum_{j=1}^{\infty} r_j z^{j-1}, \qquad z \in \mathbb{D}.$$

The associated causal linear predictor is

$$\widehat{x}_{t+1}^{(R)} := \sum_{j=1}^{\infty} r_j x_{t+1-j}.$$

Absolute summability guarantees that the series converges in L^2, since $\sup_t \|x_t\|_2^2 = \gamma(0)$.

Proposition 7.5 (Innovation-space image of a linear predictor)**.** *Let R be a causal linear predictor symbol with absolutely summable coefficients. Then*

$$U_t \widehat{x}_{t+1}^{(R)} = RD \quad \text{in } H^2.$$

Proof. From the Wold representation,

$$x_{t+1-j} = \sum_{k=0}^{\infty} c_k \varepsilon_{t+1-j-k}.$$

Therefore,

$$\widehat{x}_{t+1}^{(R)} = \sum_{j=1}^{\infty} r_j \sum_{k=0}^{\infty} c_k \varepsilon_{t+1-j-k}.$$

Since (r_j) and (c_k) are absolutely summable, the double series is absolutely convergent in L^2 and may be rearranged:

$$\widehat{x}_{t+1}^{(R)} = \sum_{m=0}^{\infty} \Big(\sum_{j=1}^{m+1} r_j c_{m+1-j} \Big) \varepsilon_{t-m}.$$

The coefficient of ε_{t-m} is exactly the coefficient of z^m in the product

$$R(z)D(z) = \Big(\sum_{j=1}^{\infty} r_j z^{j-1} \Big) \Big(\sum_{k=0}^{\infty} c_k z^k \Big).$$

Applying the unitary map U_t therefore yields

$$U_t \widehat{x}_{t+1}^{(R)} = RD.$$

□

Theorem 7.6 (Exact Hardy-space formula for linear prediction defect)**.** *Let R be a causal linear predictor symbol with absolutely summable coefficients. Define its excess risk over the full-past benchmark by*

$$\mathcal{E}_{\mathrm{lin}}(R) := \mathbb{E} \big| x_{t+1} - \widehat{x}_{t+1}^{(R)} \big|^2 - \sigma^2.$$

Then

$$\mathcal{E}_{\mathrm{lin}}(R) = \mathbb{E} \big| m_t^\star - \widehat{x}_{t+1}^{(R)} \big|^2 = \sigma^2 \, \|F - RD\|_{H^2}^2.$$

Moreover,

$$\mathcal{E}_{\mathrm{lin}}(R) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |F(e^{i\theta}) - R(e^{i\theta}) D(e^{i\theta})|^2 \, \sigma^2 \, d\theta.$$

Proof. Since $\widehat{x}_{t+1}^{(R)}$ is measurable with respect to the full past, Proposition 3.1 gives

$$\mathcal{E}_{\mathrm{lin}}(R) = \mathbb{E} \big| m_t^\star - \widehat{x}_{t+1}^{(R)} \big|^2.$$

Now apply the unitary map U_t from Lemma 7.2. By Propositions 7.3 and 7.5,

$$U_t \big(m_t^\star - \widehat{x}_{t+1}^{(R)} \big) = F - RD.$$

Hence

$$\mathbb{E} \big| m_t^\star - \widehat{x}_{t+1}^{(R)} \big|^2 = \sigma^2 \|F - RD\|_{H^2}^2.$$

The boundary-integral formula is just the H^2 norm identity on the unit circle. □

Remark 7.7. Theorem 7.6 is the exact Hardy-space reduction for one-step linear prediction. It says that, after passing to innovation coordinates, every causal linear predictor is represented by the analytic function RD, while the full-past benchmark is represented by the single vector $F = (D-1)/z$. The problem of causal prediction is therefore an approximation problem in H^2.

When the inverse filter exists, the same identity can be rewritten as a weighted approximation problem for the autoregressive predictor symbol.

Proposition 7.8 (Weighted approximation formula under inverse factorization)**.** *Assume Assumption 5.4, and let*

$$P(z) := -\frac{A(z)-1}{z} = -\sum_{j=1}^{\infty} a_j z^{j-1}.$$

Then

$$F = PD.$$

Consequently, for every causal linear predictor symbol R with absolutely summable coefficients,

$$\mathcal{E}_{\mathrm{lin}}(R) = \sigma^2 \|(R-P)D\|_{H^2}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |R(e^{i\theta}) - P(e^{i\theta})|^2 w(e^{i\theta})\, d\theta.$$

Proof. Because $A(z)D(z) = 1$ and $D(0) = 1$,

$$PD = -\frac{A-1}{z}D = -\frac{1-D}{z} = \frac{D-1}{z} = F.$$

Substituting this identity into Theorem 7.6 gives

$$\mathcal{E}_{\mathrm{lin}}(R) = \sigma^2 \|(R-P)D\|_{H^2}^2.$$

Since $w = \sigma^2|D|^2$ almost everywhere on the unit circle, the weighted $L^2(w)$ formula follows immediately. □

7.3. **Stable finite-state realizations.** We now restrict to finite-dimensional causal states.

Definition 7.9 (Stable d-state LTI predictor)**.** A *stable d-state LTI predictor* is a triple (A, b, c) with

$$A \in \mathbb{R}^{d\times d}, \qquad b, c \in \mathbb{R}^d, \qquad \rho(A) < 1,$$

where $\rho(A)$ is the spectral radius of A. Its stationary causal state is

$$s_t := \sum_{j=0}^{\infty} A^j b\, x_{t-j},$$

and its one-step predictor is

$$\widehat{x}_{t+1}^{A,b,c} := c^\top s_t.$$

Equivalently, the state satisfies the recursion

$$s_t = As_{t-1} + bx_t,$$

and the predictor is $\widehat{x}_{t+1}^{A,b,c} = c^\top s_t$.

The series defining s_t converges in L^2 because $\rho(A) < 1$ implies $\|A^j\| \le C\rho^j$ for some $C < \infty$ and some $\rho \in (0, 1)$.

Proposition 7.10 (Transfer symbol of a stable finite-state predictor)**.** *Let (A, b, c) be a stable d-state LTI predictor. Then*

$$\widehat{x}_{t+1}^{A,b,c} = \sum_{j=1}^{\infty} r_j x_{t+1-j}, \qquad r_j := c^\top A^{j-1} b,$$

with absolutely summable coefficients. The corresponding predictor symbol is the strictly proper stable rational function

$$R_{A,b,c}(z) := \sum_{j=1}^{\infty} r_j z^{j-1} = c^\top (I - zA)^{-1} b.$$

Proof. By definition,

$$s_t = \sum_{j=0}^{\infty} A^j b\, x_{t-j},$$

so

$$\widehat{x}_{t+1}^{A,b,c} = \sum_{j=0}^{\infty} c^\top A^j b\, x_{t-j} = \sum_{m=1}^{\infty} c^\top A^{m-1} b\, x_{t+1-m}.$$

Thus $r_m = c^\top A^{m-1} b$. Since $\rho(A) < 1$, the coefficients decay geometrically, hence are absolutely summable. Finally,

$$\sum_{j=1}^{\infty} r_j z^{j-1} = \sum_{j=1}^{\infty} c^\top A^{j-1} b\, z^{j-1} = c^\top \Big(\sum_{j=0}^{\infty} (zA)^j \Big) b = c^\top (I - zA)^{-1} b,$$

for $|z| < 1$. □

Definition 7.11 (Best d-state LTI defect)**.** Let $\mathcal{R}_d^{\mathrm{ss}}$ denote the set of predictor symbols of the form $R_{A,b,c}$ arising from stable d-state LTI predictors. Define

$$\mathcal{E}_d^{\mathrm{LTI}} := \inf_{R \in \mathcal{R}_d^{\mathrm{ss}}} \mathcal{E}_{\mathrm{lin}}(R).$$

Corollary 7.12 (Best finite-state defect as weighted rational approximation)**.** *Under Assumption 5.1,*

$$\mathcal{E}_d^{\mathrm{LTI}} = \sigma^2 \inf_{R \in \mathcal{R}_d^{\mathrm{ss}}} \|F - RD\|_{H^2}^2.$$

If Assumption 5.4 also holds, then

$$\mathcal{E}_d^{\mathrm{LTI}} = \inf_{R \in \mathcal{R}_d^{\mathrm{ss}}} \frac{1}{2\pi} \int_{-\pi}^{\pi} |R(e^{i\theta}) - P(e^{i\theta})|^2 w(e^{i\theta})\, d\theta.$$

Proof. This is immediate from Proposition 7.10, Theorem 7.6, and Proposition 7.8. □

Remark 7.13. Corollary 7.12 is the first exact state-dimension reduction in the paper. It says that the optimal d-state linear hidden representation is not an abstract learning object: it is exactly a weighted rational approximation problem for the canonical predictor symbol.

7.4. **Shift-register realization and immediate state-dimension laws.** The finite-context predictor is itself a d-state LTI model. This gives an immediate comparison between context defect and state defect.

Proposition 7.14 (Shift-register realization of the finite-context predictor)**.** *For every $d \geq 1$, there exists a stable d-state LTI predictor whose one-step output equals the finite-context predictor $m_t^{(d)}$. Consequently,*

$$\mathcal{E}_d^{\mathrm{LTI}} \leq \mathcal{E}_{\mathrm{ctx}}(d).$$

Proof. Let

$$s_t := (x_t, x_{t-1}, \dots, x_{t-d+1})^\top \in \mathbb{R}^d.$$

Then s_t obeys the stable shift-register recursion

$$s_t = A_{\mathrm{sh}} s_{t-1} + b_{\mathrm{sh}} x_t,$$

where

$$A_{\mathrm{sh}} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \end{pmatrix}, \qquad b_{\mathrm{sh}} = e_1.$$

The matrix A_{sh} is nilpotent, hence stable. By Proposition 5.8,

$$m_t^{(d)} = \sum_{j=1}^{d} \phi_{d,j} x_{t+1-j} = c_d^\top s_t, \qquad c_d := (\phi_{d,1}, \dots, \phi_{d,d})^\top.$$

Thus $m_t^{(d)}$ is realized by a stable d-state LTI predictor. Its excess risk is exactly $\mathcal{E}_{\mathrm{ctx}}(d)$ by Proposition 4.1. Taking the infimum over all d-state LTI predictors gives

$$\mathcal{E}_d^{\mathrm{LTI}} \le \mathcal{E}_{\mathrm{ctx}}(d).$$

□

Corollary 7.15 (State-dimension laws inherited from context laws)**.** *Every upper bound proved earlier for $\mathcal{E}_{\mathrm{ctx}}(d)$ automatically yields the same upper bound for $\mathcal{E}_d^{\mathrm{LTI}}$. In particular:*

(i) *under the exponential PACF-decay hypotheses of Corollary 5.14(i),*

$$\mathcal{E}_d^{\mathrm{LTI}} \le C\rho^{2d};$$

(ii) *under the polynomial short-memory AR-tail hypotheses of Corollary 6.13,*

$$\mathcal{E}_d^{\mathrm{LTI}} \le C d^{1-2p};$$

(iii) *for causal invertible FARIMA(p, δ, q) with $0 < \delta < \frac{1}{2}$,*

$$\mathcal{E}_d^{\mathrm{LTI}} \le C d^{-1},$$

and, more precisely,

$$\mathcal{E}_d^{\mathrm{LTI}} \le \mathcal{E}_{\mathrm{ctx}}(d) \sim \sigma^2 \delta^2 d^{-1} \qquad (d \to \infty),$$

after substituting the FARIMA context law.

Proof. This is immediate from Proposition 7.14 and the previously established bounds for $\mathcal{E}_{\mathrm{ctx}}(L)$. □

Remark 7.16. The comparison $\mathcal{E}_d^{\mathrm{LTI}} \le \mathcal{E}_{\mathrm{ctx}}(d)$ is important conceptually. A learned d-dimensional state can always choose to imitate the raw shift register of length d, but it may also compress the past more efficiently than the raw last-d-lag window. Thus the context law is a universal benchmark, not necessarily the end of the story.

There is also a second, more direct d-state construction: truncate the infinite autoregressive predictor itself.

Proposition 7.17 (Truncated-AR realization bound)**.** *Assume Assumption 5.4 and suppose, in addition, that $(a_j)_{j\ge 1} \in \ell^1$. For $d \ge 1$, define the truncated autoregressive predictor symbol*

$$P_d(z) := -\sum_{j=1}^{d} a_j z^{j-1}.$$

Then P_d is realized by a stable d-state shift-register predictor, and

$$\mathcal{E}_d^{\mathrm{LTI}} \le \mathcal{E}_{\mathrm{lin}}(P_d) = \mathbb{E}\Big|\sum_{j=d+1}^{\infty} a_j x_{t+1-j}\Big|^2 \le \gamma(0)\Big(\sum_{j=d+1}^{\infty} |a_j|\Big)^2.$$

Proof. The shift-register realization is the same as in Proposition 7.14, now with output vector

$$\widetilde{c}_d := (-a_1, \dots, -a_d)^\top.$$

By Proposition 7.8, the full-past predictor is

$$m_t^\star = -\sum_{j=1}^{\infty} a_j x_{t+1-j},$$

so

$$m_t^\star - \widehat{x}_{t+1}^{(P_d)} = -\sum_{j=d+1}^{\infty} a_j x_{t+1-j}.$$

Therefore

$$\mathcal{E}_{\mathrm{lin}}(P_d) = \mathbb{E}\Big|\sum_{j=d+1}^{\infty} a_j x_{t+1-j}\Big|^2.$$

Finally,

$$\Big\|\sum_{j=d+1}^{\infty} a_j x_{t+1-j}\Big\|_2 \le \sum_{j=d+1}^{\infty} |a_j|\,\|x_{t+1-j}\|_2 = \sqrt{\gamma(0)}\sum_{j=d+1}^{\infty} |a_j|,$$

which yields the stated upper bound after squaring. □

Corollary 7.18 (Direct state-dimension bounds from AR tails)**.** *Under the assumptions of Proposition 7.17:*

(i) *if $|a_j| \le A\rho^j$ for some $A > 0$ and $\rho \in (0,1)$, then*

$$\mathcal{E}_d^{\mathrm{LTI}} \le C\rho^{2d};$$

(ii) *if $|a_j| \le Aj^{-p}$ for some $A > 0$ and $p > 1$, then*

$$\mathcal{E}_d^{\mathrm{LTI}} \le Cd^{2-2p}.$$

Proof. In the exponential case,

$$\sum_{j=d+1}^{\infty} |a_j| \le \frac{A}{1-\rho}\rho^{d+1},$$

so Proposition 7.17 yields an $O(\rho^{2d})$ bound. In the polynomial case,

$$\sum_{j=d+1}^{\infty} |a_j| \le A\sum_{j=d+1}^{\infty} j^{-p} \le \frac{A}{p-1}d^{1-p},$$

and squaring gives the $O(d^{2-2p})$ bound. □

Remark 7.19. The direct AR-truncation bound is generally weaker than the bound inherited from the exact finite-context predictor, because truncating the infinite autoregressive coefficients is only one possible d-state approximation. Nevertheless, it is useful because it is completely explicit and already exhibits the weighted-rational-approximation character of the problem.

7.5. **Interpretation and the next approximation problem.** Section 7 identifies the first exact mathematical core of state compression:

$$\boxed{\mathcal{E}_d^{\mathrm{LTI}} = \sigma^2 \inf_{R \in \mathcal{R}_d^{\mathrm{ss}}} \|F - RD\|_{H^2}^2}.$$

Equivalently, when the inverse filter is available,

$$\boxed{\mathcal{E}_d^{\mathrm{LTI}} = \inf_{R \in \mathcal{R}_d^{\mathrm{ss}}} \|R - P\|_{L^2(w)}^2}.$$

So the optimal d-state linear hidden representation problem is exactly a stable weighted rational approximation problem for the canonical predictor symbol.

Classical control and model-reduction theory strongly suggest that this problem should ultimately be organized by finite-rank approximation, balanced realizations, and Hankel singular values; modern deep state-space model compression has already started to exploit these quantities directly. The next technical target is therefore to turn the exact approximation formulas of the present section into a genuine approximation theorem, ideally one that controls $\mathcal{E}_d^{\mathrm{LTI}}$ by a canonical sequence of compressibility numbers associated with the predictor map.[14, 15, 16, 17]

8. Pole Realizations and Approximation-Transfer Principles

Section 7 reduced the best finite-state linear predictor to a stable weighted rational approximation problem. The present section pushes that reduction one step further in two directions:

(i) it gives a completely explicit hard-cutoff theorem for predictor symbols with finitely many stable poles;
(ii) it shows how any external approximation family for the canonical predictor symbol P automatically transfers into a state-dimension law for $\mathcal{E}_d^{\mathrm{LTI}}$.

The resulting statements are still elementary, but they make the approximation-theoretic content of Section 7 much more concrete.

8.1. **Exact realizations from finitely many stable poles.** We work under Assumption 5.4, so that the canonical autoregressive predictor symbol

$$P(z) = -\sum_{j=1}^{\infty} a_j z^{j-1}$$

is defined and Proposition 7.8 is available.

Theorem 8.1 (Finite-pole realization theorem)**.** *Assume Assumption 5.4. Suppose that the canonical predictor symbol has the form*

$$P(z) = \sum_{\ell=1}^{r} \frac{\alpha_\ell}{1 - \lambda_\ell z}, \qquad |\lambda_\ell| < 1,$$

for some $r \in \mathbb{N}$. Then there exists a stable r-state LTI predictor whose predictor symbol is exactly P. Consequently,

$$\mathcal{E}_r^{\mathrm{LTI}} = 0.$$

Proof. Let

$$A := \operatorname{diag}(\lambda_1, \dots, \lambda_r), \qquad b := (1, \dots, 1)^\top, \qquad c := (\alpha_1, \dots, \alpha_r)^\top.$$

Since $|\lambda_\ell| < 1$ for each ℓ, the matrix A is stable. By Proposition 7.10, the associated stable r-state LTI predictor has predictor symbol

$$R_{A,b,c}(z) = c^\top (I - zA)^{-1} b.$$

Because A is diagonal,

$$(I - zA)^{-1} = \operatorname{diag}\left(\frac{1}{1 - \lambda_1 z}, \dots, \frac{1}{1 - \lambda_r z} \right),$$

and therefore

$$R_{A,b,c}(z) = \sum_{\ell=1}^{r} \frac{\alpha_\ell}{1 - \lambda_\ell z} = P(z).$$

Now apply Proposition 7.8 with $R = P$. It gives

$$\mathcal{E}_{\mathrm{lin}}(P) = \sigma^2 \|(P - P)D\|_{H^2}^2 = 0.$$

Since $P \in \mathcal{R}_r^{\mathrm{ss}}$, taking the infimum over all stable r-state LTI predictors yields

$$\mathcal{E}_r^{\mathrm{LTI}} = 0.$$

□

Corollary 8.2 (Hard cutoff for finite pole complexity)**.** *Under the assumptions of Theorem 8.1,*

$$\mathcal{E}_d^{\mathrm{LTI}} = 0 \qquad \text{for every } d \geq r.$$

Proof. If $d \geq r$, one may embed the r-state realization from Theorem 8.1 into dimension d by adjoining $d - r$ zero modes. Hence $P \in \mathcal{R}_d^{\mathrm{ss}}$, and the same argument gives zero defect. □

Remark 8.3. Theorem 8.1 is the state-dimension analogue of the AR(p) hard-cutoff theorem from Section 4. There the hard cutoff comes from finite context length; here it comes from finite pole complexity of the canonical predictor symbol.

8.2. **Approximation-transfer principles.** The exact reduction of Corollary 7.12 shows that every approximation scheme for P immediately generates an upper bound on $\mathcal{E}_d^{\mathrm{LTI}}$. We record two useful abstract transfer principles.

Proposition 8.4 (Weighted transfer principle)**.** *Assume Assumption 5.4. Let $d \geq 1$, and let $R_d \in \mathcal{R}_d^{\mathrm{ss}}$ be any stable d-state predictor symbol. Then*

$$\mathcal{E}_d^{\mathrm{LTI}} \leq \mathcal{E}_{\mathrm{lin}}(R_d) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |R_d(e^{i\theta}) - P(e^{i\theta})|^2 w(e^{i\theta})\, d\theta.$$

Proof. Since $R_d \in \mathcal{R}_d^{\mathrm{ss}}$, the definition of $\mathcal{E}_d^{\mathrm{LTI}}$ gives

$$\mathcal{E}_d^{\mathrm{LTI}} \leq \mathcal{E}_{\mathrm{lin}}(R_d).$$

The weighted integral identity is exactly Proposition 7.8. □

Proposition 8.5 (Unweighted H^2 transfer under bounded outer factor)**.** *Assume Assumptions 5.4 and 6.1. Let $d \geq 1$, and let $R_d \in \mathcal{R}_d^{\mathrm{ss}}$ be such that $P - R_d \in H^2$. Then*

$$\mathcal{E}_d^{\mathrm{LTI}} \leq \mathcal{E}_{\mathrm{lin}}(R_d) \leq \sigma^2 \|D\|_{H^\infty}^2 \, \|P - R_d\|_{H^2}^2.$$

In particular, any H^2 approximation rate for P by stable d-state rational functions yields the same squared rate for the state defect.

Proof. By Proposition 7.8,

$$\mathcal{E}_{\mathrm{lin}}(R_d) = \sigma^2 \|(R_d - P)D\|_{H^2}^2.$$

Since Assumption 6.1 implies $(c_k) \in \ell^1$, the outer factor satisfies

$$D(z) = \sum_{k=0}^{\infty} c_k z^k \in H^\infty, \qquad \|D\|_{H^\infty} \le \sum_{k=0}^{\infty} |c_k|.$$

Therefore multiplication by D is a bounded operator on H^2, and

$$\|(R_d - P)D\|_{H^2} \le \|D\|_{H^\infty} \, \|R_d - P\|_{H^2}.$$

Combining this with the definition of $\mathcal{E}_d^{\mathrm{LTI}}$ gives the stated bound. □

8.3. **Pole-tail state-dimension laws.** A particularly transparent approximation family arises when P admits an absolutely summable expansion in stable simple poles.

Theorem 8.6 (Pole-tail upper bound)**.** *Assume Assumptions 5.4 and 6.1. Suppose that*

$$P(z) = \sum_{\ell=1}^{\infty} \frac{\alpha_\ell}{1 - \lambda_\ell z}, \qquad \sup_{\ell \ge 1} |\lambda_\ell| \le \rho < 1, \qquad \sum_{\ell=1}^{\infty} |\alpha_\ell| < \infty.$$

For $d \ge 1$, define the truncated pole sum

$$P_d(z) := \sum_{\ell=1}^{d} \frac{\alpha_\ell}{1 - \lambda_\ell z}.$$

Then $P_d \in \mathcal{R}_d^{\mathrm{ss}}$ and

$$\mathcal{E}_d^{\mathrm{LTI}} \le \sigma^2 \|D\|_{H^\infty}^2 \, \|P - P_d\|_{H^2}^2 \le \frac{\sigma^2 \|D\|_{H^\infty}^2}{1 - \rho^2} \left(\sum_{\ell=d+1}^{\infty} |\alpha_\ell| \right)^2.$$

Proof. By the same diagonal realization used in Theorem 8.1, the truncated sum P_d is realized by a stable d-state LTI predictor, so $P_d \in \mathcal{R}_d^{\mathrm{ss}}$.

By Proposition 8.5, it is enough to estimate $\|P - P_d\|_{H^2}$. Now

$$P(z) - P_d(z) = \sum_{\ell=d+1}^{\infty} \frac{\alpha_\ell}{1 - \lambda_\ell z},$$

and for each ℓ,

$$\left\| \frac{1}{1 - \lambda_\ell z} \right\|_{H^2}^2 = \sum_{n=0}^{\infty} |\lambda_\ell|^{2n} = \frac{1}{1 - |\lambda_\ell|^2} \le \frac{1}{1 - \rho^2}.$$

Hence, by the triangle inequality,

$$\|P - P_d\|_{H^2} \le \sum_{\ell=d+1}^{\infty} |\alpha_\ell| \left\| \frac{1}{1 - \lambda_\ell z} \right\|_{H^2} \le \frac{1}{\sqrt{1 - \rho^2}} \sum_{\ell=d+1}^{\infty} |\alpha_\ell|.$$

Squaring and substituting into Proposition 8.5 yields the claim. □

Corollary 8.7 (Exponential and polynomial pole-tail laws)**.** *Under the assumptions of Theorem 8.6:*

(i) *if there exist constants* $A > 0$ *and* $\eta \in (0,1)$ *such that*

$$\sum_{\ell=d+1}^{\infty} |\alpha_\ell| \le A\eta^d, \qquad d \ge 1,$$

then

$$\mathcal{E}_d^{\mathrm{LTI}} \le C\eta^{2d};$$

(ii) *if there exist constants* $A > 0$ *and* $p > 0$ *such that*

$$\sum_{\ell=d+1}^{\infty} |\alpha_\ell| \le Ad^{-p}, \qquad d \ge 1,$$

then

$$\mathcal{E}_d^{\mathrm{LTI}} \le Cd^{-2p}.$$

for suitable constants C*.*

Proof. This is immediate from Theorem 8.6. □

Remark 8.8. The pole-tail law makes the approximation-theoretic content of the state-dimension problem completely explicit. Fast decay of the canonical pole weights implies fast decay of the best finite-state defect. In this sense, pole compressibility is one concrete mechanism underlying hidden-state compressibility.

9. Matrix-Valued Extension: Block Prediction, Phase Matrices, and IPF

This section begins the genuinely multivariate extension of the draft. The main point is that the scalar context-defect formulas have precise block-matrix analogues, while the scalar phase function is replaced by a matrix-valued phase function and the scalar past/future geometry is replaced by the intersection-of-past-and-future (IPF) property. We do not yet attempt to reproduce the full multivariate phase-series formulas from the literature. Instead, we record the exact block identities that the later theory will need and isolate the places where matrix phase functions and IPF enter.

For notational convenience, we now allow the process to be complex-valued. Thus let $(X_t)_{t\in\mathbb{Z}}$ be a centered q-variate stationary Gaussian process with values in $\mathbb{C}^q$. Write

$$\Gamma(k) := \mathbb{E}[X_0 X_k^*], \qquad k \in \mathbb{Z},$$

so that $\Gamma(-k) = \Gamma(k)^*$. Assume X has matrix spectral density $W \in L^1(\mathbb{T})^{q\times q}$:

$$\Gamma(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ik\theta} W(e^{i\theta})\, d\theta.$$

Assumption 9.1 (Matrix Szegö regime)**.** The spectral density satisfies

$$W(e^{i\theta}) = W(e^{i\theta})^* > 0 \quad \text{for a.e. } \theta, \qquad \log\det W \in L^1(\mathbb{T}).$$

Assumption 9.2 (Matrix inverse-integrable regime)**.** In addition to Assumption 9.1, assume

$$W^{-1} \in L^1(\mathbb{T})^{q\times q}.$$

9.1. **Normalized outer factorization and innovations.** Under Assumption 9.1, matrix spectral factorization gives a normalized outer factor and a positive definite innovation covariance matrix.

Theorem 9.3 (Imported classical baseline: normalized matrix Wold–Szegö factorization)**.** *Assume Assumption 9.1. Then there exist a positive definite matrix $\Sigma \in \mathbb{C}^{q\times q}$ and an outer matrix-valued Hardy function*

$$H(z) = \sum_{k=0}^{\infty} C_k z^k \in H_2^{q\times q}, \qquad C_0 = I_q,$$

such that

$$W(e^{i\theta}) = H(e^{i\theta})\,\Sigma\,H(e^{i\theta})^* \qquad \text{for a.e. } \theta.$$

Consequently, there exists an innovation sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ with

$$\mathbb{E}[\varepsilon_t] = 0, \qquad \mathbb{E}[\varepsilon_t \varepsilon_s^*] = \delta_{ts}\Sigma,$$

such that

$$X_t = \sum_{k=0}^{\infty} C_k \varepsilon_{t-k} \qquad \text{in } L^2(\Omega;\mathbb{C}^q).$$

Moreover, if

$$M_t^\star := \mathbb{E}[X_{t+1} \mid \mathcal{F}_t],$$

then

$$M_t^\star = \sum_{k=1}^{\infty} C_k \varepsilon_{t+1-k}, \qquad V_\infty := \mathbb{E}[(X_{t+1} - M_t^\star)(X_{t+1} - M_t^\star)^*] = \Sigma.$$

Remark 9.4. The factorization and representation themselves are classical and are imported here; see, for instance, the multivariate prediction literature built around outer matrix factorizations and dual processes. In the present draft, what we use explicitly are the displayed formulas for the full-past predictor and its error covariance.[10]

Proof of the predictor and error formulas from the representation. From the representation,

$$X_{t+1} = \varepsilon_{t+1} + \sum_{k=1}^{\infty} C_k \varepsilon_{t+1-k}.$$

The second term belongs to the closed linear span of the past, while ε_{t+1} is orthogonal to that span. Hence

$$M_t^\star = \sum_{k=1}^{\infty} C_k \varepsilon_{t+1-k}, \qquad X_{t+1} - M_t^\star = \varepsilon_{t+1}.$$

Therefore

$$V_\infty = \mathbb{E}[\varepsilon_{t+1}\varepsilon_{t+1}^*] = \Sigma.$$

□

9.2. **Finite block prediction and defect covariances.** For $L \geq 1$, define the block Toeplitz covariance matrix

$$T_L(W) := (\Gamma(i-j))_{1\leq i,j\leq L} \in \mathbb{C}^{Lq\times Lq},$$

and the block row

$$G_L := [\Gamma(1) \ \cdots \ \Gamma(L)] \in \mathbb{C}^{q\times Lq}.$$

Lemma 9.5 (Strict positivity of the block Toeplitz covariance matrix)**.** *Under Assumption 9.1, the matrix $T_L(W)$ is Hermitian positive definite for every $L \geq 1$.*

Proof. Let $u = (u_1^\top, \dots, u_L^\top)^\top \in \mathbb{C}^{Lq}$, and define the vector polynomial

$$p_u(z) := \sum_{j=1}^{L} z^{j-1} u_j \in \mathbb{C}^q.$$

Then

$$u^* T_L(W) u = \sum_{i,j=1}^{L} u_i^* \Gamma(i-j) u_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} p_u(e^{i\theta})^* W(e^{i\theta}) p_u(e^{i\theta})\, d\theta.$$

Since $W(e^{i\theta}) > 0$ for a.e. θ, the integral is nonnegative. If it vanishes, then $p_u(e^{i\theta}) = 0$ for a.e. θ. As p_u is a vector polynomial, it must vanish identically, so $u_1 = \cdots = u_L = 0$. Therefore $T_L(W)$ is positive definite. □

For each $L \geq 1$, define the finite-past predictor

$$M_t^{(L)} := \mathbb{E}[X_{t+1} \mid \mathcal{F}_t^{(L)}], \qquad \mathcal{F}_t^{(L)} := \sigma(X_{t-L+1}, \dots, X_t).$$

Since the process is Gaussian, this is the orthogonal projection of X_{t+1} onto $\operatorname{span}\{X_t, \dots, X_{t-L+1}\}$.

Proposition 9.6 (Finite predictor matrices and block Yule–Walker equations)**.** *For every $L \geq 1$, there exists a unique block row*

$$\Phi_L = [\Phi_{L,1} \ \cdots \ \Phi_{L,L}] \in \mathbb{C}^{q \times Lq}, \qquad \Phi_{L,j} \in \mathbb{C}^{q \times q},$$

such that

$$M_t^{(L)} = \sum_{j=1}^{L} \Phi_{L,j} X_{t+1-j}.$$

Moreover,

$$\Phi_L T_L(W) = G_L, \qquad \Phi_L = G_L\, T_L(W)^{-1},$$

and the finite-prediction error covariance is

$$V_L := \mathbb{E}[(X_{t+1} - M_t^{(L)})(X_{t+1} - M_t^{(L)})^*] = \Gamma(0) - G_L T_L(W)^{-1} G_L^*.$$

Proof. Because the prediction space is finite-dimensional, there exists a unique block row Φ_L with the displayed representation of $M_t^{(L)}$.

By orthogonality of projection,

$$\mathbb{E}\Big[\Big(X_{t+1} - \sum_{j=1}^{L} \Phi_{L,j} X_{t+1-j}\Big) X_{t+1-i}^*\Big] = 0, \qquad i = 1, \dots, L.$$

Using stationarity,

$$\Gamma(i) = \sum_{j=1}^{L} \Phi_{L,j} \Gamma(i-j), \qquad i = 1, \dots, L,$$

which is exactly the block Yule–Walker system

$$\Phi_L T_L(W) = G_L.$$

Lemma 9.5 implies that $T_L(W)$ is invertible, so

$$\Phi_L = G_L T_L(W)^{-1}.$$

Now set

$$e_t^{(L)} := X_{t+1} - M_t^{(L)}.$$

Expanding the covariance and using the orthogonality condition gives

$$\begin{aligned} V_L &= \mathbb{E}[e_t^{(L)} e_t^{(L)*}] \\ &= \Gamma(0) - \sum_{j=1}^{L} \Phi_{L,j} \Gamma(j)^* \\ &= \Gamma(0) - \Phi_L G_L^* \\ &= \Gamma(0) - G_L T_L(W)^{-1} G_L^*. \end{aligned}$$

□

The matrix analogue of the scalar context defect is the positive semidefinite covariance defect

$$\Delta_{\mathrm{ctx}}(L) := V_L - V_\infty.$$

The scalar defect from Section 2 is then the trace of this matrix defect.

Proposition 9.7 (Matrix defect covariance identity)**.** *Under Assumption 9.1,*

$$\Delta_{\mathrm{ctx}}(L) = V_L - V_\infty = \mathbb{E}[(M_t^\star - M_t^{(L)})(M_t^\star - M_t^{(L)})^*] \succeq 0.$$

Consequently,

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathrm{tr}(\Delta_{\mathrm{ctx}}(L)) = \mathrm{tr}(V_L - V_\infty).$$

Proof. Write

$$X_{t+1} - M_t^{(L)} = (X_{t+1} - M_t^\star) + (M_t^\star - M_t^{(L)}).$$

The first term is orthogonal to $\mathcal{F}_t$, hence also to $\mathcal{F}_t^{(L)}$; the second term is $\mathcal{F}_t$-measurable. Therefore the two terms are orthogonal in $L^2(\Omega; \mathbb{C}^q)$, which gives

$$V_L = V_\infty + \mathbb{E}[(M_t^\star - M_t^{(L)})(M_t^\star - M_t^{(L)})^*].$$

This proves the covariance identity and the Loewner positivity.

For the scalar defect based on Euclidean loss,

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathbb{E}\|M_t^\star - M_t^{(L)}\|^2 = \mathrm{tr}\ \mathbb{E}[(M_t^\star - M_t^{(L)})(M_t^\star - M_t^{(L)})^*],$$

which is exactly $\mathrm{tr}(\Delta_{\mathrm{ctx}}(L))$. □

Corollary 9.8 (Monotonicity in Loewner order)**.** *For every $L \geq 1$,*

$$V_L \succeq V_{L+1} \succeq V_\infty, \qquad \Delta_{\mathrm{ctx}}(L) \succeq \Delta_{\mathrm{ctx}}(L+1) \succeq 0, \qquad \mathcal{E}_{\mathrm{ctx}}(L) \downarrow 0.$$

Proof. The projection spaces increase with L, so finite-prediction errors decrease in the Loewner order. The remaining claims follow from Proposition 9.7 by taking traces. □

9.3. **Phase matrices.** Under Assumption 9.2, the multivariate theory has a second outer factor and a genuinely noncommutative phase function.

Definition 9.9 (Matrix outer factors and phase matrix)**.** Under Assumption 9.2, choose outer matrix-valued Hardy functions $h, h^\sharp \in H_2^{q\times q}$ such that

$$W(e^{i\theta}) = h(e^{i\theta}) h(e^{i\theta})^* = h^\sharp(e^{i\theta})^* h^\sharp(e^{i\theta}) \qquad \text{for a.e. } \theta,$$

and such that $h^{-1}, (h^\sharp)^{-1} \in H_2^{q\times q}$. The associated *phase matrix* is

$$Q(e^{i\theta}) := h(e^{i\theta})^*\, h^\sharp(e^{i\theta})^{-1} = h(e^{i\theta})^{-1} h^\sharp(e^{i\theta})^*.$$

Following one common convention, its Fourier coefficients are

$$B_k := -\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\theta} Q(e^{i\theta})\, d\theta, \qquad k \in \mathbb{Z}.$$

Lemma 9.10 (Unitarity of the phase matrix)**.** *Under Assumption 9.2, the phase matrix satisfies*

$$Q(e^{i\theta})Q(e^{i\theta})^* = Q(e^{i\theta})^*Q(e^{i\theta}) = I_q \qquad \text{for a.e. } \theta.$$

Proof. Using the second representation of Q from Definition 9.9,

$$QQ^* = h^{-1}h^{\sharp *}h^{\sharp}(h^{-1})^* = h^{-1}W(h^{-1})^* = h^{-1}hh^*(h^{-1})^* = I_q.$$

The identity $Q^*Q = I_q$ is proved similarly. □

Remark 9.11. The scalar phase function from Section 5 is recovered when $q = 1$. In the genuinely matrix-valued case, the phase matrix is the object whose Fourier coefficients drive explicit formulas for truncated block Toeplitz inverses, finite predictor matrices, and PACF matrices. This is one of the key places where the multivariate theory becomes qualitatively richer than the scalar theory.[10, 9]

9.4. **Vector autoregressions and hard context cutoffs.** The first exact multivariate context-length theorem is the vector analogue of Theorem 4.2.

Theorem 9.12 (Exact context-length law for VAR(p))**.** *Assume*

$$X_{t+1} = A_1X_t + \cdots + A_pX_{t-p+1} + \varepsilon_{t+1},$$

where $A_1, \dots, A_p \in \mathbb{C}^{q\times q}$, the matrix polynomial

$$I_q - A_1z - \cdots - A_pz^p$$

is invertible for $|z| \le 1$, and (ε_t) is an i.i.d. centered Gaussian innovation sequence with covariance matrix $\Sigma > 0$. Then:

(i) *the full-past predictor is*

$$M_t^\star = A_1X_t + \cdots + A_pX_{t-p+1};$$

(ii) *the full-past prediction error covariance is*

$$V_\infty = \Sigma;$$

(iii) *for every $L \ge p$,*

$$\Delta_{\mathrm{ctx}}(L) = 0;$$

(iv) *for $1 \le L < p$,*

$$\Delta_{\mathrm{ctx}}(L) = \mathrm{Cov}\left(\sum_{j=L+1}^{p} A_j\big(X_{t+1-j} - \mathbb{E}[X_{t+1-j} \mid \mathcal{F}_t^{(L)}]\big)\right).$$

Proof. Because ε_{t+1} is independent of $\mathcal{F}_t$ and has mean zero,

$$M_t^\star = \mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = A_1X_t + \cdots + A_pX_{t-p+1}.$$

This proves (i). Since

$$X_{t+1} - M_t^\star = \varepsilon_{t+1},$$

we also have

$$V_\infty = \mathbb{E}[\varepsilon_{t+1}\varepsilon_{t+1}^*] = \Sigma,$$

proving (ii).

If $L \ge p$, then $X_t, \dots, X_{t-p+1} \in \mathcal{F}_t^{(L)}$, so the same formula shows

$$M_t^{(L)} = A_1X_t + \cdots + A_pX_{t-p+1} = M_t^\star.$$

Hence $\Delta_{\mathrm{ctx}}(L) = 0$, proving (iii).

For $L < p$, decompose

$$M_t^\star = \sum_{j=1}^{L} A_j X_{t+1-j} + \sum_{j=L+1}^{p} A_j X_{t+1-j}.$$

The first sum is $\mathcal{F}_t^{(L)}$-measurable, so

$$M_t^\star - M_t^{(L)} = \sum_{j=L+1}^{p} A_j \big(X_{t+1-j} - \mathbb{E}[X_{t+1-j} \mid \mathcal{F}_t^{(L)}]\big).$$

Now apply Proposition 9.7. □

Corollary 9.13 (Matrix hard cutoff for finite memory)**.** *For a stable VAR(p) process, the matrix effective context length at tolerance* 0 *is exactly p in the sense that*

$$\inf\{L \geq 1 : \Delta_{\mathrm{ctx}}(L) = 0\} = p.$$

Equivalently,

$$\inf\{L \geq 1 : \mathcal{E}_{\mathrm{ctx}}(L) = 0\} = p.$$

Proof. The first identity follows from Theorem 9.12. The second follows by taking traces and using positive semidefiniteness of $\Delta_{\mathrm{ctx}}(L)$. □

9.5. **Past–future intersection.** The matrix-valued theory has a structural ingredient that does not appear explicitly in the scalar reduction.

Definition 9.14 (IPF property)**.** For an index set $I \subseteq \mathbb{Z}$, let

$$\mathcal{M}_I(X) := \overline{\mathrm{span}}\{X_j : j \in I\} \subset L^2(\Omega; \mathbb{C}^q).$$

We say that X has the *intersection-of-past-and-future* (IPF) property if, for every $n \geq 1$,

$$\mathcal{M}_{(-\infty,-1]}(X) \cap \mathcal{M}_{[-n,\infty)}(X) = \mathcal{M}_{[-n,-1]}(X).$$

Theorem 9.15 (Imported IPF baseline)**.** *A useful sufficient spectral condition for the IPF property is that the process have maximal rank and satisfy the inverse-integrability condition* $W^{-1} \in L^1(\mathbb{T})^{q\times q}$*. Under this condition, the process has the IPF property.*[8]

Remark 9.16. The IPF property is the structural bridge between multivariate prediction geometry and the representation theory of finite predictor coefficient matrices. It is one of the main reasons the matrix extension is not a routine vectorization of the scalar case. In particular, the explicit phase-matrix formulas for finite predictor coefficients and finite prediction error covariances are naturally organized around the joint geometry of past and future subspaces.[8, 10]

9.6. **Backward block prediction, matrix PACF, and defect tails.** To pass from block Yule–Walker formulas to quantitative context laws, one needs the multivariate analogue of the Durbin–Levinson step. The correct objects are the forward and backward finite prediction errors and the corresponding canonical PACF matrices.

For $L \geq 0$, define the backward finite-past predictor by

$$\widetilde{M}_t^{(L)} := \mathbb{E}[X_{t-L} \mid \sigma(X_{t-L+1}, \dots, X_t)], \qquad \widetilde{V}_L := \mathbb{E}\big[(X_{t-L} - \widetilde{M}_t^{(L)})(X_{t-L} - \widetilde{M}_t^{(L)})^*\big].$$

By convention,

$$M_t^{(0)} := 0, \qquad \widetilde{M}_t^{(0)} := 0, \qquad V_0 = \widetilde{V}_0 = \Gamma(0).$$

For $n \geq 1$, set

$$E_t^{f,n-1} := X_{t+1} - M_t^{(n-1)}, \qquad E_t^{b,n-1} := X_{t-n+1} - \widetilde{M}_t^{(n-1)},$$

and define the forward–backward residual covariance

$$C_n := \mathbb{E}\big[E_t^{f,n-1} E_t^{b,n-1*}\big].$$

Lemma 9.17 (Projection onto an orthogonal innovation block)**.** *Let $\mathcal{K} \subset L^2(\Omega;\mathbb{C}^q)$ be a closed linear subspace, and let $Z \in L^2(\Omega;\mathbb{C}^q)$ be orthogonal to $\mathcal{K}$ in the sense that each component of Z is orthogonal to $\mathcal{K}$. Assume that*

$$K_Z := \mathbb{E}[ZZ^*] \succ 0.$$

Then for every $Y \in L^2(\Omega;\mathbb{C}^q)$,

$$\mathrm{Proj}_{\mathcal{K}\vee Z} Y = \mathrm{Proj}_{\mathcal{K}} Y + \mathbb{E}\big[(Y - \mathrm{Proj}_{\mathcal{K}} Y)Z^*\big] K_Z^{-1} Z,$$

where $\mathcal{K} \vee Z$ denotes the closed linear span of $\mathcal{K}$ together with the components of Z.

Proof. Let

$$R := Y - \mathrm{Proj}_{\mathcal{K}} Y, \qquad B := \mathbb{E}[RZ^*] K_Z^{-1} Z.$$

Because R is orthogonal to $\mathcal{K}$ and Z is also orthogonal to $\mathcal{K}$, we have $B \in \mathcal{K}^\perp$. Hence

$$\mathrm{Proj}_{\mathcal{K}} Y + B \in \mathcal{K} \vee Z.$$

It remains to show that the residual

$$Y - (\mathrm{Proj}_{\mathcal{K}} Y + B) = R - B$$

is orthogonal to $\mathcal{K} \vee Z$. Orthogonality to $\mathcal{K}$ is immediate because both R and B lie in $\mathcal{K}^\perp$. For orthogonality to the components of Z, compute

$$\mathbb{E}[(R-B)Z^*] = \mathbb{E}[RZ^*] - \mathbb{E}[RZ^*] K_Z^{-1} \mathbb{E}[ZZ^*] = \mathbb{E}[RZ^*] - \mathbb{E}[RZ^*] = 0.$$

Thus $R - B$ is orthogonal to $\mathcal{K} \vee Z$, so the displayed formula is the orthogonal projection of Y onto $\mathcal{K} \vee Z$. □

Corollary 9.18 (Forward one-step projection update)**.** *For every $n \geq 1$,*

$$M_t^{(n)} = M_t^{(n-1)} + C_n \widetilde{V}_{n-1}^{-1} E_t^{b,n-1},$$

and hence

$$V_n = V_{n-1} - C_n \widetilde{V}_{n-1}^{-1} C_n^*.$$

Proof. Let

$$\mathcal{K}_{n-1} := \overline{\mathrm{span}}\{X_t, \dots, X_{t-n+2}\} \subset L^2(\Omega;\mathbb{C}^q).$$

Then $M_t^{(n-1)} = \mathrm{Proj}_{\mathcal{K}_{n-1}} X_{t+1}$ by definition. Also,

$$X_{t-n+1} = \widetilde{M}_t^{(n-1)} + E_t^{b,n-1},$$

where $\widetilde{M}_t^{(n-1)} \in \mathcal{K}_{n-1}$ and $E_t^{b,n-1} \perp \mathcal{K}_{n-1}$. Therefore

$$\overline{\mathrm{span}}\{X_t, \dots, X_{t-n+1}\} = \mathcal{K}_{n-1} \vee E_t^{b,n-1}.$$

Applying Lemma 9.17 with $Y = X_{t+1}$ and $Z = E_t^{b,n-1}$ gives

$$M_t^{(n)} = M_t^{(n-1)} + \mathbb{E}[E_t^{f,n-1} E_t^{b,n-1*}] \widetilde{V}_{n-1}^{-1} E_t^{b,n-1},$$

which is the first formula.

Subtracting this identity from X_{t+1} yields

$$E_t^{f,n} = E_t^{f,n-1} - C_n \widetilde{V}_{n-1}^{-1} E_t^{b,n-1}.$$

Taking covariance matrices and using

$$\mathbb{E}[E_t^{f,n-1} E_t^{b,n-1*}] = C_n, \qquad \mathbb{E}[E_t^{b,n-1} E_t^{b,n-1*}] = \widetilde{V}_{n-1},$$

we obtain

$$V_n = V_{n-1} - C_n \widetilde{V}_{n-1}^{-1} C_n^*. \qquad \square$$

Proposition 9.19 (Last block coefficient and canonical matrix PACF)**.** *For every $n \geq 1$,*

$$C_n = \Phi_{n,n} \widetilde{V}_{n-1}.$$

Consequently, the canonical matrix PACF is given by

$$\alpha_n := V_{n-1}^{-1/2} C_n \widetilde{V}_{n-1}^{-1/2} = V_{n-1}^{-1/2} \Phi_{n,n} \widetilde{V}_{n-1}^{1/2}.$$

Proof. As in the proof of Corollary 9.18, let

$$\mathcal{K}_{n-1} = \overline{\operatorname{span}}\{X_t, \dots, X_{t-n+2}\}.$$

Write the finite predictor of order n as

$$M_t^{(n)} = \sum_{j=1}^{n-1} \Phi_{n,j} X_{t+1-j} + \Phi_{n,n} X_{t-n+1}.$$

Since

$$X_{t-n+1} = \widetilde{M}_t^{(n-1)} + E_t^{b,n-1} \qquad \text{with} \qquad \widetilde{M}_t^{(n-1)} \in \mathcal{K}_{n-1},$$

we may rewrite $M_t^{(n)}$ as

$$M_t^{(n)} = Y_t + \Phi_{n,n} E_t^{b,n-1},$$

for some $Y_t \in \mathcal{K}_{n-1}$. Since $M_t^{(n-1)}$ is the orthogonal projection of X_{t+1} onto $\mathcal{K}_{n-1}$, the difference $M_t^{(n)} - M_t^{(n-1)}$ belongs to the orthogonal complement of $\mathcal{K}_{n-1}$ inside $\mathcal{K}_{n-1} \vee E_t^{b,n-1}$ and hence must equal

$$\Phi_{n,n} E_t^{b,n-1}.$$

Comparing this with Corollary 9.18, we obtain

$$\Phi_{n,n} E_t^{b,n-1} = C_n \widetilde{V}_{n-1}^{-1} E_t^{b,n-1}.$$

Taking covariance with $E_t^{b,n-1}$ gives

$$\Phi_{n,n} \widetilde{V}_{n-1} = C_n.$$

The formula for α_n is then immediate. $\square$

Corollary 9.20 (Matrix PACF recursion and contractivity)**.** *For every $n \geq 1$,*

$$V_n = V_{n-1}^{1/2} (I_q - \alpha_n \alpha_n^*) V_{n-1}^{1/2}.$$

By applying the same argument to the time-reversed process,

$$\widetilde{V}_n = \widetilde{V}_{n-1}^{1/2} (I_q - \alpha_n^* \alpha_n) \widetilde{V}_{n-1}^{1/2}.$$

In particular,

$$\alpha_n \alpha_n^* \preceq I_q, \qquad \alpha_n^* \alpha_n \preceq I_q, \qquad \|\alpha_n\| \leq 1.$$

Proof. The first identity is Corollary 9.18 rewritten using Proposition 9.19:

$$V_n = V_{n-1} - V_{n-1}^{1/2} \alpha_n \alpha_n^* V_{n-1}^{1/2} = V_{n-1}^{1/2} (I_q - \alpha_n \alpha_n^*) V_{n-1}^{1/2}.$$

The backward identity is the same statement applied to the time-reversed process. Since $V_n \succeq 0$ and $\widetilde{V}_n \succeq 0$, both $I_q - \alpha_n \alpha_n^*$ and $I_q - \alpha_n^* \alpha_n$ are positive semidefinite, which implies the contractivity bounds. $\square$

Theorem 9.21 (Exact matrix PACF tail representation of the defect covariance)**.** *For every* $L \geq 0$,

$$\Delta_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} (V_{n-1} - V_n) = \sum_{n=L+1}^{\infty} V_{n-1}^{1/2} \alpha_n \alpha_n^* V_{n-1}^{1/2},$$

with convergence in the Loewner order. Consequently,

$$\lambda_{\min}(V_\infty) \sum_{n=L+1}^{\infty} \|\alpha_n\|_F^2 \leq \mathcal{E}_{\mathrm{ctx}}(L) \leq \lambda_{\max}(\Gamma(0)) \sum_{n=L+1}^{\infty} \|\alpha_n\|_F^2.$$

Proof. By Proposition 9.7,

$$\Delta_{\mathrm{ctx}}(L) = V_L - V_\infty.$$

By Corollary 9.8, the sequence V_n decreases to V_∞ in the Loewner order, so the telescoping identity

$$V_L - V_\infty = \sum_{n=L+1}^{\infty} (V_{n-1} - V_n)$$

holds in the Loewner order. Corollary 9.20 gives

$$V_{n-1} - V_n = V_{n-1}^{1/2} \alpha_n \alpha_n^* V_{n-1}^{1/2},$$

which proves the first formula.

Taking traces and using cyclicity,

$$\mathrm{tr}(V_{n-1}^{1/2} \alpha_n \alpha_n^* V_{n-1}^{1/2}) = \mathrm{tr}(\alpha_n^* V_{n-1} \alpha_n).$$

Since

$$V_\infty \preceq V_{n-1} \preceq \Gamma(0),$$

we obtain

$$\lambda_{\min}(V_\infty)\, \mathrm{tr}(\alpha_n^* \alpha_n) \leq \mathrm{tr}(\alpha_n^* V_{n-1} \alpha_n) \leq \lambda_{\max}(\Gamma(0))\, \mathrm{tr}(\alpha_n^* \alpha_n).$$

Summing over $n \geq L+1$ yields the stated bounds because $\mathrm{tr}(\alpha_n^* \alpha_n) = \|\alpha_n\|_F^2$ and $\mathcal{E}_{\mathrm{ctx}}(L) = \mathrm{tr}(\Delta_{\mathrm{ctx}}(L))$. □

Corollary 9.22 (Norm-based multivariate context laws)**.** *The following implications hold.*

(i) *If there exist constants* $C > 0$ *and* $\rho \in (0,1)$ *such that*

$$\|\alpha_n\|_F \leq C\rho^n \qquad (n \geq 1),$$

then

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq \frac{\lambda_{\max}(\Gamma(0)) C^2}{1-\rho^2}\, \rho^{2L+2}.$$

(ii) *If there exist constants* $C > 0$ *and* $s > \frac{1}{2}$ *such that*

$$\|\alpha_n\|_F \leq C n^{-s} \qquad (n \geq 1),$$

then there exists $C_s > 0$ *such that*

$$\mathcal{E}_{\mathrm{ctx}}(L) \leq C_s L^{1-2s} \qquad (L \geq 1).$$

Proof. Both parts are immediate from the upper bound in Theorem 9.21 together with the geometric-series formula in case (i) and the integral test in case (ii). □

9.7. **Imported phase-matrix representation package.** The precise link from the Fourier coefficients of the phase matrix to the finite predictor coefficient matrices, the finite prediction error covariance matrices, and the canonical matrix PACF is already available in the multivariate prediction literature. We record the part of that theory that interfaces directly with the present draft.

Theorem 9.23 (Imported phase-matrix representation package)**.** *Assume the hypotheses* (A) *and* (IPF) *of Inoue–Kasahara–Pourahmadi. Let B_k denote the Fourier coefficients of the matrix phase function from Definition 9.9. Then there exist explicitly defined alternating-projection matrices*

$$\{b_{n,j}^{2k}\}_{k,j\ge 0}, \qquad \{\widetilde{b}_{n,j}^{2k}\}_{k,j\ge 0},$$

built recursively from the phase coefficients, such that the finite predictor coefficient matrices, the finite prediction error covariance matrices, and the canonical PACF matrices admit explicit series representations in terms of these quantities. In particular:

(i) *the last forward coefficient matrix $\Phi_{n,n}$ admits an explicit phase-series representation built from B_{n+j};*
(ii) *the forward and backward finite prediction error covariance matrices V_n and $\widetilde{V}_n$ admit explicit series representations in terms of even alternating-projection coefficients;*
(iii) *the forward–backward residual covariance and hence the canonical PACF matrix α_n admit explicit phase-series representations in terms of odd alternating-projection coefficients;*
(iv) *under the stronger minimality assumption* (M)*, every finite predictor coefficient matrix $\Phi_{n,j}$ and $\widetilde{\Phi}_{n,j}$ admits an explicit phase-series representation.*

These are the matrix-valued analogues of the scalar phase-series formulas from Section 6.[9]

Remark 9.24. Theorem 9.23 is the exact point at which matrix phase coefficients feed into multivariate context laws. Once the imported phase formulas yield bounds on α_n, Theorem 9.21 turns them immediately into bounds on the defect covariance $\Delta_{\mathrm{ctx}}(L)$ and on the scalar Euclidean-loss defect $\mathcal{E}_{\mathrm{ctx}}(L)$.

9.8. **Imported multivariate FARIMA law.** The clearest multivariate long-memory baseline presently available is the common-order FARIMA case.

Theorem 9.25 (Imported multivariate FARIMA asymptotics)**.** *Let (X_t) be a q-variate FARIMA process with common fractional differencing order $d \in (-\frac{1}{2}, \frac{1}{2}) \setminus \{0\}$ and spectral density of the form described in Inoue–Kasahara–Pourahmadi. Then, as $n \to \infty$,*

$$V_n = V_\infty + \frac{d^2}{n} V_\infty + O(n^{-2}), \qquad \widetilde{V}_n = \widetilde{V}_\infty + \frac{d^2}{n} \widetilde{V}_\infty + O(n^{-2}),$$

and

$$\alpha_n = \frac{d}{n} U + O(n^{-2}),$$

where $U \in \mathbb{C}^{q\times q}$ is unitary and the $O(\cdot)$ terms are taken in any fixed matrix norm.[9]

Corollary 9.26 (Multivariate FARIMA context law)**.** *Under the assumptions of Theorem 9.25,*

$$\Delta_{\mathrm{ctx}}(L) = V_L - V_\infty = \frac{d^2}{L} V_\infty + O(L^{-2}),$$

and therefore

$$\mathcal{E}_{\mathrm{ctx}}(L) = \frac{d^2}{L} \operatorname{tr}(V_\infty) + O(L^{-2}) \qquad (L \to \infty).$$

In particular, multivariate common-order FARIMA processes satisfy the same L^{-1} context law as in the scalar long-memory case, with the innovation covariance replacing the scalar innovation variance.

Proof. By Proposition 9.7,

$$\Delta_{\mathrm{ctx}}(L) = V_L - V_\infty.$$

Substituting the first asymptotic formula from Theorem 9.25 gives

$$\Delta_{\mathrm{ctx}}(L) = \frac{d^2}{L} V_\infty + O(L^{-2}).$$

Taking traces yields

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathrm{tr}(\Delta_{\mathrm{ctx}}(L)) = \frac{d^2}{L}\,\mathrm{tr}(V_\infty) + O(L^{-2}). \qquad \square$$

9.9. **Summary of the multivariate extension.**

Proposition 9.27 (Summary formulas in the matrix case)**.** *Under Assumption 9.1, the multivariate finite-past predictor is determined by the block Yule–Walker equations*

$$\Phi_L T_L(W) = G_L, \qquad V_L = \Gamma(0) - G_L T_L(W)^{-1} G_L^*.$$

Moreover, the matrix context-defect covariance is

$$\Delta_{\mathrm{ctx}}(L) = V_L - V_\infty = \mathbb{E}[(M_t^\star - M_t^{(L)})(M_t^\star - M_t^{(L)})^*] \succeq 0,$$

and the scalar defect from Euclidean loss is

$$\mathcal{E}_{\mathrm{ctx}}(L) = \mathrm{tr}(\Delta_{\mathrm{ctx}}(L)).$$

If α_n denotes the canonical matrix PACF defined in Proposition 9.19, then

$$V_n = V_{n-1}^{1/2}(I_q - \alpha_n \alpha_n^*) V_{n-1}^{1/2}, \qquad \Delta_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} V_{n-1}^{1/2} \alpha_n \alpha_n^* V_{n-1}^{1/2},$$

and therefore

$$\lambda_{\min}(V_\infty) \sum_{n=L+1}^{\infty} \|\alpha_n\|_F^2 \le \mathcal{E}_{\mathrm{ctx}}(L) \le \lambda_{\max}(\Gamma(0)) \sum_{n=L+1}^{\infty} \|\alpha_n\|_F^2.$$

Under Assumption 9.2, the associated phase matrix is

$$Q = h^*(h^\sharp)^{-1} = h^{-1} h^{\sharp *},$$

which is unitary on $\mathbb{T}$; under the additional IPF/minimality hypotheses from the imported multivariate theory, its Fourier coefficients control the finite predictor matrices, the error covariances, and the PACF matrices through explicit phase-series formulas.

Proof. The block Yule–Walker formulas are Proposition 9.6. The defect-covariance identity is Proposition 9.7. The PACF recursion and tail representation are Corollary 9.20 and Theorem 9.21. The phase-matrix statement is Definition 9.9 together with Lemma 9.10 and Theorem 9.23. $\square$

10. Multivariate Linear State Compression and Hankel-Type State-Dimension Laws

The multivariate context-defect theory of Section 9 explains how finite lag windows lose predictive information. To turn that theory into a genuine hidden-state theory, we now repeat the scalar reduction of Section 7 in the matrix-valued setting.

The outcome is an exact matrix-weighted rational approximation formula for the best stable multivariable linear state of a given dimension. This is the correct bridge from block prediction to model reduction, balanced realizations, and Hankel singular value compressibility.

10.1. **Whitened innovations and the matrix prediction symbol.** Fix the normalized Wold–Szegö factorization of Theorem 9.3:

$$W(e^{i\theta}) = H(e^{i\theta})\,\Sigma\,H(e^{i\theta})^*, \qquad H(z) = \sum_{k=0}^{\infty} C_k z^k, \qquad C_0 = I_q.$$

Define the whitened innovations

$$e_t := \Sigma^{-1/2}\varepsilon_t.$$

Then

$$\mathbb{E}[e_t] = 0, \qquad \mathbb{E}[e_t e_s^*] = \delta_{ts} I_q,$$

and the process admits the representation

$$X_t = \sum_{k=0}^{\infty} B_k e_{t-k}, \qquad B_k := C_k \Sigma^{1/2}.$$

Let

$$\widetilde{H}(z) := H(z)\Sigma^{1/2} = \sum_{k=0}^{\infty} B_k z^k.$$

Then

$$W(e^{i\theta}) = \widetilde{H}(e^{i\theta})\widetilde{H}(e^{i\theta})^* \qquad \text{for a.e. } \theta.$$

Definition 10.1 (Matrix Hardy space)**.** Let $H_2^{q\times q}$ denote the space of analytic matrix-valued functions

$$G(z) = \sum_{k=0}^{\infty} G_k z^k, \qquad G_k \in \mathbb{C}^{q\times q},$$

such that

$$\|G\|_{H_2^{q\times q}}^2 := \sum_{k=0}^{\infty} \|G_k\|_F^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\big(G(e^{i\theta})G(e^{i\theta})^*\big)\,d\theta < \infty.$$

Definition 10.2 (Matrix prediction symbol)**.** Define the *matrix prediction symbol*

$$F_{\mathrm{mat}}(z) := \sum_{j=0}^{\infty} B_{j+1} z^j = \frac{\widetilde{H}(z) - \Sigma^{1/2}}{z} = \frac{H(z) - I_q}{z}\,\Sigma^{1/2}.$$

Lemma 10.3 (Matrix innovation-space isometry)**.** *Let*

$$\mathscr{H}_t^- := \overline{\left\{ \sum_{j=0}^{N} M_j e_{t-j} : N \ge 0,\ M_j \in \mathbb{C}^{q\times q} \right\}} \subset L^2(\Omega;\mathbb{C}^q).$$

Define

$$U_t\left(\sum_{j=0}^{N} M_j e_{t-j}\right) := \sum_{j=0}^{N} M_j z^j.$$

Then U_t extends uniquely to a unitary map from $\mathscr{H}_t^-$ onto $H_2^{q\times q}$.

Proof. Because the whitened innovations are orthonormal in time and satisfy $\mathbb{E}[e_t e_t^*] = I_q$,

$$\begin{aligned}\mathbb{E}\left\|\sum_{j=0}^{N} M_j e_{t-j}\right\|^2 &= \sum_{i,j=0}^{N} \mathbb{E}\left[e_{t-j}^* M_j^* M_i e_{t-i}\right] \\ &= \sum_{j=0}^{N} \operatorname{tr}(M_j^* M_j) \\ &= \sum_{j=0}^{N} \|M_j\|_F^2 \\ &= \left\|\sum_{j=0}^{N} M_j z^j\right\|_{H_2^{q\times q}}^2.\end{aligned}$$

Thus U_t is an isometry on a dense subspace of $\mathscr{H}_t^-$. Surjectivity onto $H_2^{q\times q}$ is immediate from the definition, so the extension is unitary. □

Proposition 10.4 (The full-past predictor as a matrix Hardy-space vector)**.** *Under the unitary map U_t,*

$$U_t M_t^\star = F_{\mathrm{mat}}.$$

Equivalently,

$$M_t^\star = \sum_{j=0}^{\infty} B_{j+1} e_{t-j}.$$

Proof. Theorem 9.3 gives

$$X_{t+1} = B_0 e_{t+1} + \sum_{j=0}^{\infty} B_{j+1} e_{t-j}, \qquad B_0 = \Sigma^{1/2}.$$

The second term is measurable with respect to the full past, while $B_0 e_{t+1}$ is orthogonal to it. Hence

$$M_t^\star = \sum_{j=0}^{\infty} B_{j+1} e_{t-j}.$$

Applying Lemma 10.3 yields

$$U_t M_t^\star = \sum_{j=0}^{\infty} B_{j+1} z^j = F_{\mathrm{mat}}.$$

□

10.2. **General causal matrix linear predictors.**

Definition 10.5 (Causal matrix predictor kernel)**.** Let $(R_j)_{j\geq 1}$ be a sequence of matrices in $\mathbb{C}^{q\times q}$ such that

$$\sum_{j=1}^{\infty} \|R_j\| < \infty$$

for some fixed matrix norm. Its *predictor symbol* is

$$R(z) := \sum_{j=1}^{\infty} R_j z^{j-1}, \qquad z \in \mathbb{D}.$$

The associated causal matrix linear predictor is

$$\widehat{X}_{t+1}^{(R)} := \sum_{j=1}^{\infty} R_j X_{t+1-j}.$$

Absolute summability implies convergence in $L^2(\Omega;\mathbb{C}^q)$ because $\sup_t \mathbb{E}\|X_t\|^2 < \infty$.

Proposition 10.6 (Innovation-space image of a matrix linear predictor)**.** *Let R be a causal matrix predictor symbol with absolutely summable coefficients. Then*

$$U_t \widehat{X}_{t+1}^{(R)} = R\,\widetilde{H} \quad \text{in } H_2^{q\times q}.$$

Proof. Using the Wold representation,

$$X_{t+1-j} = \sum_{k=0}^{\infty} B_k e_{t+1-j-k}.$$

Therefore,

$$\widehat{X}_{t+1}^{(R)} = \sum_{j=1}^{\infty} R_j \sum_{k=0}^{\infty} B_k e_{t+1-j-k}.$$

Since (R_j) is absolutely summable and (B_k) is square summable, the series converges in L^2 and may be grouped by powers of z:

$$\widehat{X}_{t+1}^{(R)} = \sum_{m=0}^{\infty} \left(\sum_{j=1}^{m+1} R_j B_{m+1-j} \right) e_{t-m}.$$

The coefficient of e_{t-m} is exactly the coefficient of z^m in the product

$$R(z)\widetilde{H}(z) = \left(\sum_{j=1}^{\infty} R_j z^{j-1} \right) \left(\sum_{k=0}^{\infty} B_k z^k \right).$$

Applying U_t proves the claim. □

Theorem 10.7 (Exact matrix Hardy-space formula for linear prediction defect)**.** *Let R be a causal matrix predictor symbol with absolutely summable coefficients. Define its excess risk over the full-past benchmark by*

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) := \mathbb{E}\|X_{t+1} - \widehat{X}_{t+1}^{(R)}\|^2 - \mathrm{tr}(\Sigma).$$

Then

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \mathbb{E}\|M_t^\star - \widehat{X}_{t+1}^{(R)}\|^2 = \|F_{\mathrm{mat}} - R\widetilde{H}\|_{H_2^{q\times q}}^2.$$

Equivalently,

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\Big(\big(F_{\mathrm{mat}}(e^{i\theta}) - R(e^{i\theta})\widetilde{H}(e^{i\theta})\big)\big(F_{\mathrm{mat}}(e^{i\theta}) - R(e^{i\theta})\widetilde{H}(e^{i\theta})\big)^{*}\Big)\, d\theta.$$

Proof. Since $\widehat{X}_{t+1}^{(R)}$ is measurable with respect to the full past, the orthogonal decomposition

$$X_{t+1} - \widehat{X}_{t+1}^{(R)} = (X_{t+1} - M_t^{\star}) + (M_t^{\star} - \widehat{X}_{t+1}^{(R)})$$

implies

$$\mathbb{E}\|X_{t+1} - \widehat{X}_{t+1}^{(R)}\|^2 = \mathbb{E}\|X_{t+1} - M_t^{\star}\|^2 + \mathbb{E}\|M_t^{\star} - \widehat{X}_{t+1}^{(R)}\|^2.$$

By Theorem 9.3,

$$\mathbb{E}\|X_{t+1} - M_t^{\star}\|^2 = \mathrm{tr}(\Sigma),$$

so

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \mathbb{E}\|M_t^{\star} - \widehat{X}_{t+1}^{(R)}\|^2.$$

Now apply the unitary map U_t. By Proposition 10.4 and Proposition 10.6,

$$U_t(M_t^{\star} - \widehat{X}_{t+1}^{(R)}) = F_{\mathrm{mat}} - R\widetilde{H}.$$

Lemma 10.3 gives

$$\mathbb{E}\|M_t^{\star} - \widehat{X}_{t+1}^{(R)}\|^2 = \|F_{\mathrm{mat}} - R\widetilde{H}\|_{H_2^{q\times q}}^2,$$

and the integral formula is the definition of the $H_2^{q\times q}$ norm. □

Remark 10.8. Theorem 10.7 is the multivariate analogue of Theorem 7.6. In innovation coordinates, the entire problem of causal matrix linear prediction is an analytic approximation problem in $H_2^{q\times q}$.

Lemma 10.9 (Imported inverse factorization baseline)**.** *Under Assumption 9.2, the normalized outer factor H in Theorem 9.3 satisfies*

$$H^{-1} \in H_2^{q\times q}.$$

Remark 10.10. Lemma 10.9 is classical matrix Hardy-space theory under inverse-integrability; it is imported here only to justify the autoregressive-side symbol below.

Proposition 10.11 (Matrix weighted approximation formula under inverse factorization)**.** *Assume Assumption 9.2, and let*

$$A(z) := H(z)^{-1} = \sum_{k=0}^{\infty} A_k z^k, \qquad A_0 = I_q,$$

and

$$P_{\mathrm{mat}}(z) := \frac{I_q - A(z)}{z} = -\sum_{j=1}^{\infty} A_j z^{j-1}.$$

Then

$$F_{\mathrm{mat}} = P_{\mathrm{mat}}\,\widetilde{H}.$$

Consequently, for every causal matrix predictor symbol R,

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \|(R - P_{\mathrm{mat}})\widetilde{H}\|_{H_2^{q\times q}}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\Big((R - P_{\mathrm{mat}})W(R - P_{\mathrm{mat}})^{*}\Big)\, d\theta.$$

Proof. Since $A(z)H(z) = I_q$,

$$P_{\mathrm{mat}}(z)\widetilde{H}(z) = \frac{I_q - A(z)}{z} H(z)\Sigma^{1/2} = \frac{H(z) - I_q}{z}\Sigma^{1/2} = F_{\mathrm{mat}}(z).$$

Substituting this into Theorem 10.7 yields

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \|(R - P_{\mathrm{mat}})\widetilde{H}\|^2_{H_2^{q\times q}}.$$

Since

$$W = \widetilde{H}\,\widetilde{H}^*,$$

the integral formula follows from the matrix Hardy-space norm identity. □

10.3. **Stable multivariable finite-state predictors.**

Definition 10.12 (Stable r-state MIMO predictor)**.** A *stable r-state MIMO predictor* is a triple (A, B, C) with

$$A \in \mathbb{C}^{r\times r}, \qquad B \in \mathbb{C}^{r\times q}, \qquad C \in \mathbb{C}^{q\times r}, \qquad \rho(A) < 1.$$

Its stationary causal state is

$$s_t := \sum_{j=0}^{\infty} A^j B X_{t-j},$$

and its one-step predictor is

$$\widehat{X}_{t+1}^{A,B,C} := Cs_t.$$

Equivalently,

$$s_t = As_{t-1} + BX_t, \qquad \widehat{X}_{t+1}^{A,B,C} = Cs_t.$$

Proposition 10.13 (Transfer symbol of a stable MIMO predictor)**.** *Let (A, B, C) be a stable r-state MIMO predictor. Then*

$$\widehat{X}_{t+1}^{A,B,C} = \sum_{j=1}^{\infty} R_j X_{t+1-j}, \qquad R_j := CA^{j-1}B,$$

with absolutely summable matrix coefficients. The corresponding transfer symbol is the strictly proper stable rational matrix

$$R_{A,B,C}(z) := \sum_{j=1}^{\infty} R_j z^{j-1} = C(I_r - zA)^{-1}B.$$

Proof. The proof is the same as in Proposition 7.10. From the definition of s_t,

$$\widehat{X}_{t+1}^{A,B,C} = \sum_{j=0}^{\infty} CA^j B\, X_{t-j} = \sum_{m=1}^{\infty} CA^{m-1}B\, X_{t+1-m}.$$

Thus $R_m = CA^{m-1}B$. Since $\rho(A) < 1$, the coefficients decay geometrically and are absolutely summable. Summing the geometric series gives

$$R_{A,B,C}(z) = C(I_r - zA)^{-1}B.$$

□

Definition 10.14 (Best r-state MIMO defect)**.** Let $\mathcal{R}_r^{\mathrm{MIMO}}$ denote the set of all transfer symbols of the form $R_{A,B,C}$ arising from stable r-state MIMO predictors. Define

$$\mathcal{E}_r^{\mathrm{MIMO}} := \inf_{R\in\mathcal{R}_r^{\mathrm{MIMO}}} \mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R).$$

Corollary 10.15 (Best multivariable finite-state defect as matrix-weighted rational approximation)**.** *Under Assumption 9.1,*

$$\mathcal{E}_r^{\mathrm{MIMO}} = \inf_{R\in\mathcal{R}_r^{\mathrm{MIMO}}} \|F_{\mathrm{mat}} - R\widetilde{H}\|^2_{H_2^{q\times q}}.$$

Under Assumption 9.2,

$$\mathcal{E}_r^{\mathrm{MIMO}} = \inf_{R\in\mathcal{R}_r^{\mathrm{MIMO}}} \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\Big((R-P_{\mathrm{mat}})W(R-P_{\mathrm{mat}})^*\Big)\,d\theta.$$

Proof. Combine Proposition 10.13, Theorem 10.7, and Proposition 10.11. □

Remark 10.16. Corollary 10.15 is the exact multivariable state-compression reduction. The best stable r-state hidden representation is exactly a matrix-weighted rational approximation problem for the canonical predictor symbol.

10.4. **Block shift-register realizations and immediate multivariate state laws.**

Proposition 10.17 (Block shift-register realization of the finite-context predictor)**.** *For every $L \geq 1$, there exists a stable Lq-state MIMO predictor whose one-step output equals the finite-context predictor $M_t^{(L)}$. Consequently,*

$$\mathcal{E}_{Lq}^{\mathrm{MIMO}} \leq \mathcal{E}_{\mathrm{ctx}}(L).$$

More generally, for every $r \geq q$,

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq \mathcal{E}_{\mathrm{ctx}}\big(\lfloor r/q\rfloor\big).$$

Proof. Define the block shift-register state

$$s_t := (X_t^\top, \dots, X_{t-L+1}^\top)^\top \in \mathbb{C}^{Lq}.$$

Then

$$s_t = A_{\mathrm{blk}} s_{t-1} + B_{\mathrm{blk}} X_t,$$

where A_{blk} is the nilpotent block shift matrix and B_{blk} injects X_t into the first block. By Proposition 9.6,

$$M_t^{(L)} = \sum_{j=1}^{L} \Phi_{L,j} X_{t+1-j} = C_L s_t,$$

with

$$C_L = [\Phi_{L,1} \ \cdots \ \Phi_{L,L}].$$

Thus $M_t^{(L)}$ is realized by a stable Lq-state MIMO predictor, and its excess risk is exactly $\mathcal{E}_{\mathrm{ctx}}(L)$ by Proposition 9.7. Hence

$$\mathcal{E}_{Lq}^{\mathrm{MIMO}} \leq \mathcal{E}_{\mathrm{ctx}}(L).$$

For a general $r \geq q$, let $L = \lfloor r/q\rfloor$. Pad the above Lq-state realization by $r - Lq$ unused stable coordinates. This yields an r-state realization with the same output, so

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq \mathcal{E}_{\mathrm{ctx}}(L).$$

□

Corollary 10.18 (Multivariate state-dimension laws inherited from context laws)**.** *Every upper bound proved earlier for $\mathcal{E}_{\mathrm{ctx}}(L)$ yields an immediate upper bound for $\mathcal{E}_r^{\mathrm{MIMO}}$ after substituting $L = \lfloor r/q\rfloor$. In particular:*

(i) *if there exist constants* $C > 0$ *and* $\rho \in (0,1)$ *such that*

$$\|\alpha_n\|_F \le C\rho^n \qquad (n \ge 1),$$

then

$$\mathcal{E}_r^{\mathrm{MIMO}} \le C_1 \rho^{2\lfloor r/q \rfloor};$$

(ii) *under the multivariate FARIMA assumptions of Corollary 9.26,*

$$\mathcal{E}_r^{\mathrm{MIMO}} \le \frac{d^2}{\lfloor r/q \rfloor} \operatorname{tr}(V_\infty) + O(r^{-2}) \qquad (r \to \infty),$$

where $d \in (-\frac{1}{2}, \frac{1}{2}) \setminus \{0\}$ *is the common fractional differencing order.*

Proof. Combine Proposition 10.17 with Theorem 9.21 and Corollary 9.26. □

10.5. **Exact realizations from finitely many stable matrix poles.**

Theorem 10.19 (Exact realization from finitely many stable matrix poles)**.** *Assume Assumption 9.2. Suppose that the canonical predictor symbol has the form*

$$P_{\mathrm{mat}}(z) = \sum_{\ell=1}^{m} \frac{M_\ell}{1 - \lambda_\ell z}, \qquad M_\ell \in \mathbb{C}^{q\times q}, \qquad |\lambda_\ell| < 1.$$

Then there exists a stable mq*-state MIMO predictor whose transfer symbol is exactly* P_{mat}*. Consequently,*

$$\mathcal{E}_{mq}^{\mathrm{MIMO}} = 0.$$

Proof. Set

$$A := \operatorname{diag}(\lambda_1 I_q, \dots, \lambda_m I_q) \in \mathbb{C}^{mq\times mq}, \qquad B := \begin{pmatrix} I_q \\ \vdots \\ I_q \end{pmatrix} \in \mathbb{C}^{mq\times q},$$

and

$$C := [M_1 \ \cdots \ M_m] \in \mathbb{C}^{q\times mq}.$$

Then $\rho(A) = \max_\ell |\lambda_\ell| < 1$, so (A, B, C) is stable. Moreover,

$$C(I - zA)^{-1}B = \sum_{\ell=1}^{m} M_\ell (I_q - \lambda_\ell z I_q)^{-1} = \sum_{\ell=1}^{m} \frac{M_\ell}{1 - \lambda_\ell z} = P_{\mathrm{mat}}(z).$$

Thus $P_{\mathrm{mat}} \in \mathcal{R}_{mq}^{\mathrm{MIMO}}$. By Corollary 10.15,

$$\mathcal{E}_{mq}^{\mathrm{MIMO}} \le \mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(P_{\mathrm{mat}}) = 0.$$

Since defects are nonnegative, equality follows. □

Remark 10.20. Theorem 10.19 is the multivariable analogue of the scalar finite-pole realization theorem from Section 8. It gives a hard state-dimension cutoff whenever the predictor symbol has finite McMillan structure.

10.6. **Balanced truncation, Hankel singular values, and compressibility laws.** The exact weighted-approximation formula above opens the door to imported model-reduction theorems.

Theorem 10.21 (Imported balanced-truncation / Hankel baseline)**.** *Let P_{mat} be a strictly proper, stable rational $q \times q$ transfer matrix of McMillan degree N, and let*

$$s_1 \geq s_2 \geq \cdots \geq s_N > 0$$

be its Hankel singular values. Then for each $0 \leq r < N$, classical model-reduction theory provides a stable rational $q \times q$ transfer matrix R_r of degree at most r such that

$$\|P_{\mathrm{mat}} - R_r\|_{H_\infty} \leq 2 \sum_{k=r+1}^{N} s_k.$$

Moreover, optimal Hankel-norm approximation achieves error s_{r+1}.[14, 15]

Remark 10.22. Theorem 10.21 is imported from classical control/model-reduction theory. It is the exact point at which balanced realizations and Hankel singular values enter prediction geometry.

Corollary 10.23 (Hankel-singular-value state-dimension law)**.** *Assume Assumption 9.2, assume in addition that*

$$\|W\|_{L^\infty_{\mathrm{op}}} := \operatorname*{ess\,sup}_{\theta} \|W(e^{i\theta})\|_{\mathrm{op}} < \infty,$$

and suppose that P_{mat} is strictly proper, stable, and rational with Hankel singular values $(s_k)_{k=1}^N$. Then for every $0 \leq r < N$,

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq 4q \, \|W\|_{L^\infty_{\mathrm{op}}} \left(\sum_{k=r+1}^{N} s_k \right)^2.$$

Proof. Let R_r be the degree-r approximant from Theorem 10.21. By Corollary 10.15,

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{tr}\Big((R_r - P_{\mathrm{mat}}) W (R_r - P_{\mathrm{mat}})^* \Big) \, d\theta.$$

Using the operator norm bound on W,

$$\mathrm{tr}\Big((R_r - P_{\mathrm{mat}}) W (R_r - P_{\mathrm{mat}})^* \Big) \leq \|W\|_{L^\infty_{\mathrm{op}}} \, \|R_r - P_{\mathrm{mat}}\|_F^2.$$

Therefore

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq \|W\|_{L^\infty_{\mathrm{op}}} \, \|R_r - P_{\mathrm{mat}}\|_{H_2^{q\times q}}^2.$$

Since normalized arc measure on $\mathbb{T}$ has total mass 1,

$$\|R_r - P_{\mathrm{mat}}\|_{H_2^{q\times q}}^2 \leq \|R_r - P_{\mathrm{mat}}\|_{L^\infty_F}^2 \leq q \, \|R_r - P_{\mathrm{mat}}\|_{H_\infty}^2.$$

Applying Theorem 10.21 gives

$$\mathcal{E}_r^{\mathrm{MIMO}} \leq q \, \|W\|_{L^\infty_{\mathrm{op}}} \left(2 \sum_{k=r+1}^{N} s_k \right)^2,$$

which is the stated bound. □

Corollary 10.24 (Compressibility laws from Hankel singular value tails)**.** *Under the assumptions of Corollary 10.23:*

(i) *if* $s_k \le C\rho^k$ *for some* $C > 0$ *and* $\rho \in (0,1)$*, then*

$$\mathcal{E}_r^{\mathrm{MIMO}} \le C_1 \rho^{2r};$$

(ii) *if* $s_k \le Ck^{-p}$ *for some* $C > 0$ *and* $p > 1$*, then*

$$\mathcal{E}_r^{\mathrm{MIMO}} \le C_2 r^{2-2p}.$$

Proof. If $s_k \le C\rho^k$, then

$$\sum_{k=r+1}^{\infty} s_k \le \frac{C}{1-\rho}\rho^{r+1},$$

and Corollary 10.23 yields an $O(\rho^{2r})$ bound. If $s_k \le Ck^{-p}$ with $p > 1$, then

$$\sum_{k=r+1}^{\infty} s_k \le C \sum_{k=r+1}^{\infty} k^{-p} \le \frac{C}{p-1} r^{1-p},$$

and squaring gives an $O(r^{2-2p})$ law. □

10.7. **Interpretation.** Section 10 adds a sixth exact backbone to the draft:

$$\boxed{\mathcal{E}_r^{\mathrm{MIMO}} = \inf_{R \in \mathcal{R}_r^{\mathrm{MIMO}}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{tr}\Big((R - P_{\mathrm{mat}}) W (R - P_{\mathrm{mat}})^*\Big)\, d\theta}.$$

This is the matrix-valued hidden-state analogue of the scalar weighted rational approximation formula. It shows that multivariate hidden-state compression is governed simultaneously by:

(i) the matrix spectral geometry encoded by W and the phase/IPF machinery;
(ii) the matrix predictor symbol P_{mat};
(iii) stable multivariable rational approximation;
(iv) classical compressibility quantities such as Hankel singular values.

In particular, the framework now contains both a *lag-window route* to state-dimension laws (through block shift registers and context defects) and a *system-theoretic route* (through balanced truncation and Hankel-singular-value tails). The two routes need not coincide, and their comparison is itself a promising new problem.

11. Finite Hankel Lower Bounds, Optimality Gaps, and Exact Thresholds

The multivariate state-compression section above gives a rich family of upper bounds. We now complement those estimates with rigorous lower bounds. The key point is that every stable r-state MIMO predictor induces a finite-section block Hankel matrix of rank at most r. Combined with the exact weighted approximation formula from Section 10, this yields computable lower bounds for $\mathcal{E}_r^{\mathrm{MIMO}}$ in terms of singular values of finite Hankel sections of the canonical predictor symbol.

11.1. **A mild ellipticity regime.** For lower bounds it is convenient to strengthen Assumption 9.2 slightly.

Assumption 11.1 (Uniform spectral ellipticity)**.** Assume Assumption 9.2 and, in addition, that there exists a constant $m_W > 0$ such that

$$W(e^{i\theta}) \succeq m_W I_q \qquad \text{for a.e. } \theta \in [-\pi, \pi].$$

Equivalently,

$$m_W := \operatorname*{ess\,inf}_{\theta} \lambda_{\min}(W(e^{i\theta})) > 0.$$

Under Assumption 11.1, the weighted matrix defect dominates the unweighted Hardy-space error:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\operatorname{tr}\Big((R-P_{\mathrm{mat}})W(R-P_{\mathrm{mat}})^{*}\Big)\,d\theta \geq m_W\,\|R-P_{\mathrm{mat}}\|_{H_2^{q\times q}}^2.$$

11.2. **Finite block Hankel sections.** Write the canonical matrix predictor symbol as

$$P_{\mathrm{mat}}(z)=\sum_{j=1}^{\infty}P_j z^{j-1}, \qquad P_j\in\mathbb{C}^{q\times q}.$$

For $K\geq 1$, define the Kth *block Hankel section*

$$\Gamma_K(P_{\mathrm{mat}}) := \big[P_{i+j-1}\big]_{i,j=1}^{K}\in\mathbb{C}^{Kq\times Kq}.$$

Thus the (i,j) block of $\Gamma_K(P_{\mathrm{mat}})$ is P_{i+j-1}.

Lemma 11.2 (Finite Hankel rank bound for stable realizations)**.** *Let*

$$R(z)=C(I-zA)^{-1}B=\sum_{j=1}^{\infty}R_j z^{j-1}$$

be the transfer symbol of a stable r-state MIMO predictor. Then for every $K\geq 1$,

$$\operatorname{rank}\Gamma_K(R)\leq r.$$

Proof. Since

$$R_j=CA^{j-1}B \qquad (j\geq 1),$$

the (i,j) block of $\Gamma_K(R)$ is

$$R_{i+j-1}=CA^{i+j-2}B.$$

Define the block observability and controllability matrices

$$\mathcal{O}_K(A,C):=\begin{pmatrix} C\\ CA\\ \vdots\\ CA^{K-1}\end{pmatrix}\in\mathbb{C}^{Kq\times r}, \qquad \mathcal{C}_K(A,B):=\begin{pmatrix} B & AB & \cdots & A^{K-1}B\end{pmatrix}\in\mathbb{C}^{r\times Kq}.$$

Then

$$\Gamma_K(R)=\mathcal{O}_K(A,C)\,\mathcal{C}_K(A,B).$$

Hence

$$\operatorname{rank}\Gamma_K(R)\leq\min\{\operatorname{rank}\mathcal{O}_K(A,C),\operatorname{rank}\mathcal{C}_K(A,B)\}\leq r.$$

□

Lemma 11.3 (Hankel-section norm is controlled by Hardy error)**.** *Let*

$$E(z)=\sum_{j=1}^{\infty}E_j z^{j-1}\in H_2^{q\times q}.$$

Then for every $K\geq 1$,

$$\|\Gamma_K(E)\|_F^2\leq K\,\|E\|_{H_2^{q\times q}}^2.$$

Equivalently,

$$\|E\|_{H_2^{q\times q}}^2\geq\frac{1}{K}\,\|\Gamma_K(E)\|_F^2.$$

Proof. By definition,

$$\|\Gamma_K(E)\|_F^2 = \sum_{i=1}^{K}\sum_{j=1}^{K} \|E_{i+j-1}\|_F^2.$$

For each $m \in \{1, \dots, 2K-1\}$, let

$$\nu_m := \#\{(i,j) \in \{1,\dots,K\}^2 : i+j-1 = m\}.$$

Then $1 \le \nu_m \le K$ and

$$\|\Gamma_K(E)\|_F^2 = \sum_{m=1}^{2K-1} \nu_m \|E_m\|_F^2 \le K \sum_{m=1}^{2K-1} \|E_m\|_F^2 \le K \sum_{m=1}^{\infty} \|E_m\|_F^2 = K \|E\|_{H_2^{q\times q}}^2.$$

This is the claim. □

11.3. **Finite-Hankel lower bounds for state defect.** The next theorem is the main lower-bound mechanism.

Theorem 11.4 (Finite-Hankel lower bound for multivariate state defect)**.** *Assume Assumption 11.1. Let $K \ge 1$, and let*

$$\sigma_1\big(\Gamma_K(P_{\mathrm{mat}})\big) \ge \sigma_2\big(\Gamma_K(P_{\mathrm{mat}})\big) \ge \cdots \ge 0$$

be the singular values of the $Kq \times Kq$ block Hankel matrix $\Gamma_K(P_{\mathrm{mat}})$. Then for every $r \ge 0$,

$$\mathcal{E}_r^{\mathrm{MIMO}} \ge \frac{m_W}{K} \sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2.$$

Proof. Fix any $R \in \mathcal{R}_r^{\mathrm{MIMO}}$, and write

$$E := P_{\mathrm{mat}} - R.$$

By Corollary 10.15 and Assumption 11.1,

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\Big(EWE^*\Big)\, d\theta \ge m_W \|E\|_{H_2^{q\times q}}^2.$$

By Lemma 11.3,

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) \ge \frac{m_W}{K} \|\Gamma_K(E)\|_F^2.$$

Now

$$\Gamma_K(E) = \Gamma_K(P_{\mathrm{mat}}) - \Gamma_K(R),$$

and Lemma 11.2 gives

$$\operatorname{rank} \Gamma_K(R) \le r.$$

Therefore, by the Eckart–Young theorem for best rank-r approximation in Frobenius norm,

$$\|\Gamma_K(E)\|_F^2 \ge \sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2.$$

Combining the last three displays yields

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{mat}}(R) \ge \frac{m_W}{K} \sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2.$$

Since this bound is uniform over all $R \in \mathcal{R}_r^{\mathrm{MIMO}}$, taking the infimum proves the theorem. □

Corollary 11.5 (Positive-defect criterion from finite Hankel rank)**.** *Assume Assumption 11.1. If for some $K \geq 1$ one has*

$$\operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) > r,$$

then

$$\mathcal{E}_r^{\mathrm{MIMO}} > 0.$$

More quantitatively,

$$\mathcal{E}_r^{\mathrm{MIMO}} \geq \frac{m_W}{K} \sigma_{r+1}\big(\Gamma_K(P_{\mathrm{mat}})\big)^2.$$

Proof. If $\operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) > r$, then

$$\sigma_{r+1}\big(\Gamma_K(P_{\mathrm{mat}})\big) > 0.$$

Apply Theorem 11.4. □

11.4. **Exact zero-defect thresholds.** The preceding corollary shows that finite Hankel rank creates an unavoidable lower bound. On the other hand, Section 10 already contains exact realizability results. Together they imply a sharp threshold theorem on every class for which the finite Hankel rank equals the minimal realization dimension.

Definition 11.6 (Finite Hankel complexity)**.** Define the *finite Hankel complexity* of the canonical predictor symbol by

$$\mathfrak{h}(P_{\mathrm{mat}}) := \sup_{K \geq 1} \operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) \in \mathbb{N} \cup \{\infty\}.$$

Proposition 11.7 (Zero defect forces finite Hankel sections of rank at most r)**.** *Assume Assumption 11.1. If $\mathcal{E}_r^{\mathrm{MIMO}} = 0$, then for every $K \geq 1$,*

$$\operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) \leq r.$$

In particular,

$$\mathfrak{h}(P_{\mathrm{mat}}) \leq r.$$

Proof. If $\mathcal{E}_r^{\mathrm{MIMO}} = 0$, then Theorem 11.4 implies

$$\sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2 = 0$$

for every $K \geq 1$. Hence all singular values beyond the rth vanish, so

$$\operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) \leq r.$$

The last assertion is immediate. □

The converse requires a realization theorem. Rather than proving the full Kronecker theorem here, we record the needed part as a standard imported baseline from realization theory.

Theorem 11.8 (Imported rational-realization baseline)**.** *Let P_{mat} be a strictly proper stable rational $q \times q$ transfer matrix. Then its McMillan degree equals the rank of its infinite block Hankel operator, equivalently*

$$\deg_{\mathrm{McM}}(P_{\mathrm{mat}}) = \mathfrak{h}(P_{\mathrm{mat}}),$$

and there exists a stable realization of dimension $\deg_{\mathrm{McM}}(P_{\mathrm{mat}})$. This is a standard consequence of classical realization theory / Kronecker's theorem.[14, 15]

Theorem 11.9 (Exact zero-defect threshold for rational predictor symbols)**.** *Assume Assumption 11.1, and suppose that P_{mat} is strictly proper, stable, and rational of McMillan degree N. Then*

$$\mathcal{E}_r^{\mathrm{MIMO}} = 0 \quad \Longleftrightarrow \quad r \geq N.$$

Equivalently,

$$\mathcal{E}_r^{\mathrm{MIMO}} > 0 \quad \text{for every } r < N, \qquad \mathcal{E}_N^{\mathrm{MIMO}} = 0.$$

Proof. If $r \geq N$, Theorem 11.8 provides a stable realization of P_{mat} with N states. Hence $P_{\mathrm{mat}} \in \mathcal{R}_r^{\mathrm{MIMO}}$, and Corollary 10.15 gives $\mathcal{E}_r^{\mathrm{MIMO}} = 0$.

Conversely, if $r < N$, then by Theorem 11.8

$$\mathfrak{h}(P_{\mathrm{mat}}) = N > r.$$

Therefore there exists K such that

$$\operatorname{rank} \Gamma_K(P_{\mathrm{mat}}) > r.$$

Corollary 11.5 now implies $\mathcal{E}_r^{\mathrm{MIMO}} > 0$. □

11.5. **A two-sided optimality-gap sandwich.** The upper bounds from Section 10 and the finite-Hankel lower bounds above combine into a clean sandwich estimate.

Corollary 11.10 (Finite-Hankel / balanced-truncation sandwich)**.** *Assume Assumption 11.1, and suppose that P_{mat} is strictly proper, stable, and rational. Then for every $r \geq 0$ and every $K \geq 1$,*

$$\frac{m_W}{K} \sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2 \leq \mathcal{E}_r^{\mathrm{MIMO}} \leq 4q \, \|W\|_{L^\infty_{\mathrm{op}}} \left(\sum_{k=r+1}^{N} s_k \right)^2,$$

where $N = \deg_{\mathrm{McM}}(P_{\mathrm{mat}})$ and $s_1 \geq \cdots \geq s_N > 0$ are the Hankel singular values of P_{mat}.

Proof. The lower bound is Theorem 11.4. The upper bound is Corollary 10.23. □

Remark 11.11. Corollary 11.10 is the first genuine *optimality-gap* statement in the draft. The left-hand side is a finite-section obstruction coming from the first $2K-1$ Markov parameters of the canonical predictor symbol; the right-hand side is a system-theoretic compressibility bound coming from balanced truncation. Matching these two sides sharply, even on concrete model classes, is already a substantial open problem.

11.6. **Interpretation.** Section 11 adds an eighth backbone to the draft:

$$\boxed{\mathcal{E}_r^{\mathrm{MIMO}} \geq \frac{m_W}{K} \sum_{j=r+1}^{Kq} \sigma_j\big(\Gamma_K(P_{\mathrm{mat}})\big)^2 \qquad (K \geq 1)}.$$

This says that state compression is obstructed not only by phase/PACF tails and Hankel singular values, but already by the low-rank approximability of finite block Hankel sections built from the earliest predictor coefficients. In particular:

(i) finite-context upper bounds alone do not tell the whole story;
(ii) balanced-truncation upper bounds alone do not tell the whole story;
(iii) the finite-Hankel lower bounds provide a concrete bridge between predictor coefficients, realization complexity, and unavoidable state defect.

12. Matched Two-Sided Laws on Solvable Structured Classes

The lower bounds from Section 11 show that unrestricted multivariable state compression has genuine obstructions, but they do not by themselves give asymptotically sharp formulas on concrete model classes. In this section we isolate two structured regimes in which upper and lower mechanisms can be matched completely or up to exact rate. The first is the block shift-register class, where state compression is identical to finite-context prediction. The second is a diagonal multichannel memory-allocation class, where one can solve the optimal allocation problem exactly to first order.

12.1. **Exact state laws for block shift-register classes.** Let $(x_t)_{t\in\mathbb{Z}}$ be a centered q-variate stationary Gaussian process. For $L \geq 1$ define the block shift-register state

$$s_t^{(L)} := \begin{bmatrix} x_t^\top & x_{t-1}^\top & \cdots & x_{t-L+1}^\top \end{bmatrix}^\top \in \mathbb{R}^{qL}.$$

Consider the structured predictor class

$$\mathfrak{S}_{qL}^{\mathrm{shift}} := \{\hat{x}_{t+1} = B s_t^{(L)} : \ B \in \mathbb{R}^{q\times qL}\}.$$

Define the best defect within this class by

$$\mathcal{E}_{qL}^{\mathrm{shift}} := \inf_{B\in\mathbb{R}^{q\times qL}} \mathbb{E}\|x_{t+1} - B s_t^{(L)}\|^2 - \mathbb{E}\|x_{t+1} - m_t^\star\|^2.$$

Theorem 12.1 (Exact shift-register state law)**.** *For every $L \geq 1$,*

$$\mathcal{E}_{qL}^{\mathrm{shift}} = \mathcal{E}_{\mathrm{ctx}}(L).$$

If, in addition, the process is purely nondeterministic so that the matrix PACF representation from Section 9 applies, then

$$\mathcal{E}_{qL}^{\mathrm{shift}} = \operatorname{tr}\Delta_{\mathrm{ctx}}(L) = \sum_{n=L+1}^{\infty} \operatorname{tr}\big(V_{n-1}^{1/2}\alpha_n\alpha_n^* V_{n-1}^{1/2}\big).$$

Proof. The state $s_t^{(L)}$ is simply a linear re-encoding of the finite block $(x_t, \dots, x_{t-L+1})$, so the linear space generated by its coordinates is exactly the finite-past block prediction space. Therefore the minimum over B is attained by the orthogonal projection of x_{t+1} onto that space, namely by the finite-past linear predictor $M_t^{(L)}$. By Proposition 9.7, the resulting excess risk is exactly

$$\operatorname{tr}\Delta_{\mathrm{ctx}}(L) = \mathcal{E}_{\mathrm{ctx}}(L).$$

The second identity is Theorem 9.21. □

Corollary 12.2 (Exact structured state laws inherited from context laws)**.** *Suppose that for some positive function Φ one has*

$$\mathcal{E}_{\mathrm{ctx}}(L) \asymp \Phi(L).$$

Then along multiples of q,

$$\mathcal{E}_r^{\mathrm{shift}} \asymp \Phi(r/q), \qquad r \in q\mathbb{N},$$

where $\mathcal{E}_r^{\mathrm{shift}} := \mathcal{E}_r^{\mathrm{shift}}(x)$ denotes the optimal defect in the block shift-register class of dimension r.

Proof. Apply Theorem 12.1 with $r = qL$. □

Corollary 12.3 (Exact FARIMA shift-register law)**.** *Assume the multivariate FARIMA-type asymptotic from Corollary 9.26:*

$$\mathcal{E}_{\mathrm{ctx}}(L) = \frac{d^2}{L}\operatorname{tr}(V_\infty) + O(L^{-2}).$$

Then, for $r = qL$,

$$\mathcal{E}_r^{\mathrm{shift}} = \frac{qd^2}{r}\,\mathrm{tr}(V_\infty) + O(r^{-2}).$$

Proof. Substitute $L = r/q$ into Theorem 12.1. □

12.2. **Diagonal multichannel memory allocation.** We now consider a second structured class in which the total memory budget can be distributed unevenly across channels.

Assumption 12.4 (Orthogonal scalar channels)**.** The process is of the form

$$x_t = \big(x_t^{(1)}, \dots, x_t^{(q)}\big)^\top,$$

where each $(x_t^{(i)})_{t\in\mathbb{Z}}$ is a centered scalar stationary Gaussian process, and different channels are pairwise orthogonal in L^2:

$$\mathbb{E}[x_s^{(i)} x_t^{(j)}] = 0 \qquad \text{for all } s,t \in \mathbb{Z} \text{ and } i \neq j.$$

Under Assumption 12.4, the full-past predictor splits coordinatewise:

$$m_t^\star = \big(m_t^{\star,(1)}, \dots, m_t^{\star,(q)}\big)^\top, \qquad m_t^{\star,(i)} := \mathbb{E}[x_{t+1}^{(i)} \mid \mathcal{F}_t^{(i)}],$$

where $\mathcal{F}_t^{(i)} := \sigma(x_s^{(i)} : s \le t)$. For each channel define the scalar context defect

$$\mathcal{E}_i(L) := \mathbb{E}\Big[|x_{t+1}^{(i)} - \mathbb{E}[x_{t+1}^{(i)} \mid x_t^{(i)}, \dots, x_{t-L+1}^{(i)}]|^2\Big] - \mathbb{E}\Big[|x_{t+1}^{(i)} - m_t^{\star,(i)}|^2\Big].$$

For an integer allocation vector

$$\mathbf{L} = (L_1, \dots, L_q) \in \mathbb{N}^q,$$

consider the diagonal shift-register state

$$s_t(\mathbf{L}) := \big(x_t^{(1)}, \dots, x_{t-L_1+1}^{(1)};\ \dots\ ; x_t^{(q)}, \dots, x_{t-L_q+1}^{(q)}\big)^\top \in \mathbb{R}^{L_1+\cdots+L_q}$$

and the corresponding diagonal predictor class

$$\mathfrak{S}_{\mathbf{L}}^{\mathrm{diag}} := \Big\{\hat{x}_{t+1}^{(i)} = b_i^\top s_t^{(i)}(L_i),\ i = 1, \dots, q\Big\},$$

where $s_t^{(i)}(L_i)$ is the length-L_i scalar lag vector of channel i.

Proposition 12.5 (Exact defect decomposition under orthogonal channels)**.** *Under Assumption 12.4, the optimal defect in the class $\mathfrak{S}_{\mathbf{L}}^{\mathrm{diag}}$ is exactly*

$$\mathcal{D}(\mathbf{L}) := \inf_{\hat{x}\in\mathfrak{S}_{\mathbf{L}}^{\mathrm{diag}}} \mathbb{E}\|x_{t+1} - \hat{x}_{t+1}\|^2 - \mathbb{E}\|x_{t+1} - m_t^\star\|^2 = \sum_{i=1}^q \mathcal{E}_i(L_i).$$

Proof. By orthogonality of the channels,

$$\mathbb{E}\|x_{t+1} - \hat{x}_{t+1}\|^2 = \sum_{i=1}^q \mathbb{E}\big|x_{t+1}^{(i)} - \hat{x}_{t+1}^{(i)}\big|^2,$$

and similarly

$$\mathbb{E}\|x_{t+1} - m_t^\star\|^2 = \sum_{i=1}^q \mathbb{E}\big|x_{t+1}^{(i)} - m_t^{\star,(i)}\big|^2.$$

Moreover, the predictor constraints are coordinatewise and independent across channels. Therefore the optimization splits into q scalar problems, and the optimal defect is the sum of the scalar channel defects. □

Define the optimal diagonal structured state defect at total memory budget r by

$$\mathcal{D}_r^{\mathrm{diag}} := \min\Big\{\mathcal{D}(\mathbf{L}) : \ \mathbf{L} \in \mathbb{N}^q, \ \sum_{i=1}^q L_i = r\Big\}.$$

12.3. **Power-law channels: exact first-order asymptotics.**

Theorem 12.6 (Power-law memory allocation theorem)**.** *Assume Assumption 12.4. Suppose that for some $\beta > 0$ and constants $c_i > 0$,*

$$\mathcal{E}_i(L) = c_i L^{-\beta} + O(L^{-\beta-1}) \qquad (L \to \infty), \quad i = 1, \dots, q.$$

Let

$$S_\beta := \sum_{i=1}^q c_i^{1/(\beta+1)}.$$

Then

$$\mathcal{D}_r^{\mathrm{diag}} = S_\beta^{\beta+1} r^{-\beta} + O(r^{-\beta-1}).$$

Moreover:

(i) *there exists an asymptotically optimal integer allocation $\mathbf{L}^\sharp(r)$ such that*

$$L_i^\sharp(r) = \frac{c_i^{1/(\beta+1)}}{S_\beta}\, r + O(1), \qquad i = 1, \dots, q;$$

(ii) *if $\mathbf{L}^\star(r)$ is any sequence of asymptotically optimal allocations, then*

$$L_i^\star(r) = \frac{c_i^{1/(\beta+1)}}{S_\beta}\, r + o(r), \qquad i = 1, \dots, q.$$

Proof. Let

$$f(x_1, \dots, x_q) := \sum_{i=1}^q c_i x_i^{-\beta} \qquad (x_i > 0).$$

The continuous constrained minimization problem

$$\min\{f(x) : \ x_i > 0, \ \sum_i x_i = r\}$$

has a unique minimizer because each map $x \mapsto c_i x^{-\beta}$ is strictly convex. The Lagrange equations are

$$-\beta c_i x_i^{-\beta-1} + \lambda = 0,$$

so the minimizer is

$$x_i^{\mathrm{cont}}(r) = \frac{c_i^{1/(\beta+1)}}{S_\beta}\, r, \qquad S_\beta = \sum_{j=1}^q c_j^{1/(\beta+1)}.$$

Substituting back gives the continuous minimum

$$\min f = S_\beta^{\beta+1} r^{-\beta}.$$

Choose an integer allocation $\mathbf{L}^\sharp(r)$ with

$$L_i^\sharp(r) = x_i^{\mathrm{cont}}(r) + O(1) \qquad \text{and} \qquad \sum_{i=1}^q L_i^\sharp(r) = r.$$

Such an allocation is obtained by rounding the continuous optimizer and adjusting finitely many coordinates by at most 1. Since each $x_i^{\mathrm{cont}}(r) \asymp r$, Taylor expansion gives

$$f(\mathbf{L}^\sharp(r)) = f(x^{\mathrm{cont}}(r)) + O(r^{-\beta-1}) = S_\beta^{\beta+1} r^{-\beta} + O(r^{-\beta-1}).$$

Using the assumed asymptotic for each channel,

$$\mathcal{D}(\mathbf{L}^\sharp(r)) = f(\mathbf{L}^\sharp(r)) + O(r^{-\beta-1}) = S_\beta^{\beta+1} r^{-\beta} + O(r^{-\beta-1}),$$

so

$$\mathcal{D}_r^{\mathrm{diag}} \le S_\beta^{\beta+1} r^{-\beta} + O(r^{-\beta-1}).$$

This already proves part (i).

For the matching lower bound, fix for each i constants $0 < c_i^- < c_i < c_i^+$ and an integer L_0 such that

$$c_i^- L^{-\beta} \le \mathcal{E}_i(L) \le c_i^+ L^{-\beta} \qquad (L \ge L_0).$$

Let $\mathbf{L}^\star(r)$ be any minimizing allocation. Since the upper bound above shows

$$\mathcal{D}_r^{\mathrm{diag}} = O(r^{-\beta}) \to 0,$$

each coordinate of $\mathbf{L}^\star(r)$ must tend to infinity; hence for all large r, every $L_i^\star(r) \ge L_0$. For such r,

$$c_i^- \big(L_i^\star(r)\big)^{-\beta} \le \mathcal{E}_i(L_i^\star(r)) \le \mathcal{D}(\mathbf{L}^\star(r)) = \mathcal{D}_r^{\mathrm{diag}} \le C_0 r^{-\beta}$$

for some constant $C_0 > 0$. Therefore

$$L_i^\star(r) \ge \left(\frac{c_i^-}{C_0}\right)^{1/\beta} r,$$

so every minimizing allocation satisfies

$$L_i^\star(r) \asymp r \qquad (i = 1, \dots, q)$$

uniformly in r. Consequently the channelwise remainder estimate is uniform on minimizing allocations, and

$$\mathcal{D}(\mathbf{L}^\star(r)) = f(\mathbf{L}^\star(r)) + O(r^{-\beta-1}).$$

Since $f(\mathbf{L}^\star(r)) \ge \min f = S_\beta^{\beta+1} r^{-\beta}$, we obtain

$$\mathcal{D}_r^{\mathrm{diag}} \ge S_\beta^{\beta+1} r^{-\beta} + O(r^{-\beta-1}).$$

Together with the upper bound, this proves the stated asymptotic formula.

Finally, let $\mathbf{L}^\star(r)$ be any asymptotically optimal sequence. Set

$$u_i(r) := \frac{L_i^\star(r)}{r}.$$

The lower bound above shows that each $u_i(r)$ stays away from 0 for large r, so after passing to a subsequence we may assume $u(r) \to u$ in the simplex. Dividing the identity

$$\mathcal{D}(\mathbf{L}^\star(r)) = f(\mathbf{L}^\star(r)) + O(r^{-\beta-1})$$

by $r^{-\beta}$ gives

$$r^\beta \mathcal{D}(\mathbf{L}^\star(r)) = \sum_{i=1}^q c_i u_i(r)^{-\beta} + o(1).$$

Asymptotic optimality implies that the left-hand side converges to $S_\beta^{\beta+1}$, so the limit point u minimizes

$$\sum_{i=1}^q c_i u_i^{-\beta}$$

over the simplex. By strict convexity, the minimizer is unique and equals

$$u_i = \frac{c_i^{1/(\beta+1)}}{S_\beta}.$$

Hence every convergent subsequence has the same limit, so in fact

$$u_i(r) \to \frac{c_i^{1/(\beta+1)}}{S_\beta},$$

which proves part (ii). □

Corollary 12.7 (Diagonal multichannel FARIMA law)**.** *Under Assumption 12.4, suppose channel i satisfies the scalar FARIMA context law*

$$\mathcal{E}_i(L) = c_i L^{-1} + O(L^{-2}).$$

Then

$$\mathcal{D}_r^{\mathrm{diag}} = \frac{\left(\sum_{i=1}^q \sqrt{c_i}\right)^2}{r} + O(r^{-2}),$$

and an asymptotically optimal allocation is

$$L_i^\star(r) = \frac{\sqrt{c_i}}{\sum_{j=1}^q \sqrt{c_j}}\, r + O(1).$$

In particular, if $c_i = d_i^2 v_{\infty,i}$, then

$$\mathcal{D}_r^{\mathrm{diag}} = \frac{\left(\sum_{i=1}^q |d_i|\sqrt{v_{\infty,i}}\right)^2}{r} + O(r^{-2}).$$

Proof. Apply Theorem 12.6 with $\beta = 1$. □

12.4. **Exponentially decaying channels: a harmonic-bottleneck law.**

Theorem 12.8 (Exponential memory allocation theorem)**.** *Assume Assumption 12.4. Suppose there exist constants $a_i > 0$ and $0 < c_i \le C_i < \infty$ such that for all sufficiently large L,*

$$c_i e^{-a_i L} \le \mathcal{E}_i(L) \le C_i e^{-a_i L}, \qquad i = 1, \dots, q.$$

Set

$$H_a := \sum_{i=1}^q \frac{1}{a_i}.$$

Then there exist constants $0 < c_\star \le C_\star < \infty$ such that

$$c_\star e^{-r/H_a} \le \mathcal{D}_r^{\mathrm{diag}} \le C_\star e^{-r/H_a} \qquad (r \gg 1).$$

Thus the optimal structured defect decays exponentially at rate e^{-r/H_a}, where the effective bottleneck is the harmonic sum H_a of the channel exponents.

Proof. For the lower bound, let $\mathbf{L} = (L_1, \dots, L_q)$ be any allocation with $\sum_i L_i = r$. Then

$$\mathcal{D}(\mathbf{L}) \ge \sum_{i=1}^q c_i e^{-a_i L_i}.$$

By the weighted arithmetic-geometric mean inequality with weights

$$w_i := \frac{1/a_i}{H_a}, \qquad \sum_i w_i = 1,$$

we have

$$\sum_{i=1}^{q} c_i e^{-a_i L_i} = H_a \sum_{i=1}^{q} w_i\, a_i c_i e^{-a_i L_i} \geq H_a \prod_{i=1}^{q} (a_i c_i e^{-a_i L_i})^{w_i}.$$

Since

$$\sum_{i=1}^{q} w_i a_i L_i = \frac{1}{H_a} \sum_{i=1}^{q} L_i = \frac{r}{H_a},$$

this gives

$$\mathcal{D}(\mathbf{L}) \geq H_a \left(\prod_{i=1}^{q} (a_i c_i)^{1/a_i} \right)^{1/H_a} e^{-r/H_a}.$$

Taking the minimum over all $\mathbf{L}$ proves the lower bound with

$$c_\star := H_a \left(\prod_{i=1}^{q} (a_i c_i)^{1/a_i} \right)^{1/H_a}.$$

For the upper bound, choose an allocation with

$$L_i^\sharp(r) = \left\lfloor \frac{r}{H_a a_i} \right\rfloor, \qquad i = 1, \dots, q-1,$$

and let $L_q^\sharp(r) := r - \sum_{i=1}^{q-1} L_i^\sharp(r)$. Then each $L_i^\sharp(r) = \frac{r}{H_a a_i} + O(1)$, so

$$\mathcal{E}_i(L_i^\sharp(r)) \leq C_i e^{-a_i L_i^\sharp(r)} \leq C_i' e^{-r/H_a}$$

for suitable constants C_i'. Summing over i and using Proposition 12.5 gives

$$\mathcal{D}_r^{\mathrm{diag}} \leq \sum_{i=1}^{q} \mathcal{E}_i(L_i^\sharp(r)) \leq \left(\sum_{i=1}^{q} C_i' \right) e^{-r/H_a}.$$

This proves the upper bound. □

Remark 12.9 (Interpretation for memory allocation). Theorem 12.6 says that in heterogeneous long-memory channels, an optimal memory budget should be split in proportion to $c_i^{1/(\beta+1)}$. Theorem 12.8 says that in heterogeneous short-memory channels, the overall exponential compression rate is controlled by the harmonic sum of the individual rates. These are the first matched two-sided state-dimension laws in the draft that are simultaneously explicit, nontrivial, and fully proved.

13. Constant Channel Mixing and Transferred Laws

The solvable classes of Section 12 were formulated in coordinates in which the channels decouple. We now show that their quantitative laws survive under any constant invertible mixing of the observation coordinates. This produces the first rigorous results in the draft for observed multivariate processes whose coordinates are genuinely coupled.

13.1. **Invertible mixing of the observed channels.** Let $(Y_t)_{t\in\mathbb{Z}}$ be a centered q-variate stationary Gaussian process, and let

$$S \in \mathrm{GL}_q(\mathbb{R}).$$

Define a new observed process by

$$X_t := SY_t.$$

Write $\mathcal{F}_{Y,t}$ and $\mathcal{F}_{X,t}$ for the full-past sigma-algebras generated by $(Y_s)_{s\leq t}$ and $(X_s)_{s\leq t}$, and similarly $\mathcal{F}_{Y,t}^{(L)}$ and $\mathcal{F}_{X,t}^{(L)}$ for the length-L context sigma-algebras.

Proposition 13.1 (Predictor transport under invertible mixing)**.** *For every t and every $L \geq 1$,*

$$\mathcal{F}_{X,t} = \mathcal{F}_{Y,t}, \qquad \mathcal{F}^{(L)}_{X,t} = \mathcal{F}^{(L)}_{Y,t},$$

and the corresponding predictors satisfy

$$M^\star_{X,t} = SM^\star_{Y,t}, \qquad M^{(L)}_{X,t} = SM^{(L)}_{Y,t}.$$

Proof. Because $X_s = SY_s$ and $Y_s = S^{-1}X_s$ for each s, the families $\{X_s : s \leq t\}$ and $\{Y_s : s \leq t\}$ generate the same sigma-algebra, and likewise for the finite windows of length L. Therefore

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_{X,t}] = \mathbb{E}[SY_{t+1} \mid \mathcal{F}_{Y,t}] = S\mathbb{E}[Y_{t+1} \mid \mathcal{F}_{Y,t}] = SM^\star_{Y,t}.$$

The finite-context identity is identical:

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L)}_{X,t}] = S\mathbb{E}[Y_{t+1} \mid \mathcal{F}^{(L)}_{Y,t}] = SM^{(L)}_{Y,t}.$$

□

Let

$$s_{\min}(S) := \sigma_{\min}(S), \qquad s_{\max}(S) := \sigma_{\max}(S)$$

be the smallest and largest singular values of S.

Theorem 13.2 (Context-defect quasi-invariance under constant mixing)**.** *For every $L \geq 1$,*

$$s_{\min}(S)^2\, \mathcal{E}^Y_{\mathrm{ctx}}(L) \leq \mathcal{E}^X_{\mathrm{ctx}}(L) \leq s_{\max}(S)^2\, \mathcal{E}^Y_{\mathrm{ctx}}(L),$$

where $\mathcal{E}^Y_{\mathrm{ctx}}(L)$ and $\mathcal{E}^X_{\mathrm{ctx}}(L)$ denote the context defects of the processes Y and X, respectively. If S is orthogonal, then

$$\mathcal{E}^X_{\mathrm{ctx}}(L) = \mathcal{E}^Y_{\mathrm{ctx}}(L) \qquad \textit{for every } L \geq 1.$$

Proof. By Proposition 13.1,

$$M^\star_{X,t} - M^{(L)}_{X,t} = S\big(M^\star_{Y,t} - M^{(L)}_{Y,t}\big).$$

Hence

$$\mathcal{E}^X_{\mathrm{ctx}}(L) = \mathbb{E}\big\|S(M^\star_{Y,t} - M^{(L)}_{Y,t})\big\|^2.$$

The singular-value bounds

$$s_{\min}(S)^2\|u\|^2 \leq \|Su\|^2 \leq s_{\max}(S)^2\|u\|^2$$

now give

$$s_{\min}(S)^2\, \mathbb{E}\|M^\star_{Y,t} - M^{(L)}_{Y,t}\|^2 \leq \mathcal{E}^X_{\mathrm{ctx}}(L) \leq s_{\max}(S)^2\, \mathbb{E}\|M^\star_{Y,t} - M^{(L)}_{Y,t}\|^2.$$

By Proposition 9.7, the outer expectations are exactly $\mathcal{E}^Y_{\mathrm{ctx}}(L)$. If S is orthogonal, then $\|Su\| = \|u\|$ and equality holds throughout. □

13.2. **Full MIMO state-dimension laws under mixing.** For a process $Z = (Z_t)$, write $\mathcal{E}^{\mathrm{MIMO}}_r(Z)$ for the optimal r-state MIMO defect defined in Section 10.

Theorem 13.3 (MIMO state-defect quasi-invariance under constant mixing)**.** *For every $r \geq 0$,*

$$s_{\min}(S)^2\, \mathcal{E}^{\mathrm{MIMO}}_r(Y) \leq \mathcal{E}^{\mathrm{MIMO}}_r(X) \leq s_{\max}(S)^2\, \mathcal{E}^{\mathrm{MIMO}}_r(Y).$$

If S is orthogonal, then

$$\mathcal{E}^{\mathrm{MIMO}}_r(X) = \mathcal{E}^{\mathrm{MIMO}}_r(Y) \qquad \textit{for every } r \geq 0.$$

Proof. Let (A, B, C) be any stable r-state MIMO predictor for Y:

$$s_t = As_{t-1} + BY_t, \qquad \widehat{Y}_{t+1} = Cs_t.$$

Since $Y_t = S^{-1}X_t$, define the transported predictor for X by

$$s_t = As_{t-1} + BS^{-1}X_t, \qquad \widehat{X}_{t+1} = SCs_t = S\widehat{Y}_{t+1}.$$

This is again a stable r-state MIMO predictor. Proposition 13.1 gives

$$M^\star_{X,t} - \widehat{X}_{t+1} = S(M^\star_{Y,t} - \widehat{Y}_{t+1}),$$

so

$$\mathbb{E}\|M^\star_{X,t} - \widehat{X}_{t+1}\|^2 \le s_{\max}(S)^2\, \mathbb{E}\|M^\star_{Y,t} - \widehat{Y}_{t+1}\|^2.$$

Using the exact defect identity from Section 10 and taking the infimum over all r-state predictors for Y yields

$$\mathcal{E}^{\mathrm{MIMO}}_r(X) \le s_{\max}(S)^2\, \mathcal{E}^{\mathrm{MIMO}}_r(Y).$$

Applying the same argument with S^{-1} in place of S gives

$$\mathcal{E}^{\mathrm{MIMO}}_r(Y) \le \|S^{-1}\|^2_{\mathrm{op}}\, \mathcal{E}^{\mathrm{MIMO}}_r(X) = \frac{1}{s_{\min}(S)^2}\, \mathcal{E}^{\mathrm{MIMO}}_r(X),$$

which rearranges to the lower bound. If S is orthogonal, then both inequalities become equalities. □

Corollary 13.4 (Transfer of full-state laws under constant mixing)**.** *Any upper or lower state-dimension law proved for Y transfers to $X = SY$ with the same rate and with constants modified by at most the factors $s_{\min}(S)^2$ and $s_{\max}(S)^2$. In particular, exact zero-defect thresholds are invariant under orthogonal mixing.*

Proof. Apply Theorem 13.3. □

13.3. **Mixed structured classes.** Assume now that Y satisfies Assumption 12.4. For an allocation vector $\mathbf{L} = (L_1, \dots, L_q)$, define the transported structured predictor class on $X = SY$ by

$$\mathfrak{S}^{\mathrm{mix}}_{S,\mathbf{L}}(X) := \big\{\hat{X}_{t+1} = S\hat{Y}_{t+1} :\ \hat{Y}_{t+1} \in \mathfrak{S}^{\mathrm{diag}}_{\mathbf{L}}(Y)\big\}.$$

Let

$$\mathcal{D}^{\mathrm{mix}}_S(\mathbf{L}; X) := \inf_{\hat{X} \in \mathfrak{S}^{\mathrm{mix}}_{S,\mathbf{L}}(X)} \mathbb{E}\|X_{t+1} - \hat{X}_{t+1}\|^2 - \mathbb{E}\|X_{t+1} - M^\star_{X,t}\|^2.$$

Finally, define the optimal mixed structured defect at total budget r by

$$\mathcal{D}^{\mathrm{mix}}_{r,S}(X) := \min\Big\{\mathcal{D}^{\mathrm{mix}}_S(\mathbf{L}; X) :\ \mathbf{L} \in \mathbb{N}^q,\ \sum_{i=1}^{q} L_i = r\Big\}.$$

For clarity, write

$$\mathcal{D}(\mathbf{L}; Y) := \mathcal{D}(\mathbf{L}), \qquad \mathcal{D}^{\mathrm{diag}}_r(Y) := \mathcal{D}^{\mathrm{diag}}_r,$$

for the diagonal structured defects associated with the reference process Y.

Proposition 13.5 (Transport of diagonal structured defects)**.** *Assume Assumption 12.4. Then for every allocation vector $\mathbf{L}$,*

$$s_{\min}(S)^2\, \mathcal{D}(\mathbf{L}; Y) \le \mathcal{D}^{\mathrm{mix}}_S(\mathbf{L}; X) \le s_{\max}(S)^2\, \mathcal{D}(\mathbf{L}; Y),$$

where $\mathcal{D}(\mathbf{L}; Y)$ denotes the diagonal structured defect from Proposition 12.5. Consequently, for every $r \ge q$,

$$s_{\min}(S)^2\, \mathcal{D}^{\mathrm{diag}}_r(Y) \le \mathcal{D}^{\mathrm{mix}}_{r,S}(X) \le s_{\max}(S)^2\, \mathcal{D}^{\mathrm{diag}}_r(Y).$$

If S is orthogonal, then equality holds in both displays.

Proof. Fix $\mathbf{L}$ and any predictor $\hat{Y} \in \mathfrak{S}_{\mathbf{L}}^{\mathrm{diag}}(Y)$. Set $\hat{X} := S\hat{Y}$. By Proposition 13.1,

$$M^{\star}_{X,t} - \hat{X}_{t+1} = S(M^{\star}_{Y,t} - \hat{Y}_{t+1}),$$

so the same singular-value comparison as in Theorem 13.2 gives

$$s_{\min}(S)^2\,\mathbb{E}\|M^{\star}_{Y,t} - \hat{Y}_{t+1}\|^2 \le \mathbb{E}\|M^{\star}_{X,t} - \hat{X}_{t+1}\|^2 \le s_{\max}(S)^2\,\mathbb{E}\|M^{\star}_{Y,t} - \hat{Y}_{t+1}\|^2.$$

Taking the infimum over $\hat{Y} \in \mathfrak{S}_{\mathbf{L}}^{\mathrm{diag}}(Y)$ proves the first display, and minimizing over $\sum_i L_i = r$ proves the second. Orthogonal mixing gives equality. □

Corollary 13.6 (Power-law and exponential laws for mixed observed channels)**.** *Assume Assumption 12.4, and let $X_t = SY_t$ with $S \in \mathrm{GL}_q(\mathbb{R})$.*

(i) *If the channel defects of Y satisfy the power-law hypothesis of Theorem 12.6, then*

$$\mathcal{D}^{\mathrm{mix}}_{r,S}(X) \asymp r^{-\beta}.$$

If S is orthogonal, then the exact first-order asymptotic is preserved:

$$\mathcal{D}^{\mathrm{mix}}_{r,S}(X) = S_{\beta}^{\beta+1} r^{-\beta} + O(r^{-\beta-1}).$$

(ii) *If the channel defects of Y satisfy the exponential hypothesis of Theorem 12.8, then*

$$\mathcal{D}^{\mathrm{mix}}_{r,S}(X) \asymp e^{-r/H_a}.$$

If S is orthogonal, then the optimal mixed structured defect equals the diagonal structured defect of Y exactly.

Proof. Part (i) follows from Proposition 13.5 together with Theorem 12.6; the orthogonal case is the equality statement in Proposition 13.5. Part (ii) is identical, using Theorem 12.8 in place of Theorem 12.6. □

Remark 13.7. Section 13 is the first point in the draft where we can move beyond coordinate-decoupled observations without giving up rigorous control. The resulting observed spectral density

$$W_X(e^{i\theta}) = S\,W_Y(e^{i\theta})\,S^{\top}$$

typically has nonzero off-diagonal entries at every frequency, yet the context and state exponents survive unchanged; only the multiplicative constants are distorted, and even those distortions disappear under orthogonal mixing.

14. Frequency-Dependent Finite-Memory Mixing and Transported Laws

We now move beyond constant channel mixing to genuinely frequency-dependent couplings. To keep all proofs exact, we restrict to *causal finite-memory invertible filters*. Let $(Y_t)_{t\in\mathbb{Z}}$ be a centered q-variate stationary Gaussian process, and let

$$X_t = \sum_{k=0}^{m} S_k Y_{t-k}, \qquad Y_t = \sum_{\ell=0}^{n} T_\ell X_{t-\ell},$$

where $S_0, \dots, S_m, T_0, \dots, T_n \in \mathbb{R}^{q\times q}$. Thus the two observed processes are linked by mutually inverse causal finite-memory filters. In particular,

$$T_0 S_0 = S_0 T_0 = I_q,$$

so S_0 and T_0 are invertible and $T_0 = S_0^{-1}$. If we define the matrix polynomial

$$S(z) := \sum_{k=0}^{m} S_k z^k,$$

then the spectral densities satisfy

$$W_X(e^{i\theta}) = S(e^{-i\theta})W_Y(e^{i\theta})S(e^{-i\theta})^*.$$

Unless $m = 0$, this is a genuinely frequency-dependent coupling of the channels.

Write $\mathcal{F}_{Y,t}, \mathcal{F}_{X,t}$ for the full-past sigma-algebras and $\mathcal{F}_{Y,t}^{(L)}, \mathcal{F}_{X,t}^{(L)}$ for the corresponding finite-context sigma-algebras of the processes Y and X.

Lemma 14.1 (Monotonicity under information enlargement)**.** *Let Z_{t+1} be a square-integrable $\mathbb{R}^q$-valued random vector, let $\mathcal{F}_t^Z := \sigma(Z_s : s \le t)$, and let $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{F}_t^Z$ be sub-sigma-algebras. Then*

$$\mathbb{E}\big\|\mathbb{E}[Z_{t+1} \mid \mathcal{H}_2] - \mathbb{E}[Z_{t+1} \mid \mathcal{H}_1]\big\|^2 \le \mathbb{E}\big\|Z_{t+1} - \mathbb{E}[Z_{t+1} \mid \mathcal{H}_1]\big\|^2 - \mathbb{E}\big\|Z_{t+1} - \mathbb{E}[Z_{t+1} \mid \mathcal{H}_2]\big\|^2,$$

and in particular

$$\mathbb{E}\big\|\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t^Z] - \mathbb{E}[Z_{t+1} \mid \mathcal{H}_2]\big\|^2 \le \mathbb{E}\big\|\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t^Z] - \mathbb{E}[Z_{t+1} \mid \mathcal{H}_1]\big\|^2.$$

Proof. Apply Proposition 3.1 twice, first with $\mathcal{H} = \mathcal{H}_1$ and then with $\mathcal{H} = \mathcal{H}_2$, and subtract the resulting identities. Since conditional expectation is the orthogonal projection in L^2, the second displayed inequality is just the statement that projection error decreases when the sigma-algebra is enlarged. □

Proposition 14.2 (Full-past equality and finite-context inclusions)**.** *For every $t \in \mathbb{Z}$ and every $L \ge 1$,*

$$\mathcal{F}_{X,t} = \mathcal{F}_{Y,t}, \qquad \mathcal{F}_{X,t}^{(L)} \subseteq \mathcal{F}_{Y,t}^{(L+m)}, \qquad \mathcal{F}_{Y,t}^{(L)} \subseteq \mathcal{F}_{X,t}^{(L+n)}.$$

Proof. Because each X_s is a measurable function of $(Y_s, \dots, Y_{s-m})$, we have $\mathcal{F}_{X,t} \subseteq \mathcal{F}_{Y,t}$. Because each Y_s is a measurable function of $(X_s, \dots, X_{s-n})$, the reverse inclusion also holds, so $\mathcal{F}_{X,t} = \mathcal{F}_{Y,t}$.

For the finite-context inclusions, fix $L \ge 1$. Each variable X_{t-j} with $0 \le j \le L-1$ is measurable with respect to

$$\sigma(Y_{t-j}, Y_{t-j-1}, \dots, Y_{t-j-m}) \subseteq \mathcal{F}_{Y,t}^{(L+m)},$$

because the smallest index appearing is $t-(L-1)-m = t-L-m+1$. Hence $\mathcal{F}_{X,t}^{(L)} \subseteq \mathcal{F}_{Y,t}^{(L+m)}$. The second inclusion is identical after swapping the roles of X and Y and replacing m by n. □

Proposition 14.3 (Finite-memory predictor transport)**.** *Let $\mathcal{H} \subseteq \mathcal{F}_{Y,t} = \mathcal{F}_{X,t}$ be a sub-sigma-algebra.*

(i) *If $Y_t, \dots, Y_{t-m+1}$ are $\mathcal{H}$-measurable, then*

$$\mathbb{E}[X_{t+1} \mid \mathcal{H}] = S_0\mathbb{E}[Y_{t+1} \mid \mathcal{H}] + \sum_{k=1}^{m} S_k Y_{t+1-k},$$

and consequently

$$M_{X,t}^\star - \mathbb{E}[X_{t+1} \mid \mathcal{H}] = S_0\big(M_{Y,t}^\star - \mathbb{E}[Y_{t+1} \mid \mathcal{H}]\big).$$

(ii) *If $X_t, \dots, X_{t-n+1}$ are $\mathcal{H}$-measurable, then*

$$\mathbb{E}[Y_{t+1} \mid \mathcal{H}] = T_0\mathbb{E}[X_{t+1} \mid \mathcal{H}] + \sum_{\ell=1}^{n} T_\ell X_{t+1-\ell},$$

and consequently

$$M_{Y,t}^\star - \mathbb{E}[Y_{t+1} \mid \mathcal{H}] = T_0\big(M_{X,t}^\star - \mathbb{E}[X_{t+1} \mid \mathcal{H}]\big).$$

Proof. For part (i), write

$$X_{t+1} = S_0 Y_{t+1} + \sum_{k=1}^{m} S_k Y_{t+1-k}.$$

By hypothesis, the lagged terms $Y_t, \dots, Y_{t-m+1}$ are $\mathcal{H}$-measurable, so taking conditional expectation yields

$$\mathbb{E}[X_{t+1} \mid \mathcal{H}] = S_0 \mathbb{E}[Y_{t+1} \mid \mathcal{H}] + \sum_{k=1}^{m} S_k Y_{t+1-k}.$$

Taking $\mathcal{H} = \mathcal{F}_{Y,t} = \mathcal{F}_{X,t}$ gives

$$M^\star_{X,t} = S_0 M^\star_{Y,t} + \sum_{k=1}^{m} S_k Y_{t+1-k},$$

and subtraction proves the second identity.

Part (ii) is identical after swapping the roles of X and Y and replacing $(S_k)_{k=0}^m$ by $(T_\ell)_{\ell=0}^n$. □

Let

$$s_{\min}(S_0) := \sigma_{\min}(S_0), \qquad s_{\max}(S_0) := \sigma_{\max}(S_0)$$

be the smallest and largest singular values of S_0.

Theorem 14.4 (Context-defect comparison under finite-memory dynamic mixing)**.** *For every $L \geq 1$,*

$$s_{\min}(S_0)^2 \, \mathcal{E}^Y_{\mathrm{ctx}}(L+m) \leq \mathcal{E}^X_{\mathrm{ctx}}(L).$$

If moreover $L \geq m+n$, then

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \leq s_{\max}(S_0)^2 \, \mathcal{E}^Y_{\mathrm{ctx}}(L-n).$$

Therefore the context defects of X and Y have the same asymptotic decay exponent, up to a finite shift of the context length.

Proof. Fix $L \geq 1$. By Proposition 14.2,

$$\mathcal{F}^{(L)}_{X,t} \subseteq \mathcal{F}^{(L+m)}_{Y,t} \subseteq \mathcal{F}_{Y,t} = \mathcal{F}_{X,t}.$$

Hence Lemma 14.1 gives

$$\mathcal{E}^X_{\mathrm{ctx}}(L) = \mathbb{E}\|M^\star_{X,t} - \mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L)}_{X,t}]\|^2 \geq \mathbb{E}\|M^\star_{X,t} - \mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L+m)}_{Y,t}]\|^2.$$

Because $L+m \geq m$, the sigma-algebra $\mathcal{F}^{(L+m)}_{Y,t}$ contains $Y_t, \dots, Y_{t-m+1}$. Proposition 14.3(i) therefore yields

$$M^\star_{X,t} - \mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L+m)}_{Y,t}] = S_0\big(M^\star_{Y,t} - M^{(L+m)}_{Y,t}\big).$$

Taking norms and expectations gives

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \geq \mathbb{E}\big\|S_0(M^\star_{Y,t} - M^{(L+m)}_{Y,t})\big\|^2 \geq s_{\min}(S_0)^2 \, \mathcal{E}^Y_{\mathrm{ctx}}(L+m),$$

which proves the lower bound.

Now assume $L \geq m+n$. Then $L-n \geq m$ and Proposition 14.2 gives

$$\mathcal{F}^{(L-n)}_{Y,t} \subseteq \mathcal{F}^{(L)}_{X,t}.$$

Lemma 14.1 therefore gives

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \leq \mathbb{E}\|M^\star_{X,t} - \mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L-n)}_{Y,t}]\|^2.$$

Since $L-n \geq m$, Proposition 14.3(i) again applies and yields

$$M^\star_{X,t} - \mathbb{E}[X_{t+1} \mid \mathcal{F}^{(L-n)}_{Y,t}] = S_0\big(M^\star_{Y,t} - M^{(L-n)}_{Y,t}\big).$$

Hence

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \le \mathbb{E}\big\|S_0(M^\star_{Y,t} - M^{(L-n)}_{Y,t})\big\|^2 \le s_{\max}(S_0)^2\, \mathcal{E}^Y_{\mathrm{ctx}}(L-n),$$

as claimed. □

Corollary 14.5 (Transfer of polynomial, exponential, and FARIMA context laws)**.** *Assume the hypotheses of Theorem 14.4.*

(i) *If* $\mathcal{E}^Y_{\mathrm{ctx}}(L) \asymp L^{-\beta}$ *for some* $\beta > 0$*, then*

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \asymp L^{-\beta}.$$

More precisely, for all sufficiently large L*,*

$$s_{\min}(S_0)^2 c_-(L+m)^{-\beta} \le \mathcal{E}^X_{\mathrm{ctx}}(L) \le s_{\max}(S_0)^2 c_+(L-n)^{-\beta}$$

whenever $c_- L^{-\beta} \le \mathcal{E}^Y_{\mathrm{ctx}}(L) \le c_+ L^{-\beta}$.

(ii) *If* $\mathcal{E}^Y_{\mathrm{ctx}}(L) \asymp e^{-aL}$ *for some* $a > 0$*, then*

$$\mathcal{E}^X_{\mathrm{ctx}}(L) \asymp e^{-aL}.$$

(iii) *If* $\mathcal{E}^Y_{\mathrm{ctx}}(L) = CL^{-1} + O(L^{-2})$*, then*

$$s_{\min}(S_0)^2 CL^{-1} + O(L^{-2}) \le \mathcal{E}^X_{\mathrm{ctx}}(L) \le s_{\max}(S_0)^2 CL^{-1} + O(L^{-2}).$$

In particular, if S_0 *is orthogonal, then*

$$\mathcal{E}^X_{\mathrm{ctx}}(L) = CL^{-1} + O(L^{-2}).$$

Proof. All three claims are immediate from Theorem 14.4, since replacing L by $L \pm c$ with fixed c does not change the leading decay exponent, and in part (iii) one has

$$(L \pm c)^{-1} = L^{-1} + O(L^{-2}).$$

If S_0 is orthogonal, then $s_{\min}(S_0) = s_{\max}(S_0) = 1$, so the lower and upper bounds in part (iii) have the same leading term. □

14.1. **State-dimension transport with finite overhead.** We now transport full MIMO state laws. Set

$$M := m + n.$$

The key point is that the finite-memory inverse filter allows us to recover all lagged variables needed for the dynamic mixing from a block shift register of length M.

Theorem 14.6 (Finite-overhead transport of MIMO state laws)**.** *For every* $r \ge 0$*,*

$$\mathcal{E}^{\mathrm{MIMO}}_{r+qM}(X) \le s_{\max}(S_0)^2\, \mathcal{E}^{\mathrm{MIMO}}_r(Y).$$

By symmetry,

$$\mathcal{E}^{\mathrm{MIMO}}_{r+qM}(Y) \le s_{\max}(T_0)^2\, \mathcal{E}^{\mathrm{MIMO}}_r(X) = \frac{1}{s_{\min}(S_0)^2}\, \mathcal{E}^{\mathrm{MIMO}}_r(X).$$

Consequently, for every $r \ge qM$*,*

$$s_{\min}(S_0)^2\, \mathcal{E}^{\mathrm{MIMO}}_{r+qM}(Y) \le \mathcal{E}^{\mathrm{MIMO}}_r(X) \le s_{\max}(S_0)^2\, \mathcal{E}^{\mathrm{MIMO}}_{r-qM}(Y).$$

Proof. Fix $r \ge 0$ and let

$$s_t = As_{t-1} + BY_t, \qquad \widehat{Y}_{t+1} = Cs_t,$$

be any stable r-state MIMO predictor for Y. We construct a stable $(r+qM)$-state predictor for X.

Let

$$\nu_t := \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-M+1} \end{bmatrix} \in \mathbb{R}^{qM}$$

be the M-block shift-register state driven by X. Thus there exist fixed matrices $J \in \mathbb{R}^{qM\times qM}$ and $K \in \mathbb{R}^{qM\times q}$, with J nilpotent, such that

$$\nu_t = J\nu_{t-1} + KX_t.$$

Because $M = m + n$, for each $k = 0, 1, \dots, m$ the variable

$$Y_{t+1-k} = \sum_{\ell=0}^{n} T_\ell X_{t+1-k-\ell}$$

is a linear function of ν_t. Hence there exist matrices $L_k \in \mathbb{R}^{q\times qM}$ such that

$$Y_{t+1-k} = L_k\nu_t, \qquad k = 0, 1, \dots, m.$$

In particular, $Y_t = L_0\nu_t$.

Now define the augmented predictor state

$$\xi_t := \begin{bmatrix} s_t \\ \nu_t \end{bmatrix} \in \mathbb{R}^{r+qM}.$$

Since $\nu_t = J\nu_{t-1} + KX_t$, the original Y-predictor recursion becomes

$$s_t = As_{t-1} + BL_0\nu_t = As_{t-1} + BL_0J\nu_{t-1} + BL_0KX_t.$$

Therefore ξ_t satisfies a linear state recursion

$$\xi_t = \widetilde{A}\,\xi_{t-1} + \widetilde{B}\,X_t$$

with

$$\widetilde{A} = \begin{bmatrix} A & BL_0J \\ 0 & J \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} BL_0K \\ K \end{bmatrix}.$$

The matrix $\widetilde{A}$ is stable because it is block upper triangular with diagonal blocks A and the nilpotent shift-register matrix J. Define the predictor output by

$$\widehat{X}_{t+1} := S_0Cs_t + \sum_{k=1}^{m} S_kL_k\nu_t.$$

Since $\widehat{Y}_{t+1} = Cs_t$ and $L_k\nu_t = Y_{t+1-k}$ for $1 \le k \le m$, we have

$$\widehat{X}_{t+1} = S_0\widehat{Y}_{t+1} + \sum_{k=1}^{m} S_kY_{t+1-k}.$$

By Proposition 14.3(i) applied with $\mathcal{H} = \mathcal{F}_{X,t} = \mathcal{F}_{Y,t}$,

$$M^\star_{X,t} = S_0M^\star_{Y,t} + \sum_{k=1}^{m} S_kY_{t+1-k}.$$

Subtracting gives

$$M^\star_{X,t} - \widehat{X}_{t+1} = S_0(M^\star_{Y,t} - \widehat{Y}_{t+1}).$$

Therefore

$$\mathbb{E}\|M^\star_{X,t} - \widehat{X}_{t+1}\|^2 \le s_{\max}(S_0)^2\,\mathbb{E}\|M^\star_{Y,t} - \widehat{Y}_{t+1}\|^2.$$

Taking the infimum over all stable r-state predictors for Y proves

$$\mathcal{E}^{\mathrm{MIMO}}_{r+qM}(X) \le s_{\max}(S_0)^2\, \mathcal{E}^{\mathrm{MIMO}}_{r}(Y).$$

The second displayed inequality follows by swapping the roles of X and Y and replacing (S_k) by (T_ℓ). Since $T_0 = S_0^{-1}$, one has $s_{\max}(T_0) = 1/s_{\min}(S_0)$. Finally, replacing r by $r - qM$ in the first inequality yields the upper bound in the last display for $r \ge qM$, while the lower bound is just a rearrangement of the symmetric inequality. □

Corollary 14.7 (Transfer of state exponents under finite-memory dynamic mixing)**.** *Assume the hypotheses of Theorem 14.6.*

(i) *If* $\mathcal{E}^{\mathrm{MIMO}}_{r}(Y) \asymp r^{-\beta}$ *for some* $\beta > 0$*, then*

$$\mathcal{E}^{\mathrm{MIMO}}_{r}(X) \asymp r^{-\beta}.$$

(ii) *If* $\mathcal{E}^{\mathrm{MIMO}}_{r}(Y) \asymp e^{-ar}$ *for some* $a > 0$*, then*

$$\mathcal{E}^{\mathrm{MIMO}}_{r}(X) \asymp e^{-ar}.$$

(iii) *The same conclusions hold for the exact shift-register class, because*

$$\mathcal{E}^{\mathrm{shift}}_{qL}(X) = \mathcal{E}^{X}_{\mathrm{ctx}}(L)$$

by Theorem 12.1.

Proof. An additive shift of the state dimension does not change either a polynomial decay exponent or an exponential decay exponent. Hence parts (i) and (ii) follow immediately from Theorem 14.6. Part (iii) is the exact shift-register law together with Corollary 14.5. □

Remark 14.8. Section 14 is the first point in the draft where nonconstant frequency coupling is handled rigorously. The price of passing from Y to X is no longer a mere condition-number distortion, as in Section 13; one must also pay a finite context shift and a finite state-dimension overhead reflecting the filter memory. But the core exponents—polynomial or exponential—survive unchanged. This is precisely the kind of robustness one would want in a prediction-theoretic theory of learned sequence models.

15. Next Technical Targets

The present draft now contains eleven quantitative backbones:

(1) a scalar context-length backbone, in which $\mathcal{E}_{\mathrm{ctx}}(L)$ is reduced to finite-prediction error variances and PACF tails;
(2) a scalar phase/PACF backbone, in which phase coefficients, AR tails, and long-memory asymptotics are converted into explicit context laws;
(3) a scalar linear hidden-state backbone, in which $\mathcal{E}^{\mathrm{LTI}}_{d}$ is reduced to stable weighted rational approximation of the canonical predictor symbol;
(4) a pole-compressibility backbone, in which explicit pole expansions of P imply explicit scalar state-dimension laws;
(5) a multivariate block-prediction backbone, in which $\mathcal{E}_{\mathrm{ctx}}(L)$ is identified with the trace of a positive semidefinite defect covariance $V_L - V_\infty$;
(6) a multivariate PACF/phase backbone, in which $\Delta_{\mathrm{ctx}}(L)$ is reduced to a Loewner-order tail sum of canonical PACF matrices and the phase matrix/IPF machinery is connected to concrete context laws;
(7) a multivariate hidden-state backbone, in which $\mathcal{E}^{\mathrm{MIMO}}_{r}$ is reduced to matrix-weighted rational approximation and then bounded by either block context laws or Hankel-singular-value tails;

(8) a multivariate lower-bound backbone, in which finite block Hankel sections and realization complexity produce unavoidable state-defect lower bounds and exact zero-defect thresholds;
(9) a structured exact-law backbone, in which block shift registers and diagonal memory allocation admit matched two-sided formulas;
(10) a constant-mixing backbone, in which context and state laws are shown to be invariant or quasi-invariant under invertible channel mixing;
(11) a dynamic-mixing backbone, in which frequency-dependent bi-causal finite-state couplings preserve context laws up to finite shifts and state laws up to finite additive overheads.

The next mathematically natural steps are:

(1) move beyond the bi-causal transducer classes of Sections 13 and 14 to match lower and upper bounds for genuinely frequency-coupled multivariate predictor symbols that are not obtained by finite-state conjugacy from simpler models;
(2) formulate matrix-valued analogues of the scalar pole-tail transfer principles that compare phase-based, lag-based, and Hankel-based compressibility mechanisms;
(3) sharpen the lower-bound machinery into asymptotically matched two-sided laws on concrete multivariate symbol classes;
(4) pass from oracle/model-based linear states to observation-driven learned state-space models;
(5) formulate and analyze innovation-aware training objectives using prediction defect and state defect as geometric regularizers.

These are the points at which the present framework should reconnect both to classical multivariate prediction theory and to modern state-space architectures in machine learning.

References

[1] N. H. Bingham, *Prediction theory*, arXiv:2111.08560, 2021.
[2] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, *HiPPO: Recurrent memory with optimal polynomial projections*, NeurIPS, 2020.
[3] A. Gu, K. Goel, and C. Ré, *Efficiently modeling long sequences with structured state spaces*, ICLR, 2022.
[4] A. Gu and T. Dao, *Mamba: Linear-time sequence modeling with selective state spaces*, arXiv:2312.00752, 2023.
[5] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. G. van Sloun, and Y. C. Eldar, *KalmanNet: Neural network aided Kalman filtering for partially known dynamics*, IEEE Transactions on Signal Processing 70 (2022), 1532–1547.
[6] M. Pourahmadi, *On the convergence of finite linear predictors of stationary processes*, Journal of Multivariate Analysis 30 (1989), 90–103.
[7] A. Poltoratski, *The type problem for measures*, Annals of Mathematics 178 (2013), 1043–1110.
[8] A. Inoue, Y. Kasahara, and M. Pourahmadi, *The intersection of past and future for multivariate stationary processes*, Proceedings of the American Mathematical Society 144 (2016), 1779–1786.
[9] A. Inoue, Y. Kasahara, and M. Pourahmadi, *Baxter's inequality for finite predictor coefficients of multivariate long-memory stationary processes*, Bernoulli 24 (2018), no. 2, 1202–1232.
[10] A. Inoue, *Explicit formulas for the inverses of Toeplitz matrices, with applications*, arXiv:2105.01165, 2022.
[11] N. H. Bingham, A. Inoue, and Y. Kasahara, *An explicit representation of Verblunsky coefficients*, Statistics & Probability Letters 82 (2012), no. 2, 403–410.
[12] A. Inoue, *AR and MA representation of partial autocorrelation functions, with applications*, Probability Theory and Related Fields 140 (2008), 523–551.
[13] S. Dégérine, *Canonical partial autocorrelation function of a multivariate time series*, Annals of Statistics 18 (1990), no. 2, 961–971.
[14] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Transactions on Automatic Control 26 (1981), no. 1, 17–32.

[15] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, International Journal of Control 39 (1984), no. 6, 1115–1193.
[16] M. Forgione, M. Mejari, and D. Piga, *Model order reduction of deep structured state-space models: A system-theoretic approach*, arXiv:2403.14833, 2024.
[17] P. Schwerdtner, J. Berman, and B. Peherstorfer, *Hankel singular value regularization for highly compressible state space models*, NeurIPS 2025.