Canonical Predictive Quotients for Prediction under Hidden-State Uncertainty

Hiroyuki Kasai
Waseda University, Tokyo, Japan
hiroyuki.kasai@waseda.jp

Abstract

Bayes-optimal prediction does not always require knowing the full hidden state. We ask which differences between hidden states must be retained to achieve the same risk as an oracle that knows the full state. We study this question when the predictor already observes an instance. We introduce the canonical predictive quotient (CPQ), which groups hidden states whose selected prediction rules agree for almost every observable instance. Under suitable conditions, the CPQ preserves oracle risk. However, the state differences captured by the CPQ do not alone determine how much prediction improves. We show that the same grouping of states can lead to different oracle gains, including zero. Under additional assumptions, we also show that recovering the CPQ is necessary for preserving oracle risk. We then study whether information sufficient at one stage remains sufficient as observations accumulate. Under squared loss and query-linear responses, we characterize the query directions for which a summary is sufficient. We also give exact conditions for later sufficiency and quantify the resulting excess risk. In linear-Gaussian models, two explicit conditions characterize one-step persistence.

1 Introduction

Optimal prediction does not always require knowing the full hidden state. Two hidden states may give the same Bayes-optimal prediction for one observable instance but different predictions for another. We ask which state differences are needed for optimal prediction and when knowing them reduces risk. These differences may change as observations accumulate, even for a fixed query.

Bayesian decision theory [6] and the comparison of experiments [5, 4] study the value of information. Classical prediction-sufficiency theory studies what information can be removed without changing predictive risk [14, 15]. Bayes-sufficient representation theory studies representations that preserve Bayes-optimal decisions and Bayes quotients under suitable conditions [12]. In our setting, the predictor already has an observable instance. We ask what it still needs to know about the hidden state to achieve the same risk as a full-state oracle. We study this additional information through a state summary that depends only on the hidden state.

We first ask which differences between hidden states are needed for prediction. We compare a selected family of prediction rules that is optimal under the realized joint law. The canonical predictive quotient (CPQ) groups states whose selected rules agree for almost every observable instance. When this grouping can be represented exactly and used with the observable instance to recover the selected oracle act, the CPQ preserves oracle risk. However, a summary may preserve oracle risk without recovering the CPQ. Under additional uniqueness and domination assumptions, recovering the CPQ is also necessary. We next ask when this state information improves prediction. We show that the same grouping of hidden states can give different oracle gains, including zero, even when the state and observable-instance marginals and response kernel remain the same. Thus, knowing

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

which states are grouped together does not tell us how much full-state information improves prediction. We also give conditions for when missing state information leads to positive excess risk and when full-state information gives a positive gain. We also study how sufficiency changes as observations accumulate. A state summary may be sufficient at one stage but not at a later stage. Motivated by in-context learning (ICL) [18, 11, 19] and sequential prediction, we study squared loss and responses linear in the query. We determine the query directions for which a fixed state summary is sufficient. We also give an exact condition for sufficiency to remain valid and measure the excess risk when it is lost. For linear-Gaussian models, we give two explicit coefficient conditions for one-step persistence.

Our results concern the state information needed for Bayes-optimal prediction. They do not show whether a trained predictor can represent or use this information. Previous work studies transfer between loss functions under a fixed data law [13] and prediction error in in-context learning [8, 16]. Our study differs in its focus on state information that preserves oracle risk and whether its sufficiency remains valid as observations accumulate. Detailed comparisons appear in Section A.

2 Prediction Problem and Bayes Predictors

We specify the prediction problem and the information available to competing predictors. Their risk ordering provides a baseline for assessing the value of additional state information.

2.1 Prediction Problem and Hidden-State Information

We define the observable instance, hidden state, and loss, together with the regularity needed for conditional-risk comparisons. All random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Write $\mathrm{Law}(X)$ for the law of X and $\mathrm{Law}(X|H)$ for a regular conditional law.

Definition 2.1 (Prediction problem)**. (i) Random elements.** Let ν denote the *observable instance*, taking values in the instance space $\mathcal{O}$; let Z take values in the *hidden state* space $\mathcal{Z}$; and let y_{q} be the *query response*, taking values in the response space $\mathcal{Y}$. In the specializations below, the query is encoded in the observable instance (fixed in Section 4 and random in Section F); no separate query variable is required in this general definition.

(ii) Joint law and conditional response kernel. The prediction problem specifies a joint probability law for (ν, Z, y_{q}) on $\mathcal{O}\times\mathcal{Z}\times\mathcal{Y}$ and a *prescribed conditional response kernel* $\mathsf{K}(\cdot \mid v, z)$ for every $(v, z) \in \mathcal{O} \times \mathcal{Z}$. For each pair (v, z), this is a probability measure on $\mathcal{Y}$; for every measurable $B \subseteq \mathcal{Y}$, the map $(v, z) \mapsto \mathsf{K}(B \mid v, z)$ is measurable. The prescribed conditional response kernel specifies the response law at each instance–state pair. Its consistency with the realized conditional law is required in Assumption 2.2. On $\mathrm{Law}(\nu, Z)$-null sets, the values of K are prescribed rather than determined by the joint law.

(iii) Loss and risk. Let $\mathcal{A}$ denote the action space. The *loss* is a measurable map $L : \mathcal{A}\times\mathcal{Y} \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R}\cup\{-\infty, +\infty\}$; an *action* $a \in \mathcal{A}$ incurs loss $L(a, y_{\mathrm{q}})$. For any measurable $\mathcal{A}$-valued random action A' with well-defined extended-real expected loss (possibly infinite, with no indeterminate $+\infty - \infty$), define its *risk* by $R[A'] := \mathbb{E}[L(A', y_{\mathrm{q}})]$. The *state-conditional risk* is $\rho(a; v, z) := \int L(a, y)\mathsf{K}(dy \mid v, z)$ whenever the integral is well defined.

(iv) State summary and predictors. A *state summary* is $S := f(Z)$ for a fixed measurable *state-only* map $f : \mathcal{Z} \to \mathcal{S}$. The map f takes the hidden state as its sole argument: it does not use the values of ν or y_{q}. Thus S is hypothetical supplementary information about the hidden state, not a representation computed from ν. Neither injectivity of f nor preservation of the risk attainable from (ν, Z) by S is assumed; which state distinctions matter for prediction is studied below. For a random element X, write $\sigma(X)$ for the sigma-field generated by X; for random elements X, Y, write $\sigma(X, Y) := \sigma((X, Y))$. Because $S = f(Z)$ is determined by Z, $\sigma(\nu, S) \subseteq \sigma(\nu, Z)$: observing the summary alongside ν cannot reveal more than observing the full state. Predictors may use ν alone, (ν, S), or (ν, Z). For measurable predictors φ, write $R[\varphi(\nu)]$, $R[\varphi(\nu, S)]$, and $R[\varphi(\nu, Z)]$. Lowercase v, z, s, y denote generic values of $\nu, Z, S, y_{\mathrm{q}}$.

Assumption 2.2 (Standard-Borel regularity and kernel consistency)**.** The instance, state, action, and response spaces, as well as the value space of each state summary $S = f(Z)$, are standard Borel. The prescribed conditional response kernel $\mathsf{K}(dy \mid v, z)$ from Definition 2.1 satisfies *kernel consistency*: $\mathsf{K}(\cdot \mid \nu, Z) = \mathrm{Law}(y_{\mathrm{q}} \mid \nu, Z)$ almost surely, as an identity of random probability measures.

Measure and conditioning conventions. Let $\mu \coloneqq \mathrm{Law}(Z)$ be the *hidden-state law*, $\lambda \coloneqq \mathrm{Law}(\nu)$ the *observable-instance law*, $\pi \coloneqq \mathrm{Law}(\nu, Z)$ the *realized joint law*, and $\pi_\otimes \coloneqq \lambda \otimes \mu$ the *product law*, formed by pairing the marginals independently. The statewise comparisons below concern entire prediction rules across observable instances up to λ-almost-sure equality, not merely at realized instance–state pairs. For measures ξ_1, ξ_2 on the same measurable space, $\xi_1 \ll \xi_2$ means absolute continuity (every ξ_2-null measurable set is ξ_1-null). The directions $\pi_\otimes \ll \pi$ and $\pi \ll \pi_\otimes$ are distinct; the former is called *product-to-joint domination*. We write $\mathbb{P}$ for probability under the realized law on all random elements, whose (ν, Z)-marginal is π; unqualified "almost surely" refers to this law or its relevant marginal. When well defined, expressions such as $\mathbb{E}[H \mid \nu, S]$ and $\mathrm{Law}(H \mid \nu, S)$ condition on random elements, whereas $\mathbb{E}[H \mid \nu = v, S = s]$ and $\mathrm{Law}(H \mid \nu = v, S = s)$ evaluate chosen conditional expectation functions or kernels at (v, s). Different choices may disagree on $\mathrm{Law}(\nu, S)$-null sets. An equality such as $Z = k$ indicates a fixed conditioning value; all other conditioning variables remain random.

Lemma 2.3 (Conditional-mixture and conditional-risk identities)**.** *Under Assumption 2.2, for every state summary $S = f(Z)$, the following hold.*

(i) Conditional mixture. *For every measurable $B \subseteq \mathcal{Y}$,*

$$\mathbb{P}(y_{\mathrm{q}} \in B \mid \nu, S) = \mathbb{E}[\mathsf{K}(B \mid \nu, Z) \mid \nu, S] = \int_{\mathcal{Z}} \mathsf{K}(B \mid \nu, z)\, \mathbb{P}(Z \in dz \mid \nu, S). \tag{1}$$

This holds almost surely *as an identity of random probability measures.*

(ii) Conditional risk. *For every $\sigma(\nu, Z)$-measurable $\mathcal{A}$-valued random action A' and measurable predictor φ based on (ν, S),*

$$\mathbb{E}[L(A', y_{\mathrm{q}}) \mid \nu, Z] = \rho(A'; \nu, Z), \tag{2}$$

$$\mathbb{E}[L(\varphi(\nu, S), y_{\mathrm{q}}) \mid \nu, S] = \mathbb{E}[\rho(\varphi(\nu, S); \nu, Z) \mid \nu, S], \tag{3}$$

almost surely*, whenever the conditional expectations are defined. Hence $R[\varphi(\nu, S)] = \mathbb{E}[\rho(\varphi(\nu, S); \nu, Z)]$ whenever this risk is defined. The identities also apply to $S = Z$.*

The mixture in (i) uses $\mathrm{Law}(Z \mid \nu, S)$ without assuming independence of ν and Z. The risk identities in (ii) require no minimization and are used in Sections 2.2 and 3.1. The proof is in Section B.1.

2.2 Bayes Predictors and the Value of State Information

Assumption 2.4 (Existence and measurability of Bayes predictors)**.** For each state summary $S = f(Z)$ considered in a stated result, measurable Bayes predictors $\hat{y}_{\mathrm{obs}}(\nu)$, $\hat{y}_S(\nu, S)$, and $\hat{y}_{\mathrm{ora}}(\nu, Z)$ exist and minimize their respective conditional risks almost surely (Definitions 2.5–2.8). Their risks are finite. All conditional risks and risk comparisons are assumed to be well defined.

Definition 2.5 (Conditional Bayes optimality)**.** For any available-information random element $W \in \{\nu, (\nu, S), (\nu, Z)\}$, where $S = f(Z)$, the quantity $\mathbb{E}[L(a, y_{\mathrm{q}}) \mid W]$ is the *conditional risk* of an action a, and its minimizers are *Bayes acts* given W. For $W = (\nu, Z)$, this conditional risk equals $\rho(a; \nu, Z)$ almost surely, whenever defined, by (2). All actions are compared using conditional-risk versions induced by a single regular conditional law of y_{q} given W, consistently with Assumption 2.2. A measurable map φ defined on the value space of W is *conditionally Bayes-optimal* given W, and is then called a *Bayes predictor* given W, if

$$\varphi(W) \in \operatorname*{arg\,min}_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid W] \quad \text{almost surely.} \tag{4}$$

The *Bayes risk* given W is $R[\varphi(W)]$ for any Bayes predictor φ given W; it does not depend on the choice of φ.

Definition 2.6 (Observable-instance Bayes predictor)**.** The *observable-instance Bayes predictor* $\hat{y}_{\mathrm{obs}}(\nu)$ is conditionally Bayes-optimal given $W = \nu$. Its Bayes risk is $R_{\mathrm{obs}} \coloneqq R[\hat{y}_{\mathrm{obs}}(\nu)]$.

Definition 2.7 (State-summary Bayes predictor)**.** For a state summary $S = f(Z)$, the *state-summary Bayes predictor* $\hat{y}_S(\nu, S)$ is conditionally Bayes-optimal given $W = (\nu, S)$. Its Bayes risk is $R_S \coloneqq R[\hat{y}_S(\nu, S)]$.

Definition 2.8 (Oracle Bayes predictor)**.** The *oracle Bayes predictor* $\hat{y}_{\mathrm{ora}}(\nu, Z)$ is conditionally Bayes-optimal given $W = (\nu, Z)$. Its Bayes risk is $R_{\mathrm{ora}} \coloneqq R[\hat{y}_{\mathrm{ora}}(\nu, Z)]$.

Definition 2.9 (Bayes risk gain from state information)**.** For every state summary $S = f(Z)$, define the *Bayes risk gain functional* by $G(S) := R_{\mathrm{obs}} - R_S$. In particular, the *identity summary* $S = Z$ gives $R_Z = R_{\mathrm{ora}}$ and $G(Z) = R_{\mathrm{obs}} - R_{\mathrm{ora}}$ (the *oracle gain*), whereas the *constant summary* S gives $R_S = R_{\mathrm{obs}}$ and $G(S) = 0$. Here $G(Z)$ is the expected risk improvement from having access to the full hidden state Z, not a gain evaluated at a realization z. We call $R_S - R_{\mathrm{ora}} = G(Z) - G(S)$ the *excess risk* of the state summary S relative to the oracle. We say that S *preserves oracle risk* (*oracle-risk preservation*) if $R_S = R_{\mathrm{ora}}$, equivalently if its excess risk is zero or $G(S) = G(Z)$.

We say that S_2 *refines* S_1 if $S_1 = h(S_2)$ almost surely for some measurable h. Lemma 2.3(ii) expresses $R[\varphi(\nu, S)]$ as $\mathbb{E}[\rho(\varphi(\nu, S); \nu, Z)]$ without comparing $R_{\mathrm{ora}}, R_S, R_{\mathrm{obs}}$. The next lemma establishes $R_{\mathrm{ora}} \le R_S \le R_{\mathrm{obs}}$ and $R_{S_2} \le R_{S_1}$ whenever $S_1 = h(S_2)$ almost surely.

Lemma 2.10 (Conditional Bayes-risk ordering and refinement monotonicity)**.** *Under Assumptions 2.2 and 2.4, the following hold.*

(i) Conditional Bayes-risk ordering. *For every state summary $S = f(Z)$,*

$$\begin{aligned}\mathbb{E}[\rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z) \mid \nu, S] &\le \inf_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, S] \\ &= \mathbb{E}[\rho(\hat{y}_S(\nu, S); \nu, Z) \mid \nu, S] \le \mathbb{E}[\rho(\hat{y}_{\mathrm{obs}}(\nu); \nu, Z) \mid \nu, S].\end{aligned} \tag{5}$$

These inequalities hold $\mathrm{Law}(\nu, S)$-almost surely. *Consequently,*

$$R_{\mathrm{ora}} \le R_S \le R_{\mathrm{obs}}, \qquad 0 \le G(S) \le G(Z). \tag{6}$$

Moreover, every ν-measurable predictor $\varphi(\nu)$ with well-defined risk satisfies $R[\varphi(\nu)] \ge R_{\mathrm{obs}}$.

(ii) Refinement monotonicity. *For any state summaries S_1, S_2 such that S_2 refines S_1, $R_{S_2} \le R_{S_1}$.*

(2)–(3) and conditional Bayes optimality give (5); taking expectations yields $R_{\mathrm{ora}} \le R_S \le R_{\mathrm{obs}}$, hence $0 \le G(S) \le G(Z)$. Conditional Bayes optimality given ν gives $R[\varphi(\nu)] \ge R_{\mathrm{obs}}$; that given (ν, S_2) gives $R_{S_2} \le R_{S_1}$ when $S_1 = h(S_2)$ almost surely. The proof is in Section B.2.

3 Canonical Predictive Quotients and the Value of State Distinctions

We begin with a finite example illustrating which hidden-state distinctions matter for optimal prediction and how the oracle gain $G(Z)$ depends on the realized joint law. We then formalize these distinctions through a measurable state-only encoding and characterize their effect on Bayes risk.

Example 3.1 (State distinctions and optimal prediction)**.**

Let the hidden state take four values z_1, z_2, z_3, z_4, each with positive probability, and let the observable instance take values v_A, v_B, each with positive marginal probability. Take $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ and squared loss $L(a, y) = (a - y)^2$ for $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Suppose $\int_{\mathbb{R}} y^2 \, \mathsf{K}(dy \mid v, z) < \infty$ for every $(v, z) \in \mathcal{O} \times \mathcal{Z}$. The unique Bayes acts $\int_{\mathbb{R}} y \, \mathsf{K}(dy \mid v, z)$ (cf. Lemma C.4) are given in Table 1.

Table 1: Values of $\int_{\mathbb{R}} y \, \mathsf{K}(dy \mid v, z)$.

Hidden state	$\nu = v_A$	$\nu = v_B$
z_1	0	0
z_2	0	0
z_3	0	1
z_4	1	1

Although z_1 and z_2 are distinct hidden states, they require identical optimal predictions at both instances. State z_3 agrees with them at v_A but not at v_B; similarly, z_4 agrees with z_3 at v_B but not at v_A. Thus agreement at a single observable instance need not imply agreement across instances. Which hidden-state distinctions must be retained to preserve optimal prediction across observable instances? The rows of Table 1 identify these distinctions. Whether retaining them strictly reduces Bayes risk relative to using only ν (that is, $G(Z) > 0$) depends on the realized joint law of (ν, Z), as illustrated in Example 3.13.

3.1 Predictive Equivalence, CPQ Definition, and Oracle-Risk Preservation

We first specify state-dependent prediction rules and then compare them across hidden states.

Definition 3.2 (Prediction-rule family and π-optimality)**.** A *prediction-rule family* is a jointly measurable map $\alpha : \mathcal{O} \times \mathcal{Z} \to \mathcal{A}$. For each fixed state z, its *prediction-rule profile* $\alpha(\cdot, z)$ is the entire

prediction rule $v \mapsto \alpha(v, z)$ over all observable instances $v \in \mathcal{O}$. Recall that $\pi = \mathrm{Law}(\nu, Z)$ is the realized joint law of the observable instance and hidden state. The family is called π-*optimal* if

$$\alpha(v, z) \in \arg\min_{a \in \mathcal{A}} \rho(a; v, z) \quad \text{for } \pi\text{-almost every } (v, z). \tag{7}$$

Equivalently, by kernel consistency (Assumption 2.2) and the single regular conditional law specified in Definition 2.5,

$$\alpha(\nu, Z) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, Z] \quad \text{almost surely.} \tag{8}$$

Elements of $\arg\min_{a \in \mathcal{A}} \rho(a; v, z)$ are called *state-conditional Bayes acts* at (v, z).

For a selected π-optimal family α, its *selected oracle act* is $\alpha(\nu, Z)$.

Lemma 3.3 (Optimality and risk equivalence of selected oracle acts)**.** *Under Assumptions 2.2 and 2.4, the oracle Bayes predictor $\hat{y}_{\mathrm{ora}}$ of Definition 2.8 is a jointly measurable π-optimal family with finite risk. For every jointly measurable π-optimal family α with finite risk, the following hold: (i) $\alpha(\nu, Z)$ and $\hat{y}_{\mathrm{ora}}(\nu, Z)$ are state-conditional Bayes acts* almost surely*; (ii) their state-conditional risks agree, $\rho(\alpha(\nu, Z); \nu, Z) = \rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z)$* almost surely*; and (iii) their risks agree, $R[\alpha(\nu, Z)] = R_{\mathrm{ora}}$. The selected oracle acts need not coincide.*

The proof is in Section C.1.

Throughout this section, fix one jointly measurable π-optimal family α with finite risk.

Definition 3.4 (Predictive equivalence)**.** For the fixed family α, recall that $\lambda = \mathrm{Law}(\nu)$ is the observable instance law. Two hidden states $z, z' \in \mathcal{Z}$ are *predictively equivalent*, written $z \sim_{\mathrm{pred}} z'$, if their prediction-rule profiles $\alpha(\cdot, z)$ and $\alpha(\cdot, z')$ agree λ-almost everywhere. Equivalently, $\lambda\big(\{v \in \mathcal{O} : \alpha(v, z) \neq \alpha(v, z')\}\big) = 0$. Thus the comparison is across observable instances, not merely at a single instance, and is relative to the fixed family α.

Assumption 3.5 (Exact predictive-equivalence encoding and realized reconstruction)**.** For the fixed family α, consider a standard Borel space $\mathcal{T}^\star$ and a measurable candidate encoding map $\tau : \mathcal{Z} \to \mathcal{T}^\star$. The following two conditions are stated separately; each is imposed only when explicitly invoked.

(P) There is a measurable set $\mathcal{Z}_0 \subseteq \mathcal{Z}$ with $\mu(\mathcal{Z}_0) = 1$ such that for *every* $z, z' \in \mathcal{Z}_0$, $\tau(z) = \tau(z')$ if and only if $z \sim_{\mathrm{pred}} z'$ (as in Definition 3.4).

(R) There is a jointly measurable map $b_\pi : \mathcal{O} \times \mathcal{T}^\star \to \mathcal{A}$ such that $\alpha(\nu, Z) = b_\pi(\nu, \tau(Z))$ holds π-almost surely.

Condition (P) encodes predictive-equivalence classes exactly for every pair in a common μ-full set; only a μ-null set of states may be excluded (see Example C.1). Condition (R) reconstructs the selected oracle act measurably from $(\nu, \tau(Z))$ under the realized joint law π; b_π may depend on this law, the fixed α, and τ. Neither condition implies the other: (P) need not preserve oracle risk, while (R) may retain distinctions irrelevant to prediction (see Remark C.2). For each fixed jointly measurable α, Section C.2 constructs a measurable encoding into $[0, 1]^{\mathbb{N}}$ satisfying (P). This guarantees neither (R) for that encoding nor (P) for an arbitrary candidate τ.

Definition 3.6 (Canonical predictive quotient)**.** When the candidate encoding map τ introduced in Assumption 3.5 satisfies both (P) and (R), the random variable $T^\star := \tau(Z)$ is called a *canonical predictive quotient* (CPQ) relative to the selected prediction-rule family α. The candidate map τ is then called a *CPQ encoding map*; a value $t^\star \in \mathcal{T}^\star$ is a *CPQ label*; and the fibers of τ on the common μ-full set in (P) form the *CPQ partition*.

In Example 3.1, the predictive-equivalence classes are $\{z_1, z_2\}$, $\{z_3\}$, and $\{z_4\}$.

Remark 3.7 (Scope of canonicity)**.** For fixed α and λ, any two encodings satisfying (P) identify the same pairs of states on a common μ-full set—those whose prediction-rule profiles agree λ-almost everywhere. Thus they induce the same partition despite different labels or codomains; for CPQ encodings, this is the canonical CPQ partition. Dependence on α is suppressed in $\sim_{\mathrm{pred}}$, τ, $T^\star$, (P), and (R). This does not guarantee invariance under reselecting a π-optimal family: π-optimality need not constrain its values on π-null sets of instance–state pairs. Changing the prescribed conditional response kernel K on such sets leaves the partition unchanged for fixed α, but not necessarily after reselection. Nor is unconditional minimality among oracle-risk-preserving state summaries asserted; see Theorem 3.14.

From reconstruction to conditional Bayes optimality. Although (R) reconstructs the selected oracle act from $(\nu, T^\star)$, its optimality conditional on $(\nu, T^\star)$ rather than (ν, Z) requires verification.

Proposition 3.8 (Reconstruction of the selected oracle act, oracle-risk preservation, and refinement)**.** *Under Assumptions 2.2 and 2.4, let α be the fixed π-optimal family of Definition 3.2 with finite risk. Suppose the candidate encoding τ satisfies (P) and (R) of Assumption 3.5, and set $T^\star = \tau(Z)$. Then $T^\star$ is a CPQ relative to α, and the reconstruction map b_π in (R) satisfies*

$$\alpha(\nu, Z) = b_\pi(\nu, T^\star) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, T^\star] \quad \text{almost surely.}$$

This CPQ preserves oracle risk and oracle gain: $R_{T^\star} = R_{\mathrm{ora}}$ and $G(T^\star) = G(Z)$. Moreover, for every state summary $S = f(Z)$ that recovers *the CPQ, meaning $T^\star = g(S)$* almost surely *for some measurable g (so S refines $T^\star$), one has $R_S = R_{\mathrm{ora}}$ and $G(S) = G(Z)$. Neither uniqueness of the state-conditional Bayes act nor uniqueness of the Bayes act given $(\nu, T^\star)$ is required. If the latter is unique, every measurable Bayes predictor given $(\nu, T^\star)$ agrees with $b_\pi(\nu, T^\star)$* almost surely.

The proof is in Section C.3. Unlike (8), the optimality here is conditional on $(\nu, T^\star)$ rather than (ν, Z). It holds $\mathrm{Law}(\nu, T^\star)$-almost surely and is pulled back to π in the displayed chain.

3.2 Strictly Positive Excess Risk and Conditional Minimality

The CPQ and state summaries that recover it already preserve oracle risk by Proposition 3.8. We now determine when losing oracle-act information forces strictly larger risk, and when risk preservation requires recovering the CPQ. The strictly positive excess-risk question concerns any state summary $S = f(Z)$, not only a CPQ.

Definition 3.9 (Residual oracle-act ambiguity)**.** For a state summary $S = f(Z)$, let D_S denote the *non-Dirac region* of the conditional law of the selected oracle act $\alpha(\nu, Z)$ given (ν, S):

$$D_S := \big\{(v, s) : \mathrm{Law}(\alpha(\nu, Z) \mid \nu = v, S = s) \text{ is not a Dirac measure}\big\}.$$

We say that S leaves *residual oracle-act ambiguity* if $\mathbb{P}((\nu, S) \in D_S) > 0$. Thus D_S identifies the pairs (v, s) at which $\mathrm{Law}(\alpha(\nu, Z) \mid \nu = v, S = s)$ is non-Dirac, whereas residual oracle-act ambiguity means that $\mathrm{Law}(\nu, S)(D_S) > 0$. The conditional law, and hence D_S, is defined only up to $\mathrm{Law}(\nu, S)$-null sets. Under the standard-Borel assumptions, D_S is measurable.

Assumption 3.10 (Optional realized uniqueness)**.** Only when explicitly invoked, impose

(J) For π-almost every (v, z), $\arg\min_{a \in \mathcal{A}} \rho(a; v, z) = \{\alpha(v, z)\}$.

The *realized uniqueness* condition (J) strengthens the π-optimality condition in (7) by requiring the state-conditional Bayes act to be unique π-almost surely.

Theorem 3.11 (Strictly positive excess risk under residual oracle-act ambiguity)**.** *Under Assumptions 2.2 and 2.4, fix a jointly measurable π-optimal family α and assume realized uniqueness (J) of Assumption 3.10. For every state summary $S = f(Z)$, the following hold.*

(i) Local strict inequality. *On the event $\{(\nu, S) \in D_S\}$, the following inequality holds* almost surely*:*

$$\mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z) \mid \nu, S] < \inf_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, S]. \tag{9}$$

(ii) Global criterion for strictly positive excess risk. *The following conditions are equivalent:*

$$R_S > R_{\mathrm{ora}} \iff G(S) < G(Z) \iff \mathbb{P}((\nu, S) \in D_S) > 0.$$

Thus the global risk is strictly larger precisely when S leaves residual oracle-act ambiguity.

By (9), the first inequality in (5) becomes strict on $\{(\nu, S) \in D_S\}$. The state-conditional risks of $\alpha(\nu, Z)$ and $\hat{y}_{\mathrm{ora}}(\nu, Z)$ agree almost surely by Lemma 3.3(ii). Part (i) is a local result and does not require D_S to have positive probability. Neither (P), (R), nor product-to-joint domination is used; uniqueness of the Bayes act given (ν, S) is not required. The constant summary removes all additional state information. Specializing the global criterion for strictly positive excess risk in Theorem 3.11(ii) gives the precise condition under which the oracle gain $G(Z)$ is positive.

Corollary 3.12 (Sharp positivity criterion)**.** *Under Assumptions 2.2 and 2.4, assume realized uniqueness (J) of Assumption 3.10. For any selected observable-instance Bayes predictor $\hat{y}_{\mathrm{obs}}$,*

$$G(Z) > 0 \iff \mathbb{P}(\mathrm{Law}(\alpha(\nu, Z) \mid \nu) \text{ is not a Dirac measure}) > 0 \iff \mathbb{P}\big(\alpha(\nu, Z) \neq \hat{y}_{\mathrm{obs}}(\nu)\big) > 0. \tag{10}$$

Uniqueness of the observable-instance Bayes act is not required.

Under the assumptions of Corollary 3.12, in the finite-state and finite-instance setting, (10) gives $G(Z) > 0$ if and only if there exist v, z_i, z_j such that $\mathbb{P}(\nu = v, Z = z_i) > 0$, $\mathbb{P}(\nu = v, Z = z_j) > 0$, and $\alpha(v, z_i) \neq \alpha(v, z_j)$.

Example 3.13 (Realized value of state distinctions)**.** For the prediction-rule family of Example 3.1, the CPQ partition is $\{z_1, z_2\}$, $\{z_3\}$, and $\{z_4\}$. Two realized joint laws of (ν, Z) with identical state and instance marginals and conditional response kernels, but different conditional state weights $\mathbb{P}(Z \in dz \mid \nu)$, yield $G(Z) = 0$ and $G(Z) > 0$, respectively. The joint laws, numerical probability table, and risk calculations are given in Section C.7.

By the oracle-risk preservation in Proposition 3.8 and the refinement comparison in Lemma 2.10(ii), any state summary S that recovers the CPQ preserves oracle risk: $T^\star = g(S)$ almost surely implies $R_S = R_{\mathrm{ora}}$. However, Example 3.13 shows that the converse need not hold: even a constant summary can preserve oracle risk without recovering the CPQ, because some instance–state pairs with positive product-law probability never occur under the realized joint law. The following theorem establishes when the converse holds: with realized uniqueness (J) and product-to-joint domination $\pi_\otimes \ll \pi$, oracle-risk preservation requires recovery of the CPQ.

Theorem 3.14 (Conditional minimality of the CPQ)**.** *Under Assumptions 2.2 and 2.4, fix a jointly measurable π-optimal family α and a CPQ $T^\star = \tau(Z)$ satisfying conditions (P) and (R) of Assumption 3.5. Assume realized uniqueness (J) of Assumption 3.10 and product-to-joint domination $\pi_\otimes \ll \pi$. Then for every measurable state summary $S = f(Z)$,*

$$R_S = R_{\mathrm{ora}} \quad \Longleftrightarrow \quad T^\star = g(S) \quad \text{almost surely for some measurable } g.$$

In particular, $G(Z) > 0$ if and only if $T^\star$ is not almost surely *constant. The minimality assertion is conditional on the stated domination and uniqueness assumptions; it does not assert unconditional minimality of the CPQ.*

The conditional-law criterion in Corollary 3.12 uses (J); the characterization by a nonconstant CPQ here additionally uses (P), (R), and product-to-joint domination. We call (P), (R), (J), and product-to-joint domination, imposed in addition to Assumptions 2.2 and 2.4, the *CPQ minimality hypotheses*.

4 Context–Query Predictive Sufficiency and Linear-Gaussian Specialization

We extend CPQ to prescribed queries and growing contexts, characterizing predictive sufficiency, its persistence, and the risk of retaining an earlier state summary.

4.1 Context–Query Predictive Sufficiency

Fix an integer $d \geq 1$ and an $\mathbb{R}^d$-valued prediction target $\boldsymbol{w}$ with $\mathbb{E}\|\boldsymbol{w}\|^2 < \infty$. For $t \geq 1$, let $C_t := ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_t, y_t))$ be the ordered context of the first t observed input–response pairs. Assume that the contexts and the hidden state take values in standard Borel spaces, and set $\mathcal{F}_t := \sigma(C_t) \subseteq \mathcal{F}_{t+1}$. For a fixed, externally prescribed query $\boldsymbol{q} \in \mathbb{R}^d$, the observable instance is $(C_t, \boldsymbol{q})$, and the query response y_{q} of Definition 2.1 is $Y_{\boldsymbol{q}} := \boldsymbol{q}^\top \boldsymbol{w}$ under squared loss. Here t denotes the *stage* (context-prefix length), not training time. We compare nested prefixes at the same query, excluding rearrangements of fixed demonstrations. The response is linear in $\boldsymbol{q}$, although its Bayes predictor may depend nonlinearly on C_t. Neither nonlinear query responses nor adaptively selected queries are considered, and no Gaussian-model or predictor-architecture assumption is imposed.

Let $S = f(Z)$ be a standard-Borel-valued state summary. Write $\mathcal{G}_t^Z := \mathcal{F}_t \vee \sigma(Z)$ and $\mathcal{G}_t^S := \mathcal{F}_t \vee \sigma(S)$, where $\vee$ denotes the generated sigma-field. Define the *oracle posterior mean* $\boldsymbol{m}_t^Z := \mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^Z]$ and the *summary-conditioned posterior mean* $\boldsymbol{m}_t^S := \mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^S]$. For each fixed query $\boldsymbol{q}$, conditional-mean Bayes predictions follow from scalar squared-loss optimality (Lemma C.4); their scalar Bayes risks are $R_{S,t}(\boldsymbol{q}) := \mathbb{E}[(Y_{\boldsymbol{q}} - \boldsymbol{q}^\top \boldsymbol{m}_t^S)^2]$ and $R_{\mathrm{ora},t}(\boldsymbol{q}) := \mathbb{E}[(Y_{\boldsymbol{q}} - \boldsymbol{q}^\top \boldsymbol{m}_t^Z)^2]$.

For a linear query subspace $U \subseteq \mathbb{R}^d$, let P_U be its orthogonal projector. Since $Y_{\boldsymbol{q}} = \boldsymbol{q}^\top P_U \boldsymbol{w}$ for every $\boldsymbol{q} \in U$, consider the vector response $P_U \boldsymbol{w}$ under squared Euclidean loss. Its Bayes predictions, as identified in the projected-problem argument (27), are $\mathbb{E}[P_U \boldsymbol{w} \mid \mathcal{G}_t^S] = P_U \boldsymbol{m}_t^S$ and $\mathbb{E}[P_U \boldsymbol{w} \mid \mathcal{G}_t^Z] = P_U \boldsymbol{m}_t^Z$, with corresponding risks $R_{S,t}^U := \mathbb{E}\|P_U(\boldsymbol{w} - \boldsymbol{m}_t^S)\|^2$ and $R_{\mathrm{ora},t}^U := \mathbb{E}\|P_U(\boldsymbol{w} - \boldsymbol{m}_t^Z)\|^2$. We call S *stage-t U-sufficient* if $R_{S,t}(\boldsymbol{q}) = R_{\mathrm{ora},t}(\boldsymbol{q})$ for every $\boldsymbol{q} \in U$. This is equivalent to $R_{S,t}^U = R_{\mathrm{ora},t}^U$ by (24) and (25). Define the *excess-risk matrix* $\mathbf{A}_{S,t} := \mathbb{E}[(\boldsymbol{m}_t^Z -$

$\boldsymbol{m}_t^S)(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)^\top] \succeq 0$, where $\succeq 0$ denotes positive semidefiniteness. Equivalently, $\mathbf{A}_{S,t} = \mathbb{E}[\mathrm{Cov}(\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S)]$; see (19).

Call the vector prediction problem for $P_U \boldsymbol{w}$ the *stage-t projected problem*, with observable instance C_t, hidden state Z, response $P_U \boldsymbol{w}$, action space U, and squared Euclidean loss. Thus stage-t U-sufficiency is exactly oracle-risk preservation for this projected problem. Here the observable-instance law λ and realized joint law π of Section 2.1 specialize to $\mathrm{Law}(C_t)$ and $\pi_t := \mathrm{Law}(C_t, Z)$, respectively. Select a prediction-rule family $\boldsymbol{\alpha}_{t,U}$ with $\boldsymbol{\alpha}_{t,U}(C_t, Z) = P_U \boldsymbol{m}_t^Z$ almost surely. Its π_t-optimality and (J) are verified in Section D.1. If an encoding $\tau_{t,U}$ satisfies (P) and (R) for this family, its U-*CPQ* is $T_{t,U}^\star := \tau_{t,U}(Z)$, and the reconstruction map b_π is denoted by b_t. The U-CPQ compares entire C_t-indexed prediction-rule profiles, not actions at a single context value; its existence is not required for the risk identities below.

For $t < u$, define the *oracle posterior-mean increment* and *summary-conditioned posterior-mean increment* $\boldsymbol{\Delta}_{t:u}^Z := \boldsymbol{m}_u^Z - \boldsymbol{m}_t^Z$ and $\boldsymbol{\Delta}_{t:u}^S := \boldsymbol{m}_u^S - \boldsymbol{m}_t^S$, and the *projected oracle increment* $\boldsymbol{\Delta}_{t:u}^{Z,U} := P_U \boldsymbol{\Delta}_{t:u}^Z$. If a stage-$t$ U-CPQ exists, its *transfer loss* to stage $u > t$, relative to the selected family $\boldsymbol{\alpha}_{t,U}$, is defined with $S = T_{t,U}^\star$ held fixed: $\mathcal{L}_{t\to u}(U) := R_{S,u}^U - R_{\mathrm{ora},u}^U$. We now provide the main results. The proof is in Section D.1.

Theorem 4.1 (U-CPQ and context-prefix transfer)**.** *Assume the preceding nested-context setting. Fix any state summary $S = f(Z)$ and any query subspace U. The following conclusions have separate assumptions. Here conditions (P) and (R) are those of Assumption 3.5, and realized uniqueness (J) is as in Assumption 3.10, applied to the stage-t or stage-u projected problem, as appropriate. Under squared Euclidean loss, (J) holds for both projected problems by the common-full-set and strict convexity argument following (26) in Section D.1.*

(i) stage-t U-sufficiency and CPQ recovery. At each stage t, the query-wise excess risk and the criterion for stage-t U-sufficiency are

$$\begin{aligned} R_{S,t}(\boldsymbol{q}) - R_{\mathrm{ora},t}(\boldsymbol{q}) &= \boldsymbol{q}^\top \mathbf{A}_{S,t} \boldsymbol{q}, \\ \big[R_{S,t}(\boldsymbol{q}) = R_{\mathrm{ora},t}(\boldsymbol{q})\ \forall \boldsymbol{q} \in U\big] &\iff U \subseteq \ker \mathbf{A}_{S,t}. \end{aligned} \tag{11}$$

If a U-CPQ $T_{t,U}^\star$ exists (satisfying (P) and (R)), realized uniqueness (J) holds (automatically at stage t under squared Euclidean loss), and product-to-joint domination $\mathrm{Law}(C_t) \otimes \mathrm{Law}(Z) \ll \mathrm{Law}(C_t, Z)$ holds, then the stage-t U-sufficiency condition $U \subseteq \ker \mathbf{A}_{S,t}$ is further equivalent to $T_{t,U}^\star = g_t(S)$ almost surely for some measurable g_t. This last equivalence applies Theorem 3.14 to the stage-t projected problem; its Bayes-act and measurability conditions are verified in Section D.1.

(ii) Persistence of U-sufficiency and transfer loss. Suppose S is stage-t U-sufficient, equivalently $U \subseteq \ker \mathbf{A}_{S,t}$ by part (i). For any $u > t$, stage-u U-sufficiency for the same S is characterized by

$$U \subseteq \ker \mathbf{A}_{S,u} \iff P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S) = \mathbf{0} \quad \text{almost surely.} \tag{12}$$

Under the same stage-t U-sufficiency assumption, the exact stage-u excess risk and its matrix identity hold for this arbitrary S, without requiring a CPQ to exist:

$$\begin{aligned} R_{S,u}^U - R_{\mathrm{ora},u}^U &= \mathbb{E}\left\|P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S)\right\|^2 = \mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)], \\ P_U \mathbf{A}_{S,u} P_U &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)]. \end{aligned} \tag{13}$$

Here $\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)$ denotes conditional covariance and tr denotes matrix trace. In each direction $\boldsymbol{q} \in U$, the stage-u excess risk is $\mathbb{E}[\mathrm{Var}(\boldsymbol{q}^\top \boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)]$.

In particular, suppose a U-CPQ $T_{t,U}^\star$ exists (satisfying (P) and (R)) at stage t, and take $S = T_{t,U}^\star$ in the following transfer statements. Condition (R), through the stage-t projected-problem application of Proposition 3.8, ensures its stage-t U-sufficiency. The summary remains fixed, while the Bayes predictor may use C_u. For this choice, the general excess-risk identity (13) specializes to

$$\mathcal{L}_{t\to u}(U) = R_{S,u}^U - R_{\mathrm{ora},u}^U = \mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)].$$

Zero transfer loss is equivalent to $\boldsymbol{\Delta}_{t:u}^{Z,U}$ having a $\mathcal{G}_u^S$-measurable version. If a U-CPQ $T_{u,U}^\star$ exists at stage u (satisfying (P) and (R)), then $T_{u,U}^\star = g_u(S)$ almost surely for some measurable g_u implies $\mathcal{L}_{t\to u}(U) = 0$, by the stage-$u$ refinement conclusion of Proposition 3.8, without (J) or product-to-joint domination. If, in addition, realized uniqueness (J) for the stage-u projected problem (automatic under squared Euclidean loss) and product-to-joint domination

$\mathrm{Law}(C_u) \otimes \mathrm{Law}(Z) \ll \mathrm{Law}(C_u, Z)$ *hold, the converse follows from the stage-u projected-problem application of Theorem 3.14, so*

$$\mathcal{L}_{t\to u}(U) = 0 \iff T^\star_{u,U} = g_u(S) \quad \text{almost surely for some measurable } g_u. \tag{14}$$

Stage-u U-sufficiency does not imply U-sufficiency at every intermediate stage.

4.2 Linear-Gaussian Specialization

We specialize Theorem 4.1(i)–(ii) to a finite-state linear-Gaussian model. The posterior calculations and CPQ minimality hypotheses are verified in Section D.2.

Let $Z \in \{1, \dots, K\}$, $K \geq 1$, satisfy $\mathbb{P}(Z = k) = \theta_k > 0$ and $\boldsymbol{w} \mid Z = k \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_k)$, where $\mathbf{V}_k \succ 0$. Here $\mathcal{N}(m, V)$ denotes a Gaussian distribution with mean m and covariance V. For deterministic nested designs $\mathbf{X}_t \in \mathbb{R}^{t\times d}$ with rows $\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_t^\top$, observe $\boldsymbol{y}_t := (y_1, \dots, y_t)^\top = \mathbf{X}_t\boldsymbol{w} + \boldsymbol{\varepsilon}_t$. The noise vector $\boldsymbol{\varepsilon}_t := (\varepsilon_1, \dots, \varepsilon_t)^\top$ comprises the first t terms of an i.i.d. $\mathcal{N}(0, \sigma^2)$ sequence independent of $(\boldsymbol{w}, Z)$, where $\sigma^2 > 0$. Since the design is deterministic, $\mathcal{F}_t = \sigma(\boldsymbol{y}_t)$. Let $\mathbf{I}_t$ denote the $t \times t$ identity matrix, and define $\mathbf{B}_{k,t} := \mathbf{V}_k\mathbf{X}_t^\top(\mathbf{X}_t\mathbf{V}_k\mathbf{X}_t^\top + \sigma^2\mathbf{I}_t)^{-1}$. Then the oracle posterior mean is $\boldsymbol{m}_t^Z = \mathbf{B}_{Z,t}\boldsymbol{y}_t$. For a state summary $S = f(Z)$, put $\omega_{k,t} := \mathbb{P}(Z = k \mid \boldsymbol{y}_t, S)$ (these are the mixing weights in (1) for the observable context C_t; the dependence on S is suppressed, and for another summary S' we use $\omega_{k,t}^{S'}$); the summary-conditioned posterior mean is $\boldsymbol{m}_t^S = \sum_{k=1}^K \omega_{k,t}\mathbf{B}_{k,t}\boldsymbol{y}_t$. For each query subspace U, the Gaussian U-CPQ admits the representative $T^\star_{t,U} = \tau_{t,U}(Z)$, where $\tau_{t,U}(k) := P_U\mathbf{B}_{k,t}$, as in the coefficient encoding (50). For the one-step extension from t to $t+1$, append the design row $\boldsymbol{x}_{t+1}^\top$. The *state-$k$ posterior covariance* and *update gain* are $\mathbf{P}_{k,t} := \mathrm{Cov}(\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k) = (\mathbf{V}_k^{-1} + \sigma^{-2}\mathbf{X}_t^\top\mathbf{X}_t)^{-1}$ and $\boldsymbol{K}_{k,t} := \mathbf{P}_{k,t}\boldsymbol{x}_{t+1}/(\sigma^2 + \boldsymbol{x}_{t+1}^\top\mathbf{P}_{k,t}\boldsymbol{x}_{t+1})$. The corresponding state-k increment is $\boldsymbol{\Delta}_{t:t+1}^{(k)} := \boldsymbol{K}_{k,t}(y_{t+1} - \boldsymbol{x}_{t+1}^\top\mathbf{B}_{k,t}\boldsymbol{y}_t)$. We now provide the main corollary. The proof is in Section D.3.

Corollary 4.2 (Gaussian U-CPQ and one-step context-prefix transfer)**.** *Fix $S = f(Z)$ and $U \subseteq \mathbb{R}^d$ in the model above.*

(i) stage-t U-sufficiency and CPQ recovery. The excess-risk identity (11) becomes

$$R_{S,t}(\boldsymbol{q}) - R_{\mathrm{ora},t}(\boldsymbol{q}) = \boldsymbol{q}^\top\mathbf{A}_{S,t}\boldsymbol{q} = \frac{1}{2}\mathbb{E}\left[\sum_{k,\ell=1}^K \omega_{k,t}\omega_{\ell,t}\big(\boldsymbol{q}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t})\boldsymbol{y}_t\big)^2\right]. \tag{15}$$

The Gaussian stage-t U-sufficiency and CPQ-recovery criteria are

$$\begin{aligned}[R_{S,t}(\boldsymbol{q}) = R_{\mathrm{ora},t}(\boldsymbol{q})\ \forall \boldsymbol{q} \in U] &\iff U \subseteq \ker\mathbf{A}_{S,t} \\ &\iff P_U(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) = \mathbf{0}_{d\times t}\ \ \forall k, \ell : f(k) = f(\ell) \\ &\iff T^\star_{t,U} = g_t(S) \ \text{ almost surely for some measurable } g_t.\end{aligned} \tag{16}$$

The first equivalence is the risk-gap criterion in (11), and the last is the CPQ-recovery equivalence of Theorem 4.1(i).

(ii) One-step persistence of U-sufficiency and transfer loss. Assume stage-t U-sufficiency. The Gaussian posterior-mean update (62) yields $\boldsymbol{\Delta}_{t:t+1}^Z = \boldsymbol{\Delta}_{t:t+1}^{(Z)}$. At $u = t+1$, (12) gives

$$U \subseteq \ker\mathbf{A}_{S,t+1} \iff P_U(\boldsymbol{\Delta}_{t:t+1}^Z - \boldsymbol{\Delta}_{t:t+1}^S) = \mathbf{0} \quad \text{almost surely.}$$

Equivalently, persistence holds if and only if both coefficient conditions *below hold for every k, ℓ with $f(k) = f(\ell)$:*

$$P_U(\boldsymbol{K}_{k,t} - \boldsymbol{K}_{\ell,t}) = \mathbf{0}, \qquad P_U\boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) \ = \mathbf{0}_{d\times t}. \tag{17}$$

With $\mathcal{G}_{t+1}^S = \sigma(\boldsymbol{y}_{t+1}, S)$, the first equality below is (13) with $u = t+1$; the Gaussian mixture gives the second:

$$R_{S,t+1}^U - R_{\mathrm{ora},t+1}^U = \mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(\boldsymbol{\Delta}_{t:t+1}^{Z,U} \mid \mathcal{G}_{t+1}^S)] = \frac{1}{2}\mathbb{E}\left[\sum_{k,\ell=1}^K \omega_{k,t+1}\omega_{\ell,t+1}\left\|P_U(\boldsymbol{\Delta}_{t:t+1}^{(k)} - \boldsymbol{\Delta}_{t:t+1}^{(\ell)})\right\|^2\right]. \tag{18}$$

For $S = T^\star_{t,U}$, this is $\mathcal{L}_{t\to t+1}(U)$ by the CPQ specialization of (13); it vanishes iff $T^\star_{t+1,U} = g_{t+1}(S)$ almost surely *for some measurable g_{t+1} by (14).*

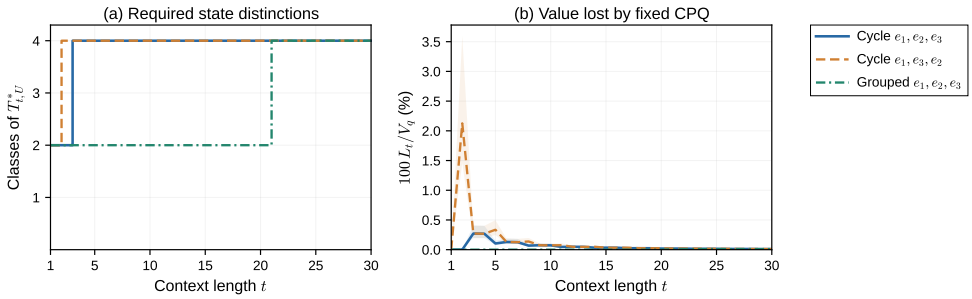

Figure 1: Gaussian illustration. (a) Number of classes of $T^{\star}_{t,U}$ for the fixed query $\boldsymbol{q}_0$. (b) Excess risk $L_t := R_{S_0,t}(\boldsymbol{q}_0) - R_{\mathrm{ora},t}(\boldsymbol{q}_0)$ of the fixed stage-1 U-CPQ, normalized by the prior query variance $V_q := \mathbb{E}[(\boldsymbol{q}_0^\top \boldsymbol{w})^2]$ (percent). The right-panel curves are medians over 48 covariance models, with between-model interquartile bands; the left panel gives the model-independent class counts.

Remark 4.3. The example in Section D.5 shows that neither condition in (17) implies the other. For $S = T^{\star}_{t,U}$, the equivalence in (17) shows that failure of either condition prevents stage-$(t+1)$ U-sufficiency. The projected risk conversion in (25), together with (11) and (18), then gives a positive one-step transfer loss. The CPQ-recovery equivalence in the corollary uses Gaussian (J) and product-to-joint domination, verified in the projected-problem argument and (49), respectively.

5 Experiments

Our results characterize predictive sufficiency and excess risk, but not the model-specific magnitudes, peak times, or persistence of oracle gain and transfer loss. We therefore examine finite-stage behavior rather than verify theorems.

We use a controlled four-state linear-Gaussian model with independent uniform binary factors $Z = (A, B)$. We compare three 30-stage context orders: periodic $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$, periodic $(\boldsymbol{e}_1, \boldsymbol{e}_3, \boldsymbol{e}_2)$, and ten consecutive observations in each coordinate direction, where $\boldsymbol{e}_i$ are the standard basis vectors of $\mathbb{R}^3$. For $\boldsymbol{q}_0 := (\boldsymbol{e}_1 + \boldsymbol{e}_2)/\sqrt{2}$ and $U := \mathrm{span}\{\boldsymbol{q}_0\}$, the stage-1 U-CPQ is represented by $S_0 := A$. The coefficient calculation and the stage-2 distinctions are given in Section E. Oracle gain compares context-only and full-state risks; excess risk compares S_0 and full-state risks. At $t = 2$, the prefix $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ preserves the sufficiency of S_0, whereas $(\boldsymbol{e}_1, \boldsymbol{e}_3)$ requires four classes and produces a median normalized excess risk of 2.126%. Under the grouped order, S_0 remains sufficient through $t = 20$, but at $t = 21$ its median normalized excess risk is only 0.0065% (Figure 1). Thus a newly relevant state distinction can have markedly different risk costs across context prefixes. Oracle gain also varies with query and context order. After $(\boldsymbol{e}_1, \boldsymbol{e}_3)$, the projected CPQs for queries $\boldsymbol{e}_1, \boldsymbol{e}_3$ are constant (see Section E), so their zero oracle gain follows from Proposition 3.8 and Definition 2.9. The positive gain for $\boldsymbol{e}_2$ is consistent with the nonconstant-CPQ criterion in Theorem 3.14 under the verified Gaussian assumptions. The median gain normalized by each query's prior variance is 10.776% for $\boldsymbol{e}_2$. The grouped order has a median model-wise peak normalized gain of 11.106% at median stage 10, versus 7.412% at stage 1 for the first periodic order; median counts of stages with maximum normalized gain at least 0.5% are 21 and 7, respectively (Section E). These observations are model-specific, not general magnitude or persistence laws.

In a separate Transformer study in Section F, a predictor supplied with CPQ labels achieves 89.4%–90.7% of the analytic oracle gain, using the observable-instance Bayes predictor's empirical loss as the baseline, versus about 91%–92% with full-state labels.

6 Limitations and Future Directions

Our risk guarantees concern prediction-relevant state information, not its recovery by a trained predictor from demonstrations. Connecting this information to ICL mechanisms involving attention, optimization, approximation, and finite-sample dynamics remains open. Moreover, CPQ minimality relies on realized uniqueness and product-to-joint domination. Developing estimable diagnostics for exact Bayes quantities and stability results for approximate or nonunique Bayes acts remains a further direction.

References

[1] Masafumi Akahira and Kei Takeuchi. A note on prediction sufficiency (adequacy) and sufficiency. *Australian Journal of Statistics*, 22(3):332–335, 1980.

[2] Waïss Azizian and Ali Hasan. How does the pretraining distribution shape in-context learning? A fundamental trade-off. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 4549–4605. PMLR, 2026.

[3] Francis Bach. *Learning Theory from First Principles*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, 2024. ISBN 9780262049443.

[4] D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

[5] David Blackwell. Comparison of experiments. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102, Berkeley, CA, 1951. University of California Press.

[6] M. H. DeGroot. *Optimal Statistical Decisions*. Wiley Classics Library. John Wiley & Sons, 2004.

[7] Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12784–12805. PMLR, 2024.

[8] H. J. Jeon, J. D. Lee, Q. Lei, and B. Van Roy. An information-theoretic analysis of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21522–21554. PMLR, 2024.

[9] Z. Lin and K. Lee. Dual operating modes of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30135–30188. PMLR, 2024.

[10] Alon Orlitsky and James R. Roche. Coding for computing. *IEEE Transactions on Information Theory*, 47(3):903–917, 2001.

[11] Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *International Conference on Learning Representations (ICLR)*, pages 49789–49843, 2024.

[12] Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning. arXiv preprint arXiv:2606.04045, 2026.

[13] Vasileios Sevetlidis. Loss-shift transfer via Bayes quotients. arXiv preprint arXiv:2606.13178, 2026.

[14] Kei Takeuchi and Masafumi Akahira. Characterizations of prediction sufficiency (adequacy) in terms of risk functions. *The Annals of Statistics*, 3(4):1018–1024, 1975.

[15] Erik N. Torgersen. Prediction sufficiency when the loss function does not depend on the unknown parameter. *The Annals of Statistics*, 5(1):155–163, 1977.

[16] Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: A generalization theory for meta-learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 124318–124351. PMLR, 2026.

[17] Hans S. Witsenhausen. The zero-error side information problem and chromatic numbers (Corresp.). *IEEE Transactions on Information Theory*, 22(5):592–593, 1976.

[18] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1684–1692. PMLR, 2025.

Supplementary Material: Canonical Predictive Quotients for Prediction under Hidden-State Uncertainty

A Related Work

Decision-theoretic and coding foundations. Blackwell's comparison of experiments [5, 4] and Bayesian decision theory [6] formalize the value of information. Classical prediction sufficiency relates information reduction to risk and conditional independence [14, 15]. Akahira and Takeuchi [1] also study sufficiency when later observations are added. Witsenhausen studies zero-error source recovery with decoder side information, while Orlitsky and Roche study communication rates for computing functions with side information [17, 10]. We instead study oracle-risk preservation by state-only summaries, not alphabet sizes or communication rates; condition (R) of Assumption 3.5 reconstructs the selected oracle act.

Bayes quotients and transfer. Sevetlidis [12] characterizes Bayes sufficiency of input representations by factorization of Bayes-optimal actions and, under uniqueness, refinement of the Bayes quotient, to which Bayes-minimal representations are informationally equivalent. For nonnegative losses, this general framework also applies to full inputs (ν, Z) and representations (ν, S). Here $S = f(Z)$ must depend only on Z. The CPQ compares selected prediction-rule profiles across observable instances under (P), and (R) reconstructs the selected oracle act under the realized joint law. A nonconstant CPQ need not yield positive oracle gain, whereas realized uniqueness and product-to-joint domination make oracle-risk preservation equivalent to CPQ recovery (Theorem 3.14 and Example 3.13). Sevetlidis [13] instead studies a fixed data law with a changing loss, relating quotient refinement to the transfer failure of source-minimal representations and, for finite-output log loss, to conditional mutual information. Our context-prefix result fixes the loss and quantifies the transfer loss of an earlier U-CPQ as the context grows (Theorem 4.1).

Bayesian in-context learning. Xie et al. [18] study latent-concept inference in a mixture of hidden Markov models; Panwar et al. [11] compare trained Transformers with Bayesian predictors across task families; and Zhang et al. [19] analyze Bayesian model averaging in ICL. Azizian and Hasan [2] characterize how the pretraining-task distribution affects task selection and generalization. Jeon et al. [8] and Wakayama and Suzuki [16] decompose prediction error and relate part of it to uncertainty about the latent task. CPQ instead characterizes which state distinctions are relevant to selected Bayes-optimal predictions.

Growing contexts. Lin and Lee [9] explain initial increases in ICL risk through task retrieval and task learning; Falck et al. [7] test Bayesian martingale properties of learned in-context predictions and find violations. Our stagewise U-sufficiency and posterior-mean-increment identities (12) and (13) instead give query-dependent conditions and exact excess risks for a fixed state summary as observations accumulate, without claims about training or model architecture.

B Prediction Problem and Bayes Predictors: Details and Proofs (Section 2)

B.1 Proof of Lemma 2.3 (Section 2.1)

We prove the conditional-mixture and conditional-risk identities.

Proof of Lemma 2.3. Let $S = f(Z)$. For every fixed measurable $B \subseteq \mathcal{Y}$, the tower property, kernel consistency, and the regular conditional law of Z given (ν, S) yield, in this order,

$$\begin{aligned}\mathbb{P}(y_{\mathrm{q}} \in B \mid \nu, S) &= \mathbb{E}[\mathbb{P}(y_{\mathrm{q}} \in B \mid \nu, Z) \mid \nu, S] \\ &= \mathbb{E}[\mathsf{K}(B \mid \nu, Z) \mid \nu, S] \\ &= \int_{\mathcal{Z}} \mathsf{K}(B \mid \nu, z)\, \mathbb{P}(Z \in dz \mid \nu, S) \quad \text{almost surely,}\end{aligned}$$

which is (1) for each B. Since $\mathcal{Y}$ is standard Borel, choose a countable generating π-system containing $\mathcal{Y}$. The equality holds outside a common null set on this system. Uniqueness of probability measures then identifies, on this common full-measure set, the selected regular conditional law of y_{q} given (ν, S) on the left-hand side with the mixture kernel on the right-hand side as probability measures, which gives (1) as an identity of random probability measures. By contrast, the middle term $\mathbb{E}[\mathsf{K}(B \mid \nu, Z) \mid \nu, S]$ is a conditional expectation, and its equality with the two sides is asserted only for each fixed B, outside a null set that may depend on B.

For (ii), the Doob–Dynkin lemma gives a measurable g with $A' = g(\nu, Z)$ almost surely. For a loss taking both signs, conditional integration is applied separately to the positive and negative parts of the loss of the action considered, each of which is nonnegative. If the expectation of at least one of the two parts is finite, the corresponding conditional integral is finite almost surely, and the difference of the two is defined as an extended real number with no indeterminate $+\infty - \infty$, as in Definition 2.1(iii). The same treatment of the positive and negative parts applies to the losses of both A' and $\varphi(\nu, S)$. Kernel consistency and conditional integration, together with the definition of ρ, yield (2) whenever the conditional risks are defined:

$$\mathbb{E}[L(A', y_{\mathrm{q}}) \mid \nu, Z] = \int L(g(\nu, Z), y)\, \mathsf{K}(dy \mid \nu, Z) = \rho(A'; \nu, Z) \quad \text{almost surely.}$$

Since $S = f(Z)$ gives $\sigma(\nu, S) \subseteq \sigma(\nu, Z)$ and $\varphi(\nu, S)$ is $\sigma(\nu, Z)$-measurable, conditioning (2) on (ν, S) and applying the tower property gives (3):

$$\begin{aligned}\mathbb{E}[L(\varphi(\nu, S), y_{\mathrm{q}}) \mid \nu, S] &= \mathbb{E}\big[\mathbb{E}[L(\varphi(\nu, S), y_{\mathrm{q}}) \mid \nu, Z] \mid \nu, S\big] \\ &= \mathbb{E}[\rho(\varphi(\nu, S); \nu, Z) \mid \nu, S] \quad \text{almost surely,}\end{aligned}$$

where the second equality applies the identity just displayed to the action $\varphi(\nu, S)$. Taking expectations gives the stated risk identity whenever the risk is well defined. The same arguments apply to $S = Z$. □

B.2 Proof of Lemma 2.10 (Section 2.2)

Proof of Lemma 2.10. Since $S = f(Z)$ is state-only, the information sets are nested:

$$\sigma(\nu) \subseteq \sigma(\nu, S) \subseteq \sigma(\nu, Z).$$

Write $A_{\mathrm{obs}} := \hat{y}_{\mathrm{obs}}(\nu)$, $A_S := \hat{y}_S(\nu, S)$, and $A_{\mathrm{ora}} := \hat{y}_{\mathrm{ora}}(\nu, Z)$. Throughout the proof, the conditional risks given each $W \in \{\nu, (\nu, S), (\nu, Z)\}$ are the versions induced by a single regular conditional law of y_{q} given W, as stipulated in Definition 2.5. Hence the minimization in (4) holds for all actions simultaneously on one event of probability one, which allows W-measurable random actions to be substituted into it below.

For (i), conditional Bayes optimality (4) given (ν, Z) and (2) imply

$$\rho(A_{\mathrm{ora}}; \nu, Z) \le \rho(A_S; \nu, Z) \quad \text{almost surely.}$$

Both terms are integrable by the finite-risk assumption of Assumption 2.4 and (2). Taking conditional expectations given (ν, S) yields only the corresponding inequality between the conditional expectations of these two state-conditional risks.

By (4) with $W = (\nu, S)$, A_S attains the minimum conditional risk. Since A_{obs} is also (ν, S)-measurable, its conditional risk is no smaller than that of A_S almost surely. Applying (3) to A_S and A_{obs} gives the equality and final inequality of (5). Combined with this equality, the inequality of the previous paragraph becomes the first inequality, which completes the proof of (5).

Taking expectations gives the risk ordering in (6); its gain form follows from Definition 2.9. For any ν-measurable predictor $\varphi(\nu)$ with well-defined risk, (4) with $W = \nu$ shows that the conditional risk of A_{obs} is no greater than that of $\varphi(\nu)$ almost surely. Averaging proves $R_{\mathrm{obs}} \le R[\varphi(\nu)]$.

For (ii), let $S_1 = h(S_2)$ almost surely with h measurable. Then the (ν, S_2)-measurable predictor $\hat{y}_{S_1}(\nu, h(S_2))$ agrees almost surely with $\hat{y}_{S_1}(\nu, S_1)$ and has finite risk R_{S_1}. Conditional Bayes optimality (4) given (ν, S_2) gives

$$R_{S_2} = \mathbb{E}[L(\hat{y}_{S_2}(\nu, S_2), y_{\mathrm{q}})] \leq \mathbb{E}[L(\hat{y}_{S_1}(\nu, h(S_2)), y_{\mathrm{q}})] = R_{S_1}.$$

□

C Canonical Predictive Quotients and the Value of State Distinctions: Details and Proofs (Section 3)

C.1 Proof of Lemma 3.3 (Section 3.1)

Proof of Lemma 3.3. By Assumption 2.4 and definitions 2.5 and 2.8 and (4), the oracle Bayes predictor $\hat{y}_{\mathrm{ora}}(\nu, Z)$ is measurable and minimizes $\mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, Z]$ almost surely. Kernel consistency in Assumption 2.2 identifies the conditional response law with $\mathsf{K}(\cdot \mid \nu, Z)$ as probability measures on a common full-measure set. Thus the corresponding conditional loss integrals agree on this set for every action for which they are defined. Consequently, $\hat{y}_{\mathrm{ora}}$ satisfies (7). Its joint measurability and finite risk follow from Assumption 2.4. Therefore $\hat{y}_{\mathrm{ora}}$ is a π-optimal family; in particular, $\alpha = \hat{y}_{\mathrm{ora}}$ is one possible choice.

For any jointly measurable π-optimal family α with finite risk, (7) gives its state-conditional Bayes optimality almost surely. Hence both selected oracle acts attain the same state-conditional minimum:

$$\rho(\alpha(\nu, Z); \nu, Z) = \inf_{a \in \mathcal{A}} \rho(a; \nu, Z) = \rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z) \quad \text{almost surely.}$$

Finally, by the risk definition in Definition 2.1, the tower property, and (2), the risk of each selected oracle act equals the expectation of its state-conditional risk, as stated in Lemma 2.3(ii) for $S = Z$. Therefore,

$$R[\alpha(\nu, Z)] = \mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z)] = \mathbb{E}[\rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z)] = R[\hat{y}_{\mathrm{ora}}(\nu, Z)] = R_{\mathrm{ora}},$$

where the last equality uses Definition 2.8. Agreement of the selected oracle acts is not required: for example, if $\mathcal{A} = \{0, 1\}$ and $L \equiv 0$, distinct constant selections attain the same minimum and risk. □

C.2 Measurable Predictive-Equivalence Encodings and Independence of (P) and (R) (Section 3.1)

We construct a measurable encoding satisfying (P) for a fixed prediction-rule family. Reconstruction (R) is a separate requirement.

Existence of an exact predictive-equivalence encoding (P). Since the action space $\mathcal{A}$ is standard Borel, choose a Borel injection $\psi : \mathcal{A} \to [0, 1]$. Since the instance space $\mathcal{O}$ is standard Borel, its Borel sigma-field is generated by a countable algebra $\mathcal{C} := \{O_n : n \geq 1\}$. For the fixed jointly measurable family α, define

$$\Phi_\alpha(z) := \left(\int_{O_n} \psi(\alpha(v, z))\, \lambda(dv) \right)_{n \geq 1} \in [0, 1]^{\mathbb{N}}.$$

Every coordinate is measurable in z by joint measurability of α and the measurability of parameter integrals of bounded Borel functions. Hence Φ_α is measurable into the standard-Borel space $[0, 1]^{\mathbb{N}}$.

For each state z, define the finite measure $\eta_z(B) := \int_B \psi(\alpha(v, z))\, \lambda(dv)$ on the Borel subsets B of $\mathcal{O}$. This measure is finite since $0 \leq \psi \leq 1$ and λ is a probability measure. If $z \sim_{\mathrm{pred}} z'$, then $\eta_z = \eta_{z'}$, so $\Phi_\alpha(z) = \Phi_\alpha(z')$. Conversely, if $\Phi_\alpha(z) = \Phi_\alpha(z')$, the measures η_z and $\eta_{z'}$ agree on every O_n. Because $\mathcal{C}$ is an algebra generating the Borel sigma-field, uniqueness of finite measures implies that they agree on every Borel set. Thus $\psi(\alpha(v, z)) = \psi(\alpha(v, z'))$ for λ-almost every v (uniqueness of densities). Injectivity of ψ yields $z \sim_{\mathrm{pred}} z'$. Therefore, for *all* $z, z' \in \mathcal{Z}$,

$$\Phi_\alpha(z) = \Phi_\alpha(z') \quad \Longleftrightarrow \quad z \sim_{\mathrm{pred}} z'.$$

Condition (P) holds with $\mathcal{Z}_0 = \mathcal{Z}$. The construction does not assert (R).

Example C.1 (The common μ-full set in (P))**.** Condition (P) requires exact encoding for *every* pair in one measurable μ-full set $\mathcal{Z}_0$, not merely for almost every pair under $\mu \otimes \mu$. To see the role of the excluded null set, take the four prediction-rule profiles in Table 1 and $\lambda(\{v_A\}), \lambda(\{v_B\}) > 0$. Only the prediction-rule table of Example 3.1 is reused here; its assumption that all four states have positive probability is not carried over. Let $\tau(z_1) \coloneqq t_1^\star$, $\tau(z_2) \coloneqq t_1^\star$ and $\tau(z_3) \coloneqq t_2^\star$, $\tau(z_4) \coloneqq t_2^\star$, where $t_1^\star \neq t_2^\star$. The candidate incorrectly identifies z_3 with z_4. If the state probabilities are $(0.4, 0.3, 0.3, 0)$, $\mathcal{Z}_0 = \{z_1, z_2, z_3\}$ is μ-full and τ satisfies (P) there. If instead they are $(0.4, 0.3, 0.2, 0.1)$, every μ-full set contains all four states, so the same τ fails (P). This concerns the given candidate, not the existence of an exact encoding. Even for nonatomic μ, all excluded states must lie in one measurable μ-null set.

Remark C.2 (Logical independence of (P) and (R))**.** The two conditions are independent, even under squared loss. First, let the observable instance be constant, the response be zero, and the latent state take two values with positive probability. An encoding that distinguishes the two states satisfies (R), since the oracle act is always zero, but violates (P), since the states are predictively equivalent.

Conversely, let ν be uniform on $[0, 1]$, let B be an independent fair Bernoulli variable, and set $Z = (\nu, B)$ and $y_{\mathrm{q}} = B$. Given an observable value v and a state (u, b), prescribe a deterministic response equal to b when $v = u$ and zero otherwise. This defines a measurable kernel consistent with the realized response law and having a unique squared-loss Bayes act. Let the selected family α assign this act to every pair: for all $(v, (u, b))$, $\alpha(v, (u, b))$ equals b when $v = u$ and zero otherwise. For each fixed state, this act is zero for λ-almost every v. Thus all states are predictively equivalent, and (P) forces the encoding to be constant μ-almost surely; that is, $\tau(Z)$ is almost surely constant. However, under the realized joint law, the oracle act equals B. Since B is independent of ν, the constant encoding has risk $1/4$, whereas the oracle risk is zero. Moreover, (R) fails for every encoding satisfying (P), as a direct argument shows: if (R) held, then $B = \alpha(\nu, Z) = b_\pi(\nu, \tau(Z))$ almost surely, and, because $\tau(Z)$ is almost surely constant, the right-hand side would be almost surely equal to a $\sigma(\nu)$-measurable random variable. This is impossible: B is independent of ν, so it would be independent of itself and hence almost surely constant, whereas B is a nondegenerate Bernoulli variable. Thus, relative to this selected family α, no encoding satisfies both conditions in this example.

C.3 Proof of Proposition 3.8 (Section 3.1)

Proof of Proposition 3.8. Write $T^\star = \tau(Z)$ and $B_\star = b_\pi(\nu, T^\star)$. By (R), $B_\star = \alpha(\nu, Z)$ almost surely. By Lemma 3.3(iii), $R[B_\star] = R[\alpha(\nu, Z)] = R_{\mathrm{ora}} < \infty$. Let $\hat{y}_{T^\star}(\nu, T^\star)$ be a Bayes predictor given $(\nu, T^\star)$, whose finite risk is guaranteed by Assumption 2.4. By (6) and (4) with $W = (\nu, T^\star)$,

$$R_{\mathrm{ora}} \leq R_{T^\star} \leq R[B_\star] = R_{\mathrm{ora}},$$

so $R_{T^\star} = R[B_\star]$. We now verify the conditional Bayes optimality of $B_\star$ directly. Using (3) with $S = T^\star$, Lemma 3.3(ii), and (5) with $S = T^\star$, we obtain

$$\begin{aligned}
\mathbb{E}[L(B_\star, y_{\mathrm{q}}) \mid \nu, T^\star] &= \mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z) \mid \nu, T^\star] \\
&= \mathbb{E}[\rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z) \mid \nu, T^\star] \\
&\leq \inf_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, T^\star] \\
&\leq \mathbb{E}[L(B_\star, y_{\mathrm{q}}) \mid \nu, T^\star].
\end{aligned}$$

The first equality uses the reconstruction identity $B_\star = \alpha(\nu, Z)$ proved above. The final inequality holds because $B_\star$ is a $(\nu, T^\star)$-measurable candidate action in the same conditional-risk minimization, with all actions compared under the conditional-law convention of Definition 2.5. Equality therefore holds throughout, and $B_\star$ is conditionally Bayes optimal given $(\nu, T^\star)$. The risk comparison established above also gives $R_{T^\star} = R_{\mathrm{ora}}$, and hence $G(T^\star) = R_{\mathrm{obs}} - R_{T^\star} = R_{\mathrm{obs}} - R_{\mathrm{ora}} = G(Z)$.

Now let $S = f(Z)$ and suppose that $T^\star = g(S)$ almost surely for a measurable g. Then S refines $T^\star$ by the definition of refinement preceding Lemma 2.10: $S_1 = T^\star$, $S_2 = S$, and $h = g$. Part (ii) of that lemma, together with (6) and the oracle-risk identity for $T^\star$ proved above, gives

$$R_{\mathrm{ora}} \leq R_S \leq R_{T^\star} = R_{\mathrm{ora}}.$$

Thus $R_S = R_{\mathrm{ora}}$ and $G(S) = G(Z)$ by Definition 2.9. If the Bayes act given $(\nu, T^\star)$ is unique almost surely, every measurable such predictor agrees with $B_\star$ almost surely. Neither this uniqueness nor uniqueness of the state-conditional Bayes act was used for risk preservation or refinement. Condition (P) enters only in calling $T^\star$ a CPQ; the risk calculations themselves use (R). □

C.4 Proof of Theorem 3.11 (Section 3.2)

We first prove the following lemma.

Lemma C.3 (Conditional-Dirac reconstruction)**.** *Let X and H be standard-Borel-valued random elements. Then* $\mathrm{Law}(X \mid H)$ *is* almost surely *a Dirac measure if and only if $X = g(H)$* almost surely *for some measurable g.*

Proof. The reverse implication is immediate. For the forward implication, let $\mathcal{X}$ be the value space of X and choose a Borel isomorphism $\psi : \mathcal{X} \to B$ onto a Borel set $B \subseteq [0, 1]$. The Dirac condition gives $\mathbb{E}[(\psi(X) - \mathbb{E}[\psi(X) \mid H])^2] = 0$. By the Doob–Dynkin lemma, $\psi(X) = \mathbb{E}[\psi(X) \mid H] = m(H)$ almost surely for some measurable m. For fixed $x_0 \in \mathcal{X}$, define $g(h) := \psi^{-1}(m(h))$ if $m(h) \in B$, and $g(h) := x_0$ otherwise. Then $X = g(H)$ almost surely. □

We now give the proof of Theorem 3.11.

Proof of Theorem 3.11. Fix $S = f(Z)$ and let $\hat{y}_S$ be a measurable state-summary Bayes predictor. By (J), $\alpha(\nu, Z) = \hat{y}_{\mathrm{ora}}(\nu, Z)$ almost surely. Thus $R[\alpha(\nu, Z)] = R_{\mathrm{ora}} < \infty$ by Assumption 2.4, and Lemma 3.3 applies to α. Define the *state-conditional excess-risk random variable*

$$\mathcal{E}_S := \rho(\hat{y}_S(\nu, S); \nu, Z) - \rho(\alpha(\nu, Z); \nu, Z).$$

By π-optimality of α, $\mathcal{E}_S \geq 0$ almost surely. Both state-conditional risk terms are integrable: $\rho(\alpha(\nu, Z); \nu, Z)$ agrees almost surely with $\rho(\hat{y}_{\mathrm{ora}}(\nu, Z); \nu, Z)$ by Lemma 3.3(ii), and the risk identities (2) and (3), together with the finite risks in Assumption 2.4, give integrability. Using the tower property and (4) with $W = (\nu, S)$,

$$\begin{aligned}\Gamma_S(\nu, S) &:= \mathbb{E}[\mathcal{E}_S \mid \nu, S] = \mathbb{E}[\rho(\hat{y}_S(\nu, S); \nu, Z) \mid \nu, S] - \mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z) \mid \nu, S] \\ &= \inf_{a \in \mathcal{A}} \mathbb{E}[L(a, y_{\mathrm{q}}) \mid \nu, S] - \mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z) \mid \nu, S].\end{aligned}$$

Here $\Gamma_S(v, s)$ denotes a measurable version evaluated at (v, s). The *summary-conditioned excess risk* Γ_S is nonnegative by (5), since α and $\hat{y}_{\mathrm{ora}}$ both minimize $\rho(\cdot; \nu, Z)$ almost surely.

For (i), recall that (J) holds for π-almost every (v, z). Its conditional version is the conditional decomposition of this statement along (ν, S), and is not a new assumption: for $\mathrm{Law}(\nu, S)$-almost every (v, s), $\arg\min_{a \in \mathcal{A}} \rho(a; v, z) = \{\alpha(v, z)\}$ for $\mathrm{Law}(Z \mid \nu = v, S = s)$-almost every z. The same conditional decomposition of $\mathcal{E}_S \geq 0$ shows that, for $\mathrm{Law}(\nu, S)$-almost every (v, s), $\rho(\hat{y}_S(v, s); v, z) - \rho(\alpha(v, z); v, z) \geq 0$ for $\mathrm{Law}(Z \mid \nu = v, S = s)$-almost every z. Moreover, for $\mathrm{Law}(\nu, S)$-almost every (v, s), the selected versions satisfy $\mathrm{Law}(\alpha(\nu, Z) \mid \nu = v, S = s) = (\alpha(v, \cdot))_{\#} \mathrm{Law}(Z \mid \nu = v, S = s)$ as probability measures, where the right-hand side is the push-forward of $\mathrm{Law}(Z \mid \nu = v, S = s)$ under $z \mapsto \alpha(v, z)$, and the regular conditional law of Z given (ν, S) represents $\Gamma_S(\nu, S) = \mathbb{E}[\mathcal{E}_S \mid \nu, S]$ as

$$\Gamma_S(v, s) = \int_{\mathcal{Z}} \big[\rho(\hat{y}_S(v, s); v, z) - \rho(\alpha(v, z); v, z)\big] \, \mathbb{P}(Z \in dz \mid \nu = v, S = s).$$

Choose a (v, s) in the $\mathrm{Law}(\nu, S)$-full set where the identities above, the conditional version of (J), the conditional nonnegativity, the pushforward identity, and the integral representation all hold, and suppose $(v, s) \in D_S$. If $\Gamma_S(v, s)$ were zero, nonnegativity of the integrand would imply that the integrand vanishes for $\mathrm{Law}(Z \mid \nu = v, S = s)$-almost every z, that is,

$$\rho(\hat{y}_S(v, s); v, z) = \rho(\alpha(v, z); v, z) \quad \text{for } \mathrm{Law}(Z \mid \nu = v, S = s)\text{-almost every } z.$$

By the conditional version of (J) at the chosen (v, s), $\hat{y}_S(v, s) = \alpha(v, z)$ for $\mathrm{Law}(Z \mid \nu = v, S = s)$-almost every z. The conditional law of $\alpha(v, Z)$ would then be the Dirac measure at $\hat{y}_S(v, s)$, contradicting $(v, s) \in D_S$. This proves (9). The argument never used uniqueness of $\hat{y}_S$.

For (ii), use the nonnegative summary-conditioned excess risk Γ_S defined above. If $\mathbb{P}((\nu, S) \in D_S) > 0$, part (i) shows that $\Gamma_S > 0$ on an event of positive probability. Consequently, the tower property and integrability give

$$\begin{aligned} R_S - R_{\mathrm{ora}} &= \mathbb{E}[\rho(\hat{y}_S(\nu, S); \nu, Z)] - \mathbb{E}[\rho(\alpha(\nu, Z); \nu, Z)] \\ &= \mathbb{E}[\mathcal{E}_S] = \mathbb{E}[\Gamma_S(\nu, S)] > 0, \end{aligned}$$

where the first equality uses Lemma 2.3(ii) and Lemma 3.3(iii), the second is the definition of $\mathcal{E}_S$ with the integrability of its two terms, and the third is the tower property. The strict inequality holds because the nonnegative Γ_S is positive on an event of positive probability. Hence $G(S) < G(Z)$.

Conversely, suppose $\mathbb{P}((\nu, S) \in D_S) = 0$. By the definition of D_S, the conditional law of $\alpha(\nu, Z)$ given (ν, S) is Dirac almost surely. Since both the conditioning and action spaces are standard Borel, Lemma C.3 gives a jointly measurable map h such that $h(\nu, S) = \alpha(\nu, Z)$ almost surely. By Lemma 3.3(iii), the (ν, S)-measurable predictor $h(\nu, S)$ attains the finite oracle risk. Hence, by (6) and (4) with $W = (\nu, S)$,

$$R_{\mathrm{ora}} \le R_S \le R[h(\nu, S)] = R_{\mathrm{ora}},$$

so $R_S = R_{\mathrm{ora}}$ and $G(S) = G(Z)$. Together with the strict direction, this proves the equivalences. No (P), (R), product-to-joint domination, or uniqueness of a Bayes act given (ν, S) was used. □

C.5 Proof of Corollary 3.12 (Section 3.2)

Proof of Corollary 3.12. For the constant summary S, one has $R_S = R_{\mathrm{obs}}$, and conditioning on (ν, S) is equivalent to conditioning on ν. Thus Theorem 3.11(ii) proves the first equivalence in (10).

For any selected observable-instance Bayes predictor $\hat{y}_{\mathrm{obs}}$, set

$$\mathcal{E}_{\mathrm{obs}} := \rho(\hat{y}_{\mathrm{obs}}(\nu); \nu, Z) - \rho(\alpha(\nu, Z); \nu, Z).$$

For a constant summary S, choosing $\hat{y}_S(\nu, S) = \hat{y}_{\mathrm{obs}}(\nu)$ makes $\mathcal{E}_{\mathrm{obs}}$ the variable $\mathcal{E}_S$ defined in the proof of Theorem 3.11. The π-optimality of α makes $\mathcal{E}_{\mathrm{obs}} \ge 0$ almost surely, and (J) ensures that $\mathcal{E}_{\mathrm{obs}} > 0$ exactly when $\hat{y}_{\mathrm{obs}}(\nu) \ne \alpha(\nu, Z)$, up to a null set. By (2) and Lemma 3.3(iii), $\mathbb{E}[\mathcal{E}_{\mathrm{obs}}] = R_{\mathrm{obs}} - R_{\mathrm{ora}} = G(Z)$. Since $\mathcal{E}_{\mathrm{obs}} \ge 0$ is integrable, $G(Z) > 0$ if and only if $\mathbb{P}(\mathcal{E}_{\mathrm{obs}} > 0) > 0$, which proves the second equivalence. No uniqueness of the observable-instance Bayes act is used. □

C.6 Squared-loss Risk Identity (Section 3.2)

For scalar squared loss with $\mathbb{E}[y_{\mathrm{q}}^2] < \infty$, the conditional means of y_{q} given ν, (ν, S), and (ν, Z) are measurable Bayes predictors with finite risk, as the proof below shows. Hence the requirements of Assumption 2.4 are met as a consequence of the setting of the following lemma, not as an additional assumption.

Lemma C.4 (Squared-loss risk identity)**.** *Under Assumption 2.2, for scalar squared loss with $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ and $\mathbb{E}[y_{\mathrm{q}}^2] < \infty$, let α be a π-optimal prediction-rule family and $S = f(Z)$ any state summary. Then $\alpha(\nu, Z) = \mathbb{E}[y_{\mathrm{q}} \mid \nu, Z]$ and $\hat{y}_S(\nu, S) = \mathbb{E}[\alpha(\nu, Z) \mid \nu, S]$* almost surely*, and*

$$R_S - R_{\mathrm{ora}} = \mathbb{E}[(\alpha(\nu, Z) - \hat{y}_S(\nu, S))^2] = \mathbb{E}[\mathrm{Var}(\alpha(\nu, Z) \mid \nu, S)].$$

In particular, $G(Z) = \mathbb{E}[\mathrm{Var}(\alpha(\nu, Z) \mid \nu)]$.

Proof of Lemma C.4. Let $S = f(Z)$, put $H := (\nu, S)$, and define $M := \mathbb{E}[y_{\mathrm{q}} \mid \nu, Z]$. We first make conditional least-squares optimality explicit. Let W be any of ν, (ν, S), and (ν, Z), and, as in Definition 2.5, compute all conditional expectations given W from a single regular conditional law of y_{q} given W. Since $\mathbb{E}[y_{\mathrm{q}}^2] < \infty$, the conditional second moment $\mathbb{E}[y_{\mathrm{q}}^2 \mid W]$ is finite almost surely. On this common full-measure event, for every $a \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}[(a - y_{\mathrm{q}})^2 \mid W] &= \mathbb{E}[(y_{\mathrm{q}} - \mathbb{E}[y_{\mathrm{q}} \mid W])^2 \mid W] - 2\,(a - \mathbb{E}[y_{\mathrm{q}} \mid W])\,\mathbb{E}[y_{\mathrm{q}} - \mathbb{E}[y_{\mathrm{q}} \mid W] \mid W] \\ &\quad + (a - \mathbb{E}[y_{\mathrm{q}} \mid W])^2 \\ &= \mathbb{E}[(y_{\mathrm{q}} - \mathbb{E}[y_{\mathrm{q}} \mid W])^2 \mid W] + (a - \mathbb{E}[y_{\mathrm{q}} \mid W])^2, \end{aligned}$$

where the cross term vanishes because $\mathbb{E}[y_{\mathrm{q}} - \mathbb{E}[y_{\mathrm{q}} \mid W] \mid W] = 0$. Hence $\mathbb{E}[y_{\mathrm{q}} \mid W]$ is the unique minimizer of the conditional risk over $a \in \mathbb{R}$; the finiteness of the resulting risks follows

from the conditional Jensen argument below. By the conditional form of π-optimality in (8) and conditional least-squares optimality, $\alpha(\nu, Z) = M$; the Bayes predictor defined by (4) satisfies $\hat{y}_S(\nu, S) = \mathbb{E}[y_q \mid H]$ almost surely. Since $S = f(Z)$, the tower property yields

$$\mathbb{E}[y_q \mid H] = \mathbb{E}[\mathbb{E}[y_q \mid \nu, Z] \mid H] = \mathbb{E}[M \mid H].$$

The assumption $\mathbb{E}[y_q^2] < \infty$ and conditional Jensen's inequality imply that M and $\mathbb{E}[M \mid H]$ are square-integrable. Thus $D := M - \mathbb{E}[M \mid H]$ and $y_q - M$ are square-integrable, making all the expectations below finite. Since $y_q - \mathbb{E}[M \mid H] = (y_q - M) + D$, expansion gives

$$R_S - R_{\text{ora}} = \mathbb{E}[(y_q - \mathbb{E}[M \mid H])^2] - \mathbb{E}[(y_q - M)^2] = 2\mathbb{E}[(y_q - M)D] + \mathbb{E}[D^2].$$

The variable D is (ν, Z)-measurable because $H = (\nu, f(Z))$, and $\mathbb{E}[y_q - M \mid \nu, Z] = 0$ almost surely. The product $(y_q - M)D$ is integrable by Cauchy–Schwarz, so conditioning on (ν, Z) proves

$$\mathbb{E}[(y_q - M)D] = \mathbb{E}\big[D\,\mathbb{E}[y_q - M \mid \nu, Z]\big] = 0.$$

Consequently,

$$R_S - R_{\text{ora}} = \mathbb{E}[D^2] = \mathbb{E}\big[\mathbb{E}[(M - \mathbb{E}[M \mid H])^2 \mid H]\big] = \mathbb{E}[\text{Var}(M \mid H)] = \mathbb{E}[\text{Var}(\alpha(\nu, Z) \mid \nu, S)].$$

For the constant summary, $R_S = R_{\text{obs}}$ and conditioning on H is equivalent to conditioning on ν. Therefore $G(Z) = R_{\text{obs}} - R_{\text{ora}} = \mathbb{E}[\text{Var}(\alpha(\nu, Z) \mid \nu)]$. □

Under the finite-second-moment and scalar squared-loss assumptions of this lemma, the conditional variance of $\alpha(\nu, Z)$ given (ν, S) vanishes precisely when its conditional law is a Dirac measure. Thus its expectation is positive precisely when the non-Dirac condition of Theorem 3.11(ii) holds (with (J) satisfied by squared-loss uniqueness); compare Definition 3.9.

C.7 Verification of Example 3.13 (Section 3.2)

Fixed prediction-rule profiles and marginals. We give an explicit numerical construction for Example 3.13. Write $\alpha(v, z_j)$ for the state-conditional Bayes act in Table 1; the prediction-rule profiles are $(0, 0)$ for z_1, z_2, $(0, 1)$ for z_3, and $(1, 1)$ for z_4. Both instances have positive marginal probability, so the CPQ partition of this fixed prediction-rule family has the three classes $\{z_1, z_2\}$, $\{z_3\}$, and $\{z_4\}$. Under both laws, fix $\mathbb{P}(Z = z_j) = 1/4$ for $j = 1, 2, 3, 4$ and $\mathbb{P}(\nu = v_A) = 3/4$, $\mathbb{P}(\nu = v_B) = 1/4$. Also prescribe the same conditional response kernel on *all* instance–state pairs, namely $\mathsf{K}(\cdot \mid v, z) = \mathcal{N}(\alpha(v, z), 1)$. Even for a pair of zero realized probability, its prescribed state-conditional Bayes act remains the value in Table 1. The kernel has finite second moments, so Lemma C.4 applies under both laws.

Two realized joint laws. Let π_{dep} be the law under which $\nu = v_A$ if $Z \in \{z_1, z_2, z_3\}$, and $\nu = v_B$ if $Z = z_4$. Let π_{ind} be the independent law with the same marginals. Table C.1 displays the joint probabilities for both laws, using numerical values throughout.

Table C.1: Two realized joint laws for Example 3.13. Each entry is $\mathbb{P}(Z = z_j, \nu = v)$; both laws have identical state and instance marginals.

	Dependent law π_{dep}		Independent law π_{ind}	
Hidden state	v_A	v_B	v_A	v_B
z_1	$1/4$	0	$3/16$	$1/16$
z_2	$1/4$	0	$3/16$	$1/16$
z_3	$1/4$	0	$3/16$	$1/16$
z_4	0	$1/4$	$3/16$	$1/16$
Instance marginal	$3/4$	$1/4$	$3/4$	$1/4$

Each row sums to $1/4$ under either law, and the column sums are $3/4$ and $1/4$. For example, under independence, $\mathbb{P}(\nu = v_A, Z = z_1) = (3/4)(1/4) = 3/16$. Thus the state and instance marginals, the prescribed conditional response kernel, and the CPQ partition are identical; only the realized joint law changes.

Oracle gain under the dependent law. Put $M := \alpha(\nu, Z) = \mathbb{E}[y_{\mathrm{q}} \mid \nu, Z]$. Under π_{dep}, the conditional state probabilities are

$$\mathbb{P}(Z = z_j \mid \nu = v_A) = \begin{cases} 1/3, & j = 1, 2, 3, \\ 0, & j = 4, \end{cases} \qquad \mathbb{P}(Z = z_4 \mid \nu = v_B) = 1.$$

Since $\alpha(v_A, z_j) = 0$ for $j = 1, 2, 3$ and $\alpha(v_B, z_4) = 1$, we have

$$\mathbb{E}[M \mid \nu = v_A] = 0, \quad \mathrm{Var}(M \mid \nu = v_A) = 0,$$
$$\mathbb{E}[M \mid \nu = v_B] = 1, \quad \mathrm{Var}(M \mid \nu = v_B) = 0.$$

Therefore Lemma C.4 gives

$$G_{\mathrm{dep}}(Z) = \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0.$$

The constant summary achieves oracle risk even though it does not recover the three-valued CPQ.

Oracle gain under the independent law. Under π_{ind}, for both instances, $\mathbb{P}(Z = z_j \mid \nu = v) = 1/4$ for $j = 1, 2, 3, 4$. From Table 1, $\alpha(v_A, Z) = \mathbf{1}_{\{Z = z_4\}}$ and $\alpha(v_B, Z) = \mathbf{1}_{\{Z \in \{z_3, z_4\}\}}$. Consequently,

$$\mathbb{E}[M \mid \nu = v_A] = \frac{1}{4}, \quad \mathrm{Var}(M \mid \nu = v_A) = \frac{1}{4} - \left(\frac{1}{4}\right)^2 = \frac{3}{16},$$
$$\mathbb{E}[M \mid \nu = v_B] = \frac{1}{2}, \quad \mathrm{Var}(M \mid \nu = v_B) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Averaging these conditional variances over the instance marginal yields

$$G_{\mathrm{ind}}(Z) = \frac{3}{4} \cdot \frac{3}{16} + \frac{1}{4} \cdot \frac{1}{4} = \frac{9}{64} + \frac{4}{64} = \frac{13}{64} > 0.$$

Under the specified unit-variance Gaussian kernel, $R_{\mathrm{ora}} = 1$ under each law. The observable-instance risks are therefore $R_{\mathrm{obs}} = 1$ under π_{dep} and $R_{\mathrm{obs}} = 77/64$ under π_{ind}.

Relation to product-to-joint domination. For the dependent law, the pair (v_B, z_1) satisfies

$$\pi_{\mathrm{dep}}(\{(v_B, z_1)\}) = 0, \qquad \pi_{\otimes}(\{(v_B, z_1)\}) = \frac{1}{16} > 0.$$

Hence $\pi_{\otimes} \not\ll \pi_{\mathrm{dep}}$. Under either law, the specified unit-variance Gaussian squared loss satisfies (J). The zero dependent-law gain is also consistent with the conditional-Dirac criterion in (10). For the independent law, $\pi_{\otimes} = \pi_{\mathrm{ind}}$. For this law, the finite-state profiles satisfy (P), and the selected oracle acts are reconstructed from their CPQ labels by (R). Thus Theorem 3.14 also predicts the strictly positive gain calculated above. The failure of product-to-joint domination under the dependent law allows prediction-rule distinctions to be confined to unrealized pairs. It is the conditional support of the realized joint law, rather than dependence itself, that explains the different oracle gains.

Dependence on the selected prediction-rule family. Under π_{dep}, the jointly measurable family $\widetilde{\alpha}(v, z) = \mathbf{1}_{\{v = v_B\}}$ agrees with the tabulated $\alpha(v, z)$ on every realized pair and is therefore also π_{dep}-optimal for the same prescribed conditional response kernel. All states have the same prediction-rule profile under $\widetilde{\alpha}$, whereas the tabulated α has three distinct predictive-equivalence classes. Thus changing a selected π-optimal family on unrealized pairs can change the predictive-equivalence partition without changing risk.

C.8 Proof of Theorem 3.14 (Section 3.2)

Proof of Theorem 3.14. By (J), $\alpha(\nu, Z) = \hat{y}_{\mathrm{ora}}(\nu, Z)$ almost surely. Hence α has finite risk by Assumption 2.4, and Lemma 3.3(iii) gives $R[\alpha(\nu, Z)] = R_{\mathrm{ora}} < \infty$. Write $\lambda_S := \mathrm{Law}(S)$. If $T^\star = g(S)$ almost surely for a measurable g, the refinement conclusion of Proposition 3.8 directly gives $R_S = R_{\mathrm{ora}}$. Provided that the selected α retains finite risk, this sufficient direction does not require (J) or $\pi_{\otimes} \ll \pi$.

For the converse, suppose $R_S = R_{\mathrm{ora}}$, and take the Bayes predictor $\hat{y}_S(\nu, S)$ given (ν, S) from Assumption 2.4 and Definition 2.7. As in the proof of Theorem 3.11, use the state-conditional excess risk

$$\mathcal{E}_S = \rho(\hat{y}_S(\nu, S); \nu, Z) - \rho(\alpha(\nu, Z); \nu, Z).$$

The π-optimality of α implies $\mathcal{E}_S \geq 0$ almost surely. By (2) and the finite risks of $\hat{y}_S$ and α, both state-conditional risk terms are integrable. Consequently,

$$\mathbb{E}[\mathcal{E}_S] = R_S - R[\alpha(\nu, Z)] = R_S - R_{\mathrm{ora}} = 0.$$

Thus $\mathcal{E}_S = 0$ almost surely. Condition (J), applied on its π-full set of singleton minimizers, gives

$$\hat{y}_S(v, f(z)) = \alpha(v, z) \qquad \text{for } \pi\text{-almost every } (v, z).$$

The disagreement set in this display is measurable because both functions are jointly measurable and $\mathcal{A}$ is standard Borel. By $\pi_\otimes \ll \pi$, the equality also holds $\pi_\otimes$-almost surely. Fubini's theorem therefore supplies a measurable set $\mathcal{Z}_1 \subseteq \mathcal{Z}$ of full μ measure such that, for every $z \in \mathcal{Z}_1$,

$$\hat{y}_S(v, f(z)) = \alpha(v, z) \qquad \text{for } \lambda\text{-almost every } v.$$

Let $\mathcal{Z}_0$ be the common μ-full set in (P), and set $A := \mathcal{Z}_0 \cap \mathcal{Z}_1$. For any $z, z' \in A$ with $f(z) = f(z')$, the preceding two λ-almost-everywhere identities yield $\alpha(v, z) = \alpha(v, z')$ for λ-almost every v. Consequently $z \sim_{\mathrm{pred}} z'$, and (P) gives

$$f(z) = f(z') \quad \Longrightarrow \quad \tau(z) = \tau(z') \qquad (z, z' \in A).$$

It remains to turn this fiberwise statement into a *measurable* reconstruction. Since $\mathcal{Z}$ and the value space of S are standard Borel, let $\kappa_s = \mathrm{Law}(Z \mid S = s)$ be a regular conditional law. As $\mu(A) = 1$ and $S = f(Z)$ almost surely, disintegration gives

$$\int \kappa_s\big(A \cap f^{-1}(\{s\})\big)\, \lambda_S(ds) = \mathbb{P}(Z \in A,\ f(Z) = S) = 1.$$

The integrand lies in $[0, 1]$; hence $\kappa_s(A \cap f^{-1}(\{s\})) = 1$ for λ_S-almost every s. On each such fiber, τ is constant by the preceding implication. Therefore its pushforward $(\tau)_{\#}\kappa_s = \mathrm{Law}(T^\star \mid S = s)$ is a Dirac measure for λ_S-almost every s. Since the CPQ label space is standard Borel, Lemma C.3 yields a measurable map g with $T^\star = g(S)$ almost surely.

Finally take S to be constant. Then $R_S = R_{\mathrm{obs}}$, and $T^\star = g(S)$ almost surely for a measurable g if and only if $T^\star$ is almost surely constant. The established equivalence gives $R_{\mathrm{obs}} = R_{\mathrm{ora}}$ if and only if $T^\star$ is almost surely constant. Since $G(Z) = R_{\mathrm{obs}} - R_{\mathrm{ora}} \geq 0$ by (6) and the risks are finite, this is equivalent to the stated $G(Z) > 0$ criterion. □

D Context–Query Predictive Sufficiency and Linear-Gaussian Specialization: Details and Proofs (Section 4)

We prove Theorem 4.1, verify the CPQ minimality hypotheses for the Gaussian model, and derive Corollary 4.2 and the example in Section D.5.

D.1 Proof of Theorem 4.1 (Section 4.1)

Proof of Theorem 4.1.

Part (i): stage-t U-sufficiency and CPQ recovery.

Excess-risk identity and null-space criterion. Fix a stage t and a deterministic query $\boldsymbol{q} \in \mathbb{R}^d$. Since $S = f(Z)$, $\sigma(S) \subseteq \sigma(Z)$, so $\mathcal{G}_t^S \subseteq \mathcal{G}_t^Z$. The conditional Jensen inequality and $\mathbb{E}\|\boldsymbol{w}\|^2 < \infty$ give $\mathbb{E}\|\boldsymbol{m}_t^Z\|^2 \leq \mathbb{E}\|\boldsymbol{w}\|^2 < \infty$ and likewise $\mathbb{E}\|\boldsymbol{m}_t^S\|^2 < \infty$. Thus the residuals and differences below are square-integrable; their products are integrable by the Cauchy–Schwarz inequality.

The excess-risk matrix also has a conditional-covariance representation. By the tower property for $\mathcal{G}_t^S \subseteq \mathcal{G}_t^Z$,

$$\mathbb{E}[\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S] = \mathbb{E}[\mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^Z] \mid \mathcal{G}_t^S] = \mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^S] = \boldsymbol{m}_t^S.$$

Consequently, the definition of conditional covariance, followed by the tower property applied entrywise, yields

$$\begin{aligned} \mathbb{E}[\mathrm{Cov}(\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S)] &= \mathbb{E}\big[\mathbb{E}[(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)^\top \mid \mathcal{G}_t^S]\big] \\ &= \mathbb{E}[(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)^\top] = \mathbf{A}_{S,t}. \end{aligned} \tag{19}$$

This is exactly the conditional-covariance identity stated before Theorem 4.1; it requires no Gaussian or CPQ assumptions.

For the fixed deterministic query $\boldsymbol{q}$, apply Lemma C.4 to the scalar squared-loss prediction problem with observable instance $(C_t, \boldsymbol{q})$, hidden state Z, action space $\mathbb{R}$, and response $Y_{\boldsymbol{q}} = \boldsymbol{q}^\top \boldsymbol{w}$. Since $\boldsymbol{q}$ is deterministic, conditioning on $(C_t, \boldsymbol{q}, Z)$ or $(C_t, \boldsymbol{q}, S)$ is equivalent to conditioning on $\mathcal{G}_t^Z$ or $\mathcal{G}_t^S$, respectively. Since $\boldsymbol{q}$ is fixed and $\mathbb{E}\|\boldsymbol{w}\|^2 < \infty$, $\mathbb{E}[Y_{\boldsymbol{q}}^2] \le \|\boldsymbol{q}\|^2 \mathbb{E}\|\boldsymbol{w}\|^2 < \infty$. By linearity of conditional expectation, the oracle and state-summary Bayes predictions are $\boldsymbol{q}^\top \boldsymbol{m}_t^Z$ and $\boldsymbol{q}^\top \boldsymbol{m}_t^S$, respectively. Write $\boldsymbol{r}_t := \boldsymbol{m}_t^Z - \boldsymbol{m}_t^S$. The lemma gives

$$R_{S,t}(\boldsymbol{q}) - R_{\mathrm{ora},t}(\boldsymbol{q}) = \mathbb{E}[(\boldsymbol{q}^\top \boldsymbol{m}_t^Z - \boldsymbol{q}^\top \boldsymbol{m}_t^S)^2] = \mathbb{E}[(\boldsymbol{q}^\top \boldsymbol{r}_t)^2] = \boldsymbol{q}^\top \mathbb{E}[\boldsymbol{r}_t \boldsymbol{r}_t^\top] \boldsymbol{q} = \boldsymbol{q}^\top \mathbf{A}_{S,t} \boldsymbol{q}. \tag{20}$$

The last equality uses the definition of $\mathbf{A}_{S,t}$. Thus (20) proves the excess-risk identity in (11) without CPQ assumptions.

For the null-space criterion, the definition of $\mathbf{A}_{S,t}$ and $\boldsymbol{r}_t$ gives, for every fixed $\boldsymbol{q} \in \mathbb{R}^d$,

$$\boldsymbol{q}^\top \mathbf{A}_{S,t} \boldsymbol{q} = \mathbb{E}[(\boldsymbol{q}^\top \boldsymbol{r}_t)^2] \ge 0. \tag{21}$$

If the quadratic form in (21) is zero, the nonnegative random variable $(\boldsymbol{q}^\top \boldsymbol{r}_t)^2$ has zero expectation and hence vanishes almost surely. Thus $\boldsymbol{q}^\top \boldsymbol{r}_t = 0$ almost surely, so $\mathbf{A}_{S,t}\boldsymbol{q} = \mathbb{E}[\boldsymbol{r}_t(\boldsymbol{r}_t^\top \boldsymbol{q})] = \mathbf{0}$. These vector expectations exist componentwise by Cauchy–Schwarz and the square-integrability of $\boldsymbol{r}_t$. Conversely, $\mathbf{A}_{S,t}\boldsymbol{q} = \mathbf{0}$ implies $\boldsymbol{q}^\top \mathbf{A}_{S,t} \boldsymbol{q} = 0$. Combining these directions with (20), we obtain

$$R_{S,t}(\boldsymbol{q}) = R_{\mathrm{ora},t}(\boldsymbol{q}) \quad \Longleftrightarrow \quad \boldsymbol{q} \in \ker \mathbf{A}_{S,t}. \tag{22}$$

Requiring (22) for every $\boldsymbol{q} \in U$ is exactly $U \subseteq \ker \mathbf{A}_{S,t}$. This proves the second assertion of (11).

Projected risk and vector–scalar risk equivalence. We next link the scalar risk condition to the projected vector prediction problem. Let $d_U := \dim U$ and choose an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{d_U}$ of U. If $U = \{\mathbf{0}\}$, use the empty basis; all sums below are then zero. The orthogonal projection has the form $P_U = \sum_{j=1}^{d_U} \boldsymbol{u}_j \boldsymbol{u}_j^\top$, where $P_U^\top = P_U$ and $P_U^2 = P_U$. For every $\boldsymbol{v} \in \mathbb{R}^d$, the orthonormality of the basis and the orthogonal projection give

$$\|P_U \boldsymbol{v}\|^2 = \left\| \sum_{j=1}^{d_U} \boldsymbol{u}_j (\boldsymbol{u}_j^\top \boldsymbol{v}) \right\|^2 = \sum_{j=1}^{d_U} (\boldsymbol{u}_j^\top \boldsymbol{v})^2. \tag{23}$$

Applying (23) first with $\boldsymbol{v} = \boldsymbol{w} - \boldsymbol{m}_t^S$ and then with $\boldsymbol{v} = \boldsymbol{w} - \boldsymbol{m}_t^Z$, using $Y_{\boldsymbol{u}_j} = \boldsymbol{u}_j^\top \boldsymbol{w}$, and taking expectations gives

$$\begin{aligned} R_{S,t}^U &= \mathbb{E}\|P_U(\boldsymbol{w} - \boldsymbol{m}_t^S)\|^2 = \mathbb{E}\left[\sum_{j=1}^{d_U} \{\boldsymbol{u}_j^\top (\boldsymbol{w} - \boldsymbol{m}_t^S)\}^2\right] \\ &= \sum_{j=1}^{d_U} \mathbb{E}[(Y_{\boldsymbol{u}_j} - \boldsymbol{u}_j^\top \boldsymbol{m}_t^S)^2] = \sum_{j=1}^{d_U} R_{S,t}(\boldsymbol{u}_j), \\ R_{\mathrm{ora},t}^U &= \mathbb{E}\|P_U(\boldsymbol{w} - \boldsymbol{m}_t^Z)\|^2 = \mathbb{E}\left[\sum_{j=1}^{d_U} \{\boldsymbol{u}_j^\top (\boldsymbol{w} - \boldsymbol{m}_t^Z)\}^2\right] \\ &= \sum_{j=1}^{d_U} \mathbb{E}[(Y_{\boldsymbol{u}_j} - \boldsymbol{u}_j^\top \boldsymbol{m}_t^Z)^2] = \sum_{j=1}^{d_U} R_{\mathrm{ora},t}(\boldsymbol{u}_j). \end{aligned} \tag{24}$$

The passage from the expectation of a finite sum to the sum of expectations uses linearity, and each last equality is the definition of the corresponding scalar risk. Subtracting the two identities in (24) and applying (20) at $\boldsymbol{q} = \boldsymbol{u}_j$ gives

$$\begin{aligned} R_{S,t}^U - R_{\mathrm{ora},t}^U &= \sum_{j=1}^{d_U} [R_{S,t}(\boldsymbol{u}_j) - R_{\mathrm{ora},t}(\boldsymbol{u}_j)] = \sum_{j=1}^{d_U} \boldsymbol{u}_j^\top \mathbf{A}_{S,t} \boldsymbol{u}_j \\ &= \mathrm{tr}(P_U \mathbf{A}_{S,t} P_U) = \mathbb{E}\|P_U \boldsymbol{r}_t\|^2. \end{aligned} \tag{25}$$

We justify the last two equalities separately. For the trace equality in (25), use cyclicity of the trace and the orthogonal projection operator identities:

$$\operatorname{tr}(P_U \mathbf{A}_{S,t} P_U) = \operatorname{tr}(\mathbf{A}_{S,t} P_U^2) = \operatorname{tr}(\mathbf{A}_{S,t} P_U) = \sum_{j=1}^{d_U} \operatorname{tr}(\mathbf{A}_{S,t} \boldsymbol{u}_j \boldsymbol{u}_j^\top) = \sum_{j=1}^{d_U} \boldsymbol{u}_j^\top \mathbf{A}_{S,t} \boldsymbol{u}_j.$$

The last step uses $\operatorname{tr}(\boldsymbol{a}\boldsymbol{b}^\top) = \boldsymbol{b}^\top \boldsymbol{a}$. For the final equality in (25), use the definition $\mathbf{A}_{S,t} = \mathbb{E}[\boldsymbol{r}_t \boldsymbol{r}_t^\top]$, the fact that P_U is deterministic, and $\operatorname{tr}(\boldsymbol{a}\boldsymbol{a}^\top) = \|\boldsymbol{a}\|^2$:

$$\operatorname{tr}(P_U \mathbf{A}_{S,t} P_U) = \mathbb{E}[\operatorname{tr}((P_U \boldsymbol{r}_t)(P_U \boldsymbol{r}_t)^\top)] = \mathbb{E}\|P_U \boldsymbol{r}_t\|^2.$$

To determine when the projected excess risk vanishes, we use specifically the *quadratic-form sum* in (25):

$$R_{S,t}^U - R_{\mathrm{ora},t}^U = \sum_{j=1}^{d_U} \boldsymbol{u}_j^\top \mathbf{A}_{S,t} \boldsymbol{u}_j.$$

By (21) with $\boldsymbol{q} = \boldsymbol{u}_j$, every summand is nonnegative. Thus this sum is zero exactly when each summand vanishes. Applying (22) and $U = \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_U}\}$ gives

$$\begin{aligned} R_{S,t}^U = R_{\mathrm{ora},t}^U &\iff \boldsymbol{u}_j^\top \mathbf{A}_{S,t} \boldsymbol{u}_j = 0 \quad (j = 1, \ldots, d_U) \\ &\iff \boldsymbol{u}_j \in \ker \mathbf{A}_{S,t} \quad (j = 1, \ldots, d_U) \\ &\iff U \subseteq \ker \mathbf{A}_{S,t} \\ &\iff R_{S,t}(\boldsymbol{q}) = R_{\mathrm{ora},t}(\boldsymbol{q}) \quad (\forall \boldsymbol{q} \in U). \end{aligned}$$

The equivalence between the basis vectors and the whole of U uses the linearity of the null space. The final equivalence again uses (22) for every fixed $\boldsymbol{q} \in U$. This also covers $U = \{\mathbf{0}\}$ by the empty-sum convention. Using a finite basis avoids taking an uncountable intersection of probability-one events over all query vectors.

Verification of the hypotheses of Theorem 3.14.

We verify the projected-problem regularity, Bayes-predictor requirements, and realized uniqueness (J) needed to invoke Theorem 3.14; conditions (P), (R), and product-to-joint domination are imposed separately when invoked. The stage-t projected problem, as defined in Section 4.1, has $\boldsymbol{Y}^U := P_U \boldsymbol{w}$ and $L(\boldsymbol{a}, \boldsymbol{y}) = \|\boldsymbol{a} - \boldsymbol{y}\|^2$ for $\boldsymbol{a}, \boldsymbol{y} \in U$. Because the context and state spaces are standard Borel and U is a finite-dimensional Euclidean (hence standard-Borel) space, the measurability requirement in Assumption 2.2 is satisfied. A regular conditional law

$$\mathsf{K}_{t,U}(d\boldsymbol{y} \mid c, z) := \mathbb{P}(\boldsymbol{Y}^U \in d\boldsymbol{y} \mid C_t = c, Z = z)$$

can be selected jointly measurably in (c, z); it is a version of the prescribed conditional response kernel under π_t. Its values on a π_t-null set may be fixed arbitrarily; as in Definition 2.1, such values are not determined by the realized joint law. Since $\mathbb{E}\|\boldsymbol{Y}^U\|^2 \le \mathbb{E}\|\boldsymbol{w}\|^2 < \infty$, define the measurable set

$$E_t := \left\{ (c, z) : \int_U \|\boldsymbol{y}\|^2 \, \mathsf{K}_{t,U}(d\boldsymbol{y} \mid c, z) < \infty, \boldsymbol{\alpha}_{t,U}(c, z) = \int_U \boldsymbol{y} \, \mathsf{K}_{t,U}(d\boldsymbol{y} \mid c, z) \right\}. \tag{26}$$

Here U is fixed throughout this verification, and the dependence on U is suppressed in the notation E_t. The map $\boldsymbol{\alpha}_{t,U}$ is the prediction-rule family selected in Section 4.1. The first condition in (26) holds π_t-almost everywhere: the conditional second moment is nonnegative and its expectation equals $\mathbb{E}\|\boldsymbol{Y}^U\|^2 < \infty$ by the tower property. The second also holds π_t-almost everywhere, since both the selected rule and the displayed kernel barycenter represent $\mathbb{E}[\boldsymbol{Y}^U \mid C_t, Z]$ under $\mathbb{P}$. Thus $\pi_t(E_t) = 1$, as the intersection of two full-measure sets. Crucially, the set in (26) does not depend on the candidate action $\boldsymbol{a}$.

We verify the realized uniqueness condition (J) *pointwise on the common full-measure set in* (26). Fix any $(c, z) \in E_t$. For an *arbitrary* $\boldsymbol{a} \in U$, the state-conditional risk is

$$\rho(\boldsymbol{a}; c, z) = \int_U \|\boldsymbol{y} - \boldsymbol{a}\|^2 \, \mathsf{K}_{t,U}(d\boldsymbol{y} \mid c, z).$$

Writing $\boldsymbol{a}_0 \coloneqq \boldsymbol{\alpha}_{t,U}(c,z)$, the Euclidean squared-norm identity and its integral give

$$\begin{aligned}\rho(\boldsymbol{a};c,z) &= \int_U \big[\|\boldsymbol{y}-\boldsymbol{a}_0\|^2 - 2(\boldsymbol{y}-\boldsymbol{a}_0)^\top(\boldsymbol{a}-\boldsymbol{a}_0) + \|\boldsymbol{a}-\boldsymbol{a}_0\|^2\big]\mathsf{K}_{t,U}(d\boldsymbol{y}\mid c,z)\\ &= \rho(\boldsymbol{a}_0;c,z) - 2\left(\int_U (\boldsymbol{y}-\boldsymbol{a}_0)\mathsf{K}_{t,U}(d\boldsymbol{y}\mid c,z)\right)^\top(\boldsymbol{a}-\boldsymbol{a}_0) + \|\boldsymbol{a}-\boldsymbol{a}_0\|^2\\ &= \rho(\boldsymbol{a}_0;c,z) + \|\boldsymbol{a}-\boldsymbol{a}_0\|^2.\end{aligned}$$

The last equality uses the barycenter identity in (26), so the integral of $\boldsymbol{y}-\boldsymbol{a}_0$ is $\mathbf{0}$. This calculation holds for *every* $\boldsymbol{a}\in U$ at the *same* $(c,z)\in E_t$, because E_t was chosen independently of $\boldsymbol{a}$. Since the squared norm is zero precisely when $\boldsymbol{a}=\boldsymbol{a}_0$, we obtain

$$\arg\min_{\boldsymbol{a}\in U}\rho(\boldsymbol{a};c,z) = \{\boldsymbol{\alpha}_{t,U}(c,z)\} \qquad \text{for every } (c,z)\in E_t.$$

Thus the selected prediction-rule family is π_t-optimal and satisfies (J), including its quantifier over all actions on one full-measure set.

The oracle and state-summary Bayes predictions, evaluated at (C_t,Z) and (C_t,S), are projected conditional means:

$$\begin{aligned}\boldsymbol{\alpha}_{t,U}(C_t,Z) &= \mathbb{E}[P_U\boldsymbol{w}\mid C_t,Z] = P_U\boldsymbol{m}_t^Z,\\ \mathbb{E}[P_U\boldsymbol{w}\mid C_t,S] &= P_U\mathbb{E}[\boldsymbol{w}\mid C_t,S] = P_U\boldsymbol{m}_t^S \qquad \text{almost surely.}\end{aligned} \tag{27}$$

The first identity uses the selected family's conditional-mean representation in (26); the second uses conditional-expectation linearity and the definition of $\boldsymbol{m}_t^S$. Standard-Borel regularity supplies jointly measurable versions of these conditional means. Since the zero vector is an available action, the Bayes risks $R^U_{S,t}$ and $R^U_{\mathrm{ora},t}$ for the predictors in (27) are bounded by $\mathbb{E}\|P_U\boldsymbol{w}\|^2 \le \mathbb{E}\|\boldsymbol{w}\|^2 < \infty$.

The observable-instance Bayes prediction is likewise the conditional mean $\mathbb{E}[P_U\boldsymbol{w}\mid C_t] = P_U\mathbb{E}[\boldsymbol{w}\mid C_t]$. Under squared Euclidean loss, each of these three conditional means minimizes the corresponding conditional risk. Since the zero vector is an available action, the observable-instance Bayes risk is also bounded by $\mathbb{E}\|P_U\boldsymbol{w}\|^2 < \infty$. Thus all three predictors required by Assumption 2.4 exist measurably and have finite risk.

Let $\mathsf{K}^S_{t,U}(d\boldsymbol{y}\mid c,s)$ be a regular version of $\mathrm{Law}(\boldsymbol{Y}^U\mid C_t=c,S=s)$. The stage-$t$ projected problem satisfies the standard-Borel and kernel-consistency assumptions verified above. Thus (1) yields, for each Borel $B\subseteq U$,

$$\mathsf{K}^S_{t,U}(B\mid C_t,S) = \int_{\mathcal{Z}} \mathsf{K}_{t,U}(B\mid C_t,z)\,\mathbb{P}(Z\in dz\mid C_t,S). \tag{28}$$

The displayed identity holds almost surely and, by the probability-kernel version argument in the proof of Lemma 2.3, holds as an equality of conditional kernels outside one null set. The mixing law is $\mathrm{Law}(Z\mid C_t,S)$, not $\mathrm{Law}(Z)$; no context–state independence is required. The summary-risk identity (3) applies to every measurable action based on (C_t,S), without a new disintegration or interchange-of-integrals argument.

The U-CPQ $T^\star_{t,U}$ is assumed to exist and to satisfy (P) and (R) for exactly the selected family $\boldsymbol{\alpha}_{t,U}$ just used. Finally, this part of the theorem separately assumes $\mathrm{Law}(C_t)\otimes\mathrm{Law}(Z)\ll\mathrm{Law}(C_t,Z)$, which is the product-to-joint domination condition required in Theorem 3.14. All hypotheses of that theorem are therefore verified, and its conclusion reads

$$R^U_{S,t} = R^U_{\mathrm{ora},t} \quad\Longleftrightarrow\quad T^\star_{t,U} = g_t(S) \text{ almost surely for some measurable } g_t.$$

Together with the vector–scalar risk equivalence and the null-space criterion above, this proves the final assertion of (i).

Part (ii): Persistence of U-sufficiency and transfer loss.

Persistence for an arbitrary state summary. Fix $t<u$ and assume $U\subseteq\ker\mathbf{A}_{S,t}$. By the null-space criterion (11) and the projected excess-risk identity (25), the assumed $U\subseteq\ker\mathbf{A}_{S,t}$ gives

$$R^U_{S,t} - R^U_{\mathrm{ora},t} = \mathbb{E}\|P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S)\|^2 = 0.$$

Since the squared norm is nonnegative, zero expectation implies that it is zero almost surely. We record the resulting stage-t projected equality for later use:

$$P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S) = \mathbf{0} \quad \text{almost surely.} \tag{29}$$

Using the definitions of the increments, expand the difference at stage u. The last equality uses (29):

$$\begin{aligned} P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S) &= P_U[(\boldsymbol{m}_t^Z + \boldsymbol{\Delta}_{t:u}^Z) - (\boldsymbol{m}_t^S + \boldsymbol{\Delta}_{t:u}^S)] \\ &= P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S) + P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S) = P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S) \quad \text{almost surely.} \end{aligned} \tag{30}$$

Applying (11) and (25), now at stage u, gives

$$\begin{aligned} U \subseteq \ker \mathbf{A}_{S,u} &\iff \mathbb{E}\|P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S)\|^2 = 0 \iff P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S) = \mathbf{0} \quad \text{almost surely} \\ &\iff P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S) = \mathbf{0} \quad \text{almost surely.} \end{aligned}$$

The second equivalence uses the nonnegative squared norm, and the last is (30). This proves (12) in both directions. No CPQ existence or CPQ minimality hypotheses are needed for this persistence equivalence.

The conditional mean of the projected oracle increment. Recall $\boldsymbol{\Delta}_{t:u}^{Z,U} = P_U \boldsymbol{\Delta}_{t:u}^Z$ and $\mathcal{G}_u^S = \mathcal{F}_u \vee \sigma(S)$. Because $S = f(Z)$, $\mathcal{G}_u^S \subseteq \mathcal{G}_u^Z$. The tower property gives

$$\mathbb{E}[\boldsymbol{m}_u^Z \mid \mathcal{G}_u^S] = \mathbb{E}[\mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_u^Z] \mid \mathcal{G}_u^S] = \mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_u^S] = \boldsymbol{m}_u^S. \tag{31}$$

For the *earlier* posterior mean, we cannot simply replace $\mathbb{E}[\boldsymbol{m}_t^Z \mid \mathcal{G}_u^S]$ by $\boldsymbol{m}_t^S$ in every coordinate. In general, $\mathcal{G}_t^Z$ and $\mathcal{G}_u^S$ are not nested, so the tower property does not give this equality; the later context may provide additional information. Instead, the stage-t projected equality (29) gives $P_U \boldsymbol{m}_t^Z = P_U \boldsymbol{m}_t^S$ almost surely. Since $\boldsymbol{m}_t^S$ is $\mathcal{G}_t^S$-measurable and P_U is deterministic, $P_U \boldsymbol{m}_t^S$ is also $\mathcal{G}_t^S$-measurable. Moreover, $\mathcal{G}_t^S \subseteq \mathcal{G}_u^S$, so it is $\mathcal{G}_u^S$-measurable. Therefore, by (29),

$$\mathbb{E}[P_U \boldsymbol{m}_t^Z \mid \mathcal{G}_u^S] = \mathbb{E}[P_U \boldsymbol{m}_t^S \mid \mathcal{G}_u^S] = P_U \boldsymbol{m}_t^S. \tag{32}$$

Using (31) and (32), together with conditional-expectation linearity and deterministic P_U, gives

$$\begin{aligned} \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S] &= \mathbb{E}[P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_t^Z) \mid \mathcal{G}_u^S] = P_U \mathbb{E}[\boldsymbol{m}_u^Z \mid \mathcal{G}_u^S] - \mathbb{E}[P_U \boldsymbol{m}_t^Z \mid \mathcal{G}_u^S] \\ &= P_U \boldsymbol{m}_u^S - P_U \boldsymbol{m}_t^S = P_U \boldsymbol{\Delta}_{t:u}^S. \end{aligned} \tag{33}$$

Thus $P_U \boldsymbol{\Delta}_{t:u}^S$ is the conditional mean, given $\mathcal{G}_u^S$, of the projected oracle increment $\boldsymbol{\Delta}_{t:u}^{Z,U}$. No identity for unprojected increments is asserted. Substituting (33) into the increment identity (30) gives the key residual representation:

$$P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S) = P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S) = \boldsymbol{\Delta}_{t:u}^{Z,U} - \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S] \quad \text{almost surely.} \tag{34}$$

Thus the prediction discrepancy at stage u is exactly the part of the projected oracle increment remaining after conditioning on the information available from (C_u, S).

Conditional covariance of the projected oracle increment. Put $\boldsymbol{e}_{t:u} := \boldsymbol{\Delta}_{t:u}^{Z,U} - \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S]$. By the definition of $\mathbf{A}_{S,u}$, the symmetry of P_U, and the residual representation (34),

$$P_U \mathbf{A}_{S,u} P_U = \mathbb{E}[(P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S))(P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S))^\top] = \mathbb{E}[\boldsymbol{e}_{t:u} \boldsymbol{e}_{t:u}^\top].$$

By definition, conditional covariance centers the random vector at its conditional mean:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S) &= \mathbb{E}\big[(\boldsymbol{\Delta}_{t:u}^{Z,U} - \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S])(\boldsymbol{\Delta}_{t:u}^{Z,U} - \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S])^\top \mid \mathcal{G}_u^S\big] \\ &= \mathbb{E}[\boldsymbol{e}_{t:u} \boldsymbol{e}_{t:u}^\top \mid \mathcal{G}_u^S]. \end{aligned} \tag{35}$$

Taking expectation entry by entry in (35), using the tower property and the residual representation (34), yields

$$\mathbb{E}[\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)] = \mathbb{E}[\boldsymbol{e}_{t:u} \boldsymbol{e}_{t:u}^\top] = P_U \mathbf{A}_{S,u} P_U. \tag{36}$$

Exact excess risk for any stage-t U-sufficient summary. For an arbitrary summary S satisfying stage-t U-sufficiency, applying the expected-norm form of (25) at stage u, followed by (30) and (34), gives

$$R_{S,u}^U - R_{\mathrm{ora},u}^U = \mathbb{E}\|P_U(\boldsymbol{m}_u^Z - \boldsymbol{m}_u^S)\|^2 = \mathbb{E}\|P_U(\boldsymbol{\Delta}_{t:u}^Z - \boldsymbol{\Delta}_{t:u}^S)\|^2 = \mathbb{E}\|\boldsymbol{e}_{t:u}\|^2. \tag{37}$$

To express the final term of (37) as a trace, use (35) and $\|\boldsymbol{e}\|^2 = \mathrm{tr}(\boldsymbol{e}\boldsymbol{e}^\top)$, conditional-expectation linearity, and the tower property:

$$\mathbb{E}\|\boldsymbol{e}_{t:u}\|^2 = \mathbb{E}[\mathrm{tr}(\boldsymbol{e}_{t:u}\boldsymbol{e}_{t:u}^\top)] = \mathbb{E}[\mathrm{tr}\,\mathbb{E}[\boldsymbol{e}_{t:u}\boldsymbol{e}_{t:u}^\top \mid \mathcal{G}_u^S]] = \mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)]. \tag{38}$$

Together with (36), (37), and (38), these identities establish (13) for any stage-t U-sufficient S. For a deterministic $\boldsymbol{q} \in U$, use the scalar identity (11) at stage u and $P_U\boldsymbol{q} = \boldsymbol{q}$. The residual representation (34) then gives

$$R_{S,u}(\boldsymbol{q}) - R_{\mathrm{ora},u}(\boldsymbol{q}) = \boldsymbol{q}^\top \mathbf{A}_{S,u}\boldsymbol{q} = \mathbb{E}[(\boldsymbol{q}^\top \boldsymbol{e}_{t:u})^2] = \mathbb{E}[\mathrm{Var}(\boldsymbol{q}^\top \boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S)].$$

For the last equality, use $\boldsymbol{q}^\top \boldsymbol{e}_{t:u} = \boldsymbol{q}^\top \boldsymbol{\Delta}_{t:u}^{Z,U} - \mathbb{E}[\boldsymbol{q}^\top \boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S]$, which follows from the residual identity (34) and deterministic $\boldsymbol{q}$. Conditioning its square on $\mathcal{G}_u^S$ gives the conditional variance, and taking expectation gives the displayed formula.

Why a stage-t U-CPQ is stage-t U-sufficient. This also follows by applying Proposition 3.8 to the stage-t projected problem; we verify it directly here using (R) and the tower property. Now take $S = T_{t,U}^\star$, assumed to exist with conditions (P) and (R), for the CPQ-specific conclusions. Unlike the CPQ recovery equivalence in Part (i), stage-t U-sufficiency follows from (R) alone and does not require product-to-joint domination. Condition (R) provides a measurable reconstruction map b_t satisfying

$$P_U\boldsymbol{m}_t^Z = b_t(C_t, S) \quad \text{almost surely.} \tag{39}$$

By (39), $P_U\boldsymbol{m}_t^Z$ admits a version measurable with respect to $\mathcal{G}_t^S = \mathcal{F}_t \vee \sigma(S)$. Therefore,

$$\mathbb{E}[P_U\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S] = \mathbb{E}[b_t(C_t, S) \mid \mathcal{G}_t^S] = b_t(C_t, S) = P_U\boldsymbol{m}_t^Z \quad \text{almost surely.}$$

On the other hand, using the tower property because $\mathcal{G}_t^S \subseteq \mathcal{G}_t^Z$,

$$\mathbb{E}[P_U\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S] = P_U\mathbb{E}[\mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^Z] \mid \mathcal{G}_t^S] = P_U\mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_t^S] = P_U\boldsymbol{m}_t^S. \tag{40}$$

Comparing the two expressions for $\mathbb{E}[P_U\boldsymbol{m}_t^Z \mid \mathcal{G}_t^S]$ gives $P_U\boldsymbol{m}_t^Z = P_U\boldsymbol{m}_t^S$ almost surely. This verifies the stage-t projected equality (29) for $S = T_{t,U}^\star$, without assuming it separately. The summary S is held fixed when the context is extended to stage u.

Transfer loss for the fixed stage-t U-CPQ. For $S = T_{t,U}^\star$, the general excess risk in (37) specializes to

$$\mathcal{L}_{t\to u}(U) = R_{S,u}^U - R_{\mathrm{ora},u}^U = \mathbb{E}\|\boldsymbol{e}_{t:u}\|^2.$$

The trace and matrix identities remain unchanged, so the CPQ transfer identities follow from the general formulas.

When is the transfer loss zero? By the final equality in (37), the transfer loss is the expectation of the integrable, nonnegative variable $\|\boldsymbol{e}_{t:u}\|^2$. Hence

$$\begin{aligned} \mathcal{L}_{t\to u}(U) = 0 &\iff \mathbb{E}\|\boldsymbol{e}_{t:u}\|^2 = 0 \iff \boldsymbol{e}_{t:u} = \mathbf{0} \quad \text{almost surely} \\ &\iff \boldsymbol{\Delta}_{t:u}^{Z,U} = \mathbb{E}[\boldsymbol{\Delta}_{t:u}^{Z,U} \mid \mathcal{G}_u^S] \quad \text{almost surely.} \end{aligned} \tag{41}$$

The conditional expectation on the right of (41) is $\mathcal{G}_u^S$-measurable. Thus the last equality says that $\boldsymbol{\Delta}_{t:u}^{Z,U}$ has a $\mathcal{G}_u^S$-measurable version. Conversely, if $\boldsymbol{\Delta}_{t:u}^{Z,U}$ has such a version, conditioning on $\mathcal{G}_u^S$ leaves it unchanged almost surely, and (41) gives $\mathcal{L}_{t\to u}(U) = 0$. This proves the zero-transfer-loss measurability criterion, with measurability interpreted up to almost-sure modification.

Recovery of the stage-u U-CPQ. Suppose a stage-u U-CPQ $T_{u,U}^\star$ exists and satisfies (P) and (R). First assume that there is a measurable g_u for which $T_{u,U}^\star = g_u(S)$ almost surely. The stage-u reconstruction condition (R) supplies a measurable reconstruction map b_u such that

$$P_U\boldsymbol{m}_u^Z = b_u(C_u, T_{u,U}^\star) = b_u(C_u, g_u(S)) \quad \text{almost surely.}$$

The last expression is $\mathcal{G}_u^S$-measurable. Since $\mathcal{G}_u^S \subseteq \mathcal{G}_u^Z$, the tower property gives

$$\mathbb{E}[P_U\boldsymbol{m}_u^Z \mid \mathcal{G}_u^S] = P_U\mathbb{E}[\boldsymbol{w} \mid \mathcal{G}_u^S] = P_U\boldsymbol{m}_u^S.$$

But $P_U\boldsymbol{m}_u^Z$ already has a $\mathcal{G}_u^S$-measurable version, so the same conditional expectation equals $P_U\boldsymbol{m}_u^Z$ almost surely. We conclude that $P_U\boldsymbol{m}_u^Z = P_U\boldsymbol{m}_u^S$ almost surely, and hence $\mathcal{L}_{t\to u}(U) = 0$ by (25)

at stage u. This direction uses (R), but not (J) or product-to-joint domination. Under the stated (P) and (R) and the stage-u projected-problem regularity verified as in Part (i), with t replaced by u, the equality $R^U_{S,u} = R^U_{\mathrm{ora},u}$ is also the refinement conclusion of Proposition 3.8, with $T^\star_{u,U} = g_u(S)$.

For the reverse implication, suppose $\mathcal{L}_{t\to u}(U) = 0$. This is exactly equality of the stage-u Bayes risks for the projected response $P_U\boldsymbol{w}$ under squared Euclidean loss. The assumptions now additionally include product-to-joint domination $\mathrm{Law}(C_u) \otimes \mathrm{Law}(Z) \ll \mathrm{Law}(C_u, Z)$. The stage-$u$ projected problem is a standard-Borel squared-loss problem with finite second moments. The Bayes-predictor existence and finite-risk requirements were verified in Part (i) and apply with t replaced by u; realized uniqueness (J) also holds under squared Euclidean loss, as verified there. Together with (P), (R), and the stated domination, all hypotheses of Theorem 3.14 hold at stage u. That theorem therefore gives a measurable g_u such that $T^\star_{u,U} = g_u(S)$ almost surely. This proves (14), with the assumptions for each direction kept separate.

U-sufficiency need not hold at intermediate stages. To justify the final assertion of the theorem, take $d = 2$, $\boldsymbol{u}_1 := (0,1)^\top$, $U = \mathrm{span}\{\boldsymbol{u}_1\}$, and the constant summary $S \equiv 0$. Let Z and ζ be independent uniform random signs, and let ξ be an independent sign with $\mathbb{P}(\xi = 1) = 3/4$ and $\mathbb{P}(\xi = -1) = 1/4$; then $\mathbb{E}[\xi] = 1/2$. Set $\boldsymbol{w} = (w_1, w_2)^\top := (\zeta, Z\zeta\xi)^\top$. Consider three demonstrations with deterministic inputs $\boldsymbol{x}_1 := \mathbf{0}$, $\boldsymbol{x}_2 := (1,0)^\top$, and $\boldsymbol{x}_3 := \boldsymbol{u}_1$, and noiseless responses $y_i = \boldsymbol{x}_i^\top \boldsymbol{w}$. Consequently, C_1 is constant, C_2 reveals ζ, and C_3 reveals both ζ and w_2. Independence and $\mathbb{E}[Z] = \mathbb{E}[\zeta] = 0$ give $\mathbb{E}[w_2 \mid Z] = 0$ and $\mathbb{E}[w_2 \mid \zeta] = 0$, whereas $\mathbb{E}[w_2 \mid Z, \zeta] = Z\zeta/2$; finally, w_2 is $\mathcal{F}_3$-measurable. Therefore,

$$P_U\boldsymbol{m}_1^Z = P_U\boldsymbol{m}_1^S = \mathbf{0}, \qquad P_U\boldsymbol{m}_2^Z = \tfrac{1}{2}Z\zeta\,\boldsymbol{u}_1, \;\; P_U\boldsymbol{m}_2^S = \mathbf{0}, \qquad P_U\boldsymbol{m}_3^Z = P_U\boldsymbol{m}_3^S = w_2\boldsymbol{u}_1.$$

By the projected excess-risk identity (25),

$$R^U_{S,1} - R^U_{\mathrm{ora},1} = 0, \qquad R^U_{S,2} - R^U_{\mathrm{ora},2} = \mathbb{E}[(Z\zeta/2)^2] = \frac{1}{4}, \qquad R^U_{S,3} - R^U_{\mathrm{ora},3} = 0.$$

Thus S is stage-1 U-sufficient and stage-3 U-sufficient, but not stage-2 U-sufficient. Moreover, S is a stage-1 U-CPQ: selecting $\boldsymbol{\alpha}_{1,U} \equiv \mathbf{0}$, a version of $P_U\boldsymbol{m}_1^Z = \mathbf{0}$, makes all states predictively equivalent (Definition 3.4), so (P) holds, and (R) holds with $b_1 \equiv \mathbf{0}$. Finally, $\mathbb{P}(Z = z, \zeta = z', w_2 = z'') = \frac{1}{4}\mathbb{P}(\xi = zz'z'') > 0$ for all $z, z', z'' \in \{-1, 1\}$, so product-to-joint domination holds at all three stages. □

D.2 Gaussian posterior formulas and the CPQ minimality hypotheses (Section 4.2)

The model is defined in Section 4.2. We first derive the state-conditional Gaussian posterior mean and covariance, followed by the posterior state probabilities given a state summary. We then verify product-to-joint domination and identify the Gaussian projected prediction-rule profile.

State-conditional Gaussian posterior. Fix a state k. The observation equation is $\boldsymbol{y}_t = \mathbf{X}_t\boldsymbol{w} + \boldsymbol{\varepsilon}_t$. Conditional on $Z = k$, $\boldsymbol{w}$ has mean $\mathbf{0}$ and covariance $\mathbf{V}_k$; $\boldsymbol{\varepsilon}_t$ has mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}_t$ and is independent of $\boldsymbol{w}$. Since $\mathbf{X}_t$ is deterministic, the state-conditional observation moments are

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}_t \mid Z = k] &= \mathbf{X}_t\mathbb{E}[\boldsymbol{w} \mid Z = k] + \mathbb{E}[\boldsymbol{\varepsilon}_t \mid Z = k] = \mathbf{0}, \\
\boldsymbol{\Sigma}_{k,t} &:= \mathrm{Cov}(\boldsymbol{y}_t \mid Z = k) \\
&= \mathbf{X}_t\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top \mid Z = k]\mathbf{X}_t^\top + \mathbf{X}_t\mathbb{E}[\boldsymbol{w}\boldsymbol{\varepsilon}_t^\top \mid Z = k] \\
&\quad + \mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{w}^\top \mid Z = k]\mathbf{X}_t^\top \quad + \mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top \mid Z = k] \\
&= \mathbf{X}_t\mathrm{Cov}(\boldsymbol{w} \mid Z = k)\mathbf{X}_t^\top + \mathrm{Cov}(\boldsymbol{\varepsilon}_t \mid Z = k) = \mathbf{X}_t\mathbf{V}_k\mathbf{X}_t^\top + \sigma^2\mathbf{I}_t.
\end{aligned} \tag{42}$$

In the covariance expansion, both cross terms vanish because $\mathbb{E}[\boldsymbol{w}\boldsymbol{\varepsilon}_t^\top \mid Z = k] = \mathbf{0}$ by independence and zero means. Linear transformations and sums of independent Gaussian vectors are Gaussian, so $\boldsymbol{y}_t \mid Z = k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{k,t})$. For any nonzero $\boldsymbol{a} \in \mathbb{R}^t$,

$$\boldsymbol{a}^\top\boldsymbol{\Sigma}_{k,t}\boldsymbol{a} = (\mathbf{X}_t^\top\boldsymbol{a})^\top\mathbf{V}_k(\mathbf{X}_t^\top\boldsymbol{a}) + \sigma^2\|\boldsymbol{a}\|^2 > 0,$$

where $\mathbf{V}_k \succ 0$ and $\sigma^2 > 0$. Thus $\boldsymbol{\Sigma}_{k,t} \succ 0$.

The same observation equation gives the state-conditional cross-covariance

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{w}, \boldsymbol{y}_t \mid Z = k) &= \mathbb{E}[\boldsymbol{w}(\mathbf{X}_t\boldsymbol{w} + \boldsymbol{\varepsilon}_t)^\top \mid Z = k] = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top \mid Z = k]\mathbf{X}_t^\top + \mathbb{E}[\boldsymbol{w}\boldsymbol{\varepsilon}_t^\top \mid Z = k] \\ &= \mathbf{V}_k\mathbf{X}_t^\top.\end{aligned}$$

To justify joint Gaussianity, observe that, given $Z = k$, $\boldsymbol{w}$ and $\boldsymbol{\varepsilon}_t$ are independent Gaussian vectors. Their concatenation $(\boldsymbol{w}, \boldsymbol{\varepsilon}_t)^\top$ is therefore jointly Gaussian, with mean zero and block-diagonal covariance $\mathrm{diag}(\mathbf{V}_k, \sigma^2\mathbf{I}_t)$. Equivalently, in block form,

$$\mathrm{Cov}\left(\begin{pmatrix}\boldsymbol{w} \\ \boldsymbol{\varepsilon}_t\end{pmatrix}\middle| Z = k\right) = \begin{pmatrix}\mathbf{V}_k & 0 \\ 0 & \sigma^2\mathbf{I}_t\end{pmatrix}.$$

The observation equation can be written as the deterministic linear map

$$\begin{pmatrix}\boldsymbol{w} \\ \boldsymbol{y}_t\end{pmatrix} = \mathbf{L}_t\begin{pmatrix}\boldsymbol{w} \\ \boldsymbol{\varepsilon}_t\end{pmatrix}, \qquad \mathbf{L}_t \coloneqq \begin{pmatrix}\mathbf{I}_d & 0 \\ \mathbf{X}_t & \mathbf{I}_t\end{pmatrix}.$$

A deterministic linear transformation preserves joint Gaussianity. Thus $(\boldsymbol{w}, \boldsymbol{y}_t)$ given $Z = k$ is jointly Gaussian with mean zero. Applying the covariance transformation identity to the deterministic map $\mathbf{L}_t$ gives

$$\mathrm{Cov}\left(\begin{pmatrix}\boldsymbol{w} \\ \boldsymbol{y}_t\end{pmatrix}\middle| Z = k\right) = \mathbf{L}_t\,\mathrm{diag}(\mathbf{V}_k, \sigma^2\mathbf{I}_t)\mathbf{L}_t^\top = \begin{pmatrix}\mathbf{V}_k & \mathbf{V}_k\mathbf{X}_t^\top \\ \mathbf{X}_t\mathbf{V}_k & \boldsymbol{\Sigma}_{k,t}\end{pmatrix}.$$

The upper-left block is the prior covariance of $\boldsymbol{w}$, the off-diagonal blocks are its cross-covariances with $\boldsymbol{y}_t$, and the lower-right block is the observation covariance in (42). For centered jointly Gaussian vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with invertible $\mathrm{Cov}(\boldsymbol{b})$, the conditional mean and covariance are

$$\begin{aligned}\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{b}] &= \mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b})\mathrm{Cov}(\boldsymbol{b})^{-1}\boldsymbol{b}, \\ \mathrm{Cov}(\boldsymbol{a} \mid \boldsymbol{b}) &= \mathrm{Cov}(\boldsymbol{a}) - \mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b})\mathrm{Cov}(\boldsymbol{b})^{-1}\mathrm{Cov}(\boldsymbol{b}, \boldsymbol{a}).\end{aligned}$$

Apply these under the conditional law given $Z=k$, taking $\boldsymbol{a}=\boldsymbol{w}$ and $\boldsymbol{b}=\boldsymbol{y}_t$, and note the transpose relation $\mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b}) = \mathrm{Cov}(\boldsymbol{b}, \boldsymbol{a})^\top$, which gives the reverse-order cross-covariance in (44). Substituting the cross-covariance and the positive-definite observation covariance above yields the posterior mean

$$\mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k] = \mathrm{Cov}(\boldsymbol{w}, \boldsymbol{y}_t \mid Z = k)\mathrm{Cov}(\boldsymbol{y}_t \mid Z = k)^{-1}\boldsymbol{y}_t = \mathbf{V}_k\mathbf{X}_t^\top\boldsymbol{\Sigma}_{k,t}^{-1}\boldsymbol{y}_t = \mathbf{B}_{k,t}\boldsymbol{y}_t. \tag{43}$$

The same conditional-Gaussian formula yields the posterior covariance,

$$\mathrm{Cov}(\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k) = \mathbf{V}_k - (\mathbf{V}_k\mathbf{X}_t^\top)\boldsymbol{\Sigma}_{k,t}^{-1}(\mathbf{V}_k\mathbf{X}_t^\top)^\top. \tag{44}$$

To express the covariance in (44) in posterior-precision form, use the Woodbury identity $(A^{-1} + C^\top R^{-1}C)^{-1} = A - AC^\top(CAC^\top + R)^{-1}CA$, for positive-definite A and R. Substituting $A = \mathbf{V}_k$, $C = \mathbf{X}_t$, and $R = \sigma^2\mathbf{I}_t$, with $\mathbf{V}_k \succ 0$ and $\sigma^2 > 0$, gives

$$\begin{aligned}\mathbf{P}_{k,t} = \mathrm{Cov}(\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k) &= \mathbf{V}_k - \mathbf{V}_k\mathbf{X}_t^\top(\mathbf{X}_t\mathbf{V}_k\mathbf{X}_t^\top + \sigma^2\mathbf{I}_t)^{-1}\mathbf{X}_t\mathbf{V}_k \\ &= (\mathbf{V}_k^{-1} + \sigma^{-2}\mathbf{X}_t^\top\mathbf{X}_t)^{-1}.\end{aligned} \tag{45}$$

These computations establish the posterior formulas used in Section 4.2 and Corollary 4.2.

Posterior state probabilities given a summary. Since $\boldsymbol{\Sigma}_{k,t} \succ 0$, its inverse exists and its determinant is positive. Consequently, the nondegenerate Gaussian law $\boldsymbol{y}_t \mid Z = k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{k,t})$ has the density

$$\phi_{k,t}(\boldsymbol{y}) \coloneqq \frac{1}{(2\pi)^{t/2}\det(\boldsymbol{\Sigma}_{k,t})^{1/2}}\exp\left(-\tfrac{1}{2}\boldsymbol{y}^\top\boldsymbol{\Sigma}_{k,t}^{-1}\boldsymbol{y}\right) > 0 \qquad (\boldsymbol{y} \in \mathbb{R}^t). \tag{46}$$

The denominator in (46) is positive and the exponential is strictly positive, so $\phi_{k,t}$ is positive at every observation value.

We next derive the state probabilities after observing both $\boldsymbol{y}_t$ and the state summary $S = f(Z)$. Fix a summary value s in the range of f, and write $I_s := \{j : f(j) = s\}$ and $\mathbb{P}(S = s) = \sum_{j \in I_s} \theta_j > 0$. Because S is determined by Z, observing $S = s$ restricts Z to the class I_s. Thus

$$\mathbb{P}(Z = k \mid S = s) = \begin{cases} \theta_k / \mathbb{P}(S = s), & k \in I_s, \\ 0, & k \notin I_s. \end{cases}$$

For $k \in I_s$, conditioning additionally on $S = s$ does not change the likelihood of $\boldsymbol{y}_t$ given $Z = k$; this likelihood is the density (46). Averaging these likelihoods with the conditional prior weights gives the density of $\boldsymbol{y}_t$ given $S = s$:

$$p_{t,s}(\boldsymbol{y}) := \sum_{j \in I_s} \mathbb{P}(Z = j \mid S = s) \phi_{j,t}(\boldsymbol{y}) = \frac{1}{\mathbb{P}(S = s)} \sum_{j \in I_s} \theta_j \phi_{j,t}(\boldsymbol{y}) > 0. \tag{47}$$

Because the density in (47) is strictly positive, Bayes' rule gives, for $k \in I_s$,

$$\begin{aligned} \mathbb{P}(Z = k \mid \boldsymbol{y}_t = \boldsymbol{y}, S = s) &= \frac{\mathbb{P}(Z = k \mid S = s) \phi_{k,t}(\boldsymbol{y})}{p_{t,s}(\boldsymbol{y})} \\ &= \frac{(\theta_k / \mathbb{P}(S = s)) \phi_{k,t}(\boldsymbol{y})}{\sum_{j \in I_s} (\theta_j / \mathbb{P}(S = s)) \phi_{j,t}(\boldsymbol{y})} = \frac{\theta_k \phi_{k,t}(\boldsymbol{y})}{\sum_{j \in I_s} \theta_j \phi_{j,t}(\boldsymbol{y})} > 0. \end{aligned} \tag{48}$$

The factor $1/\mathbb{P}(S = s)$ cancels between the numerator and the denominator. The latter is strictly positive since the nonempty class I_s consists of states with $\theta_j > 0$ and $\phi_{j,t}(\boldsymbol{y}) > 0$. When $k \notin I_s$, the posterior probability is zero. In particular, all states in the observed summary class have strictly positive weights $\omega_{k,t}$. The displayed formula specifies a version of the regular conditional probability, since $\boldsymbol{y}_t$ is continuously distributed.

Product-to-joint domination. By the conditional-density formula and (46), the joint law of $(\boldsymbol{y}_t, Z)$ has state-k density $\theta_k \phi_{k,t}(\boldsymbol{y}) > 0$. By the law of total probability, the marginal density of $\boldsymbol{y}_t$ is $\sum_{j=1}^K \theta_j \phi_{j,t}(\boldsymbol{y})$. Since the product law treats $\boldsymbol{y}_t$ and Z as independent, $\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z)$ has state-k density $\theta_k \sum_{j=1}^K \theta_j \phi_{j,t}(\boldsymbol{y}) > 0$.

We now show that $\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z) \ll \mathrm{Law}(\boldsymbol{y}_t, Z)$. Let $A \subseteq \mathbb{R}^t \times \{1, \dots, K\}$ be measurable and write $A_k := \{\boldsymbol{y} : (\boldsymbol{y}, k) \in A\}$. Suppose $\mathrm{Law}(\boldsymbol{y}_t, Z)(A) = 0$. Then

$$0 = \sum_{k=1}^K \int_{A_k} \theta_k \phi_{k,t}(\boldsymbol{y}) \, d\boldsymbol{y}.$$

Each integral is nonnegative and its integrand is strictly positive everywhere. Hence every A_k has Lebesgue measure zero. Integrating the density of the product law over these null sets therefore gives

$$(\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z))(A) = \sum_{k=1}^K \int_{A_k} \theta_k \sum_{j=1}^K \theta_j \phi_{j,t}(\boldsymbol{y}) \, d\boldsymbol{y} = 0.$$

Thus every null set of $\mathrm{Law}(\boldsymbol{y}_t, Z)$ is also a null set of $\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z)$, proving $\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z) \ll \mathrm{Law}(\boldsymbol{y}_t, Z)$.

For the deterministic design, define the measurable map $F_t(\boldsymbol{y}) := ((\boldsymbol{x}_i, y_i))_{i=1}^t$ for $\boldsymbol{y} = (y_1, \dots, y_t)^\top \in \mathbb{R}^t$. Then $C_t = F_t(\boldsymbol{y}_t)$ and $F_t \times \mathrm{id}$ transforms $(\boldsymbol{y}_t, Z)$ into (C_t, Z) without changing the state.

We use the fact that absolute continuity is preserved under measurable pushforward. If $\xi_1 \ll \xi_2$ and h is measurable, then $(h_\# \xi_2)(B) = 0$ implies $\xi_2(h^{-1}(B)) = 0$, hence $\xi_1(h^{-1}(B)) = 0$ and $(h_\# \xi_1)(B) = 0$. Thus $h_\# \xi_1 \ll h_\# \xi_2$.

Applying this property to the absolute continuity established above gives the required domination for the observed context:

$$\begin{aligned} \mathrm{Law}(C_t) \otimes \mathrm{Law}(Z) &= (F_t \times \mathrm{id})_\# [\mathrm{Law}(\boldsymbol{y}_t) \otimes \mathrm{Law}(Z)] \\ &\ll (F_t \times \mathrm{id})_\# \, \mathrm{Law}(\boldsymbol{y}_t, Z) = \mathrm{Law}(C_t, Z). \end{aligned} \tag{49}$$

The first equality holds because $F_t \times \mathrm{id}$ acts separately on the observation and state coordinates, preserving their independence under the product law. The last equality follows from $(F_t \times \mathrm{id})(\boldsymbol{y}_t, Z) = (C_t, Z)$.

Gaussian U-CPQ and the CPQ minimality hypotheses. We next identify the U-CPQ whose recovery will be tested in Section D.3. For fixed t and U, let $E_{t,U} := \{\mathbf{M} \in \mathbb{R}^{d\times t} : P_U\mathbf{M} = \mathbf{M}\}$. This closed linear subspace is standard Borel. Define the measurable encoding of the finite state space, as in Section 4.2:

$$\tau_{t,U}(k) = P_U\mathbf{B}_{k,t} \in E_{t,U}. \tag{50}$$

To verify the exact-encoding condition (P), note that the deterministic design admits the measurable response map $\mathrm{resp}(c) = (y_1, \ldots, y_t)^\top$ for $c = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_t, y_t))$; hence $\mathrm{resp}(C_t) = \boldsymbol{y}_t$. We verify (P) for the selected family $\boldsymbol{\alpha}_{t,U}(c,k) = P_U\mathbf{B}_{k,t}\,\mathrm{resp}(c)$, which satisfies $\boldsymbol{\alpha}_{t,U}(C_t, Z) = P_U\boldsymbol{m}_t^Z$ almost surely by (43). By the law of total probability and (46), the marginal density of $\boldsymbol{y}_t$ is $p_t(\boldsymbol{y}) = \sum_{j=1}^K \theta_j\phi_{j,t}(\boldsymbol{y}) > 0$ everywhere. Thus, by (43), two states induce the same projected prediction-rule profile $\mathrm{Law}(C_t)$-almost surely if and only if $P_U(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t})\boldsymbol{y} = \mathbf{0}$ for Lebesgue-almost every $\boldsymbol{y} \in \mathbb{R}^t$. The null space of a nonzero linear map is a proper, Lebesgue-null linear subspace. More explicitly, if $M = P_U(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t})$ is nonzero, choose a nonzero row $\boldsymbol{h}^\top$ of M. By (42), $\boldsymbol{h}^\top\boldsymbol{y}_t \mid Z = j$ is Gaussian with variance $\boldsymbol{h}^\top\boldsymbol{\Sigma}_{j,t}\boldsymbol{h} > 0$ for every state j. Consequently,

$$\mathbb{P}(M\boldsymbol{y}_t = \mathbf{0} \mid Z = j) \le \mathbb{P}(\boldsymbol{h}^\top\boldsymbol{y}_t = 0 \mid Z = j) = 0.$$

Since each state has positive prior probability, $M\boldsymbol{y}_t = \mathbf{0}$ cannot hold almost surely unless $M = 0$. Thus the equality of prediction-rule profiles holds if and only if the coefficient matrix vanishes. Equivalently,

$$\begin{aligned} &P_U\mathbf{B}_{k,t}\,\mathrm{resp}(c) = P_U\mathbf{B}_{\ell,t}\,\mathrm{resp}(c) \quad \text{for } \mathrm{Law}(C_t)\text{-almost every } c \\ \iff\quad &P_U\mathbf{B}_{k,t} = P_U\mathbf{B}_{\ell,t} \quad\iff\quad \tau_{t,U}(k) = \tau_{t,U}(\ell). \end{aligned} \tag{51}$$

This verifies (P) for all state pairs.

For reconstruction (R), use the measurable reconstruction map $b_t(c, \mathbf{M}) := \mathbf{M}\,\mathrm{resp}(c)$, which takes values in U whenever $\mathbf{M} \in E_{t,U}$. The encoding (50) and the posterior mean (43) give

$$b_t(C_t, \tau_{t,U}(Z)) = P_U\mathbf{B}_{Z,t}\boldsymbol{y}_t = P_U\boldsymbol{m}_t^Z \quad \text{almost surely.} \tag{52}$$

This verifies (R). Realized uniqueness (J) follows from squared Euclidean loss, as checked for all actions on a common full-measure set in the argument following (26); the required product-to-joint domination is (49). Hence (50) defines a U-CPQ for the stage-t projected problem, and all assumptions of the CPQ recovery part of Theorem 4.1(i) are satisfied. In the next subsection we compute the resulting excess risk and derive the corresponding coefficient criterion explicitly. For $U = \mathrm{span}\{\boldsymbol{q}\}$ with $\boldsymbol{q} \neq \mathbf{0}$, the encoding distinguishes states precisely through their query-specific coefficients $\boldsymbol{q}^\top\mathbf{B}_{k,t}$.

D.3 Proof of Corollary 4.2 (Section 4.2)

We use the nested designs and posterior-mean coefficients of Section 4.2; the next design row is $\boldsymbol{x}_{t+1}^\top$.

Proof of Corollary 4.2.

Part (i): stage-t U-sufficiency and CPQ recovery.

The general query-wise excess-risk identity is (11); the projected excess-risk identity is (25). Below we explicitly compute the Gaussian mixture and coefficient criteria, without repeating the abstract orthogonality proof.

Mixing weights. We first identify the weights in the conditional mixture. For each summary value s with $\mathbb{P}(S = s) > 0$, set $I_s = \{k : f(k) = s\}$ as in Section D.2. The state prior is θ_k, whereas the weight $\omega_{k,t} = \mathbb{P}(Z = k \mid \boldsymbol{y}_t, S)$ is the posterior state probability after observing both $\boldsymbol{y}_t$ and S. By (48), on $S = s$, the mixing weight $\omega_{k,t}$ is strictly positive for every $k \in I_s$. For $k \notin I_s$, it is zero because $S = f(Z)$ rules out such states. Consequently, the weights are supported on I_s, sum to one, and are strictly positive throughout the class I_s. Before conditioning, each $\omega_{k,t}$ is a random quantity because it depends on $(\boldsymbol{y}_t, S)$; after conditioning on those observations, it is the corresponding known mixture weight.

State-summary Bayes prediction as a posterior mixture. The mixing weights now determine how the statewise oracle predictions are combined when only $(\boldsymbol{y}_t, S)$ is observed. Define $\boldsymbol{a}_{k,t} := P_U \mathbf{B}_{k,t} \boldsymbol{y}_t$, the projected oracle posterior mean in state k, and $\overline{\boldsymbol{a}}_t := \sum_{k=1}^K \omega_{k,t} \boldsymbol{a}_{k,t}$.

We first relate the summary-conditioned posterior mean to the oracle posterior mean. Since $S = f(Z)$, we have $\sigma(\boldsymbol{y}_t, S) \subseteq \sigma(\boldsymbol{y}_t, Z)$. Using the definitions $\boldsymbol{m}_t^S = \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, S]$ and $\boldsymbol{m}_t^Z = \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z]$, the tower property, as in (40) at stage t, gives (here C_t is identified with $\boldsymbol{y}_t$ by the deterministic design)

$$P_U \boldsymbol{m}_t^S = P_U \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, S] = P_U \mathbb{E}[\mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z] \mid \boldsymbol{y}_t, S] = P_U \mathbb{E}[\boldsymbol{m}_t^Z \mid \boldsymbol{y}_t, S] = \mathbb{E}[P_U \boldsymbol{m}_t^Z \mid \boldsymbol{y}_t, S]. \tag{53}$$

The last equality uses the fact that P_U is a deterministic linear projection.

We next evaluate this conditional expectation. By (43), on the event $\{Z = k\}$, the oracle posterior mean satisfies $\boldsymbol{m}_t^Z = \mathbf{B}_{k,t} \boldsymbol{y}_t$. Since the events $\{Z = k\}$, $k = 1, \ldots, K$, partition the sample space, we obtain $\boldsymbol{m}_t^Z = \mathbf{B}_{Z,t} \boldsymbol{y}_t$ almost surely. Consequently, the projected oracle mean can be written as

$$P_U \boldsymbol{m}_t^Z = P_U \mathbf{B}_{Z,t} \boldsymbol{y}_t = \sum_{k=1}^K \mathbf{1}_{\{Z=k\}} P_U \mathbf{B}_{k,t} \boldsymbol{y}_t = \sum_{k=1}^K \mathbf{1}_{\{Z=k\}} \boldsymbol{a}_{k,t}.$$

For each fixed k, the vector $\boldsymbol{a}_{k,t}$ depends only on $\boldsymbol{y}_t$ and is therefore measurable with respect to $\sigma(\boldsymbol{y}_t, S)$. Consequently, conditional-expectation linearity and the rule for taking out measurable factors yield

$$\mathbb{E}[P_U \boldsymbol{m}_t^Z \mid \boldsymbol{y}_t, S] = \sum_{k=1}^K \mathbb{E}[\mathbf{1}_{\{Z=k\}} \boldsymbol{a}_{k,t} \mid \boldsymbol{y}_t, S] = \sum_{k=1}^K \boldsymbol{a}_{k,t} \mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \boldsymbol{y}_t, S] = \sum_{k=1}^K \omega_{k,t} \boldsymbol{a}_{k,t} = \overline{\boldsymbol{a}}_t,$$

where the penultimate equality follows from $\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \boldsymbol{y}_t, S] = \mathbb{P}(Z = k \mid \boldsymbol{y}_t, S) = \omega_{k,t}$.

Combining the identity (53), obtained by the tower property, with the conditional state expansion gives the finite-state mean obtained by integrating the conditional-law mixture (28):

$$P_U \boldsymbol{m}_t^S = \sum_{k=1}^K \omega_{k,t} \boldsymbol{a}_{k,t} = \overline{\boldsymbol{a}}_t. \tag{54}$$

Thus the state-summary Bayes prediction $P_U \boldsymbol{m}_t^S$ is the posterior-weighted average of the state-k projected oracle posterior means $\boldsymbol{a}_{k,t}$.

Pairwise representation of excess risk. We now quantify the predictive cost of retaining only the summary. The projected excess-risk identity (25), proved in Section D.1, states that $R_{S,t}^U - R_{\mathrm{ora},t}^U = \mathbb{E} \left\| P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S) \right\|^2$. Using (43) and (54), together with the linearity of P_U, $P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S) = P_U \boldsymbol{m}_t^Z - P_U \boldsymbol{m}_t^S = \boldsymbol{a}_{Z,t} - \overline{\boldsymbol{a}}_t$. We next specialize the conditional-covariance identity (19) to the Gaussian finite-state problem, evaluating the projected squared difference conditionally on $(\boldsymbol{y}_t, S)$. For each fixed k, $\boldsymbol{a}_{k,t} = P_U \mathbf{B}_{k,t} \boldsymbol{y}_t$ depends only on $\boldsymbol{y}_t$, while $\overline{\boldsymbol{a}}_t = \sum_{k=1}^K \omega_{k,t} \boldsymbol{a}_{k,t}$ depends only on $(\boldsymbol{y}_t, S)$. Thus both vectors are measurable with respect to $\sigma(\boldsymbol{y}_t, S)$. After conditioning on this information, their values are fixed, and only the state index Z remains random.

Since the events $\{Z = k\}$ partition the sample space, we can expand the squared difference according to the realized state:

$$\left\| P_U(\boldsymbol{m}_t^Z - \boldsymbol{m}_t^S) \right\|^2 = \left\| \boldsymbol{a}_{Z,t} - \overline{\boldsymbol{a}}_t \right\|^2 = \sum_{k=1}^K \mathbf{1}_{\{Z=k\}} \left\| \boldsymbol{a}_{k,t} - \overline{\boldsymbol{a}}_t \right\|^2.$$

Taking conditional expectation and using linearity over the finite sum gives the posterior-weighted squared discrepancy:

$$\begin{aligned}\mathbb{E}\big[\big\|P_U(\boldsymbol{m}_t^Z-\boldsymbol{m}_t^S)\big\|^2 \mid \boldsymbol{y}_t,S\big] &= \mathbb{E}\Big[\sum_{k=1}^K \mathbf{1}_{\{Z=k\}}\,\|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2 \;\Big|\; \boldsymbol{y}_t,S\Big]\\ &= \sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{\{Z=k\}}\,\|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2 \mid \boldsymbol{y}_t,S\big] = \sum_{k=1}^K \|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2\,\mathbb{E}[\mathbf{1}_{\{Z=k\}}\mid \boldsymbol{y}_t,S] \qquad (55)\\ &= \sum_{k=1}^K \omega_{k,t}\,\|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2 .\end{aligned}$$

The third equality uses the measurability of $\|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2$ with respect to $\sigma(\boldsymbol{y}_t,S)$, allowing this factor to be taken outside the conditional expectation. The final equality follows from $\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \boldsymbol{y}_t,S] = \mathbb{P}(Z=k \mid \boldsymbol{y}_t,S) = \omega_{k,t}$. Thus, conditioning on $(\boldsymbol{y}_t,S)$ replaces the uncertain state selection by its posterior probabilities. The conditional mean squared difference between the oracle and state-summary Bayes predictions is therefore expressed entirely as a finite weighted sum of state-specific squared differences, with weights $\omega_{k,t}$ determined by the observed information.

The weighted variance in (55) is not yet the pairwise formula in the corollary. To expose the discrepancies between pairs of possible hidden states, use $\sum_k \omega_{k,t}=1$ and $\overline{\boldsymbol{a}}_t=\sum_k \omega_{k,t}\boldsymbol{a}_{k,t}$ to expand the squares:

$$\sum_k \omega_{k,t}\,\|\boldsymbol{a}_{k,t}-\overline{\boldsymbol{a}}_t\|^2 = \sum_k \omega_{k,t}\,\|\boldsymbol{a}_{k,t}\|^2 - \|\overline{\boldsymbol{a}}_t\|^2 = \frac{1}{2}\sum_{k,\ell}\omega_{k,t}\omega_{\ell,t}\,\|\boldsymbol{a}_{k,t}-\boldsymbol{a}_{\ell,t}\|^2 . \qquad (56)$$

The final equality follows by expanding each pairwise square and summing the weights:

$$\begin{aligned}\frac{1}{2}\sum_{k,\ell}\omega_{k,t}\omega_{\ell,t}\,\|\boldsymbol{a}_{k,t}-\boldsymbol{a}_{\ell,t}\|^2 &= \frac{1}{2}\sum_{k,\ell}\omega_{k,t}\omega_{\ell,t}\big(\|\boldsymbol{a}_{k,t}\|^2 - 2\boldsymbol{a}_{k,t}^\top\boldsymbol{a}_{\ell,t}+\|\boldsymbol{a}_{\ell,t}\|^2\big)\\ &= \sum_k \omega_{k,t}\,\|\boldsymbol{a}_{k,t}\|^2 - \|\overline{\boldsymbol{a}}_t\|^2 .\end{aligned}$$

Here each of the two squared-norm terms sums to $\sum_k \omega_{k,t}\,\|\boldsymbol{a}_{k,t}\|^2$ because $\sum_k \omega_{k,t}=1$, and the cross term sums to $-2\,\|\overline{\boldsymbol{a}}_t\|^2$ because $\overline{\boldsymbol{a}}_t=\sum_k \omega_{k,t}\boldsymbol{a}_{k,t}$; the factor $\frac{1}{2}$ then gives the last line. Substituting $\boldsymbol{a}_{k,t}-\boldsymbol{a}_{\ell,t}=P_U(\mathbf{B}_{k,t}-\mathbf{B}_{\ell,t})\boldsymbol{y}_t$ into (56) and using (55) gives the conditional pairwise formula

$$\begin{aligned}\mathbb{E}\big[\big\|P_U(\boldsymbol{m}_t^Z-\boldsymbol{m}_t^S)\big\|^2 \mid \boldsymbol{y}_t,S\big] &= \frac{1}{2}\sum_{k,\ell=1}^K \omega_{k,t}\omega_{\ell,t}\,\|\boldsymbol{a}_{k,t}-\boldsymbol{a}_{\ell,t}\|^2\\ &= \frac{1}{2}\sum_{k,\ell=1}^K \omega_{k,t}\omega_{\ell,t}\,\|P_U(\mathbf{B}_{k,t}-\mathbf{B}_{\ell,t})\boldsymbol{y}_t\|^2 .\end{aligned} \qquad (57)$$

Thus the pairwise sum represents the conditional expectation of the squared discrepancy between the projected oracle and state-summary Bayes predictions, rather than the squared discrepancy itself. For each deterministic $\boldsymbol{q}\in\mathbb{R}^d$, (11) and (19) turn the vector conditional covariance into its scalar quadratic form. With $\boldsymbol{m}_t^Z=\mathbf{B}_{Z,t}\boldsymbol{y}_t$ by (43), the same pairwise variance identity applied to the scalar predictions $\boldsymbol{q}^\top\mathbf{B}_{k,t}\boldsymbol{y}_t$ gives

$$\begin{aligned}R_{S,t}(\boldsymbol{q})-R_{\mathrm{ora},t}(\boldsymbol{q}) = \boldsymbol{q}^\top\mathbf{A}_{S,t}\boldsymbol{q} &= \mathbb{E}\big[\mathrm{Var}(\boldsymbol{q}^\top\mathbf{B}_{Z,t}\boldsymbol{y}_t \mid \boldsymbol{y}_t,S)\big]\\ &= \frac{1}{2}\mathbb{E}\left[\sum_{k,\ell=1}^K \omega_{k,t}\omega_{\ell,t}\big(\boldsymbol{q}^\top(\mathbf{B}_{k,t}-\mathbf{B}_{\ell,t})\boldsymbol{y}_t\big)^2\right].\end{aligned}$$

The same calculation for the projected vector response, using (57), gives the corresponding projected excess-risk identity:

$$\begin{aligned}R_{S,t}^U-R_{\mathrm{ora},t}^U = \mathbb{E}\big[\big\|P_U(\boldsymbol{m}_t^Z-\boldsymbol{m}_t^S)\big\|^2\big] &= \mathbb{E}\Big[\mathbb{E}\big[\big\|P_U(\boldsymbol{m}_t^Z-\boldsymbol{m}_t^S)\big\|^2 \mid \boldsymbol{y}_t,S\big]\Big]\\ &= \frac{1}{2}\mathbb{E}\Big[\sum_{k,\ell=1}^K \omega_{k,t}\omega_{\ell,t}\,\|P_U(\mathbf{B}_{k,t}-\mathbf{B}_{\ell,t})\boldsymbol{y}_t\|^2\Big].\end{aligned} \qquad (58)$$

The projected excess risk is the trace of the same conditional covariance restricted to U, by (19) and (25). This proves the risk statements used in part (i).

The pairwise representation makes the predictive meaning of the excess risk explicit. Conditional on $S = s$, only pairs of states $k, \ell \in I_s$ contribute: the summary assigns these states the same value, even though their projected oracle predictions $P_U \mathbf{B}_{k,t} \boldsymbol{y}_t$ and $P_U \mathbf{B}_{\ell,t} \boldsymbol{y}_t$ may differ. Each squared difference measures a prediction-relevant distinction not preserved by the summary, weighted by the posterior probabilities of the two states given the observed context and summary. Thus the excess risk quantifies the predictive cost of merging states that require different oracle predictions, rather than the cost of failing to identify the hidden state itself.

Zero-excess-risk coefficient criterion and CPQ recovery. We now derive the exact coefficient criterion without repeating the general equivalence between oracle-risk preservation and CPQ recovery. The encoding $\tau_{t,U}(k) = P_U \mathbf{B}_{k,t}$ satisfies (P) by (51) and (R) by (52). Squared Euclidean loss gives realized uniqueness (J), and (49) verifies product-to-joint domination. Thus all hypotheses of the recovery conclusion of Theorem 4.1(i) are satisfied. By the projected risk identity (25) and the criterion (11), this conclusion gives

$$\begin{aligned} R^U_{S,t} = R^U_{\mathrm{ora},t} &\iff U \subseteq \ker \mathbf{A}_{S,t} \\ &\iff T^\star_{t,U} = g_t(S) \quad \text{almost surely for some measurable } g_t. \end{aligned}$$

We translate the last condition into a criterion on state-specific coefficients. Recall that $T^\star_{t,U} = \tau_{t,U}(Z)$, $S = f(Z)$, and $\mathbb{P}(Z = k) = \theta_k > 0$ for every state k. If $T^\star_{t,U} = g_t(S)$ almost surely, then on the event $\{Z = k\}$, which has positive probability, $\tau_{t,U}(k) = g_t(f(k))$. In particular, if $f(k) = f(\ell)$, the two coefficients must agree: $P_U \mathbf{B}_{k,t} = P_U \mathbf{B}_{\ell,t}$. Conversely, suppose the projected coefficients agree for every pair of states in the same summary class $I_s = \{k : f(k) = s\}$. For each s in the finite range of S, define $g_t(s)$ to be this common coefficient, and extend g_t measurably elsewhere by the zero matrix. Then $\tau_{t,U}(Z) = g_t(S)$ almost surely, so the preceding recovery equivalence applies in the reverse direction. We have proved

$$R^U_{S,t} = R^U_{\mathrm{ora},t} \quad \iff \quad P_U \mathbf{B}_{k,t} = P_U \mathbf{B}_{\ell,t} \quad \text{whenever } f(k) = f(\ell). \tag{59}$$

Unlike equality of predictions at a particular observed $\boldsymbol{y}_t$, this criterion compares their coefficients for all possible observation values. The reason that these coefficients identify prediction-rule profiles was proved in (51). The pairwise formula (58) remains valid even when the coefficients do not agree. Combining the criterion above with (11) and the scalar risk formula already established proves (16). The comparison with the general theorem appears in Section D.4.

Part (ii): One-step persistence of U-sufficiency and transfer loss.

We first derive the statewise one-step update, then use the persistence criterion to obtain the pairwise coefficient conditions. Finally, we express the resulting stage-$(t+1)$ excess risk as a conditional covariance and a posterior-weighted pairwise sum.

Fix a state k. The new response is $y_{t+1} = \boldsymbol{x}_{t+1}^\top \boldsymbol{w} + \varepsilon_{t+1}$. By (43) and (45), conditional on $(\boldsymbol{y}_t, Z = k)$ the old posterior mean and covariance are $\mathbf{B}_{k,t} \boldsymbol{y}_t$ and $\mathbf{P}_{k,t}$. To derive the update rather than merely invoke its formula, define the state-specific posterior mean, the centered target vector, and the innovation by

$$\boldsymbol{m}_{k,t} := \mathbf{B}_{k,t} \boldsymbol{y}_t, \qquad \widetilde{\boldsymbol{w}}_{k,t} := \boldsymbol{w} - \boldsymbol{m}_{k,t}, \qquad r_{k,t+1} := y_{t+1} - \boldsymbol{x}_{t+1}^\top \boldsymbol{m}_{k,t}.$$

On $\{Z = k\}$, the first quantity agrees with $\boldsymbol{m}_t^Z$. The subscripts emphasize that the centering and the innovation depend on the fixed state and the observation stage. Then

$$r_{k,t+1} = \boldsymbol{x}_{t+1}^\top \widetilde{\boldsymbol{w}}_{k,t} + \varepsilon_{t+1}, \qquad \mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t} \mid \boldsymbol{y}_t, Z = k] = \mathbf{0}, \qquad \mathrm{Cov}(\widetilde{\boldsymbol{w}}_{k,t} \mid \boldsymbol{y}_t, Z = k) = \mathbf{P}_{k,t}.$$

Because the new noise ε_{t+1} is independent of $(\boldsymbol{w}, \boldsymbol{y}_t, Z)$, its conditional mean is zero, its conditional variance is σ^2, and its conditional cross-covariance with $\widetilde{\boldsymbol{w}}_{k,t}$ vanishes. Consequently,

$$\begin{aligned} \mathbb{E}[y_{t+1} \mid \boldsymbol{y}_t, Z = k] &= \boldsymbol{x}_{t+1}^\top \boldsymbol{m}_{k,t} = \boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t} \boldsymbol{y}_t, \\ \mathrm{Cov}(\boldsymbol{w}, y_{t+1} \mid \boldsymbol{y}_t, Z = k) &= \mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t} r_{k,t+1} \mid \boldsymbol{y}_t, Z = k] \\ &= \mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t} \widetilde{\boldsymbol{w}}_{k,t}^\top \mid \boldsymbol{y}_t, Z = k] \boldsymbol{x}_{t+1} + \mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t} \varepsilon_{t+1} \mid \boldsymbol{y}_t, Z = k] \\ &= \mathbf{P}_{k,t} \boldsymbol{x}_{t+1}, \\ \mathrm{Var}(y_{t+1} \mid \boldsymbol{y}_t, Z = k) &= \mathbb{E}[r_{k,t+1}^2 \mid \boldsymbol{y}_t, Z = k] = \mathbb{E}[(\boldsymbol{x}_{t+1}^\top \widetilde{\boldsymbol{w}}_{k,t} + \varepsilon_{t+1})^2 \mid \boldsymbol{y}_t, Z = k] \\ &= \boldsymbol{x}_{t+1}^\top \mathbf{P}_{k,t} \boldsymbol{x}_{t+1} + \sigma^2. \end{aligned} \tag{60}$$

For the cross-covariance in (60), $\widetilde{\boldsymbol{w}}_{k,t}$ and $r_{k,t+1}$ are precisely the two variables centered at their conditional means. Expanding $r_{k,t+1} = \boldsymbol{x}_{t+1}^\top \widetilde{\boldsymbol{w}}_{k,t} + \varepsilon_{t+1}$ uses the matrix identity $\widetilde{\boldsymbol{w}}_{k,t}(\boldsymbol{x}_{t+1}^\top \widetilde{\boldsymbol{w}}_{k,t}) = (\widetilde{\boldsymbol{w}}_{k,t}\widetilde{\boldsymbol{w}}_{k,t}^\top)\boldsymbol{x}_{t+1}$. Because $\widetilde{\boldsymbol{w}}_{k,t}$ has conditional mean zero and conditional covariance $\mathbf{P}_{k,t}$, $\mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t}\widetilde{\boldsymbol{w}}_{k,t}^\top \mid \boldsymbol{y}_t, Z = k] = \mathbf{P}_{k,t}$. Independence and zero mean of the new noise give $\mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t}\varepsilon_{t+1} \mid \boldsymbol{y}_t, Z = k] = \mathbf{0}$. The conditional distribution of $(\boldsymbol{w}, y_{t+1})$ is jointly Gaussian: given $(\boldsymbol{y}_t, Z = k)$, the vector $\boldsymbol{w}$ is Gaussian, the new noise is independent Gaussian, and y_{t+1} is their linear combination. The moments in (60) give the following conditional mean and covariance:

$$\mathbb{E}\left[\begin{pmatrix}\boldsymbol{w}\\ y_{t+1}\end{pmatrix}\middle|\, \boldsymbol{y}_t, Z = k\right] = \begin{pmatrix}\boldsymbol{m}_{k,t}\\ \boldsymbol{x}_{t+1}^\top \boldsymbol{m}_{k,t}\end{pmatrix}$$

and

$$\mathrm{Cov}\left(\begin{pmatrix}\boldsymbol{w}\\ y_{t+1}\end{pmatrix}\middle|\, \boldsymbol{y}_t, Z = k\right) = \begin{pmatrix}\mathbf{P}_{k,t} & \mathbf{P}_{k,t}\boldsymbol{x}_{t+1}\\ \boldsymbol{x}_{t+1}^\top \mathbf{P}_{k,t} & \sigma^2 + \boldsymbol{x}_{t+1}^\top \mathbf{P}_{k,t}\boldsymbol{x}_{t+1}\end{pmatrix}. \tag{61}$$

Here the off-diagonal blocks are the cross-covariances computed above, and the lower-right entry is strictly positive since $\mathbf{P}_{k,t} \succeq 0$ and $\sigma^2 > 0$.

We next derive the posterior-mean update using the standard conditional-Gaussian mean formula [3, Sec. 1.1.3]. It describes the change from $\mathbb{E}[\boldsymbol{\gamma} \mid \mathcal{J}]$ to $\mathbb{E}[\boldsymbol{\gamma} \mid \eta, \mathcal{J}]$ when an additional scalar observation η becomes available. If $\boldsymbol{\gamma}$ and η are jointly Gaussian conditionally on $\mathcal{J}$, with $\mathrm{Var}(\eta \mid \mathcal{J}) > 0$, then

$$\mathbb{E}[\boldsymbol{\gamma} \mid \eta, \mathcal{J}] = \mathbb{E}[\boldsymbol{\gamma} \mid \mathcal{J}] + \mathrm{Cov}(\boldsymbol{\gamma}, \eta \mid \mathcal{J})\mathrm{Var}(\eta \mid \mathcal{J})^{-1}\big(\eta - \mathbb{E}[\eta \mid \mathcal{J}]\big).$$

To obtain this identity, subtract the conditional means of $\boldsymbol{\gamma}$ and η, apply the centered conditional-Gaussian formula of Section D.2, and add the mean of $\boldsymbol{\gamma}$ back. In the present model, the block covariance (61) implies that, given $(\boldsymbol{y}_t, Z = k)$, the centered pair $(\widetilde{\boldsymbol{w}}_{k,t}, r_{k,t+1})$ is jointly Gaussian with conditional mean zero. The centered formula gives

$$\mathbb{E}[\widetilde{\boldsymbol{w}}_{k,t} \mid \boldsymbol{y}_t, r_{k,t+1}, Z = k] = \frac{\mathbf{P}_{k,t}\boldsymbol{x}_{t+1}}{\sigma^2 + \boldsymbol{x}_{t+1}^\top \mathbf{P}_{k,t}\boldsymbol{x}_{t+1}}\, r_{k,t+1}.$$

Since $\boldsymbol{m}_{k,t}$ is determined by $(\boldsymbol{y}_t, Z = k)$ and $r_{k,t+1} = y_{t+1} - \boldsymbol{x}_{t+1}^\top \boldsymbol{m}_{k,t}$, conditioning additionally on $r_{k,t+1}$ is equivalent to conditioning additionally on y_{t+1}. Substituting the conditional moments from (60) into the noncentered formula yields

$$\begin{aligned}\mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, y_{t+1}, Z = k] &= \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k] + \mathrm{Cov}(\boldsymbol{w}, y_{t+1} \mid \boldsymbol{y}_t, Z = k)\\ &\qquad \cdot \mathrm{Var}(y_{t+1} \mid \boldsymbol{y}_t, Z = k)^{-1}\big(y_{t+1} - \mathbb{E}[y_{t+1} \mid \boldsymbol{y}_t, Z = k]\big)\\ &= \mathbf{B}_{k,t}\boldsymbol{y}_t + \frac{\mathbf{P}_{k,t}\boldsymbol{x}_{t+1}}{\sigma^2 + \boldsymbol{x}_{t+1}^\top \mathbf{P}_{k,t}\boldsymbol{x}_{t+1}}\big(y_{t+1} - \boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t}\boldsymbol{y}_t\big).\end{aligned}$$

Thus the coefficient of the innovation $y_{t+1} - \boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t}\boldsymbol{y}_t$ is precisely the update gain $\boldsymbol{K}_{k,t}$ defined in Section 4.2. Since $\boldsymbol{y}_{t+1}$ consists of $\boldsymbol{y}_t$ and y_{t+1}, we obtain the state-specific posterior-mean update

$$\begin{aligned}\mathbb{E}[\boldsymbol{w} \,|\, \boldsymbol{y}_{t+1}, Z = k] &= \mathbf{B}_{k,t}\boldsymbol{y}_t + \boldsymbol{K}_{k,t}\big(y_{t+1} - \boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t}\boldsymbol{y}_t\big)\\ &= \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k] + \boldsymbol{\Delta}_{t:t+1}^{(k)}.\end{aligned} \tag{62}$$

General increments and one-step specialization. By (62), the state-k increment in Section 4.2 is $\boldsymbol{\Delta}_{t:t+1}^{(k)} = \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_{t+1}, Z = k] - \mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_t, Z = k]$. Consequently, $\boldsymbol{\Delta}_{t:t+1}^{Z} = \boldsymbol{\Delta}_{t:t+1}^{(Z)}$ and $\boldsymbol{\Delta}_{t:t+1}^{Z,U} = P_U \boldsymbol{\Delta}_{t:t+1}^{Z}$. Assume stage-$t$ U-sufficiency, as in part (ii) of the corollary. Since the design is deterministic, $\mathcal{G}_{t+1}^{S} = \sigma(\boldsymbol{y}_{t+1}, S)$. The general conditional-mean identity (33), applied with $u = t + 1$, gives

$$\mathbb{E}[\boldsymbol{\Delta}_{t:t+1}^{Z,U} \mid \mathcal{G}_{t+1}^{S}] = P_U \boldsymbol{\Delta}_{t:t+1}^{S}.$$

Here the earlier posterior mean is treated using the stage-t projected equality, as shown in (32); no equality of unprojected posterior means is assumed. The general persistence and excess-risk identities (12) and (13) therefore apply with $u = t + 1$ and the same summary S.

Persistence and the two coefficient conditions. By (59), stage-t U-sufficiency implies, for every k, ℓ with $f(k) = f(\ell)$,

$$P_U(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) = \mathbf{0}_{d\times t}.$$

The same coefficient criterion at stage $t+1$ characterizes persistence by the equality of $P_U\mathbf{B}_{k,t+1}$ and $P_U\mathbf{B}_{\ell,t+1}$ within every summary class. We determine those updated coefficients from the state-wise update (62):

$$\begin{aligned}\mathbb{E}[\boldsymbol{w} \mid \boldsymbol{y}_{t+1}, Z=k] &= \mathbf{B}_{k,t}\boldsymbol{y}_t + \boldsymbol{K}_{k,t}(y_{t+1} - \boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t}\boldsymbol{y}_t) \\ &= (\mathbf{B}_{k,t} - \boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t})\boldsymbol{y}_t + \boldsymbol{K}_{k,t}y_{t+1} \\ &= \begin{bmatrix}\mathbf{B}_{k,t} - \boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t} & \boldsymbol{K}_{k,t}\end{bmatrix} \boldsymbol{y}_{t+1}.\end{aligned}$$

By (43) at stage $t+1$, the same conditional mean is $\mathbf{B}_{k,t+1}\boldsymbol{y}_{t+1}$. Since $\boldsymbol{y}_{t+1} \mid Z=k$ has a strictly positive density by (46) at stage $t+1$, equality of these two linear predictions almost surely implies equality of their coefficient matrices, by the argument of (51). Therefore

$$\mathbf{B}_{k,t+1} = \begin{bmatrix}\mathbf{B}_{k,t} - \boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t} & \boldsymbol{K}_{k,t}\end{bmatrix}. \tag{63}$$

For two states k, ℓ in the same summary class, subtracting their projected coefficient matrices gives

$$\begin{aligned}&P_U(\mathbf{B}_{k,t+1} - \mathbf{B}_{\ell,t+1}) \\ &\quad = \begin{bmatrix}P_U(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) - P_U(\boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t} - \boldsymbol{K}_{\ell,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{\ell,t}) & P_U(\boldsymbol{K}_{k,t} - \boldsymbol{K}_{\ell,t})\end{bmatrix}.\end{aligned}$$

The right block is zero exactly when $P_U(\boldsymbol{K}_{k,t} - \boldsymbol{K}_{\ell,t}) = \mathbf{0}$. Under this condition, expand the difference in the left block and use the equality of projected gains:

$$\begin{aligned}&P_U(\boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{k,t} - \boldsymbol{K}_{\ell,t}\boldsymbol{x}_{t+1}^\top \mathbf{B}_{\ell,t}) \\ &\quad = P_U\boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) + P_U(\boldsymbol{K}_{k,t} - \boldsymbol{K}_{\ell,t})\boldsymbol{x}_{t+1}^\top \mathbf{B}_{\ell,t} \\ &\quad = P_U\boldsymbol{K}_{k,t}\boldsymbol{x}_{t+1}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}).\end{aligned}$$

The stage-t projected coefficient difference is already zero. Thus the left block of the updated difference vanishes precisely when the second condition of (17) holds, and the right block vanishes precisely when the first condition holds. Conversely, both conditions make the two blocks zero. Applying (59) at stage $t+1$ proves both necessity and sufficiency for persistence. Importantly, stage-t sufficiency asserts only the vanishing of the *projected* coefficient difference; the unprojected difference need not vanish. The left-hand side of the second condition of (17) is the outer product of the column vector $P_U\boldsymbol{K}_{k,t}$ and the row vector $\boldsymbol{x}_{t+1}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t})$, a matrix of rank at most one; hence this condition holds if and only if $P_U\boldsymbol{K}_{k,t} = \mathbf{0}$ or $\boldsymbol{x}_{t+1}^\top(\mathbf{B}_{k,t} - \mathbf{B}_{\ell,t}) = \mathbf{0}^\top$.

Projected excess risk and one-step transfer loss. The general excess-risk identity (13) with $u = t+1$ gives $R^U_{S,t+1} - R^U_{\mathrm{ora},t+1} = \mathbb{E}[\operatorname{tr}\operatorname{Cov}(\boldsymbol{\Delta}^{Z,U}_{t:t+1} \mid \mathcal{G}^S_{t+1})]$. As in the finite-state conditional-mean calculation (54), now at stage $t+1$ and for increments rather than predictions, conditional on $(\boldsymbol{y}_{t+1}, S)$ the projected oracle increment takes the values $P_U\boldsymbol{\Delta}^{(k)}_{t:t+1}$ with probabilities $\omega_{k,t+1} = \mathbb{P}(Z=k \mid \boldsymbol{y}_{t+1}, S)$. These weights, given by (48) at stage $t+1$, are supported on the observed summary class and sum to one. Therefore its conditional mean is

$$\mathbb{E}[\boldsymbol{\Delta}^{Z,U}_{t:t+1} \mid \mathcal{G}^S_{t+1}] = \sum_{k=1}^K \omega_{k,t+1} P_U\boldsymbol{\Delta}^{(k)}_{t:t+1}.$$

By the definition of conditional covariance and the finite weighted-variance identity (56),

$$\begin{aligned}\operatorname{tr}\operatorname{Cov}(\boldsymbol{\Delta}^{Z,U}_{t:t+1} \mid \mathcal{G}^S_{t+1}) &= \sum_{k=1}^K \omega_{k,t+1}\Big\|P_U\boldsymbol{\Delta}^{(k)}_{t:t+1} - \sum_{j=1}^K \omega_{j,t+1}P_U\boldsymbol{\Delta}^{(j)}_{t:t+1}\Big\|^2 \\ &= \frac{1}{2}\sum_{k,\ell=1}^K \omega_{k,t+1}\omega_{\ell,t+1}\|P_U(\boldsymbol{\Delta}^{(k)}_{t:t+1} - \boldsymbol{\Delta}^{(\ell)}_{t:t+1})\|^2.\end{aligned}$$

Taking expectations yields the risk formula (18). When $S = T^\star_{t,U}$ is held fixed, this excess risk is $\mathcal{L}_{t\to t+1}(U)$ by definition. The Gaussian hypotheses verified in Section D.2 allow (14) to be applied at $u = t+1$; hence zero transfer loss is equivalent to recovery of $T^\star_{t+1,U}$ from the same S. □

D.4 Comparison of General and Linear-Gaussian Results

Table D.1 aligns the conclusions of Theorem 4.1 and Corollary 4.2 for the same state summary $S = f(Z)$ and query subspace U. The part-(ii) excess-risk formulas assume stage-t U-sufficiency. The Gaussian identities are proved in Sections D.2 and D.3.

Table D.1: Correspondence of quantities and conclusions in the general and linear-Gaussian settings.

	Theorem 4.1	Corollary 4.2
Corresponding mathematical quantities		
(i) Projected oracle posterior mean	$P_U \boldsymbol{m}_t^Z$	$P_U \mathbf{B}_{k,t} \boldsymbol{y}_t$ on $\{Z = k\}$
(i) Prediction-rule profile	C_t-indexed prediction-rule profile	$\tau_{t,U}(k) = P_U \mathbf{B}_{k,t}$ (coefficients)
(ii) Oracle posterior-mean increment	$\boldsymbol{\Delta}_{t:u}^Z$, $u > t$	$\boldsymbol{\Delta}_{t:t+1}^Z = \boldsymbol{\Delta}_{t:t+1}^{(Z)}$; state-$k$: $\boldsymbol{\Delta}_{t:t+1}^{(k)}$
Quantitative characterization of excess risk		
(i) Query-wise excess risk	(11)	(15)
(ii) Later-stage excess risk	(13)	(18)
Structural characterization of predictive sufficiency		
(i) stage-t U-sufficiency and CPQ recovery	(11) and Theorem 4.1(i)	(16)
(ii) Persistence criterion	(12)	(17)
(ii) Zero transfer loss and CPQ recovery	(14)	Corollary 4.2(ii)

The Gaussian coefficient matrix $\tau_{t,U}(k) = P_U \mathbf{B}_{k,t}$ represents the prediction-rule profile $c \mapsto P_U \mathbf{B}_{k,t}\, \mathrm{resp}(c)$, not merely an action at one context. Its one-step update is the projection of (63) and has the block form

$$\tau_{t+1,U}(k) = P_U \mathbf{B}_{k,t+1} = P_U \big[\, (\mathbf{I}_d - \boldsymbol{K}_{k,t} \boldsymbol{x}_{t+1}^\top) \mathbf{B}_{k,t} \quad \boldsymbol{K}_{k,t} \,\big].$$

Under stage-t U-sufficiency, equality of these updated profiles within each summary class yields the two coefficient conditions in (17).

The general persistence and excess-risk identities permit any $u > t$, also in the Gaussian model. The Gaussian corollary focuses on $u = t + 1$ because the one-step update yields explicit necessary and sufficient coefficient conditions. Its pairwise later-stage excess-risk formula applies to every stage-t U-sufficient S; only with $S = T_{t,U}^\star$ does this quantity become the transfer loss $\mathcal{L}_{t \to t+1}(U)$. For CPQ recovery, the general theorem requires the additional hypotheses in Theorem 4.1(i). Here (P) is verified by (51), (R) by (52), and product-to-joint domination by (49); the projected squared-loss uniqueness (J) is verified in the common-full-set argument following (26). Thus the coefficient recovery criterion is applicable in this Gaussian model.

D.5 Three-state example separating the two coefficient conditions (Section 4.2)

Take $K = d = 3$, $\theta_k = 1/3$, $\sigma^2 = 1$, $\mathbf{X}_1 = (1, 0, 0)$, $\boldsymbol{q} = (1, 1, 0)^\top$, and $U = \mathrm{span}\{\boldsymbol{q}\}$. Set

$$\mathbf{V}_1 := \mathbf{I}_3, \qquad \mathbf{V}_2 := \begin{pmatrix} 2 & -1/2 & 0 \\ -1/2 & 5/4 & 1/30 \\ 0 & 1/30 & 1 \end{pmatrix}, \qquad \mathbf{V}_3 := 2\mathbf{I}_3.$$

Compare the three appended design vectors $\boldsymbol{x}_2^{\mathrm{A}} := (1, 1, 0)^\top$, $\boldsymbol{x}_2^{\mathrm{B}} := (1, 0, 1)^\top$, and $\boldsymbol{x}_2^{\mathrm{C}} := (1, 1, 1)^\top$. We verify the prior assumptions, stage-1 U-CPQ, two independent coefficient conditions, stage-2 U-CPQs, and exact transfer-loss formulas using the posterior identities of Section 4.2.

Positive definiteness and the stage-1 U-CPQ. The leading principal minors of $\mathbf{V}_2$ are 2, $9/4$, and $2023/900$, all strictly positive. Thus $\mathbf{V}_2 \succ 0$ by Sylvester's criterion; the other covariance matrices

are plainly positive definite. For $\mathbf{X}_1 = (1,0,0)$ we obtain

$$\mathbf{B}_{1,1} = \begin{pmatrix} 1/2 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{B}_{2,1} = \begin{pmatrix} 2/3 \\ -1/6 \\ 0 \end{pmatrix}, \qquad \mathbf{B}_{3,1} = \begin{pmatrix} 2/3 \\ 0 \\ 0 \end{pmatrix}.$$

Since $\boldsymbol{q} = (1,1,0)^\top$ and $P_U = \boldsymbol{q}\boldsymbol{q}^\top/2$, multiplying by $\boldsymbol{q}^\top$ gives $\boldsymbol{q}^\top\mathbf{B}_{1,1} = \boldsymbol{q}^\top\mathbf{B}_{2,1} = 1/2$ and $\boldsymbol{q}^\top\mathbf{B}_{3,1} = 2/3$. By the coefficient encoding (50) and profile equivalence (51), these projected prediction-rule profiles induce exactly the partition $\{1,2\} \mid \{3\}$. Choose $S = T^\star_{1,U} = \tau_{1,U}(Z)$, and denote $s_{12} \coloneqq \tau_{1,U}(1) = \tau_{1,U}(2)$ and $s_3 \coloneqq \tau_{1,U}(3)$. These are distinct CPQ labels, so $\{S = s_{12}\} = \{Z \in \{1,2\}\}$ and $\{S = s_3\} = \{Z = 3\}$. The stage-1 U-CPQ is therefore recovered exactly from S, and S is sufficient for U by Corollary 4.2(i).

To verify that this is a genuinely nontrivial distinction, let S_0 be constant, and write $\omega^{S_0}_{k,1} \coloneqq \mathbb{P}(Z = k \mid y_1)$ for the posterior weights. The state-1 and state-2 scalar predictions $\boldsymbol{q}^\top\mathbf{B}_{k,1}y_1$ are both $y_1/2$ ($k = 1,2$), while state 3 predicts $2y_1/3$. The two distinct statewise predictions, $y_1/2$ and $2y_1/3$, have conditional probabilities $1 - \omega^{S_0}_{3,1}$ and $\omega^{S_0}_{3,1}$. The two-point specialization of (56) gives

$$\mathrm{Var}\big(\mathbb{E}[\boldsymbol{q}^\top\boldsymbol{w} \mid y_1, Z] \mid y_1\big) = \omega^{S_0}_{3,1}(1-\omega^{S_0}_{3,1})\left(\frac{2y_1}{3} - \frac{y_1}{2}\right)^2 = \frac{\omega^{S_0}_{3,1}(1-\omega^{S_0}_{3,1})y_1^2}{36}.$$

Since $P_U = \boldsymbol{q}\boldsymbol{q}^\top/2$ and $\|\boldsymbol{q}\|^2 = 2$, $\|P_U\boldsymbol{h}\|^2 = (\boldsymbol{q}^\top\boldsymbol{h})^2/2$ for any $\boldsymbol{h} \in \mathbb{R}^3$. Thus Corollary 4.2(i) gives the scalar and projected excess risks:

$$R_{S_0,1}(\boldsymbol{q}) - R_{\mathrm{ora},1}(\boldsymbol{q}) = \frac{1}{36}\mathbb{E}[\omega^{S_0}_{3,1}(1-\omega^{S_0}_{3,1})y_1^2] > 0,$$
$$R^U_{S_0,1} - R^U_{\mathrm{ora},1} = \frac{1}{72}\mathbb{E}[\omega^{S_0}_{3,1}(1-\omega^{S_0}_{3,1})y_1^2] > 0.$$

Indeed, positive priors and the everywhere-positive Gaussian likelihoods imply $0 < \omega^{S_0}_{3,1} < 1$ almost surely, while $\mathbb{P}(y_1 = 0) = 0$. Thus the unequal coefficients $1/2 \neq 2/3$ fail the zero-risk condition in (16) for a constant summary; the displayed variance calculation shows directly why the two CPQ classes cannot be merged without strictly positive excess risk.

The two independent coefficient conditions. The stage-1 posterior covariances for states 1 and 2, obtained by substituting the first design row into (45), are

$$\mathbf{P}_{1,1} = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}_{2,1} = \begin{pmatrix} 2/3 & -1/6 & 0 \\ -1/6 & 7/6 & 1/30 \\ 0 & 1/30 & 1 \end{pmatrix}.$$

The only nontrivial state pair with $f(k) = f(\ell)$ is $(1,2)$. The first coefficient condition in (17) is equivalent, for our one-dimensional U, to $\boldsymbol{q}^\top\boldsymbol{K}_{1,1} = \boldsymbol{q}^\top\boldsymbol{K}_{2,1}$. The second condition in the same equation is equivalent to $\boldsymbol{q}^\top\boldsymbol{K}_{1,1}\boldsymbol{x}_2^\top(\mathbf{B}_{1,1} - \mathbf{B}_{2,1}) = 0$. If $\boldsymbol{x}_2 = (x_{2,1}, x_{2,2}, x_{2,3})^\top$, the last row product is

$$\boldsymbol{x}_2^\top(\mathbf{B}_{1,1} - \mathbf{B}_{2,1}) = (x_{2,2} - x_{2,1})/6.$$

Substituting the displayed posterior covariances into $\boldsymbol{K}_{k,1} = \mathbf{P}_{k,1}\boldsymbol{x}_2/(1 + \boldsymbol{x}_2^\top\mathbf{P}_{k,1}\boldsymbol{x}_2)$ gives

$$\boldsymbol{q}^\top\boldsymbol{K}_{1,1} = \frac{x_{2,1}/2 + x_{2,2}}{1 + x_{2,1}^2/2 + x_{2,2}^2 + x_{2,3}^2},$$
$$\boldsymbol{q}^\top\boldsymbol{K}_{2,1} = \frac{x_{2,1}/2 + x_{2,2} + x_{2,3}/30}{1 + 2x_{2,1}^2/3 - x_{2,1}x_{2,2}/3 + 7x_{2,2}^2/6 + x_{2,2}x_{2,3}/15 + x_{2,3}^2}.$$

Evaluating these expressions at the three appended designs, and using $\boldsymbol{x}_2^\top(\mathbf{B}_{1,1} - \mathbf{B}_{2,1}) = (x_{2,2} - x_{2,1})/6$, gives

	A	B	C
$\boldsymbol{q}^\top\boldsymbol{K}_{1,1}$	3/5	1/5	3/7
$\boldsymbol{q}^\top\boldsymbol{K}_{2,1}$	3/5	1/5	46/107
$\boldsymbol{q}^\top\boldsymbol{K}_{1,1}\boldsymbol{x}_2^\top(\mathbf{B}_{1,1} - \mathbf{B}_{2,1})$	0	−1/30	0

In A both conditions hold; in B only the first holds; in C only the second holds. In particular, in C the first projected-gain difference is $46/107 - 3/7 = 1/749 \neq 0$. In case B the projected update gains agree but $\boldsymbol{x}_2^\top \mathbf{B}_{1,1} = 1/2$ and $\boldsymbol{x}_2^\top \mathbf{B}_{2,1} = 2/3$ differ. In case C these two predictions of the new response coincide at $1/2$, while the projected update gains differ. Thus each separate condition is indispensable within this single Gaussian model with fixed stage-1 design and query subspace.

Stage-2 prediction-rule profiles. The one-step posterior update gives $\mathbf{B}_{k,2} = [(\mathbf{I}_3 - \boldsymbol{K}_{k,1}\boldsymbol{x}_2^\top)\mathbf{B}_{k,1}\ \boldsymbol{K}_{k,1}]$. Multiplying by $\boldsymbol{q}^\top$ yields

$$\boldsymbol{q}^\top \mathbf{B}_{k,2} = \big(\boldsymbol{q}^\top \mathbf{B}_{k,1} - (\boldsymbol{q}^\top \boldsymbol{K}_{k,1})(\boldsymbol{x}_2^\top \mathbf{B}_{k,1}),\ \boldsymbol{q}^\top \boldsymbol{K}_{k,1}\big)\,.$$

For $\boldsymbol{x}_2 = (x_{2,1}, x_{2,2}, x_{2,3})^\top$, the initial coefficients satisfy

$$\boldsymbol{x}_2^\top \mathbf{B}_{1,1} = x_{2,1}/2, \qquad \boldsymbol{x}_2^\top \mathbf{B}_{2,1} = 2x_{2,1}/3 - x_{2,2}/6, \qquad \boldsymbol{x}_2^\top \mathbf{B}_{3,1} = 2x_{2,1}/3.$$

For state 3, $\mathbf{P}_{3,1} = \mathrm{diag}(2/3, 2, 2)$, so

$$\boldsymbol{q}^\top \boldsymbol{K}_{3,1} = \frac{2x_{2,1}/3 + 2x_{2,2}}{1 + 2x_{2,1}^2/3 + 2x_{2,2}^2 + 2x_{2,3}^2}.$$

For example, in case C, the state-2 coefficient row is $\big(1/2 - (46/107)(1/2),\ 46/107\big) = (61/214,\ 46/107)$. Evaluating the same formula for every state and design gives

	$\boldsymbol{q}^\top \mathbf{B}_{1,2}$	$\boldsymbol{q}^\top \mathbf{B}_{2,2}$	$\boldsymbol{q}^\top \mathbf{B}_{3,2}$
A	$(1/5,\ 3/5)$	$(1/5,\ 3/5)$	$(2/11,\ 8/11)$
B	$(2/5,\ 1/5)$	$(11/30,\ 1/5)$	$(6/11,\ 2/11)$
C	$(2/7,\ 3/7)$	$(61/214,\ 46/107)$	$(6/17,\ 8/17)$

Using (50) and (51), the two rows for states 1 and 2 coincide only in A, while the row for state 3 remains distinct from those for states $1, 2$ in all three cases. Thus the stage-2 partitions are $\{1, 2\} \mid \{3\}$ for A and $\{1\} \mid \{2\} \mid \{3\}$ for B and C. Subtracting the coefficients for states 1 and 2 yields

$$\boldsymbol{q}^\top(\mathbf{B}_{2,2} - \mathbf{B}_{1,2})\boldsymbol{y}_2 = \begin{cases} 0, & \text{A}, \\ -y_1/30, & \text{B}, \\ (-y_1 + 2y_2)/1498, & \text{C}. \end{cases}$$

These are also the differences $\boldsymbol{q}^\top(\boldsymbol{\Delta}_{1:2}^{(2)} - \boldsymbol{\Delta}_{1:2}^{(1)})$ between the projected oracle increments: by the stage-2 posterior-mean update (62), the states' identical stage-1 projected prediction-rule profiles cancel in the subtraction.

Exact transfer loss and strict positivity. For the fixed two-class summary, the projected excess-risk matrix can also be obtained directly from the stage-2 prediction-rule profiles:

$$P_U \mathbf{A}_{S,2} P_U = \mathbb{E}\big[\mathbf{1}_{\{S=s_{12}\}}\omega_{1,2}\omega_{2,2}(P_U(\mathbf{B}_{2,2} - \mathbf{B}_{1,2})\boldsymbol{y}_2)(P_U(\mathbf{B}_{2,2} - \mathbf{B}_{1,2})\boldsymbol{y}_2)^\top\big]. \tag{64}$$

This matrix identity is the conditional-covariance form of (13), evaluated for the two-state mixture on $S = s_{12}$: the states have mixing weights $\omega_{1,2}$ and $\omega_{2,2}$, while $S = s_3$ identifies state 3 and contributes zero covariance. Since the stage-1 projected predictions of states $1, 2$ agree, their stage-2 projected prediction difference equals the difference of their projected oracle increments. On $S = s_{12}$, $\omega_{1,2} + \omega_{2,2} = 1$. The two-point specialization of (56) now gives the scalar trace (with $\|\boldsymbol{q}\|^2 = 2$):

$$\begin{aligned} \mathrm{tr}\,\mathrm{Cov}(\boldsymbol{\Delta}_{1:2}^{Z,U} \mid \mathcal{G}_2^S) &= \mathbf{1}_{\{S=s_{12}\}}\omega_{1,2}\omega_{2,2}\big\|P_U(\boldsymbol{\Delta}_{1:2}^{(2)} - \boldsymbol{\Delta}_{1:2}^{(1)})\big\|^2 \\ &= \frac{1}{2}\mathbf{1}_{\{S=s_{12}\}}\omega_{1,2}\omega_{2,2}\big\{\boldsymbol{q}^\top(\boldsymbol{\Delta}_{1:2}^{(2)} - \boldsymbol{\Delta}_{1:2}^{(1)})\big\}^2. \end{aligned}$$

By (13), the transfer loss $\mathcal{L}_{1\to2}(U)$ is the expectation of this conditional variance. Equivalently, it is the trace of (64). Substituting the three previously calculated statewise differences yields

$$\begin{aligned} \mathcal{L}_{1\to2}^{\mathrm{A}}(U) &= 0, \\ \mathcal{L}_{1\to2}^{\mathrm{B}}(U) &= \frac{1}{1800}\mathbb{E}_{\mathrm{B}}\big[\mathbf{1}_{\{S=s_{12}\}}\omega_{1,2}\omega_{2,2}y_1^2\big] > 0, \\ \mathcal{L}_{1\to2}^{\mathrm{C}}(U) &= \frac{1}{2\cdot 1498^2}\mathbb{E}_{\mathrm{C}}\big[\mathbf{1}_{\{S=s_{12}\}}\omega_{1,2}\omega_{2,2}(-y_1 + 2y_2)^2\big] > 0. \end{aligned}$$

Here $\mathbb{E}_B$ and $\mathbb{E}_C$ denote expectations under the corresponding appended designs, with the mixing weights evaluated under the same design. On $S = s_{12}$, both mixing weights are strictly positive because the two Gaussian likelihoods and their prior masses are positive, as also seen from (48). Moreover, $\mathbb{P}(S = s_{12}) = 2/3$, and each nonzero linear form y_1 or $-y_1 + 2y_2$ vanishes only on a null set under the everywhere-positive Gaussian-mixture density. Hence the losses in B and C are strictly positive. In A the projected increments of states $1, 2$ coincide, so the transfer loss is zero. This establishes the independent roles of the two coefficient conditions in persistence and transfer loss.

E Linear-Gaussian Experiment on Oracle Gain and Transfer Loss (Section 5)

We evaluate how oracle gain depends on the query and how retaining a fixed stage-1 U-CPQ affects excess risk as the context grows.

Experimental setting. We consider the linear-Gaussian model of Section 4.2 with $Z = (A, B)$, where A, B are independent uniform signs. We generate 48 independent covariance models by drawing α and β independently from $\mathrm{Unif}[0.25, 0.75]$ (model seed 20261102). With $d = 3$, $\sigma^2 = 1$, and the coordinate basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, we set

$$\mathbf{V}_{a,b} = \mathbf{I}_3 + \alpha(\boldsymbol{e}_1 + a\boldsymbol{e}_2)(\boldsymbol{e}_1 + a\boldsymbol{e}_2)^\top + \beta(\boldsymbol{e}_3 + b\boldsymbol{e}_2)(\boldsymbol{e}_3 + b\boldsymbol{e}_2)^\top.$$

In particular, $\mathbf{V}_{a,b} \succeq \mathbf{I}_3 \succ 0$ for all $a, b \in \{-1, +1\}$ and $\alpha, \beta > 0$. We use three deterministic 30-stage context orders, each beginning with $\boldsymbol{e}_1$: periodic $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$, periodic $(\boldsymbol{e}_1, \boldsymbol{e}_3, \boldsymbol{e}_2)$, and ten consecutive observations along each of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, in that order. Each order defines its own nested prefix sequence; we compare information available at a given stage, not permutations within a fixed context. At each prefix, $\boldsymbol{y}_t = \mathbf{X}_t\boldsymbol{w} + \boldsymbol{\varepsilon}_t$ as in Section 4.2.

Information conditions. For $\boldsymbol{q}_0 = (\boldsymbol{e}_1 + \boldsymbol{e}_2)/\sqrt{2}$ and $U = \mathrm{span}\{\boldsymbol{q}_0\}$, the stage-1 U-CPQ has two classes, encoded by the fixed state-only summary $S_0 = A$. For $\boldsymbol{x}_1 = \boldsymbol{e}_1$, $\sigma^2 = 1$, and $\boldsymbol{q}_0 = (\boldsymbol{e}_1 + \boldsymbol{e}_2)/\sqrt{2}$, the coefficient is

$$\boldsymbol{q}_0^\top \mathbf{B}_{(a,b),1} = \frac{1 + \alpha(1 + a)}{\sqrt{2}\,(2 + \alpha)}.$$

It is independent of b and takes distinct values for $a = \pm 1$ since $\alpha > 0$, giving the stated CPQ partition by (50) and (51). We compare three information conditions: the observable instance alone, the observable instance with S_0, and the observable instance with Z. Let $R_{\mathrm{obs},t}(\boldsymbol{q})$ denote the observable-instance Bayes risk (equivalently, the Bayes risk for a constant summary S), and put $Q_t = R_{\mathrm{obs},t}(\boldsymbol{q}_0) - R_{\mathrm{ora},t}(\boldsymbol{q}_0)$, the oracle gain of the stage-t projected problem, and $L_t = R_{S_0,t}(\boldsymbol{q}_0) - R_{\mathrm{ora},t}(\boldsymbol{q}_0)$, the excess risk of the fixed stage-1 U-CPQ. Since $\|\boldsymbol{q}_0\| = 1$, scalar and one-dimensional projected risks agree by (24). As $S_0 = A$ encodes $T^\star_{1,U}$ up to bijective relabeling, it recovers $T^\star_{1,U}$, so Proposition 3.8 applied to the stage-1 projected problem gives $L_1 = 0$; the same relabeling gives $L_t = \mathcal{L}_{1\to t}(U)$ for $t \geq 2$.

Evaluation protocol. For the fixed query, we report the U-CPQ class count, Q_t/V_q, L_t/V_q, and L_t/Q_t, where $V_q = \boldsymbol{q}_0^\top \bar{\mathbf{V}} \boldsymbol{q}_0 = 1 + \alpha + \beta/2$ and $\bar{\mathbf{V}} = \mathbb{E}[\mathbf{V}_Z]$ for each covariance model. In all 48 covariance models, the computed U-CPQ has two or four classes at every reported stage (panel (a) of Figure E.1). Hence $Q_t > 0$ by (16) for a constant summary and the coefficient characterization (50). Moreover, $0 \leq L_t \leq Q_t$ by (6), so $L_t/Q_t \in [0, 1]$. For query dependence, we evaluate the maximum normalized oracle gain $[R_{\mathrm{obs},t}(\boldsymbol{q}) - R_{\mathrm{ora},t}(\boldsymbol{q})]/(\boldsymbol{q}^\top \bar{\mathbf{V}} \boldsymbol{q})$ over nonzero $\boldsymbol{q}$, and the normalized gains for $\boldsymbol{q} \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{q}_0\}$.

We compute risk expectations using 24-point-per-dimension Gauss–Hermite quadrature in the at-most-three-dimensional sufficient statistics: writing $N_t = \mathrm{diag}(n_{1,t}, n_{2,t}, n_{3,t})$ for the coordinate counts and $\boldsymbol{s}_t = \mathbf{X}_t^\top \boldsymbol{y}_t$ for the response sums, $\boldsymbol{s}_t \mid Z = k \sim \mathcal{N}(\mathbf{0}, N_t \mathbf{V}_k N_t + N_t)$, with zero-count coordinates omitted in quadrature. For three representative models (indices 0, 23, and 47), 24- and 32-point quadrature differed by at most 2.655×10^{-9} in normalized gain across all context orders and stages. We independently check the results using 8192 Monte Carlo realizations per model (sample seeds 20261103 plus the zero-based model index); this check is not the source of the reported values.

We compute U-CPQ classes from $P_U \mathbf{B}_{k,t}$ using (50) and (51), and numerical risks from (15). We do not directly compute the one-step update coefficients.

Results. *Transfer of the fixed stage-1 U-CPQ.* Figure E.1 shows, for $\boldsymbol{q}_0$, the U-CPQ class count, Q_t/V_q, L_t/V_q, and L_t/Q_t in panels (a)–(d), respectively. For the prefix $(\boldsymbol{e}_1, \boldsymbol{e}_2)$, the Gaussian coefficient formula (43) gives

$$\mathbf{X}_2 \mathbf{V}_{a,b} \mathbf{X}_2^\top + \mathbf{I}_2 = \begin{pmatrix} 2+\alpha & \alpha a \\ \alpha a & 2+\alpha+\beta \end{pmatrix},$$

$$\boldsymbol{q}_0^\top \mathbf{V}_{a,b} \mathbf{X}_2^\top = \frac{1}{\sqrt{2}}(1+\alpha+\alpha a,\ 1+\alpha+\beta+\alpha a).$$

Both expressions are independent of b because $a^2 = b^2 = 1$. Consequently, the query coefficients depend on a alone, and $S_0 = A$ preserves oracle risk at stage 2. For the prefix $(\boldsymbol{e}_1, \boldsymbol{e}_3)$, direct substitution gives

$$\mathbf{B}_{(a,b),2} = \begin{pmatrix} \dfrac{1+\alpha}{2+\alpha} & 0 \\ \dfrac{\alpha a}{2+\alpha} & \dfrac{\beta b}{2+\beta} \\ 0 & \dfrac{1+\beta}{2+\beta} \end{pmatrix}.$$

By (50) and (51), the coefficients are constant across states for queries $\boldsymbol{e}_1, \boldsymbol{e}_3$ but distinguish all four states for $\boldsymbol{e}_2, \boldsymbol{q}_0$. In particular, for fixed a, the two states differing in b have coefficient difference

$$\boldsymbol{q}_0^\top (\mathbf{B}_{(a,1),2} - \mathbf{B}_{(a,-1),2}) = \left(0, \frac{\sqrt{2}\beta}{2+\beta}\right).$$

On $\{S_0 = a\}$, since $\|\boldsymbol{q}_0\| = 1$, the pairwise formula (58) gives the conditional variance underlying the scalar risk gap:

$$\omega_{(a,1),2}\omega_{(a,-1),2} \frac{2\beta^2}{(2+\beta)^2} y_2^2.$$

Both posterior weights are strictly positive by (48), and $y_2 = 0$ has probability zero under the non-degenerate Gaussian mixture, as in the argument leading to (51). Thus the fixed-summary transfer loss is strictly positive for this prefix.

For the prefix $(\boldsymbol{e}_1, \boldsymbol{e}_3)$, the median $100L_2/V_q$ is 2.126%. For the grouped context order, S_0 remains sufficient through $t = 20$, but the median $100L_{21}/V_q$ is positive (0.0065%) at $t = 21$, although nearly invisible on the main figure's linear scale. The medians of model-wise peak excess risks are 0.271%, 2.126%, and 0.0101% at stages 3, 2, and 30, respectively (the last is a horizon endpoint). The median counts of stages with $L_t/V_q \geq 0.005$ are 0, 1, and 0, respectively. Here B determines $\mathrm{Cov}(w_2, w_3 \mid Z) = \beta B$, so an observation along $\boldsymbol{e}_3$ can make the previously omitted distinction in B relevant to the fixed query. The measured risk penalty nevertheless depends strongly on the preceding context. When S_0 is stage-$(t-1)$ U-sufficient, its stage-t U-sufficiency and one-step excess risk are characterized by (17) and (18), respectively; for later prefixes we apply (15) to S_0 held fixed.

Query dependence of oracle gain. Figure E.2 shows the query dependence. For the prefix $(\boldsymbol{e}_1, \boldsymbol{e}_3)$, the coefficient matrix computed above gives constant U-CPQs for queries $\boldsymbol{e}_1, \boldsymbol{e}_3$, so Proposition 3.8 yields zero oracle gain in those directions. It separates all four states for $\boldsymbol{e}_2, \boldsymbol{q}_0$, consistently with the positive-gain criterion of Theorem 3.14 under the verified Gaussian conditions. At $t = 2$, the median normalized oracle gains for $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{q}_0$ are 0, 10.776%, 0, and 6.040%, respectively. For the three context orders listed above, the medians of model-wise maximum normalized oracle gains are 7.412%, 10.776%, and 11.106%; the median peak stages are 1, 2, and 10, respectively. The first context order peaks at $t = 1$ for 25 models and at $t = 2$ for 23; the grouped context order peaks at $t = 10$ for 27 and at $t = 20$ for 21 models. The median numbers of stages with maximum normalized oracle gain at least 0.5% are 7, 8, and 21. The maximum is recomputed for each model and stage; these counts do not measure persistence of gain for a fixed query.

Scope. The query-dependent results illustrate Corollary 4.2(i), and the transfer results illustrate part (ii), for the sampled covariance models and prescribed context orders. No predictor is trained; the numerical magnitudes do not establish general persistence or gain bounds.

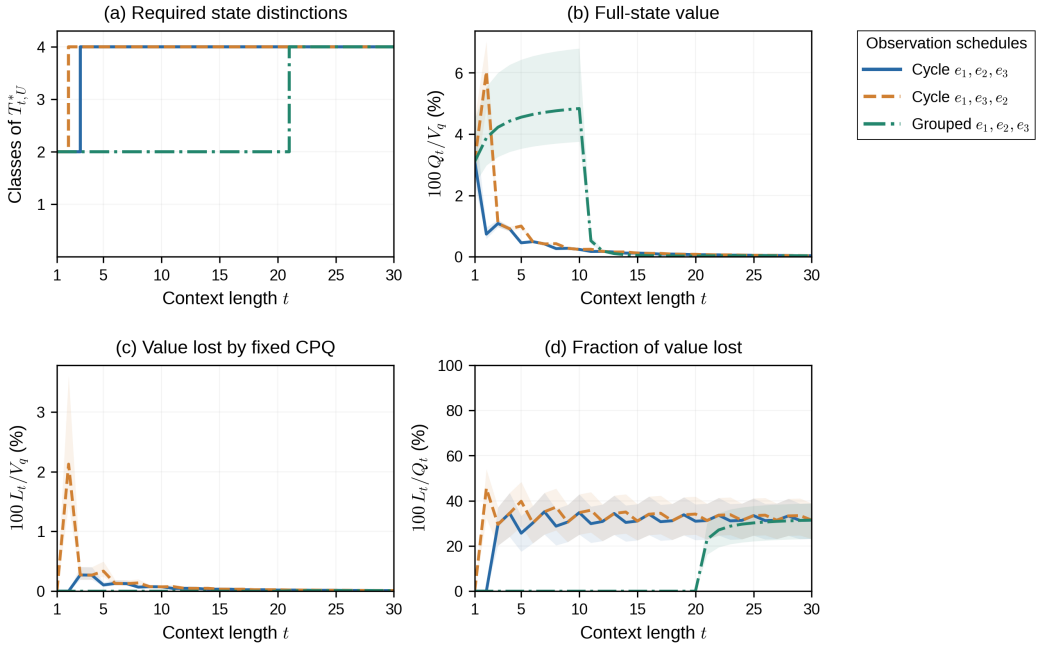

Figure E.1: Fixed-summary results for $U = \mathrm{span}\{\boldsymbol{q}_0\}$: (a) number of U-CPQ classes; (b) oracle gain $100Q_t/V_q$; (c) excess risk of the fixed stage-1 U-CPQ $100L_t/V_q$; (d) fraction $100L_t/Q_t$. Risk curves show medians with between-model interquartile bands, not confidence intervals. The main-text figure repeats panels (a) and (c) without changing their data.

F Transformer Experiment with Supplied State Information

We evaluate how effectively a trained in-context predictor uses supplied prediction-relevant state information. The CPQ theory provides Bayes-risk identities but does not determine the risk attained by a finite Transformer.

Experimental setting. For each episode, we independently sample $Z = (Z_1, Z_2)$, where Z_1, Z_2 are independent and uniform on $\{0, 1\}$. For a realized state $Z = (c, r)$, the latent coefficient has law $\boldsymbol{w} \mid Z = (c, r) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{c,r})$, where $\mathbf{V}_{c,r} = \mathrm{diag}(v_c, u_r)$ with $(v_0, v_1) = (0.05, 2.0)$ and $(u_0, u_1) = (0.2, 1.8)$. The context $C = ((\boldsymbol{x}_i, y_i))_{i=1}^4$ contains two inputs along $\boldsymbol{e}_1$ and two along $\boldsymbol{e}_2$, each with its own independently sampled standard-normal scalar coefficient; the four observation pairs are randomly ordered, and $y_i = \boldsymbol{x}_i^\top \boldsymbol{w} + \varepsilon_i$ with independent $\varepsilon_i \sim \mathcal{N}(0, 0.6^2)$. For an independent uniform sign $\gamma \in \{-1, +1\}$, the query is $\boldsymbol{x}_\mathrm{q} = \gamma \boldsymbol{e}_1$. The input coefficients, observation noises, and query sign are mutually independent and independent of $(Z, \boldsymbol{w})$. The observable instance is $\nu = (C, \boldsymbol{x}_\mathrm{q})$, and the query response of Definition 2.1 is the noiseless signal $y_\mathrm{q} = \boldsymbol{x}_\mathrm{q}^\top \boldsymbol{w} = \gamma w_1$.

Information conditions. We define $s(C) = \sum_{i:x_{i2}=0} x_{i1}^2$ and $t(C) = \sum_{i:x_{i2}=0} x_{i1} y_i$. Scalar squared-loss optimality (Lemma C.4) identifies the oracle prediction with the conditional mean; the Gaussian posterior formula for this experiment gives

$$\mathbb{E}[y_\mathrm{q} \mid \nu, Z] = \gamma \frac{v_{Z_1} t(C)}{0.6^2 + v_{Z_1} s(C)}.$$

The oracle Bayes prediction depends on Z_1 but not Z_2, whereas Z_2 changes the law of the second-coordinate observations. This identifies the reconstruction condition (R) for Z_1. For (P), the selected prediction rules in states with distinct values of Z_1 differ by

$$\gamma \frac{0.6^2 (v_0 - v_1) t(C)}{(0.6^2 + v_0 s(C))(0.6^2 + v_1 s(C))}.$$

The denominators are positive and $v_0 \neq v_1$. The continuous input coefficients give $s(C) > 0$ almost surely; conditional on the inputs and state, $t(C)$ has positive Gaussian variance, so $\mathrm{Law}(\nu)(t(C) =$

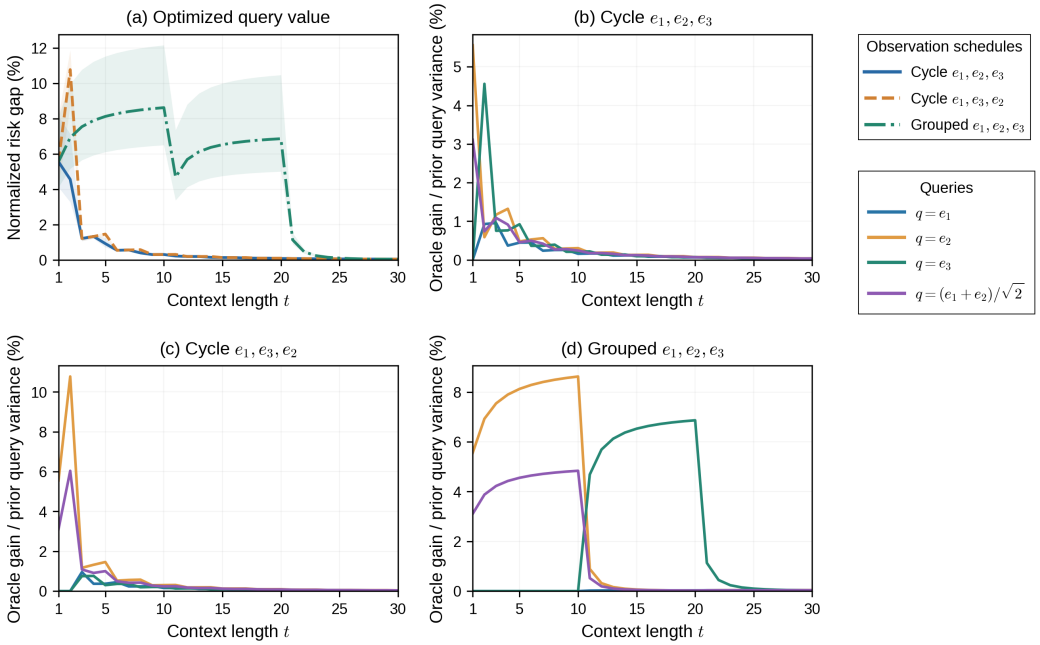

Figure E.2: Query dependence of oracle gain from (15) with a constant summary. (a) Maximum normalized oracle gain over queries; (b)–(d) normalized gains for four prescribed query directions under the three context orders. Curves show the 48-model medians; shading in (a) shows the interquartile range.

$0) = 0$. States with the same Z_1 have identical prediction rules, and those with different Z_1 have different rules $\mathrm{Law}(\nu)$-almost everywhere. All four states have positive mass, so (P) holds on the common full set. Thus $T^\star = Z_1$ is a CPQ with two classes by Definition 3.6. We compare three information conditions: $j = 0$ (the observable instance alone), $j = 1$ (the observable instance with Z_1 supplied), and $j = 2$ (the observable instance with Z supplied). For these conditions, let S_0 be constant, $S_1 = Z_1$, and $S_2 = Z$, and write $R_j := R_{S_j}$ for their Bayes risks ($j = 0, 1, 2$), so $R_0 = R_{\mathrm{obs}}$. The corresponding Bayes risks, computed from the exact mixture posterior, are $R_0 \simeq 0.26429$ and $R_1 = R_2 = R_{\mathrm{ora}} \simeq 0.20924$; $R_1 = R_{\mathrm{ora}}$ follows from the CPQ reconstruction in Proposition 3.8, whereas $S_2 = Z$ gives $R_2 = R_{\mathrm{ora}}$ by the oracle-risk definition. The oracle gain is $G(Z) = R_0 - R_{\mathrm{ora}} \simeq 0.05505$. These are Bayes-risk benchmarks, not empirical conclusions about the trained model.

Evaluation protocol. We report empirical squared-error risk, achieved oracle gain, and sensitivity to prediction-irrelevant observations. We train one causal Transformer shared across all three information conditions: four pre-normalized layers, four attention heads, embedding width 64, learned absolute position embeddings, and a final layer normalization followed by a linear readout at the query token. Four demonstration input–response pairs are interleaved before the query token. The input includes two separate state-label fields, each encoding either an undisclosed value or its binary value; the Z_1 field is identical for $j = 1$ and $j = 2$, while only the Z_2 field changes. Training samples one of the three conditions uniformly for each episode, using squared prediction loss, Adam, 256 episodes per update, and 30,000 updates with a 1,000-step learning-rate warm-up to 10^{-3} and gradient-norm clipping at 1. We repeat training with three initial-weight seeds 11, 23, and 37, holding the training episode sequence fixed across seeds. All conditions are evaluated on the same independent 20,000 episodes, with model parameters fixed at inference.

For condition j, let f_j and b_j denote the trained and Bayes predictor maps based on (ν, S_j), respectively. For evaluation episode i, write $S_{j,i}$ for the realized summary; the corresponding predictions are $f_{j,i} = f_j(\nu_i, S_{j,i})$ and $b_{j,i} = b_j(\nu_i, S_{j,i})$. We measure achieved gain relative to the *observable-*

instance Bayes predictor by

$$\widehat{d}_{\mathrm{state},j} = \frac{1}{N}\sum_{i=1}^{N}\big[(y_{\mathrm{q},i} - b_{0,i})^2 - (y_{\mathrm{q},i} - f_{j,i})^2\big], \quad j = 1, 2,$$

with $N = 20{,}000$. For a fresh episode, define $d_{\mathrm{state},j} \coloneqq (y_{\mathrm{q}} - b_0(\nu, S_0))^2 - (y_{\mathrm{q}} - f_j(\nu, S_j))^2$ for $j = 1, 2$. For fixed learned weights, let $E_j \coloneqq \mathbb{E}[(f_j(\nu, S_j) - b_j(\nu, S_j))^2] \geq 0$ denote the *Bayes gap* [16] at information condition j, namely the difference between the trained predictor's risk and the Bayes risk with the same information (ν, S_j). The final layer normalization (with a positive stabilizer) followed by a linear readout makes f_j bounded and measurable for fixed weights; hence $f_j \in L^2$ and $E_j < \infty$. Since $b_j(\nu, S_j) = \mathbb{E}[y_{\mathrm{q}} \mid \nu, S_j]$,

$$\mathbb{E}[(y_{\mathrm{q}} - b_j(\nu, S_j))(f_j(\nu, S_j) - b_j(\nu, S_j))] = 0.$$

Expanding the squared errors therefore gives

$$\mathbb{E}[(y_{\mathrm{q}} - f_j(\nu, S_j))^2] = R_j + E_j = R_{\mathrm{ora}} + E_j \quad (j = 1, 2),$$
$$\mathbb{E}[(y_{\mathrm{q}} - b_0(\nu, S_0))^2] = R_0 = R_{\mathrm{obs}},$$

where $R_1 = R_2 = R_{\mathrm{ora}}$ was established above. Consequently, the population expectation satisfies $\mathbb{E}[d_{\mathrm{state},j}] = G(Z) - E_j$. The Bayes gap E_j is distinct from the *excess risk* $R_S - R_{\mathrm{ora}}$ due to withholding state information. We report $\widehat{d}_{\mathrm{state},j}/G(Z)$ as the fraction of available oracle gain achieved in the evaluation sample; it need not equal the population fraction. The numerator uses the empirical loss of b_0 (0.2664624) on the same episodes, not the population risk R_0. We obtain 95% percentile intervals from 2,000 paired episode-level bootstrap resamples, separately for each fixed trained model.

To probe a different property of the learned predictor, we preserve the first-coordinate demonstrations, all context inputs, the query, and the supplied labels, but independently regenerate the second-coordinate latent coefficient and its two noisy responses conditional on the realized second state component $Z_{2,i}$, obtaining a context C_i'. The Bayes-optimal predictions remain unchanged. Indeed, Z_1 and Z_2 are independent and $\boldsymbol{w} \mid Z \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(v_{Z_1}, u_{Z_2}))$, so (Z_1, w_1) and (Z_2, w_2) are independent; the observation noises are independent of both. Hence, given the input coefficients, coordinate assignments, ordering, γ, and S_j $(j = 0, 1, 2)$, the pair consisting of w_1 and the first-coordinate responses is conditionally independent of the second-coordinate responses, so $\mathbb{E}[y_{\mathrm{q}} \mid \nu, S_j]$ with $y_{\mathrm{q}} = \gamma w_1$ does not depend on the latter.

We measure

$$\widehat{A}_{\mathrm{nuis},j} = \frac{1}{N}\sum_{i=1}^{N}\big[f_j((C_i', \boldsymbol{x}_{\mathrm{q},i}), S_{j,i}) - f_j((C_i, \boldsymbol{x}_{\mathrm{q},i}), S_{j,i})\big]^2.$$

This quantity measures sensitivity to irrelevant *context observations*, not sensitivity to the irrelevant state component Z_2 itself.

Results. *Achieved oracle gain and prediction risks.* Table F.1 reports the empirical prediction risks alongside the Bayes-risk benchmarks, and Figure F.1(a) shows achieved oracle gain. Across the three training seeds, the CPQ condition attains 89.4%–90.7% of $G(Z)$, versus approximately 91%–92% with the full state. The measured CPQ risk is below the analytic observable-instance Bayes risk R_0 in all three runs, consistent with effective use of the supplied state information. The measured risk difference between the CPQ and full-state conditions is 0.00035–0.00095, reflecting differences in learned approximation rather than a difference in the Bayes risks $R_1 = R_2 = R_{\mathrm{ora}}$ established above. The paired 95% interval for this difference contains zero for seed 37 but not for seeds 11 and 23; no equivalence of trained-model risks is inferred.

Sensitivity to prediction-irrelevant observations. Figure F.1(b) shows nonzero measured sensitivity in all nine cases: it ranges from 0.00076 to 0.00173 with no state label, from 0.00052 to 0.00102 with the CPQ label, and from 0.00037 to 0.00069 with the full state. As the Bayes predictions are unchanged by this resampling, the nonzero values show residual dependence of the trained predictors on prediction-irrelevant observations. These are descriptive diagnostics, not inferential comparisons between conditions.

Scope. Sufficiency of the supplied information does not imply that a trained predictor attains oracle risk or ignores irrelevant observations. These results assess the use of supplied CPQ or full-state

Table F.1: Mean squared prediction errors on 20,000 shared evaluation episodes. Each seed denotes one model trained across all three conditions; Bayes risks are analytic benchmarks.

Information supplied	Bayes	Seed 11	Seed 23	Seed 37
None ($j = 0$)	0.2643	0.2718	0.2722	0.2721
CPQ Z_1 ($j = 1$)	0.2092	0.2173	0.2169	0.2165
Full Z ($j = 2$)	0.2092	0.2163	0.2161	0.2162

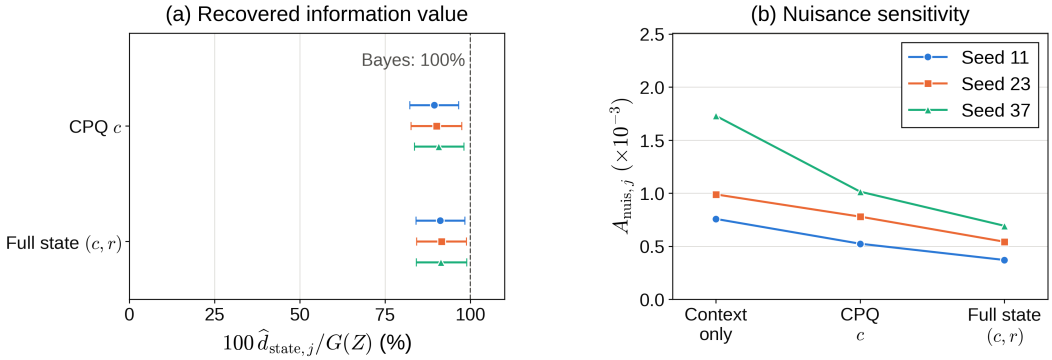

Figure F.1: Two properties of the trained predictor. (a) Achieved gain $100\widehat{d}_{\mathrm{state},j}/G(Z)$ with 95% paired-bootstrap intervals over evaluation episodes, conditional on fixed trained weights; the dashed line is the Bayes limit. (b) Squared prediction change (in 10^{-3}) after regenerating second-coordinate context responses, which leave Bayes predictions unchanged. Colored lines identify the three initial-weight seeds.

labels, not their learning or recovery from demonstrations. The nuisance factor Z_2 is irrelevant to the full conditional query-response law in this example; distinct response laws sharing a Bayes act are not examined. Bootstrap intervals are conditional on fixed trained weights and do not measure variation over initializations. We do not evaluate context-prefix transfer or the recovery of an internal CPQ representation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction present the CPQ as a state-only encoding of selected prediction-rule profiles, with oracle-risk preservation under the stated measurability and reconstruction conditions. They distinguish the realized value of state information from the CPQ partition, state the additional uniqueness and domination assumptions for conditional minimality, and describe context-prefix sufficiency and its linear-Gaussian specialization. Claims about recovery by trained predictors are explicitly excluded.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 distinguishes decision-theoretic information requirements from CPQ recovery or internal mechanisms of learned predictors. It identifies the assumptions needed for CPQ minimality and the need for estimable diagnostics and stability theory for approximate or nonunique Bayes acts. The limited scope of the supplied-label Transformer experiment is also stated in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions for the general prediction problem, CPQ construction, conditional minimality, and context–query sufficiency are stated in Sections 2, 3, and 4. The proofs, measurable-encoding constructions, and Gaussian calculations are provided in Appendices B, C, and D. The conditional minimality claim explicitly requires realized uniqueness and product-to-joint domination.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The two synthetic experiments and their principal numerical protocols are described in Section 5 and Appendices E and F. The linear-Gaussian construction is specified in detail, but the Transformer experiment does not provide executable code or all architecture and implementation choices needed for exact replication, such as its feed-forward sublayer specification and the seed of the fixed training-episode sequence.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: At camera-ready submission, the experimental source code and executable reproduction instructions are not publicly available. Both experiments use synthetic data, with their generation and evaluation protocols described in Appendices E and F. The author plans to release the source code after camera-ready submission.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E specifies the linear-Gaussian covariance family, three 30-stage context orders, 48 sampled models, queries, normalization, and numerical evaluation. Appendix F specifies the four-demonstration synthetic task, three information conditions, four-layer/four-head Transformer with width 64, Adam optimization, batch size 256, 30,000 updates, warm-up, clipping, three initialization seeds, and 20,000 shared evaluation episodes. Experimental settings were determined in advance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The linear-Gaussian plots show medians and between-model interquartile bands over 48 covariance models; these are not confidence intervals (Appendix E). The Transformer experiment uses 2,000 paired episode-level bootstrap resamples to obtain 95% intervals for fixed trained weights, and reports three initial-weight seeds separately (Appendix F). The intervals do not quantify variation over training initializations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Both experiments were small-scale, and each was completed within one hour, including all runs. All computations used only the CPU of a 2025 Apple Mac Studio with an Apple M4 Max chip and 36 GB of unified memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The work develops prediction theory and evaluates synthetic data-generating models, including a small Transformer trained for the study. It reports no human participants, personal-data collection, or release of a high-risk pretrained model. The claims about trained predictors are restricted to the stated synthetic setting; compliance with the Code of Ethics remains the author's responsibility.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is foundational decision-theoretic research and does not propose a deployed decision system, collect personal data, or establish a direct application with identifiable societal effects. Consequently, it does not claim specific positive or negative societal impacts; downstream applications are outside the investigated setting.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk model, scraped dataset, or deployable system. Its experiments use synthetic data and a small Transformer trained for the study, so safeguards for releasing such high-risk assets are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No third-party research code, datasets, or pretrained models were used in either experiment. The experiments use synthetic data and models trained specifically for this study.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No dataset, pretrained model, benchmark, or software package is presented or released as a new reusable research asset. The synthetic experimental data and reported results support the paper's theoretical analysis.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The experiments use synthetic observations and computational models only. There is no crowdsourcing, user study, human-subject experiment, or collection of data from human participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The research does not involve human subjects or crowdsourcing, so institutional review board approval or equivalent human-subject review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The research questions, the CPQ framework, and the central ideas of the main theorems were conceived by the author. Generative AI tools were used in developing and refining the mathematical formulations, claims, and proof arguments, including proposing candidate arguments and proof steps. They also assisted with literature review, manuscript preparation, figures, and supporting code. The author independently rederived and verified the final mathematical claims and proofs, checked the cited sources and bibliographic information, reviewed the figures and supporting code for consistency with the reported results, and takes full responsibility for the correctness and originality of the paper. Generative AI is not a component of the CPQ framework itself and was not used as a computational component in the reported experiments; its use is declared because it contributed to the development of the core theoretical results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.