Contextual Observability and Grammar Singularity for Compositional Task Families

Manoj Saravanan [1] **Rohit Kumar Salla** [1] **Shrikar Reddy Kota** [1]

Abstract

We study Bayesian meta-learning when tasks are sampled from a latent compositional family rather than treated independently. A task family is modeled as a stochastic grammar over reusable linear modules, and grouped datasets are observed across many tasks. The target is the latent program law together with the shared module library, modulo explicit language symmetries. We prove four results. First, an exact sufficient-statistic reduction turns grouped compositional meta-learning into a finite mixture of matrix-normal laws. Second, local quotient identifiability holds under positive linearized observability and component separation, while in the single-occurrence regime zero contextual observability yields exact non-identifiability. Third, the posterior for the grouped predictive law contracts at the explicit non-i.i.d. rate

$$\delta_{m,n}^2 = \frac{K_{\text{eff}} \log m + rd^2 \log(mn)}{m},$$

and structural contraction occurs at the inverse exponent κ: $\kappa = 1$ in regular families but $\kappa = 2$ in an explicit duplicated-module singular family. Fourth, matching minimax lower bounds identify the same hardness parameters. These quantities yield a theory-native benchmark with provable phase transitions in observability, singularity, and anchor exposure.

1. Introduction

Modern evaluation is increasingly organized around *families* of related tasks rather than isolated test points. In compositional regimes, however, benchmark scores remain fundamentally retrospective: they say how a model performed on a fixed suite, but not when it should transfer to *new compositions* of known primitives, when posterior uncertainty about reusable structure should persist, or which benchmark cells are statistically most informative. In such settings the natural statistical object is not a single task but a *task-generating family*. We study that object directly. Each task first draws a latent program Z_i from a law $\pi^\star$ on an admissible language $\mathcal{L}$, then composes a shared module library $\Theta^\star$ into an operator A_{Z_i}, and finally generates within-task data. The target is therefore

$$\eta^\star = (\pi^\star, \Theta^\star),$$

defined only up to explicit language symmetries. This grouped-data perspective puts transfer, compositional generalization, and uncertainty over latent task identity into one model: the learner must infer both the *distribution over task programs* and the *reusable module library* from many related datasets.

Recent theory isolates important pieces of this picture, but not the whole object. Special autoregressive task families can permit dramatic extrapolation from small support to exponentially larger compositional families (Abedsoltan et al., 2025). Modular solution discovery and connected-support conditions clarify when reusable modules can be identified compositionally (Schug et al., 2024). Coverage and path-ambiguity analyses show that compositional generalization can fail for structural rather than purely optimization-related reasons (Chang et al., 2025). On the statistical side, posterior contraction for non-i.i.d. experiments (Ghosal & van der Vaart, 2007; Ghosal et al., 2000), singular learning theory (Watanabe, 2009), weakly identifiable mixture geometry (Ho & Nguyen, 2016; 2019), and classical finite-mixture identifiability (Teicher, 1963; Yakowitz & Spragins, 1968) provide the right asymptotic tools. To our knowledge, these threads have not yet been assembled into a finite-sample Bayesian theory of *latent task-program learning* that simultaneously treats identifiability, predictive versus structural contraction, singularity, and theorem-native benchmark design.

The statistical hardness of this problem is organized by three quantities:

$$K_{\text{eff}}, \qquad \Delta_{\text{ctx}}, \qquad \kappa,$$

[1]Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. Correspondence to: Manoj Saravanan <manoj663@vt.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

namely effective grammar complexity, contextual observability, and grammar singularity order. Here K_{eff} controls the difficulty of learning the program law, Δ_{ctx} measures whether primitives are distinguishable in the contexts actually generated by the family, and κ is the inverse exponent that converts predictive proximity into structural proximity. Two auxiliary moduli, λ_{lin} and γ_{comp}, determine when positive observability upgrades to local quotient identifiability. The main message is that *prediction and structure can contract at different rates*: regular families have $\kappa = 1$, while singular families can have $\kappa > 1$.

Our first result is an exact sufficient-statistic reduction: conditional on the within-task designs, least-squares task statistics are sufficient, and grouped meta-learning becomes a finite mixture of matrix-normal laws indexed by latent programs. Our second result is an observability theory: zero contextual observability produces exact observational collapse in single-occurrence languages, whereas $\lambda_{\mathrm{lin}} > 0$ and $\gamma_{\mathrm{comp}} > 0$ imply local quotient identifiability. Our third result is a predictive-versus-structural contraction theory: the grouped predictive law contracts at rate $\delta^2_{m,n} = (K_{\mathrm{eff}} \log m + rd^2 \log(mn))/m$, and structural contraction occurs at the inverse exponent κ. In an explicit duplicated-module family the first-order geometry cancels, so structure only contracts at $O_\Pi(\delta^{1/2}_{m,n}/\sqrt{n})$. Our final result is a matching minimax theory, including grammar-law, anchored-module, zero-observability, and singular lower bounds. These same parameters define a theorem-native benchmark whose cells are labeled by $(K_{\mathrm{eff}}, \Delta_{\mathrm{ctx}}, \lambda_{\mathrm{lin}}, \gamma_{\mathrm{comp}}, \kappa, s_{\mathrm{anc}})$.

Section 2 introduces stochastic linear task grammars and the quotient parameterization. Section 3 states the main identifiability, contraction, and minimax theorems. Section 4 summarizes the proof architecture. Section 5 presents the theorem-native benchmark and empirical validation. All proofs are deferred to the appendix.

2. Stochastic Linear Task Grammars

We work with a finite-language stochastic grammar over reusable linear modules. This is a theorem-friendly specialization of recent compositional task-family models in which tasks are generated by latent programs built from reusable components (Schug et al., 2024; Abedsoltan et al., 2025). The restriction to bounded-depth linear modules is deliberate: it is the smallest setting in which one can simultaneously prove quotient identifiability, predictive posterior contraction, singular structural rates, and matching lower bounds.

Latent programs and grouped task data. Fix integers $r \geq 2$, $d \geq 1$, and $L \geq 1$, let $\mathcal{Z}_{L,r} := \bigcup_{t=1}^{L} [r]^t$, and let $\mathcal{L} \subseteq \mathcal{Z}_{L,r}$ be a finite nonempty admissible language. A program $z = (z_1, \ldots, z_t) \in \mathcal{L}$ composes a module library $\Theta = (A_1, \ldots, A_r) \in \mathbb{A}_B^r$, where $\mathbb{A}_B := \{A \in \mathbb{R}^{d\times d} : \|A\|_{\mathrm{op}} \leq B\}$, via

$$A_z := A_{z_t} \cdots A_{z_1}.$$

A stochastic linear task grammar is

$$\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L}) := \mathfrak{P}^\circ_{\mathcal{L}} \times \mathbb{A}_B^r,$$
$$\mathfrak{P}^\circ_{\mathcal{L}} := \{\pi \in (0,1)^{\mathcal{L}} : \sum_{z\in\mathcal{L}} \pi_z = 1\}.$$

For each task i, draw $Z_i \sim \pi$, then $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and

$$y_{ij} = A_{Z_i} x_{ij} + \varepsilon_{ij},$$
$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d), \qquad j = 1, \ldots, n.$$

With $X_i = [x_{i1}, \ldots, x_{in}]$, $Y_i = [y_{i1}, \ldots, y_{in}]$, and $D_i = (X_i, Y_i)$, the one-task grouped law is

$$Q^{(n)}_\eta = \sum_{z\in\mathcal{L}} \pi_z \, P^{(n)}_{A_z}, \tag{1}$$

and the meta-dataset satisfies $D_1, \ldots, D_m \overset{\text{i.i.d.}}{\sim} Q^{(n)}_\eta$. Throughout, the truth is $\eta^\star = (\pi^\star, \Theta^\star) \in \Xi(B, \mathcal{L})$ with $\pi^\star_z \geq \pi_{\min} > 0$.

Quotient parameterization. A statistical task family is unchanged by relabeling primitive symbols whenever the admissible language is preserved. The relevant symmetry group is

$$\mathrm{Aut}(\mathcal{L}) := \{\sigma \in S_r : \sigma \cdot \mathcal{L} = \mathcal{L}\},$$

acting on parameters by $(\sigma \cdot \pi)_u = \pi_{\sigma^{-1}\cdot u}$ and $(\sigma \cdot \Theta)_a = A_{\sigma^{-1}(a)}$. Structural error is measured by the quotient metric

$$d_q(\eta, \eta') := \min_{\sigma\in\mathrm{Aut}(\mathcal{L})} \Big(\|\pi - \sigma\cdot\pi'\|_1 + \sum_{a=1}^{r} \|A_a - (\sigma\cdot\Theta')_a\|_F \Big). \tag{2}$$

This quotient removes only explicit symbol symmetries. Any additional equality of grouped laws that is *not* induced by an automorphism is treated as genuine statistical non-identifiability rather than as notational redundancy. This distinction becomes essential in the low-observability regime.

Proposition 2.1 (Exact reduced experiment)**.** *Assume $n \geq d + 1$. Define the task-level least-squares statistic*

$$\widehat{W}_i := Y_i X_i^\top (X_i X_i^\top)^{-1}, \qquad i = 1, \ldots, m,$$

which is well defined almost surely under the Gaussian design. Then, conditional on X_i and $Z_i = z$,

$$\widehat{W}_i \sim \mathrm{MN}_{d,d}\big(A_z, \sigma^2 I_d, (X_i X_i^\top)^{-1}\big),$$

where $\mathrm{MN}_{d,d}(M,U,V)$ *denotes the matrix-normal law with mean* M*, row covariance* U*, and column covariance* V*. Consequently, conditional on the designs* $X_{1:m}$*, the reduced observations satisfy*

$$Z_i \overset{\text{i.i.d.}}{\sim} \pi, \qquad \widehat{W}_i = A_{Z_i} + \Xi_i, \tag{3}$$
$$\Xi_i \mid X_{1:m} \overset{\text{ind}}{\sim} \mathrm{MN}_{d,d}\big(0, \sigma^2 I_d, (X_i X_i^\top)^{-1}\big).$$

Moreover, under any prior on η*,*

$$\Pi(\cdot \mid X_{1:m}, Y_{1:m}) = \Pi(\cdot \mid X_{1:m}, \widehat{W}_{1:m}) \qquad \textit{a.s.}$$

Theorem 2.1 is the technical pivot of the paper: Bayesian meta-learning over latent task grammars becomes an exact grouped finite-mixture problem over the composed operators $\{A_z\}_{z\in\mathcal{L}}$. All posterior theory and all experiments in the paper are carried out in this reduced experiment.

3. Main Results

At a high level, the theory separates into two layers. The first is a *predictive* layer: after exact sufficient-statistic reduction, Bayesian meta-learning over latent task grammars becomes a non-i.i.d. grouped mixture problem, and posterior contraction is governed by the reduced predictive geometry. The second is a *structural* layer: converting predictive concentration into localization of the latent grammar requires a local inverse inequality, and the resulting exponent is regular ($\kappa = 1$) or singular ($\kappa > 1$) depending on the local geometry. The novelty is that both layers are derived for a stochastic *task-family grammar*, not for an ordinary finite mixture or a single compositional function.

Throughout, $\Pi_m(\cdot)$ denotes the posterior given the grouped data $(X_i, Y_i)_{i=1}^m$, equivalently given the reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$ by Theorem 2.1. We write

$$\delta_{m,n}^2 := \frac{K_{\text{eff}} \log m + rd^2 \log(mn)}{m}, \qquad K_{\text{eff}} := |\mathcal{L}| - 1. \tag{4}$$

Hardness quantities. Besides K_{eff}, four structural quantities govern the problem:

$$\Delta_{\text{ctx}}(\eta) := \min_{a\neq b} \sum_{(u,v)\in\mathcal{C}_{\mathcal{L}}} \nu_\pi(u,v)\, \|A_v(A_a - A_b)A_u\|_F^2, \tag{5}$$

$$\lambda_{\text{lin}}(\eta) := \inf_{\|H\|_{\text{lib}}=1} \|D\Phi_\Theta(H)\|_\pi^2,$$

$$\Phi(\Theta) := (A_z(\Theta))_{z\in\mathcal{L}}, \tag{6}$$

$$\gamma_{\text{comp}}(\eta) := \min_{z\neq z'} \|A_z - A_{z'}\|_F. \tag{7}$$

The quantity Δ_{ctx} is the exact context-averaged information carried by one-hole substitutions, λ_{lin} is the weighted smallest singular value of the Jacobian of the program-operator map, and γ_{comp} is the minimum separation between composed operators. The *grammar singularity order* $\kappa \geq 1$ is the smallest local exponent such that $\bar{h}_X^2(\eta,\eta^\star) \gtrsim d_q(\eta,\eta^\star)^{2\kappa}$. Regular families have $\kappa = 1$; singular families have $\kappa > 1$.

3.1. Identifiability and predictive contraction

Theorem 3.1 (Local quotient identifiability)**.** *Suppose*

$$\lambda_{\text{lin}}(\eta^\star) > 0 \qquad \textit{and} \qquad \gamma_{\text{comp}}(\eta^\star) > 0.$$

Then there exists $\rho_{\text{id}} > 0$ *such that for every* $n \geq d+1$*,*

$$d_q(\eta,\eta^\star) \leq \rho_{\text{id}} \quad \textit{and} \quad Q_\eta^{(n)} = Q_{\eta^\star}^{(n)} \implies d_q(\eta,\eta^\star) = 0.$$

Thus pairwise visibility and full local invertibility are distinct: Δ_{ctx} measures substitution-level information, whereas λ_{lin} controls all infinitesimal perturbations of the library.

Proposition 3.2 (Exact non-identifiability at zero contextual observability)**.** *Assume that* $\mathcal{L}$ *is* single-occurrence*, meaning that no symbol appears more than once in any program* $z \in \mathcal{L}$*. If there exist* $a \neq b$ *such that*

$$\Delta_{\text{ctx}}(a,b;\eta^\star) = 0,$$

then there exists $\widetilde{\eta} \not\sim \eta^\star$ *such that*

$$Q_{\widetilde{\eta}}^{(n)} = Q_{\eta^\star}^{(n)} \qquad \textit{for every } n \geq 1.$$

So $\Delta_{\text{ctx}} = 0$ is not merely a difficult regime: in single-occurrence languages it yields exact observational collapse.

Let

$$\bar{h}_X^2(\eta,\eta^\star) := \frac{1}{m}\sum_{i=1}^m h^2\big(q_\eta(\cdot \mid X_i), q_{\eta^\star}(\cdot \mid X_i)\big)$$

denote the empirical conditional Hellinger distance in the reduced experiment.

Theorem 3.3 (Predictive posterior contraction)**.** *Assume the model and prior conditions of Appendix A, and suppose the design regularity regime*

$$me^{-c_{d,\delta}n} \to 0 \qquad \textit{for some fixed } 0 < \delta < 1.$$

Then there exists $M_{\text{pred}} \in (0,\infty)$ *such that for every* $M \geq M_{\text{pred}}$*,*

$$\Pi_m\big(\eta : \bar{h}_X(\eta,\eta^\star) > M\delta_{m,n}\big) \longrightarrow 0 \tag{8}$$

in $P_{\eta^\star}$*-probability.*

The grouped predictive radius is therefore governed by the finite-dimensional program-law complexity K_{eff} and the ambient module dimension rd^2.

Proposition 3.4 (Oracle latent-program recovery)**.** *Assume $\gamma_{\mathrm{comp}}(\eta^\star) > 0$. Then on the high-probability design event $\mathcal{G}_{m,n}(\delta)$, for every $i \in [m]$,*

$$\mathbb{E}_{\eta^\star}\left[1 - \Pi_{\eta^\star}\left(Z_i = Z_i^\star \mid X_i, \widehat{W}_i\right) \,\middle|\, X_i, Z_i^\star\right] \le C_{\mathrm{lat}} \exp\left(-\frac{(1-\delta)n\,\gamma_{\mathrm{comp}}(\eta^\star)^2}{8\sigma^2}\right), \tag{9}$$

and the same bound holds after averaging over $i = 1, \dots, m$.

This isolates the within-task information scale for latent-program recovery: $n\gamma_{\mathrm{comp}}(\eta^\star)^2$.

3.2. Structural localization and minimax lower bounds

Predictive contraction does not by itself localize the latent grammar. The missing ingredient is a local inverse inequality from predictive distance back to quotient parameter distance.

Theorem 3.5 (Structural contraction from a local inverse inequality)**.** *Assume the design regime of Theorem 3.3. Suppose there exist constants*

$$\kappa \ge 1, \qquad \rho_\star > 0, \qquad c_\star > 0, \qquad \underline{h}_\star > 0$$

such that, on $\mathcal{G}_{m,n}(\delta)$,

$$d_q(\eta, \eta^\star) \le \rho_\star \Longrightarrow \bar{h}_X^2(\eta, \eta^\star) \ge c_\star d_q(\eta, \eta^\star)^{2\kappa}, \tag{10}$$

$$d_q(\eta, \eta^\star) \ge \rho_\star \Longrightarrow \bar{h}_X(\eta, \eta^\star) \ge \underline{h}_\star. \tag{11}$$

Then there exists $M_{\mathrm{str}} \in (0, \infty)$ such that for every $M \ge M_{\mathrm{str}}$,

$$\Pi_m\left(\eta : d_q(\eta, \eta^\star) > M\delta_{m,n}^{1/\kappa}\right) \longrightarrow 0 \tag{12}$$

in $P_{\eta^\star}$-probability.

Corollary 3.6 (Regular structural contraction)**.** *Suppose*

$$\lambda_{\mathrm{lin}}(\eta^\star) > 0, \qquad \gamma_{\mathrm{comp}}(\eta^\star) > 0,$$

and assume the global separation condition Equation (11). Then the local inverse exponent is $\kappa = 1$, and there exists $M_{\mathrm{reg}} \in (0, \infty)$ such that

$$\Pi_m(\eta : d_q(\eta, \eta^\star) > M\delta_{m,n}) \longrightarrow 0 \tag{13}$$

for every $M \ge M_{\mathrm{reg}}$.

Proposition 3.7 (Explicit quadratic singularity)**.** *Consider the duplicated family*

$$\eta_t = \left(\left(\tfrac{1}{2}, \tfrac{1}{2}\right), (1+t,\, 1-t)\right), \qquad t \ge 0,$$

with truth $t^\star = 0$. Then

$$d_q(\eta_t, \eta_0) = 2t, \qquad \bar{h}_X^2(\eta_t, \eta_0) \asymp n^2 d_q(\eta_t, \eta_0)^4$$

locally around $t = 0$. Consequently, there exists $M_{\mathrm{sing}} \in (0, \infty)$ such that

$$\Pi_m\left(\eta_t : d_q(\eta_t, \eta_0) > M\,\frac{\delta_{m,n}^{1/2}}{\sqrt{n}}\right) \longrightarrow 0 \tag{14}$$

for every $M \ge M_{\mathrm{sing}}$.

This is the sharp structural gap of the paper: predictive learning can remain regular while grammar recovery is singular.

For a parameter class $\mathcal{H} \subseteq \Xi(B, \mathcal{L})$, write

$$\mathfrak{R}_{m,n}(\mathcal{H}; d_q) := \inf_{\widehat{\eta}} \sup_{\eta \in \mathcal{H}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)]\,.$$

Theorem 3.8 (Minimax lower bounds)**.** *The following lower bounds hold.*

1. ***Grammar-law complexity.*** *There exists a finite submodel $\mathcal{H}_{\mathrm{gram}} \subseteq \Xi(B, \mathcal{L})$ with a fixed module library such that*

$$\mathfrak{R}_{m,n}(\mathcal{H}_{\mathrm{gram}}; d_q) \gtrsim \sqrt{\frac{K_{\mathrm{eff}}}{m}}. \tag{15}$$

2. ***Anchored modules.*** *If $\mathcal{L}$ contains an s-anchor set, then there exists a finite anchored submodel $\mathcal{H}_{\mathrm{anc}}$ such that*

$$\mathfrak{R}_{m,n}(\mathcal{H}_{\mathrm{anc}}; d_q) \gtrsim \sigma d\sqrt{\frac{s}{mn}}. \tag{16}$$

3. ***Zero observability.*** *If there exist η_0, η_1 with*

$$d_q(\eta_0, \eta_1) = \Delta_0 > 0, \qquad Q_{\eta_0}^{(n)} = Q_{\eta_1}^{(n)} \quad \text{for all } n, \tag{17}$$

then

$$\mathfrak{R}_{m,n}(\{\eta_0, \eta_1\}; d_q) \ge \frac{\Delta_0}{4}. \tag{18}$$

4. ***Singular duplicated family.*** *In the duplicated family of Theorem 3.7,*

$$\mathfrak{R}_{m,n}(\{\eta_0, \eta_{t_{m,n}}\}; d_q) \gtrsim \frac{\sigma}{m^{1/4}\sqrt{n}}, \qquad t_{m,n} \asymp \frac{\sigma}{m^{1/4}\sqrt{n}}. \tag{19}$$

The four parts of Theorem 3.8 expose the statistical phase diagram of the problem. The grammar law contributes an unavoidable $\sqrt{K_{\mathrm{eff}}/m}$ term. Explicit anchor tasks recover the classical $\sqrt{1/(mn)}$ shared-parameter scaling. Zero contextual observability creates an exact non-identifiability floor. And overfitted duplicated-module families are singular, with the sharp $m^{-1/4}n^{-1/2}$ structural rate. These are precisely the hardness axes used to build the benchmark in Section 5.

4. Proof Architecture

The proof architecture has two logically distinct layers:

$$\begin{aligned}(X_i, Y_i)_{i=1}^m &\equiv (X_i, \widehat{W}_i)_{i=1}^m \\ &\Longrightarrow \bar{h}_X(\eta, \eta^\star) \lesssim \delta_{m,n} \\ &\Longrightarrow d_q(\eta, \eta^\star) \lesssim \delta_{m,n}^{1/\kappa}.\end{aligned}$$

The first implication is an exact reduction to a grouped non-i.i.d. mixture experiment; the second is a local inverse problem whose exponent is regular or singular. This is the same conceptual split that appears in singular statistical models and weakly identifiable mixtures, but here the latent object is a *task-family grammar* rather than an ordinary mixture component.

Reduction and identifiability. The reduction theorem shows that, conditional on the design X_i and latent program $Z_i = z$, the least-squares statistic $\widehat{W}_i = Y_i X_i^\top (X_i X_i^\top)^{-1}$ is matrix normal around the composed operator A_z, while the residual is conditionally independent of η. Thus the posterior given grouped data is exactly the posterior given $(X_i, \widehat{W}_i)_{i=1}^m$. Identifiability then has two layers: the grouped law first determines the induced mixing measure on composed operators by common-covariance kernel injectivity, and $\lambda_{\mathrm{lin}}(\eta^\star) > 0$ with $\gamma_{\mathrm{comp}}(\eta^\star) > 0$ upgrades that statement to local quotient identifiability of the grammar itself.

Predictive contraction and decoding. Conditional on $X_{1:m}$, the reduced observations are independent but non-identically distributed, so the relevant predictive metric is the empirical conditional Hellinger distance $\bar{h}_X$. The predictive theorem is proved by verifying the Ghosal–van der Vaart conditions in this reduced experiment: exponentially consistent tests for $\bar{h}_X$-separated alternatives, entropy bounds at radius $\delta_{m,n}$, and local prior thickness of Kullback–Leibler neighborhoods. Once the true task-family parameter is fixed, latent-program recovery reduces to responsibilities in a finite Gaussian mixture, and the decoding error decays at the within-task information scale $n\gamma_{\mathrm{comp}}(\eta^\star)^2$.

Inverse geometry and lower bounds. Structural contraction is an inverse problem. In the regular regime, deconvolved Fourier witnesses produce a local inverse inequality of the form $\bar{h}_X^2(\eta, \eta^\star) \gtrsim d_q(\eta, \eta^\star)^2$. In the duplicated family, by contrast, the first-order term cancels exactly and the reduced χ^2-divergence admits a closed-form quartic expansion, yielding $\bar{h}_X^2(\eta_t, \eta_0) \asymp n^2 d_q(\eta_t, \eta_0)^4$ and hence $\kappa = 2$. The lower bounds then follow from explicit subexperiments: Fano packings for grammar-law complexity, oracle anchored submodels for exposed modules, an exact Le Cam two-point argument for zero observability, and a product-χ^2 calculation for the duplicated singular family.

5. Theory-Native Benchmark and Experiments

Our evaluation philosophy is *theorem-native*: the benchmark is not an external testbed to which theory is retrofitted, but a direct instantiation of the stochastic task-grammar model analyzed above. Each benchmark cell is generated from a known latent task family and is released with exact hardness labels

$$\big(K_{\mathrm{eff}},\ \Delta_{\mathrm{ctx}},\ \lambda_{\mathrm{lin}},\ \gamma_{\mathrm{comp}},\ \kappa,\ s_{\mathrm{anc}}\big),$$

computed from the definitions in Section 3 rather than inferred post hoc. In particular, the benchmark generator is designed so that the hardness axes correspond exactly to the statistical phase structure isolated by the theorems.

All experiments are run in the reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$, which is Blackwell equivalent to the full grouped-data experiment by Theorem 2.1. This keeps the empirical section aligned with the proved model and removes optimization confounders unrelated to the statistical questions studied here. Posterior inference is exact by low-dimensional grid integration in the observability and singular suites, and MAP plus exact-Hessian Laplace approximation in the theorem-matched regular and anchored suites. The primary metrics are $R_{\mathrm{pred},0.9}$, $R_{\mathrm{str},0.9}$, $R_{\mathrm{anc},0.9}$, and $M_Z := m^{-1} \sum_{i=1}^m \Pi_m(Z_i = Z_i^\star)$; curves show medians over 32 replications.

To conserve main-paper space, we place the regular-family sanity checks in Appendix I. Those figures verify that, in the regular regime, normalization by the theoretical predictive and structural scales substantially stabilizes the posterior radii, and that the $q = 4$ family exhibits the predicted local inverse law $\bar{h}_X(\eta, \eta^\star)^2 \asymp d_q(\eta, \eta^\star)^2$. The main text instead concentrates on the three benchmark regimes that most sharply diagnose the new hardness parameters: observability, singularity, and anchor exposure.

Observability phase transition. We use

$$\mathcal{L}_{\mathrm{obs}} = \{(1,3), (2,3)\}, \qquad \pi^\star = \left(\tfrac{1}{2}, \tfrac{1}{2}\right),$$

with

$$A_1^\star = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right], \qquad A_2^\star = \left[\begin{smallmatrix} 0 & 0 \\ 2 & 0 \end{smallmatrix}\right], \qquad A_3^\star(\rho) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & \rho \end{smallmatrix}\right].$$

Then $\Delta_{\mathrm{ctx}}(\eta_\rho^\star) = \rho^2/2$ and $\gamma_{\mathrm{comp}}(\eta_\rho^\star) = \rho$. Figure 1 shows the predicted collapse of the latent-program error $1 - M_Z$ against the theorem scale $n\Delta_{\mathrm{ctx}}$. At $\rho = 0$, both $R_{\mathrm{str},0.9}$ and $1 - M_Z$ remain bounded away from zero as n increases, matching Theorems 3.2 and 3.8.

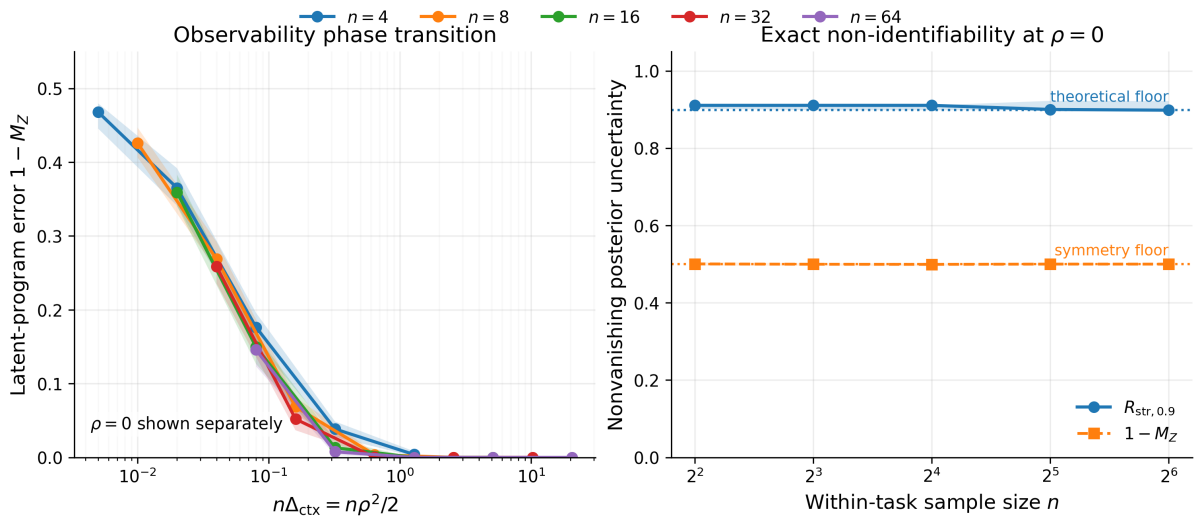

Figure 1. **Observability phase transition.** Left: latent-program error collapses when plotted against the theorem scale $n\Delta_{\mathrm{ctx}} = n\rho^2/2$, confirming that the within-task information budget is the relevant observable quantity. Right: at $\rho = 0$, both $R_{\mathrm{str},0.9}$ and $1 - M_Z$ remain bounded away from zero, confirming exact non-identifiability. Curves show medians; bands are 10th–90th percentiles.

Singularity gap: prediction is regular, structure is singular. We next evaluate the duplicated-module family

$$\mathcal{L}_{\mathrm{sing}} = \{(1), (2)\}, \qquad \pi^\star = \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \\ A_1(t) = 1 + t, \quad A_2(t) = 1 - t$$

with truth $t^\star = 0$. This is the explicit $\kappa = 2$ family of Theorem 3.7. The theory predicts

$$R_{\mathrm{pred},0.9} = O(\delta_{m,n}), \qquad R_{\mathrm{str},0.9} = O\left(\frac{\delta_{m,n}^{1/2}}{\sqrt{n}}\right).$$

Figure 2 confirms both claims and directly validates the quartic inverse law $\bar{h}_X(\eta_t, \eta_0)^2 \asymp n^2 d_q(\eta_t, \eta_0)^4$.

Anchors as benchmark design interventions. Finally, consider

$$\mathcal{L}_s^{\mathrm{anc}} = \{(1), \ldots, (s)\} \cup \{(s+1), (s+1, s+2)\},$$

where the first s programs expose anchor modules in isolation, and the matched no-anchor family

$$\mathcal{L}_s^{\mathrm{noanc}} = \{(1, s+1), \ldots, (s, s+1)\} \\ \cup \{(s+1), (s+1, s+2)\},$$

where the same modules appear only through a shared context. By Theorem 3.8, the natural normalized anchor error is

$$\frac{R_{\mathrm{anc},0.9}\sqrt{mn}}{d\, s^{3/2}}.$$

The left panel of Figure 3 shows that this quantity stays order one across the anchor-count grid. The right panel isolates the benchmark-design effect: explicit anchor tasks materially reduce anchor-module posterior radius relative to matched no-anchor cells.

Taken together, the main-text experiments validate three theorem-guided hardness mechanisms: observability controls the within-task phase transition for latent-program recovery, singularity separates predictive and structural learning, and anchor exposure changes which latent modules are directly identifiable from benchmark cells. The benchmark is therefore not merely compatible with the theory; it is organized by the theory.

These three suites also isolate genuinely different statistical resources. In the observability family the grammar law is fixed and only the within-task information budget $n\Delta_{\mathrm{ctx}}$ is varied, so the phase transition cannot be mistaken for poor cross-task coverage. In the singular family the grouped predictive experiment is regular while the local inverse map is quartic, so larger m improves predictive fit faster than it resolves latent structure. In the anchor family the ambient noise level and module class are held fixed while only exposure pattern changes. That separation is why the plots should be read as benchmark-design interventions on identifiability, not merely as descriptive summaries of synthetic data.

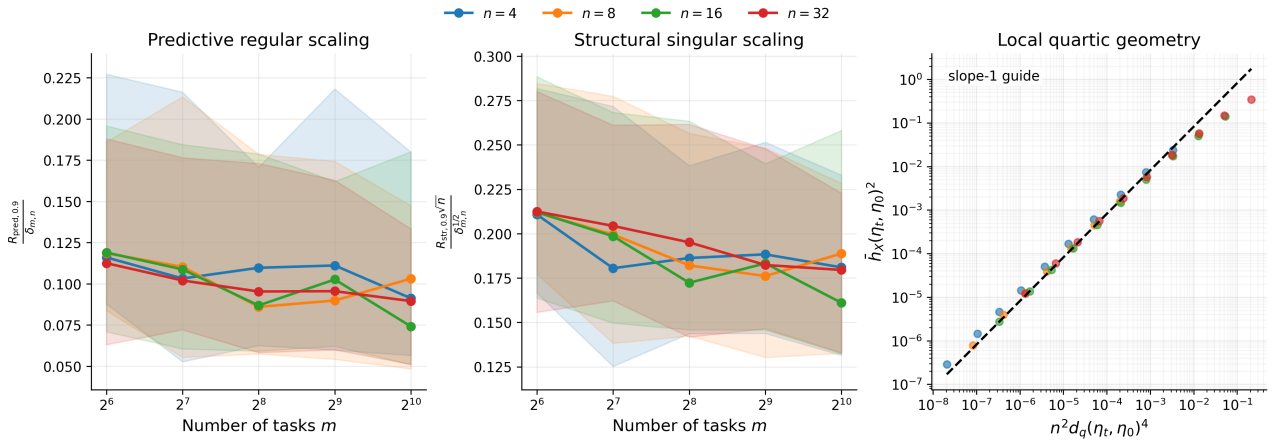

Figure 2. **Singularity gap in the duplicated-module family.** *Left:* after normalization by the regular predictive scale $\delta_{m,n}$, the predictive posterior radius $R_{\mathrm{pred},0.9}$ is approximately stable across m, consistent with regular predictive learning. *Middle:* after normalization by the singular structural scale $\delta_{m,n}^{1/2}/\sqrt{n}$, the structural posterior radius $R_{\mathrm{str},0.9}$ is approximately stable across m, confirming slower structural contraction. *Right:* direct local-geometry validation of the quartic law $\bar{h}_X^2 \asymp n^2 d_q^4$ underlying $\kappa = 2$. Curves show medians; bands are 10th–90th percentiles.

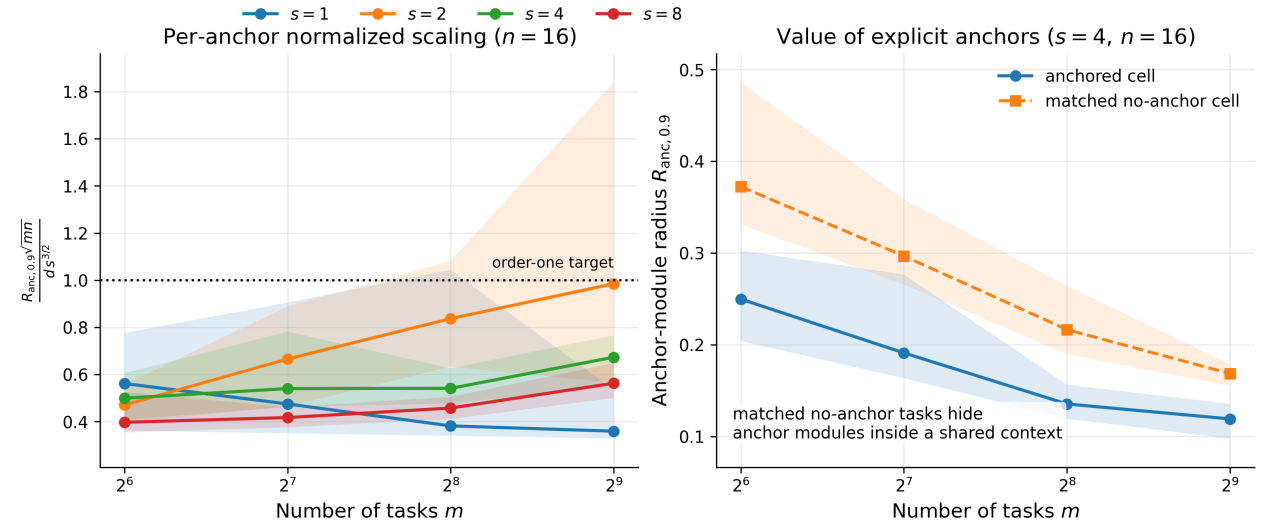

Figure 3. **Anchored-module scaling and the value of explicit anchors.** *Left:* after per-anchor normalization, the anchor-module radius remains order one across the anchor-count grid, consistent with the predicted $d\, s^{3/2}/\sqrt{mn}$ scaling for the summed anchor-module error. *Right:* benchmark cells with explicit anchor tasks yield substantially smaller anchor-module posterior radius than matched no-anchor cells, in which anchor modules appear only inside a shared context. Curves show medians; bands are interquartile ranges.

6. Discussion

The main conceptual shift in this paper is to treat the *task-generating family*, rather than the individual task, as the statistical object of interest. Once tasks are modeled as draws from a latent compositional grammar, transfer across related tasks, uncertainty over latent task identity, module reuse, and the effect of benchmark design on identifiability become part of one posterior problem. In that formulation, effective grammar complexity, contextual observability, and singularity are not metaphors; they are explicit hardness parameters that govern upper and lower bounds.

A second message is that *prediction and structure are not the same learning problem*. In regular families, predictive contraction and structural localization occur at the same order. In singular families, the inverse map from predictive laws back to latent task structure flattens, so structural uncertainty can decay strictly more slowly than predictive uncertainty. For evaluation, this means that strong held-out performance on related tasks does not by itself certify rapid recovery of the reusable compositional structure.

A third message is methodological: benchmark design should be viewed as a *statistical intervention*. In our analysis, observability, anchor exposure, and component separation are not cosmetic properties of a benchmark suite; they determine identifiability and rates. This viewpoint is complementary to recent work on benchmark compression and score prediction (Procaccia et al., 2025; Zhao et al., 2024): rather than selecting representative metrics or imputing missing scores after a benchmark is fixed, we expose ex ante hardness knobs that can be dialed deliberately when constructing benchmark cells.

Our results are proved for finite-language stochastic linear grammars, and the experiments validate the theory inside theorem-matched synthetic subfamilies. We therefore do not claim that present-day foundation models are exactly described by the model class studied here. The point is more structural: if one wants guarantees for transfer across compositional task families, then observability, inverse geometry, and benchmark-cell design must enter the theory explicitly.

From that perspective, the benchmark generator should be read as a controlled diagnostic instrument. Grammar-law complexity determines how many distinct tasks are needed before the latent program law can be learned. Observability and component separation determine whether within-task data are even informative enough to decode latent programs. Singularity determines whether predictive success can be converted into structural certainty. Anchor exposure determines which parts of the library are directly visible to the learner. Separating these mechanisms matters in practice: they correspond to different failure modes, different remedies, and different claims that an evaluation suite can responsibly support.

The most important next step is to extend this theorem-benchmark cycle beyond linear grammars without losing the quotient and inverse-geometry viewpoint. Natural directions include bounded-arity tree grammars, nonlinear module classes, controlled model misspecification, and larger benchmark suites whose cells are labeled ex ante by statistical hardness rather than only scored ex post by predictive accuracy. Even when the present model is only an approximation, the distinction between predictive and structural uncertainty, and the role of benchmark design in mediating that distinction, should remain central.

The rates also separate cross-task and within-task sample budgets. The term $K_{\mathrm{eff}} \log m/m$ is a grammar-law burden that only additional tasks can reduce, whereas observability and singularity operate through the within-task experiment and the local inverse geometry. A suite can therefore be rich in examples per task yet underpowered for recovering the law over programs, or conversely span many tasks while still revealing too little within-task information to decode latent structure. For theorem-native evaluation, m and n should be treated as distinct resources rather than collapsed into a single total sample size.

One practical implication is that theorem-native benchmark cells can disaggregate failure modes that are otherwise collapsed into a single held-out score. Poor performance may reflect insufficient cross-task coverage of the grammar law, insufficient within-task evidence to decode latent programs, hidden singularity in the inverse map from predictive behavior back to structure, or the simple absence of anchor exposure for the modules one hopes to identify. Those

cases call for different responses. More tasks help the first, redesigned cells help the second and fourth, and the third requires acknowledging persistent structural uncertainty even when predictive fit is strong.

Finally, the quotient formulation clarifies what successful recovery can mean. Even in favorable regular regimes the target is an equivalence class under $\mathrm{Aut}(\mathcal{L})$, not a preferred symbolic labeling of modules. Conversely, when grouped laws coincide for reasons not induced by an automorphism, the ambiguity is substantive and should be reported as non-identifiability. The $\rho = 0$ observability cell is informative precisely because it separates unavoidable statistical collapse from mere relabeling symmetry.

More broadly, we view the present linear setting as an existence proof that benchmark construction can itself be placed on a statistical footing. The lesson is not that realistic foundation-model pipelines are linear, but that benchmark suites should ideally be accompanied by formal statements about what their cells reveal and what they cannot reveal. In that sense, hardness labels play the role of experimental metadata: they make evaluation results easier to interpret, easier to compare across cells, and harder to overstate.

Impact Statement

This paper develops a statistical theory of compositional task families and uses that theory to construct theorem-native benchmark cells with explicit hardness labels. The main positive impact is methodological. Our results suggest that benchmark design should expose the structural quantities that determine what can and cannot be learned—for example observability, anchor exposure, and singularity—rather than relying only on aggregate held-out accuracy. In settings where one wants guarantees for transfer across related tasks, this can lead to more informative evaluation suites, finer-grained failure analysis, and better calibrated uncertainty about latent reusable structure. Because the benchmark families studied here are synthetic and theorem-matched, the work does not directly increase model capabilities or provide a recipe for scaling a deployed system.

The main risk is *over-interpretation*. Our guarantees are proved for finite-language stochastic linear grammars and experimentally validated only within theorem-matched synthetic regimes. If the results were treated as applying unchanged to realistic foundation-model pipelines, they could create false confidence about identifiability, transfer, or uncertainty decomposition in settings that violate the model assumptions. A related risk is that theorem-native benchmark generators, if used carelessly, could themselves become targets for narrow benchmark-specific optimization rather than broader scientific understanding. We therefore view the intended use of this work as *diagnostic* rather than *certificatory*: hardness labels should guide evaluation design and hypothesis testing, not substitute for robustness checks, external validation, or real-world safety assessment.

Finally, the paper may help reduce a common failure mode in evaluation: conflating predictive success with structural understanding. Our singular-family analysis shows that a model can achieve regular predictive contraction while remaining substantially uncertain about the underlying compositional structure. Making that distinction explicit is, in our view, socially beneficial, because it pushes benchmark practice away from overclaiming what a finite suite of successful evaluations actually certifies.

References

Abedsoltan, A., Zhang, H., Wen, K., Lin, H., Zhang, J., and Belkin, M. Task generalization with autoregressive compositional structure: Can learning from D tasks generalize to D^T tasks? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 154–173. PMLR, 2025. URL https://proceedings.mlr.press/v267/abedsoltan25a.html.

Birgé, L. Robust tests for model selection. In Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V., and Maathuis, M. H. (eds.), *From Probability to Statistics and Back: High-Dimensional Models and Processes*, volume 9 of *IMS Collections*, pp. 47–64. Institute of Mathematical Statistics, 2013. doi: 10.1214/12-IMSCOLL905.

Chang, H., Park, J., Cho, H., Yang, S., Ko, M., Hwang, H., Won, S., Lee, D., Ahn, Y., and Seo, M. The coverage principle: A framework for understanding compositional generalization. *arXiv preprint arXiv:2505.20278*, 2025. doi: 10.48550/arXiv.2505.20278.

Ghosal, S. and van der Vaart, A. W. Convergence rates of posterior distributions for non-i.i.d. observations. *The Annals of Statistics*, 35(1):192–223, 2007. doi: 10.1214/009053606000001172.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000. doi: 10.1214/aos/1016218228.

Ho, N. and Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755, 2016. doi: 10.1214/16-AOS1444.

Ho, N. and Nguyen, X. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019. doi: 10.1137/18M122947X.

Procaccia, A., Schiffer, B., Wang, S., and Zhang, S. Metritocracy: Representative metrics for lite benchmarks. *arXiv preprint arXiv:2506.09813*, 2025. doi: 10.48550/arXiv.2506.09813.

Schug, S., Kobayashi, S., Akram, Y., Wołczyk, M., Proca, A., von Oswald, J., Pascanu, R., Sacramento, J., and Steger, A. Discovering modular solutions that generalize compositionally. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=H98CVcX1eh.

Teicher, H. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963. doi: 10.1214/aoms/1177703862.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511800474.

Yakowitz, S. J. and Spragins, J. D. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968. doi: 10.1214/aoms/1177698520.

Zhao, Q., Xu, M., Gupta, K., Asthana, A., Zheng, L., and Gould, S. Can we predict performance of large models across vision-language tasks? *arXiv preprint arXiv:2410.10112*, 2024. doi: 10.48550/arXiv.2410.10112.

A. Additional Notation, Assumptions, and Quotient Geometry

This appendix freezes the exact parameter space, group action, and hardness functionals used throughout the paper. No later section changes these definitions.

A.1. Parameter space and basic notation

Fix an admissible language $\mathcal{L} \subseteq \mathcal{Z}_{L,r}$, where

$$\mathcal{Z}_{L,r} = \bigcup_{t=1}^{L} [r]^t, \qquad [r] = \{1, \dots, r\}, \qquad [r]^0 = \{\emptyset\}.$$

For $z = (z_1, \dots, z_t) \in \mathcal{Z}_{L,r}$, let $|z| = t$. If $u = (u_1, \dots, u_s) \in [r]^s$ and $v = (v_1, \dots, v_q) \in [r]^q$, we write uav for the concatenated string

$$(u_1, \dots, u_s, a, v_1, \dots, v_q) \in [r]^{s+1+q}.$$

The grammar parameter space is

$$\mathfrak{P}_{\mathcal{L}}^{\circ} = \Big\{ \pi = (\pi_z)_{z \in \mathcal{L}} \in (0,1)^{\mathcal{L}} : \sum_{z \in \mathcal{L}} \pi_z = 1 \Big\},$$

and the bounded module space is

$$\mathbb{A}_B = \{ A \in \mathbb{R}^{d \times d} : \|A\|_{\mathrm{op}} \le B \}.$$

The full parameter space is

$$\Xi(B, \mathcal{L}) = \mathfrak{P}_{\mathcal{L}}^{\circ} \times \mathbb{A}_B^r.$$

If $\Theta = (A_1, \dots, A_r) \in \mathbb{A}_B^r$, then for any nonempty string $z = (z_1, \dots, z_t)$ we define

$$A_z = A_{z_t} \cdots A_{z_1}, \qquad A_{\emptyset} = I_d.$$

For any $W \in \mathbb{R}^{d \times d}$, $P_W^{(n)}$ denotes the law of one task dataset $D = (X, Y)$ under

$$x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \qquad y_j = W x_j + \varepsilon_j, \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d),$$

with the Gaussian design and noise independent.

For dominated probability laws P and Q with densities p and q, respectively, we write

$$h^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\mu, \qquad \mathrm{KL}(P \| Q) = \int p \log \frac{p}{q} \, d\mu,$$

where μ is any common dominating measure.

A.2. Standing model and prior conditions

Assumption A.1 (Sampling model and truth)**.** Throughout, the admissible language $\mathcal{L} \subseteq \mathcal{Z}_{L,r}$, the dimensions (d, r, L), the operator bound B, and the noise level $\sigma > 0$ are fixed. The true parameter is $\eta^\star = (\pi^\star, \Theta^\star) \in \Xi(B, \mathcal{L})$, with $\pi_z^\star \ge \pi_{\min} > 0$ for every $z \in \mathcal{L}$. For each $z \in \mathcal{L}$, let $A_z^\star$ denote the composed operator induced by $\Theta^\star$. For each task $i \in [m]$, the latent program satisfies $Z_i \sim \pi^\star$, the design vectors satisfy $x_{i1}, \dots, x_{in} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ with $n \ge d + 1$, and, conditional on $Z_i = z$, the responses obey

$$y_{ij} = A_z^\star x_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d),$$

with the Gaussian designs, latent programs, and noises mutually independent across tasks.

Assumption A.2 (Prior regularity)**.** The prior factors as $\Pi = \Pi_\pi \otimes \Pi_\Theta$ on $\Xi(B, \mathcal{L})$. The grammar-law marginal Π_π admits a density with respect to the intrinsic Lebesgue measure on $\mathfrak{P}_{\mathcal{L}}^{\circ}$ that is continuous and strictly positive in a neighborhood of $\pi^\star$. The library marginal $\Pi_\Theta = \bigotimes_{a=1}^r \Pi_a$ admits a density on $\mathbb{A}_B^r$ that is continuous and strictly positive in a neighborhood of $\Theta^\star$.

A.3. Automorphism group and quotient metric

The correct symmetry group is the automorphism group of the admissible language.

Definition A.3 (Automorphism group)**.** The automorphism group of $\mathcal{L}$ is

$$\mathrm{Aut}(\mathcal{L}) = \{\sigma \in S_r : \sigma \cdot \mathcal{L} = \mathcal{L}\},$$

where $\sigma \cdot z = (\sigma(z_1), \ldots, \sigma(z_t))$ for every string $z = (z_1, \ldots, z_t)$.

For $\sigma \in \mathrm{Aut}(\mathcal{L})$ and $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$, define

$$(\sigma \cdot \pi)_u := \pi_{\sigma^{-1} \cdot u}, \qquad (\sigma \cdot \Theta)_a := A_{\sigma^{-1}(a)}, \qquad \sigma \cdot \eta := (\sigma \cdot \pi, \sigma \cdot \Theta).$$

The base distance on $\Xi(B, \mathcal{L})$ is

$$d_{\mathrm{par}}\big((\pi, \Theta), (\pi', \Theta')\big) = \|\pi - \pi'\|_1 + \sum_{a=1}^{r} \|A_a - A'_a\|_F.$$

The quotient distance is

$$d_q(\eta, \eta') = \min_{\sigma \in \mathrm{Aut}(\mathcal{L})} d_{\mathrm{par}}(\eta, \sigma \cdot \eta').$$

The next lemma shows that the automorphism action preserves the grouped law exactly.

Lemma A.4 (Permutation invariance of the grouped law)**.** *For every $\eta \in \Xi(B, \mathcal{L})$, every $\sigma \in \mathrm{Aut}(\mathcal{L})$, and every $n \geq 1$,*

$$Q^{(n)}_{\sigma \cdot \eta} = Q^{(n)}_{\eta}.$$

Proof. Write $\eta = (\pi, \Theta)$. For any $z = (z_1, \ldots, z_t) \in \mathcal{L}$,

$$A_{\sigma \cdot z}(\sigma \cdot \Theta) = (\sigma \cdot \Theta)_{\sigma(z_t)} \cdots (\sigma \cdot \Theta)_{\sigma(z_1)} = A_{z_t} \cdots A_{z_1} = A_z(\Theta).$$

Since $\sigma \in \mathrm{Aut}(\mathcal{L})$, the map $z \mapsto \sigma \cdot z$ is a bijection of $\mathcal{L}$. Therefore,

$$Q^{(n)}_{\sigma \cdot \eta} = \sum_{u \in \mathcal{L}} (\sigma \cdot \pi)_u \, P^{(n)}_{A_u(\sigma \cdot \Theta)} = \sum_{z \in \mathcal{L}} \pi_z \, P^{(n)}_{A_{\sigma \cdot z}(\sigma \cdot \Theta)} = \sum_{z \in \mathcal{L}} \pi_z \, P^{(n)}_{A_z(\Theta)} = Q^{(n)}_{\eta}.$$

□

Lemma A.5 (The quotient distance is well defined)**.** *The function d_q induces a metric on the quotient space $\Xi(B, \mathcal{L})/\mathrm{Aut}(\mathcal{L})$.*

Proof. We first note that the base distance is invariant under the simultaneous group action:

$$d_{\mathrm{par}}(\sigma \cdot \eta, \sigma \cdot \eta') = d_{\mathrm{par}}(\eta, \eta') \qquad \text{for all } \sigma \in \mathrm{Aut}(\mathcal{L}).$$

Indeed, permutation of the coordinates of π preserves the ℓ_1-norm, and permutation of the module labels preserves $\sum_{a=1}^{r} \|A_a - A'_a\|_F$.

Nonnegativity of d_q is immediate. If $d_q(\eta, \eta') = 0$, then because $\mathrm{Aut}(\mathcal{L})$ is finite, the minimum is attained by some $\sigma \in \mathrm{Aut}(\mathcal{L})$, and

$$0 = d_{\mathrm{par}}(\eta, \sigma \cdot \eta')$$

implies $\eta = \sigma \cdot \eta'$. Hence η and η' represent the same quotient class.

For symmetry, using invariance of d_{par} and the fact that $\sigma^{-1} \in \mathrm{Aut}(\mathcal{L})$ whenever $\sigma \in \mathrm{Aut}(\mathcal{L})$,

$$d_q(\eta, \eta') = \min_{\sigma} d_{\mathrm{par}}(\eta, \sigma \cdot \eta') = \min_{\sigma} d_{\mathrm{par}}(\sigma^{-1} \cdot \eta, \eta') = d_q(\eta', \eta).$$

For the triangle inequality, let $\eta, \eta', \eta'' \in \Xi(B, \mathcal{L})$, and fix $\sigma, \tau \in \mathrm{Aut}(\mathcal{L})$. Then

$$d_{\mathrm{par}}(\eta, \sigma\tau \cdot \eta'') \leq d_{\mathrm{par}}(\eta, \sigma \cdot \eta') + d_{\mathrm{par}}(\sigma \cdot \eta', \sigma\tau \cdot \eta'') = d_{\mathrm{par}}(\eta, \sigma \cdot \eta') + d_{\mathrm{par}}(\eta', \tau \cdot \eta'').$$

Taking the minimum first over σ, then over τ, yields

$$d_q(\eta, \eta'') \leq d_q(\eta, \eta') + d_q(\eta', \eta'').$$

Hence d_q is a metric on the quotient space. □

A.4. Context distribution and contextual observability

The contextual observability functional is defined using the distribution of one-hole contexts induced by the latent task-family law.

Definition A.6 (One-hole contexts)**.** The set of one-hole contexts associated with $\mathcal{L}$ is

$$\mathcal{C}_{\mathcal{L}} = \Big\{(u, v) : u \in \bigcup_{s=0}^{L-1} [r]^s,\ v \in \bigcup_{q=0}^{L-1} [r]^q,\ \exists a \in [r] \text{ such that } uav \in \mathcal{L}\Big\}.$$

For $z = (z_1, \dots, z_t) \in \mathcal{L}$ and $\ell \in \{1, \dots, t\}$, define

$$\mathrm{ctx}_\ell(z) = \big((z_1, \dots, z_{\ell-1}), (z_{\ell+1}, \dots, z_t)\big) \in \mathcal{C}_{\mathcal{L}}.$$

Lemma A.7 (The context law is a probability mass function)**.** *For every $\pi \in \mathfrak{P}^\circ_{\mathcal{L}}$, the function*

$$\nu_\pi(u, v) = \sum_{z \in \mathcal{L}} \pi_z \frac{1}{|z|} \sum_{\ell=1}^{|z|} \mathbf{1}\{\mathrm{ctx}_\ell(z) = (u, v)\}, \qquad (u, v) \in \mathcal{C}_{\mathcal{L}},$$

defines a probability mass function on $\mathcal{C}_{\mathcal{L}}$.

Proof. Nonnegativity is immediate. Summing over $(u, v) \in \mathcal{C}_{\mathcal{L}}$ and exchanging the order of summation gives

$$\sum_{(u,v) \in \mathcal{C}_{\mathcal{L}}} \nu_\pi(u, v) = \sum_{z \in \mathcal{L}} \pi_z \frac{1}{|z|} \sum_{\ell=1}^{|z|} \sum_{(u,v) \in \mathcal{C}_{\mathcal{L}}} \mathbf{1}\{\mathrm{ctx}_\ell(z) = (u, v)\}.$$

For each fixed z and ℓ, exactly one context is selected, so the innermost sum equals 1. Therefore,

$$\sum_{(u,v) \in \mathcal{C}_{\mathcal{L}}} \nu_\pi(u, v) = \sum_{z \in \mathcal{L}} \pi_z \frac{1}{|z|} \sum_{\ell=1}^{|z|} 1 = \sum_{z \in \mathcal{L}} \pi_z = 1.$$

□

For $\eta = (\pi, \Theta)$, $\Theta = (A_1, \dots, A_r)$, and $a \neq b$, define

$$\Delta_{\mathrm{ctx}}(a, b; \eta) = \sum_{(u,v) \in \mathcal{C}_{\mathcal{L}}} \nu_\pi(u, v)\, \|A_v(A_a - A_b)A_u\|_F^2, \qquad \Delta_{\mathrm{ctx}}(\eta) = \min_{a \neq b} \Delta_{\mathrm{ctx}}(a, b; \eta).$$

Lemma A.8 (Continuity of contextual observability)**.** *For each pair $a \neq b$, the map*

$$\eta \mapsto \Delta_{\mathrm{ctx}}(a, b; \eta)$$

is continuous on $\Xi(B, \mathcal{L})$. Consequently, $\eta \mapsto \Delta_{\mathrm{ctx}}(\eta)$ is continuous on $\Xi(B, \mathcal{L})$.

Proof. Because $\mathcal{L}$ is finite, $\mathcal{C}_{\mathcal{L}}$ is finite. For fixed (u, v), the map

$$(\pi, \Theta) \mapsto \nu_\pi(u, v)\, \|A_v(A_a - A_b)A_u\|_F^2$$

is continuous: $\nu_\pi(u, v)$ is linear in π, and $A_v(A_a - A_b)A_u$ is a polynomial expression in the entries of the module matrices. Hence $\Delta_{\mathrm{ctx}}(a, b; \eta)$, being a finite sum of continuous functions, is continuous. Since $\Delta_{\mathrm{ctx}}(\eta)$ is the minimum over finitely many continuous functions indexed by (a, b) with $a \neq b$, it is also continuous. □

Lemma A.9 (Quotient invariance of contextual observability)**.** *For every $\eta \in \Xi(B, \mathcal{L})$ and every $\sigma \in \mathrm{Aut}(\mathcal{L})$,*

$$\Delta_{\mathrm{ctx}}(\sigma \cdot \eta) = \Delta_{\mathrm{ctx}}(\eta).$$

Proof. Fix distinct $a, b \in [r]$. For every u, v, the map $(u, v) \mapsto (\sigma \cdot u, \sigma \cdot v)$ is a bijection of $\mathcal{C}_{\mathcal{L}}$, because $\sigma \in \mathrm{Aut}(\mathcal{L})$. Moreover,

$$\nu_{\sigma\cdot\pi}(\sigma \cdot u, \sigma \cdot v) = \nu_{\pi}(u, v),$$

since relabeling the symbols of a latent program and then choosing a uniformly random position leaves the induced context weights unchanged after reindexing. Also,

$$A_{\sigma\cdot v}(\sigma \cdot \Theta)\left((\sigma \cdot \Theta)_{\sigma(a)} - (\sigma \cdot \Theta)_{\sigma(b)}\right) A_{\sigma\cdot u}(\sigma \cdot \Theta) = A_v(\Theta)(A_a(\Theta) - A_b(\Theta))A_u(\Theta).$$

Therefore,

$$\Delta_{\mathrm{ctx}}(\sigma(a), \sigma(b); \sigma \cdot \eta) = \Delta_{\mathrm{ctx}}(a, b; \eta).$$

Taking the minimum over all distinct pairs yields

$$\Delta_{\mathrm{ctx}}(\sigma \cdot \eta) = \Delta_{\mathrm{ctx}}(\eta).$$

□

Remark A.10. The quantities $Q_{\eta}^{(n)}$, d_q, and $\Delta_{\mathrm{ctx}}(\eta)$ are therefore all well defined on quotient classes. Later, when the local inverse inequality is established, the singularity order $\kappa(\eta^{\star})$ will also be a quotient-invariant object.

B. Sufficient-Statistic Reduction

This appendix proves that, conditional on the random design, no inferential information about the latent task-family parameter $\eta = (\pi, \Theta)$ is lost by replacing each raw task dataset $D = (X, Y)$ with the task-level least-squares estimator $\widehat{W}$. The reduction is exact: under any prior on η, the posterior based on the full grouped data $(X_i, Y_i)_{i=1}^m$ coincides almost surely with the posterior based on the reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$. Moreover, conditional on the designs, the reduced experiment is a heteroskedastic finite mixture of matrix-normal laws centered at the composed operators $\{A_z\}_{z\in\mathcal{L}}$.

B.1. Task-level least-squares reduction

We first record the matrix-normal convention used throughout the appendix.

Definition B.1 (Matrix-normal law)**.** For integers $p, q \geq 1$, a random matrix $M \in \mathbb{R}^{p\times q}$ is said to follow the matrix-normal law

$$M \sim \mathcal{MN}_{p,q}(M_0, U, V)$$

if

$$\mathrm{vec}(M) \sim \mathcal{N}\big(\mathrm{vec}(M_0),\, V \otimes U\big),$$

where $M_0 \in \mathbb{R}^{p\times q}$, $U \in \mathbb{R}^{p\times p}$ is positive semidefinite, and $V \in \mathbb{R}^{q\times q}$ is positive semidefinite.

Fix one task $D = (X, Y)$, where

$$X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}, \qquad Y = [y_1, \ldots, y_n] \in \mathbb{R}^{d\times n}.$$

Recall from Theorem A.1 that $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and $n \geq d + 1$.

Define the empirical Gram matrix

$$G_X := XX^{\top} \in \mathbb{R}^{d\times d}.$$

Lemma B.2 (Full row rank of the Gaussian design)**.** *Under Theorem A.1, G_X is positive definite almost surely.*

Proof. It suffices to show that X has row rank d almost surely. Consider the first d columns $(x_1, \ldots, x_d)$. Since these are independent draws from an absolutely continuous distribution on $\mathbb{R}^d$, the $d \times d$ matrix $[x_1, \ldots, x_d]$ has nonzero determinant almost surely. Hence X has row rank d almost surely, and therefore $G_X = XX^{\top}$ is invertible almost surely. □

In view of Theorem B.2, all identities below hold on an event of probability one. We suppress this null set from the notation.

Define the task-level least-squares estimator

$$\widehat{W} := YX^\top G_X^{-1} \in \mathbb{R}^{d\times d}, \tag{20}$$

the design projection

$$P_X := X^\top G_X^{-1} X \in \mathbb{R}^{n\times n}, \tag{21}$$

and the residual matrix

$$R := Y(I_n - P_X) \in \mathbb{R}^{d\times n}. \tag{22}$$

Proposition B.3 (Orthogonal decomposition of the task likelihood)**.** *For every X with G_X invertible and every $Y \in \mathbb{R}^{d\times n}$, the following identities hold:*

1. $Y = \widehat{W}X + R$;
2. $RX^\top = 0$;
3. *for every $W \in \mathbb{R}^{d\times d}$,*

$$\|Y - WX\|_F^2 = \|(\widehat{W} - W)X\|_F^2 + \|R\|_F^2 = \mathrm{tr}\big((\widehat{W} - W)G_X(\widehat{W} - W)^\top\big) + \|R\|_F^2. \tag{23}$$

Proof. By definition,

$$\widehat{W}X = YX^\top G_X^{-1} X = YP_X.$$

Hence

$$\widehat{W}X + R = YP_X + Y(I_n - P_X) = Y.$$

This proves part (1).

For part (2), using $P_X X^\top = X^\top G_X^{-1} XX^\top = X^\top$,

$$RX^\top = Y(I_n - P_X)X^\top = Y(X^\top - P_X X^\top) = 0.$$

For part (3), fix $W \in \mathbb{R}^{d\times d}$. By part (1),

$$Y - WX = (\widehat{W} - W)X + R.$$

Using part (2),

$$\langle(\widehat{W} - W)X,\, R\rangle_F = \mathrm{tr}\big((\widehat{W} - W)XR^\top\big) = \mathrm{tr}\big(RX^\top(\widehat{W} - W)^\top\big) = 0.$$

Therefore

$$\|Y - WX\|_F^2 = \|(\widehat{W} - W)X\|_F^2 + \|R\|_F^2.$$

Finally,

$$\|(\widehat{W} - W)X\|_F^2 = \mathrm{tr}\big((\widehat{W} - W)XX^\top(\widehat{W} - W)^\top\big) = \mathrm{tr}\big((\widehat{W} - W)G_X(\widehat{W} - W)^\top\big).$$

□

Corollary B.4 (Exact factorization of the conditional likelihood)**.** *For $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$, the conditional density of Y given X under the grouped law $Q_\eta^{(n)}$ factorizes as*

$$p_\eta(Y \mid X) = g_X(Y)\, q_\eta(\widehat{W} \mid X), \tag{24}$$

where

$$g_X(Y) := (2\pi\sigma^2)^{-d(n-d)/2}\, |G_X|^{-d/2}\, \exp\left(-\frac{\|R\|_F^2}{2\sigma^2}\right) \tag{25}$$

is free of η, and

$$q_\eta(W \mid X) := (2\pi)^{-d^2/2}(\sigma^2)^{-d^2/2}|G_X|^{d/2} \sum_{z\in\mathcal{L}} \pi_z \exp\Big(-\frac{1}{2\sigma^2}\operatorname{tr}\big((W-A_z)G_X(W-A_z)^\top\big)\Big) \tag{26}$$

is the conditional density of $\widehat{W}$ given X under η.

In particular, conditional on X, the statistic $\widehat{W}$ is sufficient for η.

Proof. Under $\eta = (\pi, \Theta)$,

$$p_\eta(Y \mid X) = \sum_{z\in\mathcal{L}} \pi_z\, p_{A_z}(Y \mid X),$$

where

$$p_{A_z}(Y \mid X) = (2\pi\sigma^2)^{-dn/2} \exp\Big(-\frac{\|Y - A_zX\|_F^2}{2\sigma^2}\Big).$$

Applying Theorem B.3 with $W = A_z$ yields

$$\|Y - A_zX\|_F^2 = \operatorname{tr}\big((\widehat{W} - A_z)G_X(\widehat{W} - A_z)^\top\big) + \|R\|_F^2.$$

Substituting this into the display above and factoring out the term depending only on (X, Y) but not on η gives Equation (24) with g_X and q_η as in Equations (25) and (26). The sufficiency claim follows from the Neyman–Fisher factorization theorem applied conditionally on X. □

B.2. Exact conditional law of the reduced statistic

We now compute the conditional law of $\widehat{W}$ and the residual R, first under a fixed latent program and then under the full task-family law.

Proposition B.5 (Exact matrix-normal reduction)**.** *Fix $z \in \mathcal{L}$. Under the model*

$$Y = A_zX + E, \qquad E = [\varepsilon_1, \dots, \varepsilon_n], \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d),$$

the reduced statistics satisfy

$$\widehat{W} = A_z + EX^\top G_X^{-1}, \qquad R = E(I_n - P_X). \tag{27}$$

Conditional on X and $Z = z$,

$$\widehat{W} \sim \mathcal{MN}_{d,d}\big(A_z, \sigma^2 I_d, G_X^{-1}\big), \tag{28}$$

$$R \sim \mathcal{MN}_{d,n}\big(0, \sigma^2 I_d, I_n - P_X\big), \tag{29}$$

and $\widehat{W}$ is independent of R.

Proof. Starting from $Y = A_zX + E$, the definition of $\widehat{W}$ gives

$$\widehat{W} = (A_zX + E)X^\top G_X^{-1} = A_zXX^\top G_X^{-1} + EX^\top G_X^{-1} = A_z + EX^\top G_X^{-1},$$

which proves the first identity in Equation (27). Similarly,

$$R = (A_zX + E)(I_n - P_X) = A_zX(I_n - P_X) + E(I_n - P_X) = E(I_n - P_X),$$

because $X(I_n - P_X) = 0$. This proves the second identity in Equation (27).

Since $\operatorname{vec}(E) \sim \mathcal{N}(0, \sigma^2 I_{dn})$, we have

$$\operatorname{vec}(\widehat{W} - A_z) = \operatorname{vec}(EX^\top G_X^{-1}) = \big((G_X^{-1}X) \otimes I_d\big)\operatorname{vec}(E).$$

Therefore,

$$\operatorname{vec}(\widehat{W} - A_z) \mid X, Z = z \sim \mathcal{N}\big(0,\ \sigma^2\big(G_X^{-1}XX^\top G_X^{-1}\big) \otimes I_d\big) = \mathcal{N}\big(0,\ \sigma^2 G_X^{-1} \otimes I_d\big),$$

which is exactly Equation (28).

Likewise,

$$\operatorname{vec}(R) = \operatorname{vec}(E(I_n - P_X)) = \big((I_n - P_X) \otimes I_d\big) \operatorname{vec}(E),$$

so

$$\operatorname{vec}(R) \mid X, Z = z \sim \mathcal{N}\big(0,\ \sigma^2 (I_n - P_X) \otimes I_d\big)\,,$$

which gives Equation (29).

Finally, still conditional on X and $Z = z$, the pair $(\widehat{W}, R)$ is jointly Gaussian because both are linear transformations of E. Its cross-covariance is

$$\begin{aligned}\operatorname{Cov}\big(\operatorname{vec}(\widehat{W} - A_z), \operatorname{vec}(R) \mid X, Z = z\big) &= \sigma^2 \big((G_X^{-1} X)(I_n - P_X)\big) \otimes I_d \\ &= 0,\end{aligned}$$

because $X(I_n - P_X) = 0$. Hence $\widehat{W}$ and R are conditionally independent. □

Corollary B.6 (Mixture representation under the task-family law)**.** *Fix $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$. Conditional on X, the reduced statistic $\widehat{W}$ has the finite-mixture representation*

$$\mathcal{L}(\widehat{W} \mid X, \eta) = \sum_{z \in \mathcal{L}} \pi_z \ \mathcal{MN}_{d,d}\big(A_z, \sigma^2 I_d, G_X^{-1}\big). \tag{30}$$

Moreover, conditional on X and η, the residual R is independent of $\widehat{W}$, independent of the latent program Z, and has parameter-free law

$$R \mid X, \eta \sim \mathcal{MN}_{d,n}\big(0, \sigma^2 I_d, I_n - P_X\big).$$

Proof. The mixture representation Equation (30) follows immediately from Theorem B.5 by marginalizing over $Z \sim \pi$.

For the second claim, Theorem B.5 gives, for every fixed $z \in \mathcal{L}$,

$$\mathcal{L}(R \mid X, Z = z, \eta) = \mathcal{MN}_{d,n}\big(0, \sigma^2 I_d, I_n - P_X\big),$$

which does not depend on z or on η. Hence R is conditionally independent of Z given X, η. Since R is also conditionally independent of $\widehat{W}$ given X, Z, η, and its conditional law does not depend on Z, integrating out Z yields conditional independence of R and $\widehat{W}$ given X, η. □

Proposition B.7 (Exact posterior reduction)**.** *Let Π be any prior on $\Xi(B, \mathcal{L})$. For each task $i \in [m]$, define*

$$\widehat{W}_i := Y_i X_i^\top (X_i X_i^\top)^{-1}.$$

Then for every measurable $B \subseteq \Xi(B, \mathcal{L})$,

$$\Pi\big(B \mid X_{1:m}, Y_{1:m}\big) = \Pi\big(B \mid X_{1:m}, \widehat{W}_{1:m}\big) \qquad \text{a.s.} \tag{31}$$

under the joint law induced by Π and the sampling model.

Proof. By independence across tasks and Theorem B.4, the conditional likelihood satisfies

$$p_\eta(Y_{1:m} \mid X_{1:m}) = \prod_{i=1}^m g_{X_i}(Y_i) \prod_{i=1}^m q_\eta(\widehat{W}_i \mid X_i).$$

The factor $\prod_{i=1}^m g_{X_i}(Y_i)$ is free of η. Therefore Bayes' formula gives

$$\Pi(B \mid X_{1:m}, Y_{1:m}) = \frac{\int_B \prod_{i=1}^m q_\eta(\widehat{W}_i \mid X_i)\, \Pi(d\eta)}{\int_{\Xi(B,\mathcal{L})} \prod_{i=1}^m q_\eta(\widehat{W}_i \mid X_i)\, \Pi(d\eta)},$$

which is precisely the posterior based on $(X_{1:m}, \widehat{W}_{1:m})$. This proves Equation (31). □

Corollary B.8 (Exact reduced experiment)**.** *Under $\eta = (\pi, \Theta)$, conditional on the designs $X_{1:m}$, the reduced observations admit the representation*

$$Z_i \overset{\text{i.i.d.}}{\sim} \pi, \qquad \widehat{W}_i = A_{Z_i} + \Xi_i, \qquad \Xi_i \mid X_{1:m} \overset{\text{ind}}{\sim} \mathcal{MN}_{d,d}\big(0, \sigma^2 I_d, G_{X_i}^{-1}\big), \tag{32}$$

with Ξ_i conditionally independent of Z_i and independent across $i \in [m]$.

Proof. This is an immediate consequence of Theorem B.5 and independence across tasks. □

B.3. Information geometry of the reduced experiment

The previous subsection shows that the grouped meta-learning problem reduces exactly to a finite mixture of heteroskedastic Gaussian kernels centered at the composed operators A_z. The next lemma records the resulting information geometry, which will be used later to construct tests and to relate structural perturbations in operator space to Hellinger and Kullback–Leibler separation.

For X with G_X invertible and $W \in \mathbb{R}^{d\times d}$, define

$$\mathcal{N}_X^W := \mathcal{MN}_{d,d}\big(W, \sigma^2 I_d, G_X^{-1}\big).$$

Lemma B.9 (Exact conditional divergences)**.** *For any full-row-rank design X and any $W, W' \in \mathbb{R}^{d\times d}$,*

$$\begin{aligned} \mathrm{KL}\big(\mathcal{N}_X^W \,\|\, \mathcal{N}_X^{W'}\big) &= \frac{1}{2\sigma^2}\,\mathrm{tr}\big((W-W')G_X(W-W')^\top\big) \\ &= \frac{1}{2\sigma^2}\,\|(W-W')X\|_F^2, \end{aligned} \tag{33}$$

and

$$\begin{aligned} h^2\big(\mathcal{N}_X^W, \mathcal{N}_X^{W'}\big) &= 1 - \exp\left(-\frac{1}{8\sigma^2}\,\mathrm{tr}\big((W-W')G_X(W-W')^\top\big)\right) \\ &= 1 - \exp\left(-\frac{1}{8\sigma^2}\,\|(W-W')X\|_F^2\right). \end{aligned} \tag{34}$$

Proof. Let $\Delta := W - W'$. By Theorem B.1,

$$\mathrm{vec}(\mathcal{N}_X^W) \sim \mathcal{N}\big(\mathrm{vec}(W),\, \Sigma_X\big), \qquad \Sigma_X := G_X^{-1} \otimes \sigma^2 I_d.$$

The inverse covariance is

$$\Sigma_X^{-1} = G_X \otimes \sigma^{-2} I_d.$$

For Gaussian laws with common covariance, the exact formulas are

$$\mathrm{KL}\big(\mathcal{N}(m,\Sigma)\,\|\,\mathcal{N}(m',\Sigma)\big) = \frac{1}{2}(m-m')^\top \Sigma^{-1}(m-m'),$$

and

$$h^2\big(\mathcal{N}(m,\Sigma), \mathcal{N}(m',\Sigma)\big) = 1 - \exp\left(-\frac{1}{8}(m-m')^\top \Sigma^{-1}(m-m')\right).$$

Applying these formulas with $m = \mathrm{vec}(W)$, $m' = \mathrm{vec}(W')$, and $\Sigma = \Sigma_X$, we obtain

$$(\mathrm{vec}\,\Delta)^\top \Sigma_X^{-1}\,\mathrm{vec}\,\Delta = \frac{1}{\sigma^2}(\mathrm{vec}\,\Delta)^\top (G_X \otimes I_d)\,\mathrm{vec}\,\Delta = \frac{1}{\sigma^2}\,\mathrm{tr}(\Delta G_X \Delta^\top).$$

Since $G_X = XX^\top$,

$$\mathrm{tr}(\Delta G_X \Delta^\top) = \mathrm{tr}(\Delta X X^\top \Delta^\top) = \|\Delta X\|_F^2.$$

Substituting this identity into the Gaussian formulas yields Equations (33) and (34). □

B.4. Design concentration and effective noise scale

For later rate arguments, it is convenient to isolate a high-probability design event on which the reduced experiment behaves like a Gaussian location model with effective noise level $n^{-1/2}$.

For $0 < \delta < 1$, define the event

$$\mathcal{E}_n(\delta) := \{(1-\delta)nI_d \preceq G_X \preceq (1+\delta)nI_d\}. \tag{35}$$

Lemma B.10 (Gaussian design concentration)**.** *Fix $0 < \delta < 1$. There exist constants $c_{d,\delta}, C_{d,\delta} \in (0,\infty)$, depending only on d and δ, such that*

$$\Pr\big(\mathcal{E}_n(\delta)^c\big) \le C_{d,\delta}\, e^{-c_{d,\delta} n}. \tag{36}$$

The bound is uniform over $\eta \in \Xi(B,\mathcal{L})$, since the design law is independent of η.

Proof. Because the columns of X are i.i.d. $\mathcal{N}(0, I_d)$, the matrix $n^{-1}G_X$ is the sample covariance matrix of n Gaussian observations in dimension d. The exponential tail bound Equation (36) is standard; see, for example, the discussion of Gaussian covariance concentration in Vershynin (2018, Chapter 4). Since the law of X does not depend on η, the bound is automatically uniform over the parameter space. □

Corollary B.11 (Information-geometric sandwich on the good design event)**.** *Fix $0 < \delta < 1$. On the event $\mathcal{E}_n(\delta)$, for every $W, W' \in \mathbb{R}^{d\times d}$,*

$$\frac{(1-\delta)n}{2\sigma^2}\,\|W - W'\|_F^2 \le \mathrm{KL}\big(\mathcal{N}_X^W \,\|\, \mathcal{N}_X^{W'}\big) \le \frac{(1+\delta)n}{2\sigma^2}\,\|W - W'\|_F^2, \tag{37}$$

and

$$1 - \exp\!\left(-\frac{(1-\delta)n}{8\sigma^2}\|W - W'\|_F^2\right) \le h^2\big(\mathcal{N}_X^W, \mathcal{N}_X^{W'}\big) \le 1 - \exp\!\left(-\frac{(1+\delta)n}{8\sigma^2}\|W - W'\|_F^2\right). \tag{38}$$

In particular, whenever

$$\frac{(1+\delta)n}{8\sigma^2}\|W - W'\|_F^2 \le 1,$$

there exist constants $0 < c_{\delta,\sigma} \le C_{\delta,\sigma} < \infty$ such that

$$c_{\delta,\sigma}\, n\|W - W'\|_F^2 \le h^2\big(\mathcal{N}_X^W, \mathcal{N}_X^{W'}\big) \le C_{\delta,\sigma}\, n\|W - W'\|_F^2. \tag{39}$$

Proof. On $\mathcal{E}_n(\delta)$, the Gram matrix satisfies

$$(1-\delta)nI_d \preceq G_X \preceq (1+\delta)nI_d.$$

Hence for $\Delta := W - W'$,

$$(1-\delta)n\|\Delta\|_F^2 \le \mathrm{tr}(\Delta G_X \Delta^\top) \le (1+\delta)n\|\Delta\|_F^2.$$

Substituting into Equations (33) and (34) yields Equations (37) and (38).

For the local Hellinger bound, set

$$u := \frac{1}{8\sigma^2}\,\mathrm{tr}(\Delta G_X \Delta^\top).$$

If $u \le 1$, then the elementary inequalities

$$(1 - e^{-1})u \le 1 - e^{-u} \le u$$

imply

$$(1 - e^{-1})u \le h^2\big(\mathcal{N}_X^W, \mathcal{N}_X^{W'}\big) \le u.$$

Using the sandwich on u furnished by $\mathcal{E}_n(\delta)$ gives Equation (39). □

Remark B.12 (Role in later proofs). Theorems B.7 and B.8 show that every posterior statement in the paper may be proved in the reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$. Theorems B.9 and B.11 then convert operator perturbations into exact Hellinger and Kullback–Leibler separation. This is the technical bridge from latent task grammars to the testing and contraction arguments developed in the next appendices.

C. Contextual Observability and Identifiability

Recent theory has made clear that compositional generalization is controlled by structural properties of the task family rather than by raw scale alone. In particular, support conditions matter for recovering reusable modules in teacher–student meta-learning (Schug et al., 2024), special autoregressive task families can permit generalization from $\widetilde{O}(D)$ tasks to a D^T family (Abedsoltan et al., 2025), and coverage / path-ambiguity considerations constrain what pattern-matching learners can infer from input–output pairs alone (Chang et al., 2025). The purpose of this appendix is to isolate the exact operator-level observability statements needed in our grouped Bayesian model.

The appendix has two logically distinct parts. First, we show that the pairwise quantity Δ_{ctx} introduced in the main text is an exact information measure for one-hole substitutions and yields a sharp impossibility theorem in the single-occurrence regime. Second, we show that Δ_{ctx} by itself is not strong enough for full local identification of the latent grammar. For that stronger conclusion we introduce a *linearized contextual observability* modulus λ_{lin}, prove local injectivity of the program-operator map, and then derive local quotient identifiability of the full parameter.

C.1. One-hole substitutions and exact information separation

We begin by formalizing the operator-level effect of substituting one primitive module for another inside a fixed context.

For strings $u \in [r]^s$, $v \in [r]^q$, and $a \in [r]$, recall that

$$A_{uav} = A_v A_a A_u, \qquad A_\emptyset = I_d.$$

For any full-row-rank design $X \in \mathbb{R}^{d\times n}$, define

$$\mathcal{N}_X^W := \mathcal{MN}_{d,d}\big(W, \sigma^2 I_d, G_X^{-1}\big), \qquad G_X := XX^\top.$$

Lemma C.1 (Exact one-hole substitution divergences)**.** *Let $u \in [r]^s$, $v \in [r]^q$, and $a, b \in [r]$ be distinct. For every full-row-rank design X and every $n \geq d+1$,*

$$\mathrm{KL}\Big(\mathcal{N}_X^{A_{uav}} \,\Big\|\, \mathcal{N}_X^{A_{ubv}}\Big) = \frac{1}{2\sigma^2}\big\|A_v(A_a - A_b)A_u X\big\|_F^2, \tag{40}$$

$$h^2\Big(\mathcal{N}_X^{A_{uav}}, \mathcal{N}_X^{A_{ubv}}\Big) = 1 - \exp\left(-\frac{1}{8\sigma^2}\big\|A_v(A_a - A_b)A_u X\big\|_F^2\right). \tag{41}$$

Proof. Apply Theorem B.9 from Appendix B with

$$W = A_{uav} = A_v A_a A_u, \qquad W' = A_{ubv} = A_v A_b A_u.$$

Then $W - W' = A_v(A_a - A_b)A_u$, and substituting into Equations (33) and (34) yields Equations (40) and (41). □

The next proposition shows that Δ_{ctx} is not merely a heuristic notion of difficulty: it is exactly the ν_π-average information for distinguishing one primitive module from another through random contexts induced by the task-family law.

Proposition C.2 (Context-averaged information equals contextual observability)**.** *Fix $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$ and distinct $a, b \in [r]$. Let X be a full-row-rank design. Then*

$$\mathbb{E}_{(u,v)\sim\nu_\pi}\,\mathrm{KL}\Big(\mathcal{N}_X^{A_{uav}} \,\Big\|\, \mathcal{N}_X^{A_{ubv}}\Big) = \frac{1}{2\sigma^2}\sum_{(u,v)\in\mathcal{C}_\mathcal{L}} \nu_\pi(u,v)\,\big\|A_v(A_a - A_b)A_u X\big\|_F^2. \tag{42}$$

Moreover, on the good design event $\mathcal{E}_n(\delta)$ of Equation (35),

$$\frac{(1-\delta)n}{2\sigma^2}\,\Delta_{\mathrm{ctx}}(a,b;\eta) \leq \mathbb{E}_{(u,v)\sim\nu_\pi}\,\mathrm{KL}\Big(\mathcal{N}_X^{A_{uav}} \,\Big\|\, \mathcal{N}_X^{A_{ubv}}\Big) \leq \frac{(1+\delta)n}{2\sigma^2}\,\Delta_{\mathrm{ctx}}(a,b;\eta). \tag{43}$$

Finally, if in addition

$$\frac{(1+\delta)n}{8\sigma^2}\max_{(u,v)\in\mathcal{C}_\mathcal{L}}\big\|A_v(A_a - A_b)A_u\big\|_F^2 \leq 1,$$

then there exist constants $0 < c_{\delta,\sigma} \leq C_{\delta,\sigma} < \infty$ such that

$$c_{\delta,\sigma}\, n\,\Delta_{\mathrm{ctx}}(a,b;\eta) \leq \mathbb{E}_{(u,v)\sim\nu_\pi} h^2\Big(\mathcal{N}_X^{A_{uav}}, \mathcal{N}_X^{A_{ubv}}\Big) \leq C_{\delta,\sigma}\, n\,\Delta_{\mathrm{ctx}}(a,b;\eta). \tag{44}$$

Proof. The exact identity Equation (42) follows by multiplying Equation (40) by $\nu_\pi(u,v)$ and summing over $(u,v) \in \mathcal{C}_\mathcal{L}$.

On $\mathcal{E}_n(\delta)$, the Gram matrix satisfies

$$(1-\delta)nI_d \preceq G_X \preceq (1+\delta)nI_d.$$

Hence for every $(u,v) \in \mathcal{C}_\mathcal{L}$,

$$(1-\delta)n\big\|A_v(A_a - A_b)A_u\big\|_F^2 \le \big\|A_v(A_a - A_b)A_uX\big\|_F^2 \le (1+\delta)n\big\|A_v(A_a - A_b)A_u\big\|_F^2.$$

Substituting this bound into Equation (42) yields Equation (43).

For the Hellinger claim, apply Equation (41) together with the local inequality

$$(1-e^{-1})u \le 1 - e^{-u} \le u \qquad \text{for } u \in [0,1].$$

Under the displayed small-separation condition, the exponent in Equation (41) is uniformly at most 1 on $\mathcal{E}_n(\delta)$. Averaging over $(u,v) \sim \nu_\pi$ then gives Equation (44). □

Remark C.3 (What Δ_{ctx} does and does not control). Theorem C.2 shows that Δ_{ctx} is the exact pairwise information for single-site substitutions through random task-family contexts. This is the operator-level analogue of the data-centric support / coverage ideas emphasized in recent compositional-generalization theory (Chang et al., 2025; Schug et al., 2024). However, Δ_{ctx} is a *pairwise* quantity. It controls first-order distinguishability of one primitive from another, but it does not by itself guarantee full injectivity of the entire map $\Theta \mapsto (A_z)_{z\in\mathcal{L}}$. That stronger conclusion requires the linearized modulus introduced in Theorem C.9 below.

C.2. Zero contextual observability and exact collapse in the single-occurrence regime

The impossibility statement for $\Delta_{\mathrm{ctx}} = 0$ is exact in a natural and important special case: languages in which no primitive symbol appears more than once in a program. This covers many synthetic task families used to study compositional generalization, including settings where a task is an ordered composition of distinct modules.

Definition C.4 (Single-occurrence language)**.** We say that the admissible language $\mathcal{L}$ is *single-occurrence* if for every $z = (z_1, \dots, z_t) \in \mathcal{L}$ and every $a \in [r]$,

$$\#\{\ell \in \{1,\dots,t\} : z_\ell = a\} \le 1.$$

Proposition C.5 (Exact collapse at zero contextual observability in single-occurrence languages)**.** *Assume that $\mathcal{L}$ is single-occurrence, and fix $\eta = (\pi, \Theta) \in \Xi(B,\mathcal{L})$. Let $a, b \in [r]$ be distinct and suppose that*

$$\Delta_{\mathrm{ctx}}(a,b;\eta) = 0.$$

Define the modified library

$$\Theta^{(a\leftarrow b)} := (A_1, \dots, A_{b-1}, A_a, A_{b+1}, \dots, A_r),$$

and let

$$\widetilde{\eta} := (\pi, \Theta^{(a\leftarrow b)}).$$

Then

$$A_z(\Theta^{(a\leftarrow b)}) = A_z(\Theta) \qquad \textit{for every } z \in \mathcal{L}, \tag{45}$$

and therefore

$$Q_{\widetilde{\eta}}^{(n)} = Q_\eta^{(n)} \qquad \textit{for every } n \ge 1. \tag{46}$$

Proof. Since $\Delta_{\mathrm{ctx}}(a,b;\eta) = 0$ and every term in its definition is nonnegative, we must have

$$A_v(A_a - A_b)A_u = 0 \qquad \text{for every } (u,v) \in \mathcal{C}_\mathcal{L} \text{ with } \nu_\pi(u,v) > 0. \tag{47}$$

Because $\pi \in \mathfrak{P}_\mathcal{L}^\circ$, every program $z \in \mathcal{L}$ has strictly positive probability $\pi_z > 0$. Hence every actual occurrence of a symbol in every $z \in \mathcal{L}$ contributes strictly positive mass to the context law ν_π.

Fix $z \in \mathcal{L}$. If z contains no occurrence of b, then $A_z(\Theta^{(a\leftarrow b)}) = A_z(\Theta)$ trivially. If z contains one occurrence of b, then by the single-occurrence assumption it can be written uniquely as

$$z = ubv$$

for some u, v with $(u, v) = \mathrm{ctx}_\ell(z)$ at the unique position ℓ carrying b. Since $\pi_z > 0$, the context (u, v) has $\nu_\pi(u, v) > 0$, so Equation (47) yields

$$A_v(A_a - A_b)A_u = 0.$$

Therefore

$$A_z(\Theta^{(a\leftarrow b)}) - A_z(\Theta) = A_v A_a A_u - A_v A_b A_u = A_v(A_a - A_b)A_u = 0.$$

This proves Equation (45). Summing the identical component laws with the same weights π_z gives Equation (46). □

Remark C.6 (Why the single-occurrence restriction is real). Without the single-occurrence restriction, $\Delta_{\mathrm{ctx}}(a, b; \eta) = 0$ only rules out *first-order* contextual visibility of the substitution $b \mapsto a$. If a program contains multiple occurrences of b, then higher-order replacement terms can survive even when every one-hole context is individually unobservable. In repeated-symbol languages, the correct exact obstruction is therefore not the scalar Δ_{ctx} alone, but the full linearized operator introduced next.

C.3. A stronger local observability modulus

We now introduce the quantity that governs full local identifiability of the module library from the induced family of composed operators.

Definition C.7 (Program-operator map and library norm)**.** Let

$$\Phi : \mathbb{A}_B^r \to \big(\mathbb{R}^{d\times d}\big)^{\mathcal{L}}, \qquad \Phi(\Theta) := \big(A_z(\Theta)\big)_{z\in\mathcal{L}}.$$

For $H = (H_1, \dots, H_r) \in (\mathbb{R}^{d\times d})^r$, define

$$\|H\|_{\mathrm{lib}}^2 := \sum_{a=1}^{r} \|H_a\|_F^2.$$

For $U = (U_z)_{z\in\mathcal{L}} \in (\mathbb{R}^{d\times d})^{\mathcal{L}}$ and $\pi \in \mathfrak{P}_{\mathcal{L}}^{\circ}$, define

$$\|U\|_\pi^2 := \sum_{z\in\mathcal{L}} \pi_z \|U_z\|_F^2.$$

For $z = (z_1, \dots, z_t) \in \mathcal{L}$ and $\ell \in \{1, \dots, t\}$, write

$$z_{<\ell} := (z_1, \dots, z_{\ell-1}), \qquad z_{>\ell} := (z_{\ell+1}, \dots, z_t).$$

Lemma C.8 (Derivative of the program-operator map)**.** *The map Φ is polynomial, hence C^∞, and its Fréchet derivative at $\Theta = (A_1, \dots, A_r)$ is given by*

$$\big[D\Phi_\Theta(H)\big]_z = \sum_{\ell=1}^{|z|} A_{z_{>\ell}}\, H_{z_\ell}\, A_{z_{<\ell}} \qquad \textit{for every } z \in \mathcal{L}. \tag{48}$$

Proof. Fix $z = (z_1, \dots, z_t)$. The map

$$\Theta \mapsto A_z(\Theta) = A_{z_t} \cdots A_{z_1}$$

is a noncommutative polynomial of degree $t \le L$ in the matrix coordinates $(A_1, \dots, A_r)$, hence smooth. Differentiating the product in each location where a perturbation may enter yields

$$D\big(A_z\big)_\Theta(H) = \sum_{\ell=1}^{t} A_{z_t} \cdots A_{z_{\ell+1}}\, H_{z_\ell}\, A_{z_{\ell-1}} \cdots A_{z_1},$$

which is exactly Equation (48). Since $\mathcal{L}$ is finite, the statement follows coordinatewise. □

Definition C.9 (Linearized contextual observability)**.** For $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$, define the *linearized contextual observability modulus*

$$\lambda_{\mathrm{lin}}(\eta) := \inf_{\|H\|_{\mathrm{lib}}=1} \big\|D\Phi_\Theta(H)\big\|_\pi^2. \tag{49}$$

Equivalently, $\lambda_{\mathrm{lin}}(\eta)$ is the smallest singular value squared of the linear map

$$D\Phi_\Theta : \big((\mathbb{R}^{d\times d})^r, \|\cdot\|_{\mathrm{lib}}\big) \to \big((\mathbb{R}^{d\times d})^{\mathcal{L}}, \|\cdot\|_\pi\big).$$

Remark C.10 (Why λ_{lin} is stronger than Δ_{ctx}). The scalar Δ_{ctx} probes pairwise visibility of replacing one primitive by another inside a random one-hole context. By contrast, λ_{lin} controls *all* infinitesimal perturbations of the module library simultaneously. This distinction is unavoidable: full local identification of the map $\Theta \mapsto (A_z)_{z\in\mathcal{L}}$ requires ruling out arbitrary coordinated cancellations, not only pairwise substitutions.

The next lemma supplies the quadratic remainder estimate needed to convert injectivity of the derivative into local injectivity of the nonlinear program-operator map.

Lemma C.11 (Quadratic Taylor remainder)**.** *Fix $\eta^\star = (\pi^\star, \Theta^\star) \in \Xi(B, \mathcal{L})$. There exists a constant*

$$C_{\mathrm{quad}} = C_{\mathrm{quad}}(\mathcal{L}, r, d, L, B, \pi^\star) < \infty$$

such that for every $H \in (\mathbb{R}^{d\times d})^r$ with $\Theta^\star + H \in \mathbb{A}_B^r$,

$$\big\|\Phi(\Theta^\star + H) - \Phi(\Theta^\star) - D\Phi_{\Theta^\star}(H)\big\|_{\pi^\star} \le C_{\mathrm{quad}}\|H\|_{\mathrm{lib}}^2. \tag{50}$$

Proof. Each coordinate map $\Theta \mapsto A_z(\Theta)$ is a polynomial of degree at most L on the finite-dimensional Euclidean space $(\mathbb{R}^{d\times d})^r$. Therefore its second derivative is continuous. Since $\mathbb{A}_B^r$ is compact, the second derivative of each coordinate is uniformly bounded on $\mathbb{A}_B^r$. Summing over the finitely many coordinates $z \in \mathcal{L}$ and using the weighted norm $\|\cdot\|_{\pi^\star}$ shows that the Hessian of Φ is uniformly bounded on $\mathbb{A}_B^r$ as a bilinear map from $(\mathbb{R}^{d\times d})^r \times (\mathbb{R}^{d\times d})^r$ into $((\mathbb{R}^{d\times d})^{\mathcal{L}}, \|\cdot\|_{\pi^\star})$. The finite-dimensional Taylor theorem with integral remainder then yields Equation (50). □

Proposition C.12 (Local injectivity of the program-operator map)**.** *Fix $\eta^\star = (\pi^\star, \Theta^\star) \in \Xi(B, \mathcal{L})$ and suppose that*

$$\lambda_{\mathrm{lin}}(\eta^\star) > 0.$$

Then there exist constants $\rho_{\mathrm{lin}} > 0$ and $c_{\mathrm{lin}} > 0$ such that for every $\Theta \in \mathbb{A}_B^r$ with

$$\|\Theta - \Theta^\star\|_{\mathrm{lib}} \le \rho_{\mathrm{lin}},$$

one has

$$\|\Phi(\Theta) - \Phi(\Theta^\star)\|_{\pi^\star} \ge c_{\mathrm{lin}}\,\|\Theta - \Theta^\star\|_{\mathrm{lib}}. \tag{51}$$

In particular, Φ is locally injective at $\Theta^\star$.

Proof. Set

$$\lambda_\star := \lambda_{\mathrm{lin}}(\eta^\star) > 0, \qquad c_\star := \sqrt{\lambda_\star}.$$

By definition of λ_{lin},

$$\|D\Phi_{\Theta^\star}(H)\|_{\pi^\star} \ge c_\star\|H\|_{\mathrm{lib}} \qquad \text{for every } H \in (\mathbb{R}^{d\times d})^r.$$

Now let $H := \Theta - \Theta^\star$. By Theorem C.11,

$$\|\Phi(\Theta) - \Phi(\Theta^\star)\|_{\pi^\star} \ge \|D\Phi_{\Theta^\star}(H)\|_{\pi^\star} - C_{\mathrm{quad}}\|H\|_{\mathrm{lib}}^2 \ge c_\star\|H\|_{\mathrm{lib}} - C_{\mathrm{quad}}\|H\|_{\mathrm{lib}}^2.$$

Choose

$$\rho_{\mathrm{lin}} := \min\left\{1, \frac{c_\star}{2C_{\mathrm{quad}}}\right\}, \qquad c_{\mathrm{lin}} := \frac{c_\star}{2}.$$

Then for $\|H\|_{\mathrm{lib}} \le \rho_{\mathrm{lin}}$,

$$C_{\mathrm{quad}}\|H\|_{\mathrm{lib}}^2 \le \frac{c_\star}{2}\|H\|_{\mathrm{lib}},$$

whence

$$\|\Phi(\Theta) - \Phi(\Theta^\star)\|_{\pi^\star} \ge \frac{c_\star}{2}\|H\|_{\mathrm{lib}} = c_{\mathrm{lin}}\|\Theta - \Theta^\star\|_{\mathrm{lib}}.$$

This is Equation (51). □

C.4. From grouped laws to the induced mixing measure

The grouped law $Q_\eta^{(n)}$ depends on the latent grammar only through the induced discrete measure on composed operators.

Definition C.13 (Induced mixing measure and component separation)**.** For $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$, define the induced finite measure on operator space

$$\mu_\eta := \sum_{z \in \mathcal{L}} \pi_z \, \delta_{A_z(\Theta)} \qquad \text{on } \mathbb{R}^{d \times d}. \tag{52}$$

For $\eta^\star = (\pi^\star, \Theta^\star)$, define the component-separation margin

$$\gamma_{\mathrm{comp}}(\eta^\star) := \min_{z \neq z'} \|A_z^\star - A_{z'}^\star\|_F. \tag{53}$$

Remark C.14 (Mixture identifiability layer). Classical results of Teicher (1963); Yakowitz & Spragins (1968) show that finite mixtures of Gaussian kernels are identifiable. In our setting we can prove the relevant statement directly, because the reduced experiment of Appendix B is a finite mixture of matrix-normal kernels with a common covariance depending only on the design.

Lemma C.15 (Injectivity of the common-covariance matrix-normal kernel)**.** *Fix a full-row-rank design $X \in \mathbb{R}^{d \times n}$. Let μ and ν be finite signed Borel measures on $\mathbb{R}^{d \times d}$. Suppose that the corresponding mixture densities satisfy*

$$\int \varphi_X(W_{\mathrm{obs}} - W)\, \mu(dW) = \int \varphi_X(W_{\mathrm{obs}} - W)\, \nu(dW) \qquad \textit{for Lebesgue-a.e. } W_{\mathrm{obs}} \in \mathbb{R}^{d \times d},$$

where $\varphi_X(\cdot - W)$ denotes the density of $\mathcal{MN}_{d,d}(W, \sigma^2 I_d, G_X^{-1})$. Then $\mu = \nu$.

Proof. Vectorize the matrices. Let

$$p := d^2, \qquad \Sigma_X := G_X^{-1} \otimes \sigma^2 I_d \in \mathbb{R}^{p \times p},$$

which is positive definite because G_X is. Let $\bar{\mu}$ and $\bar{\nu}$ be the pushforwards of μ and ν under the map $W \mapsto \mathrm{vec}(W)$. Then the assumed equality becomes

$$\phi_{\Sigma_X} * \bar{\mu} = \phi_{\Sigma_X} * \bar{\nu} \qquad \text{a.e. on } \mathbb{R}^p,$$

where ϕ_{Σ_X} is the p-variate Gaussian density with covariance Σ_X, and $*$ denotes convolution.

Taking Fourier transforms gives

$$\widehat{\phi_{\Sigma_X}}(t)\, \widehat{\bar{\mu}}(t) = \widehat{\phi_{\Sigma_X}}(t)\, \widehat{\bar{\nu}}(t) \qquad \text{for all } t \in \mathbb{R}^p.$$

Since

$$\widehat{\phi_{\Sigma_X}}(t) = \exp\left(-\frac{1}{2} t^\top \Sigma_X t\right)$$

is strictly positive for every t, it follows that

$$\widehat{\bar{\mu}}(t) = \widehat{\bar{\nu}}(t) \qquad \text{for all } t \in \mathbb{R}^p.$$

By the uniqueness theorem for finite Borel measures on $\mathbb{R}^p$, $\bar{\mu} = \bar{\nu}$. Pushing back through the vectorization map yields $\mu = \nu$. □

Proposition C.16 (The grouped law determines the induced mixing measure)**.** *Let $\eta, \eta' \in \Xi(B, \mathcal{L})$, and let $n \geq d + 1$. If*

$$Q_\eta^{(n)} = Q_{\eta'}^{(n)}$$

as probability laws on one task dataset $D = (X, Y)$, then

$$\mu_\eta = \mu_{\eta'}$$

as finite measures on $\mathbb{R}^{d \times d}$.

Proof. Since $\widehat{W} = YX^\top(XX^\top)^{-1}$ is a measurable function of $D = (X, Y)$ on the almost-sure event $\{XX^\top \text{ invertible}\}$, equality of the grouped laws implies equality of the joint laws of $(X, \widehat{W})$ under η and η'. By disintegration, for almost every X with full row rank,

$$\mathcal{L}(\widehat{W} \mid X, \eta) = \mathcal{L}(\widehat{W} \mid X, \eta').$$

By Theorem B.6 from Appendix B,

$$\mathcal{L}(\widehat{W} \mid X, \eta) = \int \mathcal{MN}_{d,d}(W, \sigma^2 I_d, G_X^{-1})\, \mu_\eta(dW),$$

and analogously for η'. Applying Theorem C.15 for any such full-row-rank X yields $\mu_\eta = \mu_{\eta'}$. □

C.5. Local quotient identifiability

We can now state and prove the exact local identifiability theorem.

Theorem C.17 (Local quotient identifiability)**.** *Fix $\eta^\star = (\pi^\star, \Theta^\star) \in \Xi(B, \mathcal{L})$. Assume that*

$$\lambda_{\text{lin}}(\eta^\star) > 0, \qquad \gamma_{\text{comp}}(\eta^\star) > 0. \tag{54}$$

Then there exists $\rho_{\text{id}} > 0$ such that the following holds. For any $n \geq d + 1$ and any $\eta \in \Xi(B, \mathcal{L})$,

$$d_q(\eta, \eta^\star) \leq \rho_{\text{id}} \quad \textit{and} \quad Q_\eta^{(n)} = Q_{\eta^\star}^{(n)}$$

imply

$$d_q(\eta, \eta^\star) = 0.$$

Equivalently, in a sufficiently small quotient neighborhood of $\eta^\star$, the grouped law uniquely identifies the latent task-family parameter.

Proof. Let

$$\gamma_\star := \gamma_{\text{comp}}(\eta^\star) > 0.$$

By continuity of Φ, there exists $\rho_{\text{sep}} > 0$ such that whenever

$$\|\Theta - \Theta^\star\|_{\text{lib}} \leq \rho_{\text{sep}},$$

one has

$$\max_{z \in \mathcal{L}} \|A_z(\Theta) - A_z^\star\|_F < \frac{\gamma_\star}{4}. \tag{55}$$

Let $\rho_{\text{lin}} > 0$ and $c_{\text{lin}} > 0$ be given by Theorem C.12. Set

$$\rho_{\text{id}} := \min\{\rho_{\text{sep}}, \rho_{\text{lin}}\}.$$

Now fix $\eta = (\pi, \Theta)$ with $d_q(\eta, \eta^\star) \leq \rho_{\text{id}}$, and assume

$$Q_\eta^{(n)} = Q_{\eta^\star}^{(n)}.$$

By definition of d_q, there exists $\sigma \in \text{Aut}(\mathcal{L})$ such that

$$d_{\text{par}}(\sigma \cdot \eta, \eta^\star) = d_q(\eta, \eta^\star) \leq \rho_{\text{id}}.$$

Since grouped laws are quotient-invariant by Theorem A.4,

$$Q_{\sigma\cdot\eta}^{(n)} = Q_\eta^{(n)} = Q_{\eta^\star}^{(n)}.$$

Hence it suffices to prove $\sigma \cdot \eta = \eta^\star$. Replacing η by $\sigma \cdot \eta$, we may therefore assume from the outset that

$$d_{\text{par}}(\eta, \eta^\star) \leq \rho_{\text{id}}.$$

By Theorem C.16,

$$\mu_\eta = \mu_{\eta^\star}.$$

Write

$$\mu_{\eta^\star} = \sum_{z\in\mathcal{L}} \pi_z^\star \delta_{A_z^\star}, \qquad \mu_\eta = \sum_{z\in\mathcal{L}} \pi_z \delta_{A_z(\Theta)}.$$

By Equation (55), every atom $A_z(\Theta)$ lies inside the open ball

$$B_z := B\Big(A_z^\star, \frac{\gamma_\star}{4}\Big).$$

These balls are pairwise disjoint because the centers are separated by at least $\gamma_\star$. Moreover, if $z' \neq z$, then

$$\|A_{z'}(\Theta) - A_z^\star\|_F \geq \|A_{z'}^\star - A_z^\star\|_F - \|A_{z'}(\Theta) - A_{z'}^\star\|_F > \gamma_\star - \frac{\gamma_\star}{4} = \frac{3\gamma_\star}{4},$$

so $A_{z'}(\Theta) \notin B_z$. Hence B_z contains exactly one atom of μ_η, namely $A_z(\Theta)$, and exactly one atom of $\mu_{\eta^\star}$, namely $A_z^\star$.

Since $\mu_\eta = \mu_{\eta^\star}$, the two measures assign the same mass to B_z. Therefore

$$\pi_z = \mu_\eta(B_z) = \mu_{\eta^\star}(B_z) = \pi_z^\star \qquad \text{for every } z \in \mathcal{L}.$$

Moreover, because the only atom of μ_η inside B_z is $A_z(\Theta)$, and the only atom of $\mu_{\eta^\star}$ inside B_z is $A_z^\star$, equality of the two atomic measures forces

$$A_z(\Theta) = A_z^\star \qquad \text{for every } z \in \mathcal{L}.$$

Thus

$$\Phi(\Theta) = \Phi(\Theta^\star).$$

Since $\|\Theta - \Theta^\star\|_{\mathrm{lib}} \leq \rho_{\mathrm{id}} \leq \rho_{\mathrm{lin}}$, Theorem C.12 implies

$$0 = \|\Phi(\Theta) - \Phi(\Theta^\star)\|_{\pi^\star} \geq c_{\mathrm{lin}} \|\Theta - \Theta^\star\|_{\mathrm{lib}},$$

hence $\Theta = \Theta^\star$. Together with $\pi = \pi^\star$, this shows $\eta = \eta^\star$. Undoing the initial quotient alignment yields $d_q(\eta, \eta^\star) = 0$. □

Remark C.18 (Interpretation of the theorem). The theorem cleanly separates two layers of identifiability. The first layer is *mixture identifiability*: the grouped law determines the induced discrete measure μ_η on composed operators. The second is *grammar identifiability*: the operator family $(A_z)_{z\in\mathcal{L}}$ locally determines the module library Θ. The first layer is classical in spirit and follows from the common-covariance Gaussian kernel; the second is the genuinely compositional part, and it is governed by λ_{lin}, not by Δ_{ctx} alone.

D. Testing, Entropy, and Prior Mass

This appendix verifies the three technical ingredients needed for posterior contraction in the reduced experiment of Appendix B: exponentially consistent tests, metric entropy bounds, and local prior thickness. The overall architecture is standard in Bayesian nonparametrics (Ghosal et al., 2000), but the substantive content here is model-specific: we compute all bounds in the grouped stochastic-grammar experiment and make explicit how the within-task sample size n sharpens the geometry of the module parameters. For robust tests between Hellinger balls, we appeal to the Le Cam–Birgé theory in the form summarized by Birgé (2013).

D.1. Conditional product experiment and predictive radius

Fix $m \geq 1$ and deterministic designs $X_{1:m} = (X_1, \ldots, X_m)$ with each $X_i \in \mathbb{R}^{d\times n}$ full row rank. For $\eta \in \Xi(B, \mathcal{L})$, let

$$q_{\eta,i}(\cdot) := q_\eta(\cdot \mid X_i)$$

denote the conditional density of the reduced statistic $\widehat{W}_i$ given X_i, as defined in Equation (26). The conditional product law of the reduced experiment is

$$\mathbb{P}_{\eta,X}^{(m)} := \bigotimes_{i=1}^m q_{\eta,i}.$$

We write

$$\rho(P,Q) := \int \sqrt{pq}\, d\mu = 1 - h^2(P,Q)$$

for the Hellinger affinity of two dominated probability laws P and Q.

Definition D.1 (Average conditional Hellinger metric)**.** For $\eta, \eta' \in \Xi(B, \mathcal{L})$, define

$$\bar{h}_X^2(\eta, \eta') := \frac{1}{m} \sum_{i=1}^{m} h^2(q_{\eta,i}, q_{\eta',i}). \tag{56}$$

The predictive contraction scale of the grouped law is

$$\delta_{m,n}^2 := \frac{K_{\mathrm{eff}} \log m + rd^2 \log(mn)}{m}. \tag{57}$$

This is the scale at which the entropy and prior-thickness bounds balance. The sharper $1/(mn)$ scaling for the shared module library will emerge only after converting grouped-law contraction into structural error in Appendix F.

D.2. Conditional Hellinger tests

We begin with the simple-vs-simple testing bound. It is included explicitly because it is the quantitative hinge on which the later net argument turns.

Lemma D.2 (Likelihood-ratio test for the conditional product experiment)**.** *Fix $\eta_0, \eta_1 \in \Xi(B, \mathcal{L})$ and deterministic designs $X_{1:m}$. There exists a measurable test $\phi = \phi_{\eta_0,\eta_1,X} \in [0,1]$ such that*

$$\mathbb{E}_{\eta_0}[\phi \mid X_{1:m}] + \mathbb{E}_{\eta_1}[1-\phi \mid X_{1:m}] \le \rho\Big(\mathbb{P}_{\eta_0,X}^{(m)}, \mathbb{P}_{\eta_1,X}^{(m)}\Big) \le \exp\big(-m\, \bar{h}_X^2(\eta_0,\eta_1)\big)\,. \tag{58}$$

Proof. Let p_0 and p_1 denote the densities of $\mathbb{P}_{\eta_0,X}^{(m)}$ and $\mathbb{P}_{\eta_1,X}^{(m)}$ with respect to a common dominating measure on $(\mathbb{R}^{d\times d})^m$. Consider the likelihood-ratio test

$$\phi := \mathbf{1}\{p_1 \ge p_0\}.$$

Then

$$\mathbb{E}_{\eta_0}[\phi \mid X_{1:m}] + \mathbb{E}_{\eta_1}[1-\phi \mid X_{1:m}] = \int \min\{p_0, p_1\} \le \int \sqrt{p_0 p_1} = \rho\Big(\mathbb{P}_{\eta_0,X}^{(m)}, \mathbb{P}_{\eta_1,X}^{(m)}\Big)\,.$$

Because the product experiment is conditionally independent across tasks,

$$\rho\Big(\mathbb{P}_{\eta_0,X}^{(m)}, \mathbb{P}_{\eta_1,X}^{(m)}\Big) = \prod_{i=1}^{m} \rho(q_{\eta_0,i}, q_{\eta_1,i}) = \prod_{i=1}^{m} \big(1 - h^2(q_{\eta_0,i}, q_{\eta_1,i})\big).$$

Using $1 - u \le e^{-u}$ for $u \in [0,1]$ gives

$$\prod_{i=1}^{m} \big(1 - h^2(q_{\eta_0,i}, q_{\eta_1,i})\big) \le \exp\left(-\sum_{i=1}^{m} h^2(q_{\eta_0,i}, q_{\eta_1,i})\right) = \exp\big(-m\, \bar{h}_X^2(\eta_0,\eta_1)\big)\,.$$

□

The next proposition upgrades Theorem D.2 to composite alternatives by combining a finite net with the classical Le Cam–Birgé robust test between Hellinger balls.

Proposition D.3 (Composite tests from Hellinger coverings)**.** *Fix $\eta^\star \in \Xi(B, \mathcal{L})$, deterministic designs $X_{1:m}$, and $\varepsilon > 0$. Let $A \subseteq \Xi(B, \mathcal{L})$ satisfy*

$$\inf_{\eta \in A} \bar{h}_X(\eta, \eta^\star) \ge 4\varepsilon.$$

Then there exists a measurable test $\phi_{A,\varepsilon,X} \in [0,1]$ such that

$$\mathbb{E}_{\eta^\star}[\phi_{A,\varepsilon,X} \mid X_{1:m}] \le N\big(\varepsilon, A, \bar{h}_X\big)\; e^{-m\varepsilon^2}, \tag{59}$$

$$\sup_{\eta \in A} \mathbb{E}_{\eta}[1 - \phi_{A,\varepsilon,X} \mid X_{1:m}] \le e^{-m\varepsilon^2}, \tag{60}$$

where $N(\varepsilon, A, \bar{h}_X)$ denotes the ε-covering number of A under $\bar{h}_X$.

Proof. Let $\{\eta_1, \dots, \eta_N\} \subseteq A$ be an ε-net of A with $N = N(\varepsilon, A, \bar{h}_X)$. Since every center η_j satisfies $\bar{h}_X(\eta_j, \eta^\star) \geq 4\varepsilon$, the robust Hellinger-ball testing theorem of Le Cam–Birgé yields, for each j, a test ϕ_j such that

$$\sup_{\bar{h}_X(\eta,\eta^\star)\leq\varepsilon} \mathbb{E}_\eta[\phi_j \mid X_{1:m}] \leq e^{-m\varepsilon^2}, \qquad \sup_{\bar{h}_X(\eta,\eta_j)\leq\varepsilon} \mathbb{E}_\eta[1-\phi_j \mid X_{1:m}] \leq e^{-m\varepsilon^2}.$$

See Birgé (2013) for a modern statement encompassing independent non-identically distributed experiments.

Define

$$\phi_{A,\varepsilon,X} := \max_{1\leq j\leq N} \phi_j.$$

Then

$$\mathbb{E}_{\eta^\star}[\phi_{A,\varepsilon,X} \mid X_{1:m}] \leq \sum_{j=1}^{N} \mathbb{E}_{\eta^\star}[\phi_j \mid X_{1:m}] \leq N e^{-m\varepsilon^2},$$

because $\bar{h}_X(\eta^\star, \eta^\star) = 0 \leq \varepsilon$. On the other hand, if $\eta \in A$, choose j such that $\bar{h}_X(\eta, \eta_j) \leq \varepsilon$. Then

$$\mathbb{E}_\eta[1-\phi_{A,\varepsilon,X} \mid X_{1:m}] \leq \mathbb{E}_\eta[1-\phi_j \mid X_{1:m}] \leq e^{-m\varepsilon^2}.$$

This proves Equations (59) and (60). □

D.3. Entropy bounds

We next control the metric entropy of the reduced experiment on the high-probability good-design event introduced in Appendix B.

For $i \in [m]$, let

$$\mathcal{E}_n^{(i)}(\delta) := \left\{(1-\delta)nI_d \preceq X_i X_i^\top \preceq (1+\delta)nI_d\right\},$$

and define the joint good-design event

$$\mathcal{G}_{m,n}(\delta) := \bigcap_{i=1}^{m} \mathcal{E}_n^{(i)}(\delta). \tag{61}$$

Lemma D.4 (Probability of the joint good-design event)**.** *For every $0 < \delta < 1$, there exist constants $c_{d,\delta}, C_{d,\delta} \in (0,\infty)$, depending only on d and δ, such that*

$$\Pr_{\eta^\star}\big(\mathcal{G}_{m,n}(\delta)^c\big) \leq m\, C_{d,\delta}\, e^{-c_{d,\delta} n}.$$

Proof. This follows immediately from Theorem B.10 and a union bound over $i = 1, \dots, m$. □

We will repeatedly use the following discrete Hellinger distance on the simplex:

$$h^2_{\mathrm{disc}}(\pi, \pi') := \frac{1}{2} \sum_{z\in\mathcal{L}} \big(\sqrt{\pi_z} - \sqrt{\pi'_z}\big)^2.$$

Lemma D.5 (Componentwise telescoping bound for composed operators)**.** *There exists a constant*

$$C_{\mathrm{prod}} = C_{\mathrm{prod}}(L, B, r) < \infty$$

such that for every $\Theta = (A_1, \dots, A_r) \in \mathbb{A}_B^r$, every $\Theta' = (A'_1, \dots, A'_r) \in \mathbb{A}_B^r$, and every $z \in \mathcal{L}$,

$$\|A_z(\Theta) - A_z(\Theta')\|_F \leq C_{\mathrm{prod}}\, \|\Theta - \Theta'\|_{\mathrm{lib}}. \tag{62}$$

One may take $C_{\mathrm{prod}} = LB^{L-1}\sqrt{r}$.

Proof. Fix $z = (z_1, \dots, z_t) \in \mathcal{L}$, with $t \leq L$. A telescoping expansion gives

$$A_z(\Theta) - A_z(\Theta') = \sum_{\ell=1}^{t} A_{z_t} \cdots A_{z_{\ell+1}} \big(A_{z_\ell} - A'_{z_\ell}\big) A'_{z_{\ell-1}} \cdots A'_{z_1}.$$

Taking Frobenius norms, using submultiplicativity of the operator norm, and the bound $\|A_a\|_{\mathrm{op}}, \|A'_a\|_{\mathrm{op}} \le B$, we obtain

$$\|A_z(\Theta) - A_z(\Theta')\|_F \le \sum_{\ell=1}^{t} B^{t-1} \|A_{z_\ell} - A'_{z_\ell}\|_F \le LB^{L-1} \sum_{a=1}^{r} \|A_a - A'_a\|_F.$$

Finally,

$$\sum_{a=1}^{r} \|A_a - A'_a\|_F \le \sqrt{r}\, \|\Theta - \Theta'\|_{\mathrm{lib}}.$$

Combining the displays yields Equation (62). □

The next lemma is the key predictive upper bound. It shows that, on the good-design event, the reduced conditional Hellinger geometry is locally equivalent to the direct-product metric

$$(\pi, \Theta) \mapsto h_{\mathrm{disc}}(\pi, \pi^\star) + \sqrt{n}\, \|\Theta - \Theta^\star\|_{\mathrm{lib}}.$$

Lemma D.6 (Local Hellinger upper bound in the reduced experiment)**.** *Fix $0 < \delta < 1$. There exist constants $c_{\mathrm{loc}}, C_{\mathrm{loc}} \in (0, \infty)$, depending only on $(\mathcal{L}, r, d, L, B, \sigma, \delta)$, such that the following holds on $\mathcal{G}_{m,n}(\delta)$. If $\eta = (\pi, \Theta)$ and $\eta' = (\pi', \Theta')$ satisfy*

$$\|\Theta - \Theta'\|_{\mathrm{lib}} \le c_{\mathrm{loc}}\, n^{-1/2},$$

then

$$\bar{h}^2_X(\eta, \eta') \le 2h^2_{\mathrm{disc}}(\pi, \pi') + C_{\mathrm{loc}}\, n\, \|\Theta - \Theta'\|^2_{\mathrm{lib}}. \tag{63}$$

Proof. Fix a task index $i \in [m]$, and write

$$f_{z,i} := \mathcal{MN}_{d,d}\big(A_z(\Theta), \sigma^2 I_d, (X_i X_i^\top)^{-1}\big), \qquad f'_{z,i} := \mathcal{MN}_{d,d}\big(A_z(\Theta'), \sigma^2 I_d, (X_i X_i^\top)^{-1}\big).$$

Then

$$q_{\eta,i} = \sum_{z \in \mathcal{L}} \pi_z f_{z,i}, \qquad q_{\eta',i} = \sum_{z \in \mathcal{L}} \pi'_z f'_{z,i}.$$

By the triangle inequality for Hellinger distance,

$$h(q_{\eta,i}, q_{\eta',i}) \le h\left(\sum_z \pi_z f_{z,i}, \sum_z \pi_z f'_{z,i}\right) + h\left(\sum_z \pi_z f'_{z,i}, \sum_z \pi'_z f'_{z,i}\right).$$

Hence

$$h^2(q_{\eta,i}, q_{\eta',i}) \le 2h_1^2 + 2h_2^2, \tag{64}$$

where

$$h_1 := h\left(\sum_z \pi_z f_{z,i}, \sum_z \pi_z f'_{z,i}\right), \qquad h_2 := h\left(\sum_z \pi_z f'_{z,i}, \sum_z \pi'_z f'_{z,i}\right).$$

For the weights term, consider the joint laws on $\mathcal{L} \times \mathbb{R}^{d \times d}$,

$$\widetilde{q}_i(z, w) := \pi_z f'_{z,i}(w), \qquad \widetilde{q}'_i(z, w) := \pi'_z f'_{z,i}(w).$$

Marginalizing over z is a measurable map, so by data processing for Hellinger distance,

$$h_2^2 \le h^2(\widetilde{q}_i, \widetilde{q}'_i) = h^2_{\mathrm{disc}}(\pi, \pi').$$

For the component-location term, Hellinger affinity is concave in each argument, hence

$$\rho\left(\sum_z \pi_z f_{z,i}, \sum_z \pi_z f'_{z,i}\right) \ge \sum_{z \in \mathcal{L}} \pi_z\, \rho(f_{z,i}, f'_{z,i}).$$

Therefore

$$h_1^2 \le \sum_{z\in\mathcal{L}} \pi_z\, h^2(f_{z,i}, f'_{z,i}).$$

By Theorem D.5,

$$\max_{z\in\mathcal{L}} \|A_z(\Theta) - A_z(\Theta')\|_F \le C_{\mathrm{prod}} \|\Theta - \Theta'\|_{\mathrm{lib}}.$$

Hence, if $\|\Theta - \Theta'\|_{\mathrm{lib}} \le c_{\mathrm{loc}} n^{-1/2}$ and c_{loc} is chosen sufficiently small, then

$$\frac{(1+\delta)n}{8\sigma^2} \max_{z\in\mathcal{L}} \|A_z(\Theta) - A_z(\Theta')\|_F^2 \le 1.$$

On $\mathcal{G}_{m,n}(\delta)$, Theorem B.11 from Appendix B then yields

$$h^2(f_{z,i}, f'_{z,i}) \le C'_{\mathrm{loc}}\, n\, \|A_z(\Theta) - A_z(\Theta')\|_F^2 \le C'_{\mathrm{loc}}\, n\, C_{\mathrm{prod}}^2 \|\Theta - \Theta'\|_{\mathrm{lib}}^2.$$

Summing over z with weights π_z gives

$$h_1^2 \le C'_{\mathrm{loc}}\, C_{\mathrm{prod}}^2\, n\, \|\Theta - \Theta'\|_{\mathrm{lib}}^2.$$

Substituting the bounds for h_1^2 and h_2^2 into Equation (64) yields

$$h^2(q_{\eta,i}, q_{\eta',i}) \le 2h_{\mathrm{disc}}^2(\pi,\pi') + 2C'_{\mathrm{loc}} C_{\mathrm{prod}}^2\, n\, \|\Theta - \Theta'\|_{\mathrm{lib}}^2.$$

Averaging over $i = 1, \dots, m$ proves Equation (63). □

We now convert the preceding local Hellinger bound into a covering-number estimate.

Proposition D.7 (Conditional metric entropy)**.** *Fix $0 < \delta < 1$. There exist constants $\varepsilon_0 \in (0,1)$ and $C_{\mathrm{ent},1}, C_{\mathrm{ent},2}, C_{\mathrm{ent},3}, C_{\mathrm{ent},4} \in (0,\infty)$, depending only on $(\mathcal{L}, r, d, L, B, \sigma, \delta)$, such that on $\mathcal{G}_{m,n}(\delta)$, for every $0 < \varepsilon \le \varepsilon_0$,*

$$\log N\big(\varepsilon, \Xi(B,\mathcal{L}), \bar{h}_X\big) \le C_{\mathrm{ent},1}\, K_{\mathrm{eff}}\, \log\frac{C_{\mathrm{ent},2}}{\varepsilon} + C_{\mathrm{ent},3}\, rd^2\, \log\frac{C_{\mathrm{ent},4}\sqrt{n}}{\varepsilon}. \tag{65}$$

The same bound holds, up to an additive constant $\log|\operatorname{Aut}(\mathcal{L})|$, for the quotient space $\Xi(B,\mathcal{L})/\operatorname{Aut}(\mathcal{L})$.

Proof. We cover the simplex and the module library separately.

Step 1: covering the grammar weights. Since $\mathcal{L}$ is finite and

$$\mathfrak{P}_{\mathcal{L}}^{\circ} \subseteq \left\{ \pi \in [0,1]^{|\mathcal{L}|} : \sum_z \pi_z = 1 \right\},$$

the simplex has Euclidean dimension $K_{\mathrm{eff}} = |\mathcal{L}| - 1$. A standard volume argument yields

$$\log N(u, \mathfrak{P}_{\mathcal{L}}^{\circ}, \|\cdot\|_1) \le C_1 K_{\mathrm{eff}} \log\frac{C_2}{u} \qquad \text{for } 0 < u \le 1.$$

Since $h_{\mathrm{disc}}^2(\pi,\pi') \le \frac{1}{2}\|\pi - \pi'\|_1$, it follows that

$$\log N\Big(\frac{\varepsilon}{4}, \mathfrak{P}_{\mathcal{L}}^{\circ}, h_{\mathrm{disc}}\Big) \le C_1 K_{\mathrm{eff}} \log\frac{C_2}{\varepsilon}.$$

Step 2: covering the module library. The compact set $\mathbb{A}_B^r$ is contained in a Euclidean ball of radius $B\sqrt{rd}$ in the norm $\|\cdot\|_{\mathrm{lib}}$, and has ambient dimension rd^2. Hence

$$\log N(v, \mathbb{A}_B^r, \|\cdot\|_{\mathrm{lib}}) \le C_3 rd^2 \log\frac{C_4}{v} \qquad \text{for } 0 < v \le 1.$$

Step 3: combine the coverings through Theorem D.6. Choose the module covering radius

$$v_\varepsilon := \min\left\{\frac{c_{\mathrm{loc}}}{\sqrt{n}},\ \frac{\varepsilon}{4\sqrt{C_{\mathrm{loc}} n}}\right\},$$

and the simplex covering radius

$$u_\varepsilon := \frac{\varepsilon}{4}.$$

For any $\eta = (\pi, \Theta)$, select centers $\pi^\sharp$ and $\Theta^\sharp$ from the two coverings such that

$$h_{\mathrm{disc}}(\pi, \pi^\sharp) \le u_\varepsilon, \qquad \|\Theta - \Theta^\sharp\|_{\mathrm{lib}} \le v_\varepsilon.$$

Then on $\mathcal{G}_{m,n}(\delta)$, Theorem D.6 implies

$$\bar{h}_X^2\big((\pi,\Theta),(\pi^\sharp,\Theta^\sharp)\big) \le 2u_\varepsilon^2 + C_{\mathrm{loc}} n v_\varepsilon^2 \le \frac{\varepsilon^2}{8} + \frac{\varepsilon^2}{16} < \varepsilon^2.$$

Thus the Cartesian product of the two coverings is an ε-cover of $\Xi(B,\mathcal{L})$ under $\bar{h}_X$. Taking logarithms and absorbing constants yields Equation (65). Since passing to the quotient cannot increase covering numbers by more than the finite factor $|\operatorname{Aut}(\mathcal{L})|$, the quotient statement follows. □

D.4. Local prior thickness

We now verify that the prior allocates enough mass to neighborhoods whose conditional Kullback–Leibler size is $O(\varepsilon^2)$. The key simplification is that KL for the observed mixture can be bounded by KL in the augmented experiment where the latent program Z is observed.

Fix the truth $\eta^\star = (\pi^\star, \Theta^\star)$. For $X \in \mathbb{R}^{d\times n}$ full row rank and $\eta = (\pi, \Theta)$, define the augmented one-task conditional density

$$\widetilde{q}_{\eta,X}(z,w) := \pi_z\, \varphi_X\big(w; A_z(\Theta)\big), \qquad z \in \mathcal{L},\ w \in \mathbb{R}^{d\times d}, \tag{66}$$

where $\varphi_X(\cdot; W)$ is the density of $\mathcal{MN}_{d,d}(W, \sigma^2 I_d, (XX^\top)^{-1})$.

Lemma D.8 (KL upper bound via latent augmentation)**.** *For every full-row-rank design X and every $\eta = (\pi,\Theta) \in \Xi(B,\mathcal{L})$,*

$$\mathrm{KL}\big(q_{\eta^\star}(\cdot \mid X)\,\|\,q_\eta(\cdot \mid X)\big) \le \mathrm{KL}(\pi^\star, \pi) + \sum_{z\in\mathcal{L}} \pi_z^\star\ \mathrm{KL}\Big(\mathcal{N}_X^{A_z^\star}\ \Big\|\ \mathcal{N}_X^{A_z}\Big). \tag{67}$$

Proof. Consider the joint laws of $(Z, \widehat{W})$ under $\eta^\star$ and η, namely

$$\widetilde{q}_{\eta^\star,X}(z,w) = \pi_z^\star \varphi_X(w; A_z^\star), \qquad \widetilde{q}_{\eta,X}(z,w) = \pi_z \varphi_X(w; A_z).$$

Marginalizing over z maps these joint laws to the observed laws $q_{\eta^\star}(\cdot \mid X)$ and $q_\eta(\cdot \mid X)$. Since Kullback–Leibler divergence contracts under measurable maps,

$$\mathrm{KL}\big(q_{\eta^\star}(\cdot \mid X)\,\|\,q_\eta(\cdot \mid X)\big) \le \mathrm{KL}\big(\widetilde{q}_{\eta^\star,X}\,\|\,\widetilde{q}_{\eta,X}\big).$$

Because the joint density factorizes into a discrete law on z and a conditional Gaussian law on $\widehat{W}$,

$$\mathrm{KL}\big(\widetilde{q}_{\eta^\star,X}\,\|\,\widetilde{q}_{\eta,X}\big) = \sum_{z\in\mathcal{L}} \pi_z^\star \log\frac{\pi_z^\star}{\pi_z} + \sum_{z\in\mathcal{L}} \pi_z^\star\, \mathrm{KL}\Big(\mathcal{N}_X^{A_z^\star}\ \Big\|\ \mathcal{N}_X^{A_z}\Big),$$

which is exactly Equation (67). □

To describe local prior thickness, define the neighborhood

$$\mathcal{U}_n(\varepsilon) := \left\{(\pi,\Theta) \in \Xi(B,\mathcal{L}) : \|\pi - \pi^\star\|_1 \le c_\pi \varepsilon,\ \|\Theta - \Theta^\star\|_{\mathrm{lib}} \le c_\Theta \frac{\varepsilon}{\sqrt{n}}\right\}, \tag{68}$$

where $c_\pi, c_\Theta > 0$ will be chosen sufficiently small.

Lemma D.9 (Local KL control on $\mathcal{U}_n(\varepsilon)$)**.** *Fix $0 < \delta < 1$. There exist constants $c_\pi, c_\Theta, \varepsilon_0, C_{\mathrm{KL}} \in (0, \infty)$, depending only on $(\mathcal{L}, r, d, L, B, \sigma, \delta, \pi_{\min})$, such that on $\mathcal{G}_{m,n}(\delta)$, for every $0 < \varepsilon \le \varepsilon_0$ and every $\eta \in \mathcal{U}_n(\varepsilon)$,*

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{KL}\big(q_{\eta^\star}(\cdot \mid X_i) \,\|\, q_\eta(\cdot \mid X_i)\big) \le C_{\mathrm{KL}}\, \varepsilon^2. \tag{69}$$

Proof. Fix $\eta = (\pi, \Theta) \in \mathcal{U}_n(\varepsilon)$. Because $\pi^\star_z \ge \pi_{\min}$ for all $z \in \mathcal{L}$ and $\|\pi - \pi^\star\|_1 \le c_\pi \varepsilon$, choosing $c_\pi \varepsilon_0 \le \pi_{\min}/2$ ensures

$$\pi_z \ge \frac{\pi_{\min}}{2} \qquad \text{for all } z \in \mathcal{L}.$$

Using the elementary inequality $\log x \le x - 1$, one obtains

$$\mathrm{KL}(\pi^\star, \pi) = \sum_{z \in \mathcal{L}} \pi^\star_z \log \frac{\pi^\star_z}{\pi_z} \le \sum_{z \in \mathcal{L}} \frac{(\pi^\star_z - \pi_z)^2}{\pi_z} \le \frac{2}{\pi_{\min}} \|\pi - \pi^\star\|_2^2 \le \frac{2}{\pi_{\min}} \|\pi - \pi^\star\|_1^2 \le \frac{2c_\pi^2}{\pi_{\min}} \varepsilon^2.$$

Next, by Theorem D.5,

$$\max_{z \in \mathcal{L}} \|A_z - A^\star_z\|_F \le C_{\mathrm{prod}} \|\Theta - \Theta^\star\|_{\mathrm{lib}} \le C_{\mathrm{prod}} c_\Theta \frac{\varepsilon}{\sqrt{n}}.$$

On $\mathcal{G}_{m,n}(\delta)$, Equation (37) from Appendix B gives, for each $i \in [m]$ and each $z \in \mathcal{L}$,

$$\mathrm{KL}\Big(\mathcal{N}^{A^\star_z}_{X_i} \,\Big\|\, \mathcal{N}^{A_z}_{X_i}\Big) \le \frac{(1+\delta)n}{2\sigma^2} \|A^\star_z - A_z\|_F^2 \le \frac{(1+\delta) C^2_{\mathrm{prod}} c^2_\Theta}{2\sigma^2} \varepsilon^2.$$

Applying Theorem D.8 yields

$$\mathrm{KL}\big(q_{\eta^\star}(\cdot \mid X_i) \,\|\, q_\eta(\cdot \mid X_i)\big) \le \frac{2c_\pi^2}{\pi_{\min}} \varepsilon^2 + \frac{(1+\delta) C^2_{\mathrm{prod}} c^2_\Theta}{2\sigma^2} \varepsilon^2.$$

The right-hand side is independent of i, so averaging over $i = 1, \dots, m$ gives Equation (69). □

We now lower-bound the prior mass of $\mathcal{U}_n(\varepsilon)$.

Proposition D.10 (Local prior thickness)**.** *There exist constants $\varepsilon_0 \in (0, 1)$ and $c_\Pi, C_\Pi \in (0, \infty)$, depending only on $(\mathcal{L}, r, d, L, B, \sigma, \Pi, \eta^\star)$, such that for every $0 < \varepsilon \le \varepsilon_0$,*

$$\Pi\big(\mathcal{U}_n(\varepsilon)\big) \ge c_\Pi\, \varepsilon^{K_{\mathrm{eff}} + rd^2}\, n^{-rd^2/2}, \tag{70}$$

and therefore

$$-\log \Pi\big(\mathcal{U}_n(\varepsilon)\big) \le C_\Pi \Big[K_{\mathrm{eff}} \log \frac{1}{\varepsilon} + rd^2 \log \frac{\sqrt{n}}{\varepsilon}\Big]. \tag{71}$$

Proof. Because $\pi^\star$ lies in the interior of the simplex and Π_π has a density that is continuous and strictly positive near $\pi^\star$, there exist $r_\pi > 0$ and $c'_\pi > 0$ such that the prior density of Π_π is bounded below by c'_π on the ℓ_1-ball $B_1(\pi^\star, r_\pi) \cap \mathfrak{P}^\circ_{\mathcal{L}}$. Since the simplex has dimension $K_{\mathrm{eff}} = |\mathcal{L}| - 1$, the K_{eff}-dimensional volume of

$$\{\pi \in \mathfrak{P}^\circ_{\mathcal{L}} : \|\pi - \pi^\star\|_1 \le c_\pi \varepsilon\}$$

is bounded below by $c''_\pi \varepsilon^{K_{\mathrm{eff}}}$ for sufficiently small ε. Hence

$$\Pi_\pi(\|\pi - \pi^\star\|_1 \le c_\pi \varepsilon) \ge c'_\pi c''_\pi \varepsilon^{K_{\mathrm{eff}}}.$$

Similarly, because $\Pi_\Theta = \bigotimes_{a=1}^r \Pi_a$ has a continuous strictly positive density near $\Theta^\star$, there exist $r_\Theta > 0$ and $c'_\Theta > 0$ such that this density is bounded below by c'_Θ on the $\|\cdot\|_{\mathrm{lib}}$-ball $B_{\mathrm{lib}}(\Theta^\star, r_\Theta)$. The ambient dimension of $\mathbb{A}^r_B$ is rd^2, so the Euclidean volume of

$$\Big\{\Theta \in \mathbb{A}^r_B : \|\Theta - \Theta^\star\|_{\mathrm{lib}} \le c_\Theta \frac{\varepsilon}{\sqrt{n}}\Big\}$$

is bounded below by

$$c''_\Theta \left(\frac{\varepsilon}{\sqrt{n}}\right)^{rd^2}$$

for sufficiently small ε. Therefore

$$\Pi_\Theta\left(\|\Theta - \Theta^\star\|_{\mathrm{lib}} \le c_\Theta \frac{\varepsilon}{\sqrt{n}}\right) \ge c'_\Theta c''_\Theta \left(\frac{\varepsilon}{\sqrt{n}}\right)^{rd^2}.$$

Using the product structure $\Pi = \Pi_\pi \otimes \Pi_\Theta$, we conclude that

$$\Pi\big(\mathcal{U}_n(\varepsilon)\big) \ge (c'_\pi c''_\pi)(c'_\Theta c''_\Theta)\, \varepsilon^{K_{\mathrm{eff}}+rd^2}\, n^{-rd^2/2},$$

which is Equation (70). Taking negative logarithms yields Equation (71) after enlarging the constant. □

D.5. Packaging the contraction ingredients

The final proposition records the exact balance needed in Appendix E.

Proposition D.11 (Entropy and prior mass at the predictive radius)**.** *There exists $M_0 \in (0,\infty)$, depending only on the fixed model class and the prior, such that the following holds. Fix $0 < \delta < 1$, and let $\delta_{m,n}$ be as in Equation (57). For every $M \ge M_0$, if*

$$\varepsilon_{m,n} := M\,\delta_{m,n},$$

then for all sufficiently large m and all $n \ge d+1$, on the event $\mathcal{G}_{m,n}(\delta)$,

$$\log N\big(\varepsilon_{m,n}, \Xi(B,\mathcal{L}), \bar{h}_X\big) \le \frac{m\varepsilon_{m,n}^2}{8}, \tag{72}$$

$$-\log \Pi\big(\mathcal{U}_n(\varepsilon_{m,n})\big) \le \frac{m\varepsilon_{m,n}^2}{8}. \tag{73}$$

Proof. By Theorem D.7,

$$\log N\big(\varepsilon, \Xi(B,\mathcal{L}), \bar{h}_X\big) \le C_1 K_{\mathrm{eff}} \log\frac{C_2}{\varepsilon} + C_3 rd^2 \log\frac{C_4\sqrt{n}}{\varepsilon}.$$

Likewise, Theorem D.10 gives

$$-\log \Pi\big(\mathcal{U}_n(\varepsilon)\big) \le C_5 K_{\mathrm{eff}} \log\frac{1}{\varepsilon} + C_5 rd^2 \log\frac{\sqrt{n}}{\varepsilon}.$$

Substitute $\varepsilon = \varepsilon_{m,n} = M\delta_{m,n}$, with

$$m\delta_{m,n}^2 = K_{\mathrm{eff}} \log m + rd^2 \log(mn).$$

Since $\delta_{m,n} \to 0$ along any asymptotic regime with $m \to \infty$ and $n \ge d+1$, the logarithmic terms satisfy

$$\log\frac{1}{\varepsilon_{m,n}} \lesssim \log(mn)$$

for all sufficiently large m. Hence both the entropy bound and the prior-thickness bound are controlled by

$$C\big[K_{\mathrm{eff}} \log m + rd^2 \log(mn)\big] = C\, m\delta_{m,n}^2.$$

Choosing M_0 large enough so that $C \le M_0^2/8$ yields Equations (72) and (73) for every $M \ge M_0$. □

Remark D.12 (What this appendix has established). Appendix D proves the full contraction toolkit for the reduced experiment:

1. exponentially consistent conditional tests for $\bar{h}_X$-separated alternatives;
2. conditional entropy bounds at the predictive scale $\delta_{m,n}$;
3. local prior thickness in Kullback–Leibler neighborhoods of size $O(\varepsilon^2)$.

In Appendix E, these ingredients are combined with the exact reduction of Appendix B to obtain predictive posterior contraction for the grouped task-family law.

E. Predictive Posterior Contraction and Latent-Program Recovery

This appendix combines the exact reduction of Appendix B with the testing / entropy / prior-thickness bounds of Appendix D. The resulting predictive contraction theorem is proved in the natural conditional metric $\bar{h}_X$, i.e., the average Hellinger distance in the reduced non-i.i.d. experiment given the random designs. This is the correct predictive metric for the grouped task-family law after sufficient-statistic reduction. The proof follows the standard testing-based posterior-contraction program for independent non-identically distributed observations (Ghosal & van der Vaart, 2007; Ghosal et al., 2000), but every model-specific ingredient has already been established directly in the preceding appendices.

E.1. Posterior notation and design regularity regime

For the reduced experiment, define the posterior

$$\Pi_m(B) := \Pi\big(B \mid X_{1:m}, \widehat{W}_{1:m}\big) = \frac{\int_B \prod_{i=1}^m q_\eta(\widehat{W}_i \mid X_i)\,\Pi(d\eta)}{\int_{\Xi(B,\mathcal{L})} \prod_{i=1}^m q_\eta(\widehat{W}_i \mid X_i)\,\Pi(d\eta)}, \qquad B \subseteq \Xi(B,\mathcal{L}) \text{ measurable.} \tag{74}$$

By Theorem B.7 from Appendix B, this equals the posterior based on the full grouped data $(X_i, Y_i)_{i=1}^m$.

Throughout this appendix we work under the asymptotic regime

$$m\, e^{-c_{d,\delta} n} \to 0 \qquad \text{for some fixed } 0 < \delta < 1, \tag{75}$$

where $c_{d,\delta}$ is the constant from Theorem D.4. Under Equation (75), the joint good-design event

$$\mathcal{G}_{m,n}(\delta) = \bigcap_{i=1}^m \big\{(1-\delta)nI_d \preceq X_i X_i^\top \preceq (1+\delta)nI_d\big\}$$

has probability tending to one under the true law.

We also recall the predictive radius from Appendix D,

$$\delta_{m,n}^2 = \frac{K_{\text{eff}} \log m + rd^2 \log(mn)}{m}.$$

E.2. Predictive posterior contraction in the reduced experiment

The contraction theorem is most naturally stated in the empirical conditional Hellinger metric

$$\bar{h}_X^2(\eta, \eta^\star) = \frac{1}{m}\sum_{i=1}^m h^2\big(q_\eta(\cdot \mid X_i), q_{\eta^\star}(\cdot \mid X_i)\big).$$

Theorem E.1 (Predictive posterior contraction in $\bar{h}_X$)**.** *Assume Theorems A.1 and A.2, and suppose that Equation (75) holds for some fixed $0 < \delta < 1$. Then there exists a constant $M_{\text{pred}} \in (0, \infty)$, depending only on the fixed model class and the prior, such that for every $M \geq M_{\text{pred}}$,*

$$\Pi_m\big(\eta \in \Xi(B,\mathcal{L}) : \bar{h}_X(\eta, \eta^\star) > M\delta_{m,n}\big) \longrightarrow 0 \tag{76}$$

in $P_{\eta^\star}$-probability.

Proof. Fix $0 < \delta < 1$, and let $\mathcal{G}_{m,n}(\delta)$ be the joint good-design event from Equation (61). By Theorem D.4 and the regime Equation (75),

$$P_{\eta^\star}\big(\mathcal{G}_{m,n}(\delta)^c\big) \to 0.$$

It therefore suffices to prove Equation (76) on $\mathcal{G}_{m,n}(\delta)$.

Conditional on $X_{1:m}$, the reduced observations $\widehat{W}_1, \ldots, \widehat{W}_m$ are independent, though generally non-identically distributed, with conditional densities $q_{\eta,1}, \ldots, q_{\eta,m}$. Hence the general testing-based contraction theorem for independent non-i.i.d. experiments from Ghosal & van der Vaart (2007) applies once its three hypotheses are verified for the semimetric

$$d_m(\eta, \eta') := \bar{h}_X(\eta, \eta').$$

We now check those hypotheses on $\mathcal{G}_{m,n}(\delta)$.

Testing condition. For any $\varepsilon > 0$ and any measurable set $A \subseteq \Xi(B, \mathcal{L})$ satisfying

$$\inf_{\eta \in A} \bar{h}_X(\eta, \eta^\star) \geq 4\varepsilon,$$

Theorem D.3 constructs a measurable test $\phi_{A,\varepsilon,X}$ with type-I and uniform type-II error bounds of the form

$$\mathbb{E}_{\eta^\star}[\phi_{A,\varepsilon,X} \mid X_{1:m}] \leq N(\varepsilon, A, \bar{h}_X) e^{-m\varepsilon^2}, \qquad \sup_{\eta \in A} \mathbb{E}_\eta[1 - \phi_{A,\varepsilon,X} \mid X_{1:m}] \leq e^{-m\varepsilon^2}.$$

Thus the testing hypothesis holds in the metric $d_m = \bar{h}_X$.

Entropy condition. By Theorem D.7, for all sufficiently small $\varepsilon > 0$,

$$\log N\big(\varepsilon, \Xi(B, \mathcal{L}), \bar{h}_X\big) \leq C_{\mathrm{ent},1}\, K_{\mathrm{eff}}\, \log \frac{C_{\mathrm{ent},2}}{\varepsilon} + C_{\mathrm{ent},3}\, rd^2\, \log \frac{C_{\mathrm{ent},4}\sqrt{n}}{\varepsilon}.$$

Choosing $\varepsilon = \varepsilon_{m,n} := M\delta_{m,n}$ and $M \geq M_0$, where M_0 is the constant from Theorem D.11, yields

$$\log N\big(\varepsilon_{m,n}, \Xi(B, \mathcal{L}), \bar{h}_X\big) \leq \frac{m\varepsilon_{m,n}^2}{8}.$$

The same bound obviously controls the entropy of each shell

$$\big\{\eta : j\varepsilon_{m,n} < \bar{h}_X(\eta, \eta^\star) \leq 2j\varepsilon_{m,n}\big\}, \qquad j \geq 1,$$

since each shell is a subset of the full parameter space.

Prior mass of KL neighborhoods. By Theorem D.9, every $\eta \in \mathcal{U}_n(\varepsilon)$ satisfies

$$\frac{1}{m} \sum_{i=1}^{m} \mathrm{KL}\big(q_{\eta^\star}(\cdot \mid X_i) \,\|\, q_\eta(\cdot \mid X_i)\big) \leq C_{\mathrm{KL}}\, \varepsilon^2$$

on $\mathcal{G}_{m,n}(\delta)$. By Theorems D.10 and D.11, for $\varepsilon = \varepsilon_{m,n} = M\delta_{m,n}$ and M sufficiently large,

$$-\log \Pi\big(\mathcal{U}_n(\varepsilon_{m,n})\big) \leq \frac{m\varepsilon_{m,n}^2}{8}.$$

Hence the prior-thickness hypothesis of the general theorem is satisfied.

We have therefore verified all conditions of the main posterior-contraction theorem of Ghosal & van der Vaart (2007) for the semimetric $d_m = \bar{h}_X$ and the rate $\varepsilon_{m,n} = M\delta_{m,n}$. It follows that

$$\Pi_m\big(\eta : \bar{h}_X(\eta, \eta^\star) > M\delta_{m,n}\big) \to 0$$

in conditional $P_{\eta^\star}(\,\cdot \mid X_{1:m})$-probability, for every design sequence in $\mathcal{G}_{m,n}(\delta)$. Since $P_{\eta^\star}(\mathcal{G}_{m,n}(\delta)^c) \to 0$, the same conclusion holds unconditionally in $P_{\eta^\star}$-probability. □

Corollary E.2 (Predictive contraction for the full grouped-data posterior)**.** *Under the assumptions of Theorem E.1, the full-data posterior satisfies*

$$\Pi\big(\eta \in \Xi(B, \mathcal{L}) : \bar{h}_X(\eta, \eta^\star) > M\delta_{m,n} \,\big|\, X_{1:m}, Y_{1:m}\big) \longrightarrow 0 \tag{77}$$

in $P_{\eta^\star}$-probability for every $M \geq M_{\mathrm{pred}}$.

Proof. This is an immediate consequence of Theorem E.1 and the exact posterior-reduction identity Equation (31). □

E.3. The reduced predictive law of one task

The next lemma records the precise predictive object estimated by the metric $\bar{h}_X$.

Definition E.3 (Reduced one-task predictive law)**.** For $\eta \in \Xi(B, \mathcal{L})$, define the reduced one-task predictive law

$$\widetilde{Q}_\eta^{(n)}(dX, dW) := p_X(X)\, q_\eta(W \mid X)\, dX\, dW,$$

where p_X is the Gaussian design density and $q_\eta(\cdot \mid X)$ is the conditional density of $\widehat{W}$ given X.

Lemma E.4 (Exact Hellinger representation for the reduced predictive law)**.** *For every $\eta, \eta' \in \Xi(B, \mathcal{L})$,*

$$h^2\big(\widetilde{Q}_\eta^{(n)}, \widetilde{Q}_{\eta'}^{(n)}\big) = \mathbb{E}_X\big[h^2\big(q_\eta(\cdot \mid X), q_{\eta'}(\cdot \mid X)\big)\big]. \tag{78}$$

Consequently,

$$\mathbb{E}\big[\bar{h}_X^2(\eta, \eta')\big] = h^2\big(\widetilde{Q}_\eta^{(n)}, \widetilde{Q}_{\eta'}^{(n)}\big).$$

Proof. By definition,

$$\rho\big(\widetilde{Q}_\eta^{(n)}, \widetilde{Q}_{\eta'}^{(n)}\big) = \int p_X(x) \left[\int \sqrt{q_\eta(w \mid x)\, q_{\eta'}(w \mid x)}\, dw\right] dx = \mathbb{E}_X\big[\rho\big(q_\eta(\cdot \mid X), q_{\eta'}(\cdot \mid X)\big)\big].$$

Since $h^2 = 1 - \rho$, we obtain

$$h^2\big(\widetilde{Q}_\eta^{(n)}, \widetilde{Q}_{\eta'}^{(n)}\big) = 1 - \mathbb{E}_X[\rho(q_\eta, q_{\eta'})] = \mathbb{E}_X[1 - \rho(q_\eta, q_{\eta'})] = \mathbb{E}_X\big[h^2\big(q_\eta(\cdot \mid X), q_{\eta'}(\cdot \mid X)\big)\big].$$

The second claim follows because $X_1, \dots, X_m$ are i.i.d. □

Remark E.5 (Interpretation of the predictive theorem). Theorem E.1 is therefore a posterior contraction theorem for the reduced one-task predictive law $\widetilde{Q}_\eta^{(n)}$, measured through its empirical conditional Hellinger analogue $\bar{h}_X$. This is the natural predictive object after exact sufficient-statistic reduction.

E.4. Oracle latent-program recovery

We now quantify how quickly the latent task program can be recovered once the true task-family parameter is known. This is the cleanest posterior decoding statement available before the structural-localization theory of Appendix F.

For $i \in [m]$, $z \in \mathcal{L}$, and $\eta = (\pi, \Theta)$, define the component responsibility

$$\omega_{i,z}(\eta) := \Pi_\eta\Big(Z_i = z \mid X_i, \widehat{W}_i\Big) = \frac{\pi_z\, \varphi_{X_i}\big(\widehat{W}_i; A_z(\Theta)\big)}{\sum_{u \in \mathcal{L}} \pi_u\, \varphi_{X_i}\big(\widehat{W}_i; A_u(\Theta)\big)}, \tag{79}$$

where $\varphi_{X_i}(\cdot; A_z(\Theta))$ denotes the density of $\mathcal{MN}_{d,d}\big(A_z(\Theta), \sigma^2 I_d, (X_i X_i^\top)^{-1}\big)$.

Recall the true component-separation margin from Equation (53):

$$\gamma_{\mathrm{comp}}(\eta^\star) = \min_{z \neq z'} \|A_z^\star - A_{z'}^\star\|_F.$$

Proposition E.6 (Oracle latent-program recovery under component separation)**.** *Assume Theorem A.1 and suppose that*

$$\gamma_{\mathrm{comp}}(\eta^\star) > 0.$$

Fix $0 < \delta < 1$. Then on the joint good-design event $\mathcal{G}_{m,n}(\delta)$, for every $i \in [m]$,

$$\mathbb{E}_{\eta^\star}\big[1 - \omega_{i,Z_i^\star}(\eta^\star) \;\big|\; X_i, Z_i^\star\big] \le C_{\mathrm{lat}} \exp\left(-\frac{(1-\delta) n\, \gamma_{\mathrm{comp}}(\eta^\star)^2}{8\sigma^2}\right), \tag{80}$$

where one may take

$$C_{\mathrm{lat}} = \frac{|\mathcal{L}| - 1}{\sqrt{\pi_{\min}}}.$$

Consequently,

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\eta^\star}\big[1 - \omega_{i,Z_i^\star}(\eta^\star) \;\big|\; X_{1:m}, Z_{1:m}^\star\big] \le C_{\mathrm{lat}} \exp\left(-\frac{(1-\delta) n\, \gamma_{\mathrm{comp}}(\eta^\star)^2}{8\sigma^2}\right). \tag{81}$$

Proof. Fix a task index i, and condition on X_i and $Z_i^\star = z$. Under the true parameter $\eta^\star = (\pi^\star, \Theta^\star)$, the reduced observation $\widehat{W}_i$ has density

$$f_{z,i}^\star(w) = \varphi_{X_i}(w; A_z^\star).$$

For $u \neq z$, let $f_{u,i}^\star(w) := \varphi_{X_i}(w; A_u^\star)$. Then

$$1 - \omega_{i,z}(\eta^\star) = \sum_{u \neq z} \frac{\pi_u^\star f_{u,i}^\star(\widehat{W}_i)}{\sum_{v \in \mathcal{L}} \pi_v^\star f_{v,i}^\star(\widehat{W}_i)}.$$

For each $u \neq z$, set

$$t_u(\widehat{W}_i) := \frac{\pi_u^\star f_{u,i}^\star(\widehat{W}_i)}{\pi_z^\star f_{z,i}^\star(\widehat{W}_i)}.$$

Since $\sum_v \pi_v^\star f_{v,i}^\star \geq \pi_z^\star f_{z,i}^\star + \pi_u^\star f_{u,i}^\star$, we have

$$\frac{\pi_u^\star f_{u,i}^\star}{\sum_v \pi_v^\star f_{v,i}^\star} \leq \frac{t_u}{1 + t_u}.$$

Using the elementary inequality $t/(1+t) \leq \sqrt{t}$ for $t \geq 0$, we obtain

$$\frac{\pi_u^\star f_{u,i}^\star}{\sum_v \pi_v^\star f_{v,i}^\star} \leq \sqrt{\frac{\pi_u^\star}{\pi_z^\star}} \sqrt{\frac{f_{u,i}^\star}{f_{z,i}^\star}}.$$

Taking expectation under $f_{z,i}^\star$ yields

$$\mathbb{E}_{\eta^\star}\left[\frac{\pi_u^\star f_{u,i}^\star(\widehat{W}_i)}{\sum_v \pi_v^\star f_{v,i}^\star(\widehat{W}_i)} \,\middle|\, X_i, Z_i^\star = z \right] \leq \sqrt{\frac{\pi_u^\star}{\pi_z^\star}} \int \sqrt{f_{u,i}^\star(w) f_{z,i}^\star(w)}\, dw = \sqrt{\frac{\pi_u^\star}{\pi_z^\star}}\, \rho(f_{u,i}^\star, f_{z,i}^\star).$$

Summing over $u \neq z$ gives

$$\mathbb{E}_{\eta^\star}[1 - \omega_{i,z}(\eta^\star) \mid X_i, Z_i^\star = z] \leq \sum_{u \neq z} \sqrt{\frac{\pi_u^\star}{\pi_z^\star}}\, \rho(f_{u,i}^\star, f_{z,i}^\star). \tag{82}$$

Now on $\mathcal{G}_{m,n}(\delta)$, Equation (34) from Appendix B implies

$$\rho(f_{u,i}^\star, f_{z,i}^\star) = \exp\left(-\frac{1}{8\sigma^2} \|(A_u^\star - A_z^\star) X_i\|_F^2 \right) \leq \exp\left(-\frac{(1-\delta)n}{8\sigma^2} \|A_u^\star - A_z^\star\|_F^2 \right).$$

By the definition of $\gamma_{\mathrm{comp}}(\eta^\star)$,

$$\|A_u^\star - A_z^\star\|_F \geq \gamma_{\mathrm{comp}}(\eta^\star) \qquad \text{for every } u \neq z.$$

Hence

$$\rho(f_{u,i}^\star, f_{z,i}^\star) \leq \exp\left(-\frac{(1-\delta)n\, \gamma_{\mathrm{comp}}(\eta^\star)^2}{8\sigma^2} \right).$$

Also, $\pi_z^\star \geq \pi_{\min}$ and $\pi_u^\star \leq 1$, so

$$\sqrt{\frac{\pi_u^\star}{\pi_z^\star}} \leq \pi_{\min}^{-1/2}.$$

Substituting these bounds into Equation (82) gives

$$\mathbb{E}_{\eta^\star}[1 - \omega_{i,z}(\eta^\star) \mid X_i, Z_i^\star = z] \leq \frac{|\mathcal{L}| - 1}{\sqrt{\pi_{\min}}} \exp\left(-\frac{(1-\delta)n\, \gamma_{\mathrm{comp}}(\eta^\star)^2}{8\sigma^2} \right).$$

Since this bound is uniform in $z \in \mathcal{L}$, Equation (80) follows. Averaging over $i = 1, \ldots, m$ yields Equation (81). □

Remark E.7 (From oracle recovery to fully Bayesian recovery). Theorem E.6 quantifies the intrinsic within-task difficulty of decoding the latent program when the true task-family parameter is known. To pass from the oracle decoder $\omega_{i,z}(\eta^\star)$ to the full hierarchical posterior decoder

$$\Pi\Big(Z_i = z \mid X_{1:m}, \widehat{W}_{1:m}\Big) = \int \omega_{i,z}(\eta)\, \Pi_m(d\eta),$$

one needs posterior localization in a genuine parameter neighborhood of $\eta^\star$, not merely predictive contraction in $\bar{h}_X$. That structural localization is supplied later in Appendix F. Once combined with local continuity of $\eta \mapsto \omega_{i,z}(\eta)$, it yields the fully Bayesian latent-program recovery statement appearing in the main text.

F. Local Inverse Inequalities and Structural Contraction

Singular statistical models exhibit a sharp distinction between predictive accuracy and parameter localization. This phenomenon is central in singular learning theory and in the weak-identifiability analysis of finite mixtures (Watanabe, 2009; Ho & Nguyen, 2016; 2019). The purpose of this appendix is to isolate the corresponding geometry for stochastic task grammars.

The appendix has three parts. First, we introduce a finite family of deconvolved Fourier witnesses that convert predictive Hellinger control in the reduced experiment into control of a finite-dimensional feature map of the latent task-family parameter. Second, under a regularity condition, we prove a local inverse inequality of order 1 and derive regular structural contraction of the quotient parameter. Third, we construct an explicit duplicated-module family in which the first-order geometry cancels, compute its reduced divergence exactly, and show that the local inverse exponent becomes 2.

F.1. Finite deconvolved Fourier witnesses

Let

$$p := d^2.$$

For $\eta = (\pi, \Theta) \in \Xi(B, \mathcal{L})$, write

$$v_z(\Theta) := \mathrm{vec}(A_z(\Theta)) \in \mathbb{R}^p, \qquad \mu_\eta = \sum_{z \in \mathcal{L}} \pi_z\, \delta_{v_z(\Theta)}$$

for the induced mixing measure on operator space.

Definition F.1 (Characteristic map of the induced mixing measure)**.** For $t \in \mathbb{R}^p$, define

$$\psi_t(\eta) := \int e^{i\langle t, v\rangle}\, \mu_\eta(dv) = \sum_{z \in \mathcal{L}} \pi_z\, e^{i\langle t, v_z(\Theta)\rangle}. \tag{83}$$

For a finite set $T = \{t_1, \dots, t_J\} \subset \mathbb{R}^p$, define the finite witness map

$$\mathsf{T}_T(\eta) := \big(\psi_{t_1}(\eta), \dots, \psi_{t_J}(\eta)\big) \in \mathbb{C}^J. \tag{84}$$

Lemma F.2 (Quotient invariance of the witness map)**.** *For every finite $T \subset \mathbb{R}^p$, the map T_T is invariant under $\mathrm{Aut}(\mathcal{L})$. Equivalently,*

$$\mathsf{T}_T(\sigma \cdot \eta) = \mathsf{T}_T(\eta) \qquad \textit{for all } \eta \in \Xi(B, \mathcal{L}),\ \sigma \in \mathrm{Aut}(\mathcal{L}).$$

Proof. By Theorem A.4 and the definition of the induced mixing measure, the quotient action preserves μ_η. Since $\psi_t(\eta)$ is the characteristic function of μ_η evaluated at t, each coordinate of T_T is quotient-invariant. □

Fix deterministic designs $X_{1:m}$, and for each task define

$$\Sigma_i := (X_i X_i^\top)^{-1} \otimes \sigma^2 I_d \in \mathbb{R}^{p \times p}.$$

Lemma F.3 (Deconvolved Fourier identity). *Fix $t \in \mathbb{R}^p$. For each task $i \in [m]$, define the complex-valued function*

$$g_{t,i}(W) := \exp\left(i\langle t, \mathrm{vec}(W)\rangle + \frac{1}{2} t^\top \Sigma_i t \right), \qquad W \in \mathbb{R}^{d\times d}. \tag{85}$$

Then, for every $\eta \in \Xi(B, \mathcal{L})$,

$$\int g_{t,i}(W)\, q_\eta(W \mid X_i)\, dW = \psi_t(\eta). \tag{86}$$

Proof. By Theorem B.6 in Appendix B,

$$q_\eta(\cdot \mid X_i) = \sum_{z\in\mathcal{L}} \pi_z \; \mathcal{MN}_{d,d}\big(A_z(\Theta), \sigma^2 I_d, (X_i X_i^\top)^{-1}\big).$$

Fix $z \in \mathcal{L}$. If $W \sim \mathcal{MN}_{d,d}(A_z(\Theta), \sigma^2 I_d, (X_i X_i^\top)^{-1})$, then

$$\mathrm{vec}(W) \sim \mathcal{N}\big(v_z(\Theta), \Sigma_i\big).$$

Hence

$$\mathbb{E}\Big[e^{i\langle t, \mathrm{vec}(W)\rangle}\Big] = \exp\left(i\langle t, v_z(\Theta)\rangle - \frac{1}{2} t^\top \Sigma_i t \right).$$

Multiplying by $e^{\frac{1}{2} t^\top \Sigma_i t}$ gives

$$\mathbb{E}[g_{t,i}(W)] = e^{i\langle t, v_z(\Theta)\rangle}.$$

Averaging with respect to the mixing weights π_z proves Equation (86). □

Lemma F.4 (Witness stability under the predictive metric). *Fix $0 < \delta < 1$, a finite set $T = \{t_1, \dots, t_J\} \subset \mathbb{R}^p$, and the joint good-design event $\mathcal{G}_{m,n}(\delta)$. Then on $\mathcal{G}_{m,n}(\delta)$,*

$$\|\mathsf{T}_T(\eta) - \mathsf{T}_T(\eta')\|_2 \le C_T(\delta)\, \bar{h}_X(\eta, \eta') \qquad \textit{for all } \eta, \eta' \in \Xi(B, \mathcal{L}), \tag{87}$$

where one may take

$$C_T(\delta) := 2\sqrt{2J} \max_{1\le j\le J} \exp\left(\frac{\sigma^2 \|t_j\|_2^2}{2(1-\delta)(d+1)} \right). \tag{88}$$

Proof. Fix $t \in T$. By Theorem F.3,

$$\psi_t(\eta) - \psi_t(\eta') = \frac{1}{m} \sum_{i=1}^m \int g_{t,i}(W) \left(q_\eta(W \mid X_i) - q_{\eta'}(W \mid X_i) \right) dW.$$

Hence

$$|\psi_t(\eta) - \psi_t(\eta')| \le \frac{1}{m} \sum_{i=1}^m 2\|g_{t,i}\|_\infty \; \mathrm{TV}\big(q_\eta(\cdot \mid X_i), q_{\eta'}(\cdot \mid X_i)\big).$$

On $\mathcal{G}_{m,n}(\delta)$,

$$X_i X_i^\top \succeq (1-\delta) n I_d \succeq (1-\delta)(d+1) I_d,$$

so

$$\Sigma_i = (X_i X_i^\top)^{-1} \otimes \sigma^2 I_d \preceq \frac{\sigma^2}{(1-\delta)(d+1)} I_p.$$

Therefore

$$\|g_{t,i}\|_\infty \le \exp\left(\frac{1}{2} t^\top \Sigma_i t \right) \le \exp\left(\frac{\sigma^2 \|t\|_2^2}{2(1-\delta)(d+1)} \right) =: B_t(\delta).$$

Using $\mathrm{TV}(P, Q) \le \sqrt{2}\, h(P, Q)$ and Cauchy–Schwarz,

$$\frac{1}{m} \sum_{i=1}^m \mathrm{TV}(q_{\eta,i}, q_{\eta',i}) \le \sqrt{2}\, \frac{1}{m} \sum_{i=1}^m h(q_{\eta,i}, q_{\eta',i}) \le \sqrt{2}\, \bar{h}_X(\eta, \eta').$$

Hence

$$|\psi_t(\eta) - \psi_t(\eta')| \le 2\sqrt{2}\, B_t(\delta)\, \bar{h}_X(\eta, \eta').$$

Summing over $t_1, \dots, t_J$ in Euclidean norm gives Equation (87). □

F.2. Regular local inverse inequality

We now prove a local inverse theorem of order 1 under a finite-dimensional regularity condition.

Definition F.5 (Tangent space and tangent norm)**.** At the truth $\eta^\star = (\pi^\star, \Theta^\star)$, define the tangent space

$$\mathcal{T}_{\eta^\star} := \left\{ (\alpha, H) : \alpha = (\alpha_z)_{z\in\mathcal{L}} \in \mathbb{R}^{\mathcal{L}},\ \sum_{z\in\mathcal{L}} \alpha_z = 0,\ H = (H_1, \dots, H_r) \in (\mathbb{R}^{d\times d})^r \right\},$$

equipped with the norm

$$\|(\alpha, H)\|_{\tan} := \|\alpha\|_1 + \|H\|_{\mathrm{lib}}.$$

Lemma F.6 (Derivative of the characteristic map)**.** *Fix $t \in \mathbb{R}^p$. The map $\eta \mapsto \psi_t(\eta)$ is C^∞ on $\Xi(B, \mathcal{L})$. Its Fréchet derivative at $\eta^\star = (\pi^\star, \Theta^\star)$ in direction $u = (\alpha, H) \in \mathcal{T}_{\eta^\star}$ is*

$$D\psi_{t,\eta^\star}[u] = \sum_{z\in\mathcal{L}} \alpha_z\, e^{i\langle t, v_z^\star\rangle} + i \sum_{z\in\mathcal{L}} \pi_z^\star\, e^{i\langle t, v_z^\star\rangle} \big\langle t, Dv_{z,\Theta^\star}[H] \big\rangle, \tag{89}$$

where

$$v_z^\star := v_z(\Theta^\star), \qquad Dv_{z,\Theta^\star}[H] := \mathrm{vec}\big([D\Phi_{\Theta^\star}(H)]_z\big).$$

Proof. Smoothness follows because $\eta \mapsto v_z(\Theta)$ is polynomial for each $z \in \mathcal{L}$, and $\eta \mapsto \pi_z e^{i\langle t, v_z(\Theta)\rangle}$ is therefore smooth. Differentiating Equation (83) gives

$$D\psi_{t,\eta^\star}[u] = \sum_{z\in\mathcal{L}} \alpha_z e^{i\langle t, v_z^\star\rangle} + \sum_{z\in\mathcal{L}} \pi_z^\star\, i e^{i\langle t, v_z^\star\rangle} \big\langle t, Dv_{z,\Theta^\star}[H] \big\rangle,$$

which is Equation (89). □

Proposition F.7 (Injectivity of the linearized characteristic map)**.** *Assume that*

$$\gamma_{\mathrm{comp}}(\eta^\star) > 0 \qquad \textit{and} \qquad \lambda_{\mathrm{lin}}(\eta^\star) > 0. \tag{90}$$

Then the linear map

$$D\Psi_{\eta^\star} : \mathcal{T}_{\eta^\star} \to C(\mathbb{R}^p; \mathbb{C}), \qquad u \mapsto \big(t \mapsto D\psi_{t,\eta^\star}[u]\big)$$

is injective.

Proof. Fix $u = (\alpha, H) \in \mathcal{T}_{\eta^\star}$, and suppose that

$$D\psi_{t,\eta^\star}[u] = 0 \qquad \text{for all } t \in \mathbb{R}^p.$$

Define the compactly supported distribution

$$\nu_u := \sum_{z\in\mathcal{L}} \alpha_z\, \delta_{v_z^\star} - \sum_{z\in\mathcal{L}} \pi_z^\star \sum_{\ell=1}^{p} \big[Dv_{z,\Theta^\star}[H]\big]_\ell\, \partial_\ell \delta_{v_z^\star}. \tag{91}$$

With the Fourier-transform convention

$$\widehat{\delta_x}(t) = e^{i\langle t, x\rangle}, \qquad \widehat{\partial_\ell \delta_x}(t) = i t_\ell e^{i\langle t, x\rangle},$$

the identity Equation (89) shows that

$$\widehat{\nu_u}(t) = D\psi_{t,\eta^\star}[u].$$

Hence $\widehat{\nu_u} \equiv 0$, so $\nu_u = 0$ by injectivity of the Fourier transform on compactly supported distributions.

Because $\gamma_{\mathrm{comp}}(\eta^\star) > 0$, the support points $\{v_z^\star : z \in \mathcal{L}\}$ are pairwise distinct. Fix $z_0 \in \mathcal{L}$, and choose $\chi_{z_0} \in C_c^\infty(\mathbb{R}^p)$ such that

$$\chi_{z_0}(v_{z_0}^\star) = 1, \qquad \nabla\chi_{z_0}(v_{z_0}^\star) = 0, \qquad \chi_{z_0} \equiv 0 \quad \text{in a neighborhood of } v_z^\star \text{ for all } z \neq z_0.$$

Pairing ν_u with χ_{z_0} gives

$$0 = \langle \nu_u, \chi_{z_0} \rangle = \alpha_{z_0}.$$

Since z_0 was arbitrary, $\alpha_z = 0$ for all $z \in \mathcal{L}$.

Next, for each $\ell \in \{1, \dots, p\}$, define

$$\chi_{z_0,\ell}(x) := (x_\ell - v^\star_{z_0,\ell})\chi_{z_0}(x).$$

Then

$$\chi_{z_0,\ell}(v^\star_{z_0}) = 0, \qquad \nabla\chi_{z_0,\ell}(v^\star_{z_0}) = e_\ell.$$

Since $\nu_u = 0$ and $\alpha = 0$,

$$0 = \langle \nu_u, \chi_{z_0,\ell} \rangle = -\pi^\star_{z_0} \big[Dv_{z_0,\Theta^\star}[H] \big]_\ell.$$

Because $\pi^\star_{z_0} > 0$, we obtain

$$Dv_{z_0,\Theta^\star}[H] = 0.$$

As z_0 was arbitrary,

$$D\Phi_{\Theta^\star}(H) = 0.$$

Finally, $\lambda_{\mathrm{lin}}(\eta^\star) > 0$ implies

$$0 = \|D\Phi_{\Theta^\star}(H)\|^2_{\pi^\star} \geq \lambda_{\mathrm{lin}}(\eta^\star)\|H\|^2_{\mathrm{lib}},$$

hence $H = 0$. Therefore $u = (\alpha, H) = 0$, proving injectivity. □

The next proposition converts the infinite-dimensional injective witness $t \mapsto \psi_t(\eta)$ into a finite-dimensional one.

Proposition F.8 (Finite witness selection)**.** *Under Equation (90), there exist finitely many frequencies*

$$T_{\mathrm{reg}} = \{t_1, \dots, t_J\} \subset \mathbb{R}^p$$

and a constant $c_{\mathrm{freq}} > 0$ *such that*

$$\big\| D\mathsf{T}_{T_{\mathrm{reg}},\eta^\star}[u] \big\|_2 \geq c_{\mathrm{freq}} \, \|u\|_{\tan} \qquad \textit{for every } u \in \mathcal{T}_{\eta^\star}. \tag{92}$$

Proof. Consider the unit sphere

$$\mathbb{S}_{\tan} := \{u \in \mathcal{T}_{\eta^\star} : \|u\|_{\tan} = 1\},$$

which is compact because $\mathcal{T}_{\eta^\star}$ is finite-dimensional. By Theorem F.7, for each $u \in \mathbb{S}_{\tan}$ there exists $t(u) \in \mathbb{R}^p$ such that

$$|D\psi_{t(u),\eta^\star}[u]| > 0.$$

By continuity of $v \mapsto D\psi_{t(u),\eta^\star}[v]$, there exists an open neighborhood U_u of u in $\mathbb{S}_{\tan}$ such that

$$|D\psi_{t(u),\eta^\star}[v]| \geq \frac{1}{2}|D\psi_{t(u),\eta^\star}[u]| \qquad \text{for all } v \in U_u.$$

The family $\{U_u : u \in \mathbb{S}_{\tan}\}$ is an open cover of $\mathbb{S}_{\tan}$, so compactness yields a finite subcover $U_{u_1}, \dots, U_{u_J}$. Set

$$t_j := t(u_j), \qquad T_{\mathrm{reg}} := \{t_1, \dots, t_J\}.$$

Define

$$c_{\mathrm{freq}} := \min_{1 \leq j \leq J} \frac{1}{2}|D\psi_{t_j,\eta^\star}[u_j]| > 0.$$

Then for every $u \in \mathbb{S}_{\tan}$, at least one j satisfies $u \in U_{u_j}$, hence

$$|D\psi_{t_j,\eta^\star}[u]| \geq c_{\mathrm{freq}}.$$

Therefore

$$\big\| D\mathsf{T}_{T_{\mathrm{reg}},\eta^\star}[u] \big\|_2 \geq \max_{1 \leq j \leq J} |D\psi_{t_j,\eta^\star}[u]| \geq c_{\mathrm{freq}} \qquad \text{for all } u \in \mathbb{S}_{\tan}.$$

By homogeneity, Equation (92) follows for all $u \in \mathcal{T}_{\eta^\star}$. □

Theorem F.9 (Regular local inverse inequality)**.** *Assume Equation (90), and let T_{reg} be the frequency set from Theorem F.8. Fix $0 < \delta < 1$. Then there exist constants*

$$\rho_{\mathrm{reg}} > 0, \qquad c_{\mathrm{reg}} > 0, \qquad C_{\mathrm{reg}}(\delta) > 0$$

such that on $\mathcal{G}_{m,n}(\delta)$, for every $\eta \in \Xi(B, \mathcal{L})$ with

$$d_q(\eta, \eta^\star) \le \rho_{\mathrm{reg}},$$

one has

$$c_{\mathrm{reg}}\, d_q(\eta, \eta^\star) \le \left\| \mathsf{T}_{T_{\mathrm{reg}}}(\eta) - \mathsf{T}_{T_{\mathrm{reg}}}(\eta^\star) \right\|_2 \le C_{\mathrm{reg}}(\delta)\, \bar{h}_X(\eta, \eta^\star), \tag{93}$$

and consequently

$$\bar{h}_X^2(\eta, \eta^\star) \ge c_{\mathrm{inv}}(\delta)\, d_q(\eta, \eta^\star)^2 \qquad \text{for all } d_q(\eta, \eta^\star) \le \rho_{\mathrm{reg}}, \tag{94}$$

where $c_{\mathrm{inv}}(\delta) := c_{\mathrm{reg}}^2 / C_{\mathrm{reg}}(\delta)^2$.

Proof. The upper bound in Equation (93) is exactly Equation (87) applied to $T = T_{\mathrm{reg}}$.

For the lower bound, write $T := T_{\mathrm{reg}}$. Since each coordinate of T is C^∞, and $\eta^\star$ lies in the interior of the simplex, there exists a neighborhood $\mathcal{N}^\star \subset \Xi(B, \mathcal{L})$ of $\eta^\star$ and a constant $C_{\mathrm{rem}} < \infty$ such that

$$\|T(\eta^\star + u) - T(\eta^\star) - DT_{\eta^\star}[u]\|_2 \le C_{\mathrm{rem}} \|u\|_{\tan}^2 \tag{95}$$

whenever $\eta^\star + u \in \mathcal{N}^\star$.

Fix η with $d_q(\eta, \eta^\star)$ sufficiently small. By definition of the quotient metric, there exists $\sigma \in \mathrm{Aut}(\mathcal{L})$ such that

$$d_q(\eta, \eta^\star) = d_{\mathrm{par}}(\sigma \cdot \eta, \eta^\star).$$

Set

$$\bar{\eta} := \sigma \cdot \eta, \qquad u := \bar{\eta} - \eta^\star \in \mathcal{T}_{\eta^\star}.$$

Then

$$\|u\|_{\tan} = d_{\mathrm{par}}(\bar{\eta}, \eta^\star) = d_q(\eta, \eta^\star),$$

and by quotient invariance of T,

$$T(\eta) = T(\bar{\eta}).$$

Using Theorem F.8 and Equation (95),

$$\|T(\eta) - T(\eta^\star)\|_2 = \|T(\bar{\eta}) - T(\eta^\star)\|_2 \ge \|DT_{\eta^\star}[u]\|_2 - C_{\mathrm{rem}} \|u\|_{\tan}^2 \ge c_{\mathrm{freq}} \|u\|_{\tan} - C_{\mathrm{rem}} \|u\|_{\tan}^2.$$

Choose

$$\rho_{\mathrm{reg}} := \min \left\{ \mathrm{dist}(\eta^\star, \Xi(B, \mathcal{L}) \setminus \mathcal{N}^\star), \frac{c_{\mathrm{freq}}}{2 C_{\mathrm{rem}}} \right\}, \qquad c_{\mathrm{reg}} := \frac{c_{\mathrm{freq}}}{2}.$$

Then for $d_q(\eta, \eta^\star) = \|u\|_{\tan} \le \rho_{\mathrm{reg}}$,

$$\|T(\eta) - T(\eta^\star)\|_2 \ge \frac{c_{\mathrm{freq}}}{2} \|u\|_{\tan} = c_{\mathrm{reg}}\, d_q(\eta, \eta^\star).$$

This proves the lower bound in Equation (93). Combining it with the upper bound gives Equation (94). □

Remark F.10 (Regular order-1 geometry). Theorem F.9 is the stochastic-task-grammar analogue of strong identifiability in finite mixtures: the predictive geometry is locally Euclidean in the quotient parameter, hence the inverse exponent is 1. This is precisely the regular regime in the sense of singular learning theory and mixture geometry (Watanabe, 2009; Ho & Nguyen, 2016; 2019).

F.3. Structural posterior contraction under a local inverse inequality

We now isolate the exact abstract implication needed later: predictive contraction plus a local inverse inequality yields structural contraction.

Theorem F.11 (Structural contraction from predictive contraction)**.** *Fix $0 < \delta < 1$, and suppose that Equation (75) holds. Assume that there exist constants*

$$\kappa \geq 1, \qquad \rho_\star > 0, \qquad c_\star > 0, \qquad \underline{h}_\star > 0$$

such that on $\mathcal{G}_{m,n}(\delta)$, for all sufficiently large m and every $\eta \in \Xi(B, \mathcal{L})$,

$$d_q(\eta, \eta^\star) \leq \rho_\star \implies \bar{h}_X^2(\eta, \eta^\star) \geq c_\star d_q(\eta, \eta^\star)^{2\kappa}, \tag{96}$$

$$d_q(\eta, \eta^\star) \geq \rho_\star \implies \bar{h}_X(\eta, \eta^\star) \geq \underline{h}_\star. \tag{97}$$

Then there exists $M_{\mathrm{str}} \in (0, \infty)$ such that for every $M \geq M_{\mathrm{str}}$,

$$\Pi_m\left(\eta : d_q(\eta, \eta^\star) > M\delta_{m,n}^{1/\kappa}\right) \longrightarrow 0 \tag{98}$$

in $P_{\eta^\star}$-probability.

Proof. Fix $M > 0$, and let

$$A_m(M) := \left\{\eta : d_q(\eta, \eta^\star) > M\delta_{m,n}^{1/\kappa}\right\}.$$

Write

$$A_m(M) = A_m^{\mathrm{loc}}(M) \cup A_m^{\mathrm{far}}(M),$$

where

$$A_m^{\mathrm{loc}}(M) := A_m(M) \cap \{d_q(\eta, \eta^\star) \leq \rho_\star\}, \qquad A_m^{\mathrm{far}}(M) := A_m(M) \cap \{d_q(\eta, \eta^\star) > \rho_\star\}.$$

Because $\delta_{m,n} \to 0$, for every fixed M we have $M\delta_{m,n}^{1/\kappa} < \rho_\star$ eventually, so $A_m^{\mathrm{loc}}(M)$ is nonempty only inside the local inverse regime. On $\mathcal{G}_{m,n}(\delta)$, every $\eta \in A_m^{\mathrm{loc}}(M)$ satisfies by Equation (96)

$$\bar{h}_X^2(\eta, \eta^\star) \geq c_\star M^{2\kappa}\delta_{m,n}^2.$$

Hence

$$A_m^{\mathrm{loc}}(M) \subseteq \left\{\eta : \bar{h}_X(\eta, \eta^\star) \geq \sqrt{c_\star}\, M^\kappa \delta_{m,n}\right\}.$$

Likewise, by Equation (97),

$$A_m^{\mathrm{far}}(M) \subseteq \left\{\eta : \bar{h}_X(\eta, \eta^\star) \geq \underline{h}_\star\right\}.$$

Therefore on $\mathcal{G}_{m,n}(\delta)$,

$$\Pi_m(A_m(M)) \leq \Pi_m\left(\bar{h}_X(\eta, \eta^\star) \geq \sqrt{c_\star}\, M^\kappa \delta_{m,n}\right) + \Pi_m\left(\bar{h}_X(\eta, \eta^\star) \geq \underline{h}_\star\right).$$

Choose M_{str} so large that

$$\sqrt{c_\star}\, M^\kappa \geq M_{\mathrm{pred}} \qquad \text{for all } M \geq M_{\mathrm{str}},$$

where M_{pred} is the predictive-contraction constant from Theorem E.1. Then the first posterior term above converges to 0 by Theorem E.1, while the second also converges to 0 because $\underline{h}_\star$ is a fixed positive number and $\delta_{m,n} \to 0$. Since $P_{\eta^\star}(\mathcal{G}_{m,n}(\delta)^c) \to 0$, the same conclusion holds unconditionally in $P_{\eta^\star}$-probability. □

Corollary F.12 (Regular structural contraction)**.** *Assume Equation (90), the design regime Equation (75), and the empirical separation condition*

$$\inf_{\eta:\, d_q(\eta,\eta^\star) \geq \rho_{\mathrm{reg}}} \bar{h}_X(\eta, \eta^\star) \geq \underline{h}_{\mathrm{reg}} \qquad \text{on } \mathcal{G}_{m,n}(\delta) \tag{99}$$

for some $\underline{h}_{\mathrm{reg}} > 0$ and all sufficiently large m, where ρ_{reg} is the radius from Theorem F.9. Then there exists $M_{\mathrm{reg}} \in (0, \infty)$ such that for every $M \geq M_{\mathrm{reg}}$,

$$\Pi_m(\eta : d_q(\eta, \eta^\star) > M\delta_{m,n}) \longrightarrow 0 \tag{100}$$

in $P_{\eta^\star}$-probability.

Proof. Apply Theorem F.11 with $\kappa = 1$, local inverse Equation (94), and global separation Equation (99). □

Remark F.13 (On the global separation condition). The local inverse inequality is the genuinely nontrivial part of the structural theory. The separation condition Equation (99) is a compactness / identifiability requirement that prevents predictive near-collisions away from the truth. It is automatic on compact quotient classes on which the reduced predictive map is globally identifiable.

F.4. An explicit quadratic singular family

We now exhibit a concrete family in which the first-order geometry vanishes and the inverse exponent becomes 2.

Definition F.14 (Duplicated-module singular family)**.** Let

$$r = 2, \qquad \mathcal{L}_{\mathrm{dup}} := \{(1), (2)\}, \qquad \pi^\star = \left(\frac{1}{2}, \frac{1}{2}\right).$$

Fix $A^\star \in \mathbb{R}^{d\times d}$ with $\|A^\star\|_{\mathrm{op}} < B$, and define the singular center

$$\eta_0 := (\pi^\star, (A^\star, A^\star)).$$

For any $\Delta \in \mathbb{R}^{d\times d}$ such that $\|A^\star \pm \Delta\|_{\mathrm{op}} \le B$, define

$$\eta_\Delta := (\pi^\star,\ (A^\star + \Delta,\ A^\star - \Delta)). \tag{101}$$

Lemma F.15 (Quotient distance in the duplicated family)**.** *For every Δ in the duplicated family,*

$$d_q(\eta_\Delta, \eta_0) = 2\|\Delta\|_F. \tag{102}$$

Proof. The automorphism group of $\mathcal{L}_{\mathrm{dup}}$ consists of the identity and the transposition $1 \leftrightarrow 2$. Under either alignment, the weight part contributes 0, while the module part contributes

$$\|A^\star + \Delta - A^\star\|_F + \|A^\star - \Delta - A^\star\|_F = 2\|\Delta\|_F.$$

Hence the minimum in the quotient metric equals $2\|\Delta\|_F$. □

Fix a task i, and write

$$\Sigma_i := (X_i X_i^\top)^{-1} \otimes \sigma^2 I_d, \qquad \mu^\star := \mathrm{vec}(A^\star), \qquad s := \mathrm{vec}(\Delta).$$

Conditional on X_i, the reduced law under η_0 is

$$q_{0,i} = \mathcal{N}(\mu^\star, \Sigma_i),$$

while under η_Δ it is the symmetric two-point Gaussian mixture

$$q_{\Delta,i} = \frac{1}{2}\mathcal{N}(\mu^\star + s, \Sigma_i) + \frac{1}{2}\mathcal{N}(\mu^\star - s, \Sigma_i).$$

Lemma F.16 (Exact χ^2-divergence in the duplicated family)**.** *For every task $i \in [m]$,*

$$r_{i,\Delta}^2 := s^\top \Sigma_i^{-1} s = \frac{1}{\sigma^2}\,\mathrm{tr}\big(\Delta X_i X_i^\top \Delta^\top\big). \tag{103}$$

Then

$$\chi^2(q_{\Delta,i}, q_{0,i}) = \cosh\big(r_{i,\Delta}^2\big) - 1. \tag{104}$$

Proof. Let p_0 and p_Δ be the densities of $q_{0,i}$ and $q_{\Delta,i}$. Write

$$u := \Sigma_i^{-1/2}(w - \mu^\star), \qquad a := \Sigma_i^{-1/2} s, \qquad \|a\|_2^2 = r_{i,\Delta}^2.$$

A direct Gaussian calculation gives the likelihood ratio

$$\frac{p_\Delta(w)}{p_0(w)} = \frac{1}{2}\exp\left(a^\top u - \frac{1}{2}\|a\|_2^2\right) + \frac{1}{2}\exp\left(-a^\top u - \frac{1}{2}\|a\|_2^2\right) = e^{-\|a\|_2^2/2}\cosh(a^\top u).$$

Therefore

$$1 + \chi^2(q_{\Delta,i}, q_{0,i}) = \int \frac{p_\Delta(w)^2}{p_0(w)}\,dw = \mathbb{E}_{U\sim\mathcal{N}(0,I_p)}\left[e^{-\|a\|_2^2}\cosh^2(a^\top U)\right].$$

Since $a^\top U \sim \mathcal{N}(0, \|a\|_2^2)$ and

$$\cosh^2 x = \frac{\cosh(2x)+1}{2},$$

we obtain

$$1 + \chi^2(q_{\Delta,i}, q_{0,i}) = e^{-r_{i,\Delta}^2}\frac{\mathbb{E}[\cosh(2T)]+1}{2}, \qquad T \sim \mathcal{N}(0, r_{i,\Delta}^2).$$

Because $\mathbb{E}[e^{2T}] = e^{2r_{i,\Delta}^2}$, we have $\mathbb{E}[\cosh(2T)] = e^{2r_{i,\Delta}^2}$. Thus

$$1 + \chi^2(q_{\Delta,i}, q_{0,i}) = \frac{e^{r_{i,\Delta}^2} + e^{-r_{i,\Delta}^2}}{2} = \cosh(r_{i,\Delta}^2),$$

which proves Equation (104). □

Proposition F.17 (Quadratic flattening in the duplicated family)**.** *Fix $0 < \delta < 1$. There exist constants*

$$\rho_{\mathrm{sing}}(\delta) > 0, \qquad 0 < c_{\mathrm{sing}}(\delta) \le C_{\mathrm{sing}}(\delta) < \infty$$

such that on $\mathcal{G}_{m,n}(\delta)$, for every Δ satisfying

$$\sqrt{n}\,\|\Delta\|_F \le \rho_{\mathrm{sing}}(\delta),$$

one has

$$c_{\mathrm{sing}}(\delta)\,n^2\,d_q(\eta_\Delta, \eta_0)^4 \le \bar{h}_X^2(\eta_\Delta, \eta_0) \le C_{\mathrm{sing}}(\delta)\,n^2\,d_q(\eta_\Delta, \eta_0)^4. \tag{105}$$

In particular, the duplicated family has local inverse exponent 2.

Proof. Fix $i \in [m]$. On $\mathcal{G}_{m,n}(\delta)$,

$$(1-\delta)nI_d \preceq X_i X_i^\top \preceq (1+\delta)nI_d,$$

hence by Equation (103),

$$\frac{(1-\delta)n}{\sigma^2}\,\|\Delta\|_F^2 \le r_{i,\Delta}^2 \le \frac{(1+\delta)n}{\sigma^2}\,\|\Delta\|_F^2. \tag{106}$$

Choose

$$\rho_{\mathrm{sing}}(\delta) := \frac{\sigma}{\sqrt{1+\delta}},$$

so that $\sqrt{n}\,\|\Delta\|_F \le \rho_{\mathrm{sing}}(\delta)$ implies $r_{i,\Delta}^2 \le 1$ for all i.

Upper bound. For any pair of dominated laws, $h^2(P,Q) \le \chi^2(P,Q)/2$. Thus by Theorem F.16,

$$h^2(q_{\Delta,i}, q_{0,i}) \le \frac{1}{2}\big(\cosh(r_{i,\Delta}^2) - 1\big).$$

Since $r_{i,\Delta}^2 \le 1$, Taylor's theorem gives

$$\cosh(u) - 1 \le \frac{e}{2}u^2 \qquad \text{for } u \in [0,1].$$

Therefore

$$h^2(q_{\Delta,i}, q_{0,i}) \le \frac{e}{4}\,r_{i,\Delta}^4 \le \frac{e(1+\delta)^2}{4\sigma^4}\,n^2\|\Delta\|_F^4.$$

Lower bound. Let

$$a_i := \Sigma_i^{-1/2} s \in \mathbb{R}^p, \qquad r_i := \|a_i\|_2 = r_{i,\Delta}.$$

Under $q_{0,i}$, the whitened variable

$$U := \Sigma_i^{-1/2}\big(\mathrm{vec}(\widehat{W}_i) - \mu^\star\big)$$

has law $\mathcal{N}(0, I_p)$. Under $q_{\Delta,i}$, U has the symmetric mixture

$$\frac{1}{2}\mathcal{N}(a_i, I_p) + \frac{1}{2}\mathcal{N}(-a_i, I_p).$$

Define the quadratic witness

$$f_i(u) := (a_i^\top u)^2 - r_i^2.$$

Under $q_{0,i}$, $a_i^\top U \sim \mathcal{N}(0, r_i^2)$, so

$$\mathbb{E}_{q_{0,i}}[f_i(U)] = 0, \qquad \mathbb{E}_{q_{0,i}}[f_i(U)^2] = 2r_i^4.$$

Under $q_{\Delta,i}$, a direct Gaussian-moment computation yields

$$\mathbb{E}_{q_{\Delta,i}}[f_i(U)] = r_i^4, \qquad \mathbb{E}_{q_{\Delta,i}}[f_i(U)^2] = 2r_i^4 + 4r_i^6 + r_i^8 \le 7r_i^4,$$

since $r_i^2 \le 1$.

For any square-integrable f and any two dominated laws P, Q,

$$|E_P f - E_Q f| = \left|\int f(\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})\right| \le 2\sqrt{E_P f^2 + E_Q f^2}\, h(P, Q).$$

Applying this with $P = q_{\Delta,i}$, $Q = q_{0,i}$, and $f = f_i$, we obtain

$$r_i^4 = \big|\mathbb{E}_{q_{\Delta,i}}[f_i] - \mathbb{E}_{q_{0,i}}[f_i]\big| \le 2\sqrt{9r_i^4}\, h(q_{\Delta,i}, q_{0,i}) = 6r_i^2\, h(q_{\Delta,i}, q_{0,i}).$$

Hence

$$h^2(q_{\Delta,i}, q_{0,i}) \ge \frac{r_i^4}{36}.$$

Using the lower side of Equation (106),

$$h^2(q_{\Delta,i}, q_{0,i}) \ge \frac{(1-\delta)^2}{36\sigma^4}\, n^2 \|\Delta\|_F^4.$$

Averaging over $i = 1, \dots, m$ and recalling Equation (102),

$$\bar{h}_X^2(\eta_\Delta, \eta_0) = \frac{1}{m}\sum_{i=1}^m h^2(q_{\Delta,i}, q_{0,i}) \asymp n^2 \|\Delta\|_F^4 \asymp n^2 d_q(\eta_\Delta, \eta_0)^4,$$

with constants depending only on (δ, σ). This proves Equation (105). □

Corollary F.18 (Quadratic singular structural contraction on the duplicated family)**.** *Assume that the prior is supported on the duplicated family $\{\eta_\Delta : \sqrt{n}\,\|\Delta\|_F \le \rho_{\mathrm{sing}}(\delta)\}$, and that its induced density on Δ is continuous and strictly positive near $\Delta = 0$. Then there exists $M_{\mathrm{dup}} \in (0, \infty)$ such that for every $M \ge M_{\mathrm{dup}}$,*

$$\Pi_m\left(\eta_\Delta : d_q(\eta_\Delta, \eta_0) > M\,\frac{\delta_{m,n}^{1/2}}{\sqrt{n}}\right) \longrightarrow 0 \tag{107}$$

in P_{η_0}-probability.

Proof. On the support of the prior, Theorem F.17 supplies the local inverse inequality

$$\bar{h}_X^2(\eta_\Delta, \eta_0) \geq c_{\mathrm{sing}}(\delta)\, n^2\, d_q(\eta_\Delta, \eta_0)^4.$$

Hence, if

$$d_q(\eta_\Delta, \eta_0) > M\, \frac{\delta_{m,n}^{1/2}}{\sqrt{n}},$$

then

$$\bar{h}_X(\eta_\Delta, \eta_0) > \sqrt{c_{\mathrm{sing}}(\delta)}\, M^2\, \delta_{m,n}.$$

Choosing M_{dup} so that $\sqrt{c_{\mathrm{sing}}(\delta)}\, M^2 \geq M_{\mathrm{pred}}$ and applying Theorem E.1 proves Equation (107). □

Remark F.19 (Why the duplicated family matters). The duplicated family is the exact stochastic-task-grammar analogue of an overfitted symmetric finite mixture: the first-order perturbation cancels, the predictive law only changes at second order, and the structural inverse exponent becomes 2. The mechanism is classical in singular mixture geometry (Ho & Nguyen, 2016; 2019; Watanabe, 2009), but the calculation above is specific to the reduced grouped experiment induced by compositional task grammars.

G. Minimax Lower Bounds

This appendix proves four lower-bound statements. First, we establish a grammar-law lower bound of order $\sqrt{K_{\mathrm{eff}}/m}$ by reducing to a multinomial subproblem. Second, under a natural anchored-language condition, we prove a module-estimation lower bound of order $d\sqrt{s/(mn)}$ for s anchor modules. Third, we show that zero contextual observability induces an exact non-identifiability floor. Fourth, on the duplicated-module singular family of Appendix F, we derive a sharp two-point lower bound of order $m^{-1/4}n^{-1/2}$, matching the singular exponent $\kappa = 2$ up to logarithmic factors.

The lower bounds are stated for the full grouped-data experiment $(X_i, Y_i)_{i=1}^m$. Whenever convenient, we pass to a more informative oracle experiment or to the sufficient-statistic reduction of Appendix B; this is legitimate because lower bounds are monotone under Blackwell domination.

G.1. Decision-theoretic setup and generic comparison lemmas

For a parameter class $\mathcal{H} \subseteq \Xi(B, \mathcal{L})$, define the minimax structural risk

$$\mathfrak{R}_{m,n}(\mathcal{H}; d_q) := \inf_{\widehat{\eta}} \sup_{\eta \in \mathcal{H}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)]\,, \tag{108}$$

where the infimum is taken over all measurable estimators $\widehat{\eta} = \widehat{\eta}(X_{1:m}, Y_{1:m})$ with values in $\Xi(B, \mathcal{L})$.

Lemma G.1 (More informative experiments can only decrease minimax risk)**.** *Let $\mathcal{E}_1 = \{P_\eta : \eta \in \mathcal{H}\}$ and $\mathcal{E}_2 = \{Q_\eta : \eta \in \mathcal{H}\}$ be two statistical experiments on measurable spaces $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{X}, \mathcal{X})$, respectively. Suppose that there exists a Markov kernel $K(dy \mid x)$, independent of η, such that*

$$P_\eta(A) = \int_{\mathsf{X}} K(A \mid x)\, Q_\eta(dx) \qquad \text{for every } A \in \mathcal{Y},\ \eta \in \mathcal{H}.$$

Then for every loss $L : \mathcal{H} \times \Xi(B, \mathcal{L}) \to [0, \infty)$,

$$\inf_{\widehat{\eta}_1} \sup_{\eta \in \mathcal{H}} \mathbb{E}_{P_\eta}[L(\eta, \widehat{\eta}_1)] \geq \inf_{\widehat{\eta}_2} \sup_{\eta \in \mathcal{H}} \mathbb{E}_{Q_\eta}[L(\eta, \widehat{\eta}_2)]\,.$$

Proof. Fix any estimator $\widehat{\eta}_1 : \mathsf{Y} \to \Xi(B, \mathcal{L})$ for the less informative experiment $\mathcal{E}_1$. Define the randomized estimator on $\mathcal{E}_2$

$$\widehat{\eta}_2(x) \sim K(dy \mid x) \ \text{ followed by } \ \widehat{\eta}_1(y).$$

Then for every $\eta \in \mathcal{H}$,

$$\mathbb{E}_{Q_\eta}[L(\eta, \widehat{\eta}_2)] = \mathbb{E}_{P_\eta}[L(\eta, \widehat{\eta}_1)]\,.$$

Taking the supremum over η, then the infimum over $\widehat{\eta}_1$, proves the claim. □

Corollary G.2 (Two useful monotonicity reductions)**.** *For every parameter class $\mathcal{H} \subseteq \Xi(B, \mathcal{L})$:*

1. *the oracle experiment that augments the full grouped data with the latent programs $Z_{1:m}$ is more informative than the original grouped-data experiment, hence any minimax lower bound proved in the oracle experiment is valid for the original problem;*

2. *the full grouped-data experiment and the reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$ are Blackwell equivalent, because Appendix B shows that*

$$Y_i = \widehat{W}_i X_i + R_i,$$

where, conditional on $(X_i, \widehat{W}_i)$, the residual R_i has a law that does not depend on η.

Proof. Item (1) is immediate from Theorem G.1: forgetting the latent labels $Z_{1:m}$ maps the oracle experiment to the original one.

For item (2), the map $(X_i, Y_i) \mapsto (X_i, \widehat{W}_i)$ is measurable, so the reduced experiment is less informative than the full one. Conversely, by Theorems B.5 and B.6 in Appendix B, conditional on $(X_i, \widehat{W}_i)$, the residual R_i is independent of η with parameter-free law

$$R_i \mid (X_i, \widehat{W}_i) \sim \mathcal{MN}_{d,n}\big(0, \sigma^2 I_d, I_n - P_{X_i}\big).$$

Drawing R_i from this law and setting $Y_i = \widehat{W}_i X_i + R_i$ reconstructs the full conditional law of Y_i. Hence the full and reduced experiments dominate one another. □

We also use the following classical decision-theoretic tools.

Lemma G.3 (Le Cam's two-point method)**.** *Let $\eta_0, \eta_1 \in \mathcal{H}$, and let P_0, P_1 denote the corresponding laws of the observed experiment. Then*

$$\inf_{\widehat{\eta}} \sup_{\eta \in \{\eta_0, \eta_1\}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)] \geq \frac{d_q(\eta_0, \eta_1)}{4}\Big(1 - \mathrm{TV}(P_0, P_1)\Big). \tag{109}$$

Moreover,

$$\mathrm{TV}(P_0, P_1) \leq \sqrt{\frac{1}{2}\chi^2(P_1, P_0)}.$$

Proof. Let

$$r := \frac{1}{2} d_q(\eta_0, \eta_1), \qquad A := \{\widehat{\eta} : d_q(\widehat{\eta}, \eta_0) < r\}.$$

By the triangle inequality, on A we have $d_q(\widehat{\eta}, \eta_1) \geq r$, while on A^c we have $d_q(\widehat{\eta}, \eta_0) \geq r$. Therefore

$$\mathbb{E}_{\eta_0}[d_q(\widehat{\eta}, \eta_0)] + \mathbb{E}_{\eta_1}[d_q(\widehat{\eta}, \eta_1)] \geq r\Big(P_0(A^c) + P_1(A)\Big).$$

Taking the maximum of the two expectations yields

$$\sup_{\eta \in \{\eta_0, \eta_1\}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)] \geq \frac{r}{2}\Big(P_0(A^c) + P_1(A)\Big).$$

Finally,

$$\inf_A \big(P_0(A^c) + P_1(A)\big) = 1 - \mathrm{TV}(P_0, P_1),$$

which gives Equation (109). The χ^2-to-TV bound is standard:

$$\mathrm{TV}(P_0, P_1) = \frac{1}{2}\int |p_1 - p_0| \leq \frac{1}{2}\left(\int \frac{(p_1 - p_0)^2}{p_0}\right)^{1/2} = \sqrt{\frac{1}{2}\chi^2(P_1, P_0)}.$$

□

Lemma G.4 (Fano's inequality; e.g., Tsybakov (2009, Theorem 2.5))**.** *Let $\eta^0, \eta^1, \dots, \eta^M \in \mathcal{H}$, with $M \geq 2$, and suppose that*

$$d_q(\eta^j, \eta^k) \geq 2s \qquad \text{for all } 0 \leq j < k \leq M.$$

If P_j denotes the law of the observed experiment under η^j, then every estimator $\widehat{\eta}$ satisfies

$$\sup_{0 \leq j \leq M} \mathbb{E}_{\eta^j}\left[d_q(\widehat{\eta}, \eta^j)\right] \geq s\left(1 - \frac{\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_j, P_0) + \log 2}{\log M}\right). \tag{110}$$

Remark G.5 (Varshamov–Gilbert packing). We repeatedly use the standard coding-theoretic fact that for every integer $q \geq 8$, there exists a subset $V \subseteq \{-1, 1\}^q$ such that

$$\log |V| \geq \frac{q}{8} \qquad \text{and} \qquad d_H(u, v) \geq \frac{q}{8} \text{ for all distinct } u, v \in V.$$

This is the usual Varshamov–Gilbert bound; see, e.g., Tsybakov (2009, Lemma 2.9).

G.2. A generic asymmetric baseline library

The quotient metric identifies parameters up to automorphisms of the admissible language. To construct lower-bound submodels, it is convenient to freeze a baseline library whose automorphism stabilizer is trivial.

Lemma G.6 (Explicit asymmetric baseline library)**.** *Define*

$$A_a^\circ := \frac{Ba}{4r} I_d, \qquad a \in [r]. \tag{111}$$

Then $A_a^\circ \in \mathbb{A}_B$ for every $a \in [r]$, and

$$\vartheta_0 := \min_{\sigma \in \mathrm{Aut}(\mathcal{L}) \setminus \{\mathrm{id}\}} \sum_{a=1}^{r} \left\| A_a^\circ - A_{\sigma^{-1}(a)}^\circ \right\|_F > 0. \tag{112}$$

Consequently, for every submodel that keeps the grammar weights fixed and perturbs the library by at most $\vartheta_0/8$ in d_{par}, the minimizing automorphism in the quotient metric is the identity.

Proof. Since

$$\|A_a^\circ\|_{\mathrm{op}} = \frac{Ba}{4r} \leq \frac{B}{4} < B,$$

we have $A_a^\circ \in \mathbb{A}_B$. If $\sigma \neq \mathrm{id}$, then there exists at least one $a \in [r]$ with $\sigma^{-1}(a) \neq a$. Therefore

$$\left\| A_a^\circ - A_{\sigma^{-1}(a)}^\circ \right\|_F = \frac{B|a - \sigma^{-1}(a)|}{4r} \|I_d\|_F = \frac{B\sqrt{d}}{4r} |a - \sigma^{-1}(a)| > 0.$$

Summing over a and minimizing over the finite set $\mathrm{Aut}(\mathcal{L}) \setminus \{\mathrm{id}\}$ yields $\vartheta_0 > 0$.

Now let η, η' be two parameters in a fixed-weight submodel with

$$d_{\mathrm{par}}(\eta, \eta^\circ) \leq \frac{\vartheta_0}{8}, \qquad d_{\mathrm{par}}(\eta', \eta^\circ) \leq \frac{\vartheta_0}{8},$$

where η° has baseline library $\Theta^\circ = (A_1^\circ, \dots, A_r^\circ)$. If $\sigma \neq \mathrm{id}$, then by the triangle inequality,

$$d_{\mathrm{par}}(\eta, \sigma \cdot \eta') \geq \sum_{a=1}^{r} \|A_a^\circ - A_{\sigma^{-1}(a)}^\circ\|_F - d_{\mathrm{par}}(\eta, \eta^\circ) - d_{\mathrm{par}}(\eta', \eta^\circ) \geq \vartheta_0 - \frac{\vartheta_0}{4} = \frac{3\vartheta_0}{4}.$$

On the other hand, the identity alignment yields distance at most $\vartheta_0/4$. Hence the quotient minimum is attained by the identity. □

G.3. Regular grammar-law lower bound

Let

$$N := |\mathcal{L}|, \qquad K_{\text{eff}} = N - 1.$$

Assume $N \geq 2$, since the case $K_{\text{eff}} = 0$ is trivial.

Enumerate the admissible language as

$$\mathcal{L} = \{z_1, \ldots, z_N\},$$

and let

$$s := \left\lfloor \frac{N}{2} \right\rfloor.$$

Fix the asymmetric baseline library Θ° from Theorem G.6. For $\alpha \in (0, 1/4]$ and $u = (u_1, \ldots, u_s) \in \{-1, 1\}^s$, define the grammar law

$$\pi^u_{z_{2j-1}} := \frac{1 + \alpha u_j}{N}, \qquad \pi^u_{z_{2j}} := \frac{1 - \alpha u_j}{N}, \qquad j = 1, \ldots, s, \tag{113}$$

and, if N is odd, set $\pi^u_{z_N} = 1/N$. Let

$$\eta^u := (\pi^u, \Theta^\circ).$$

Theorem G.7 (Grammar-law minimax lower bound)**.** *There exists a constant $c_{\text{gram}} > 0$, depending only on $(\mathcal{L}, r, d, B)$, such that for all sufficiently large m,*

$$\mathfrak{R}_{m,n}(\{\eta^u : u \in \{-1, 1\}^s\}; d_q) \geq c_{\text{gram}} \sqrt{\frac{K_{\text{eff}}}{m}}. \tag{114}$$

Consequently, the same lower bound holds for any larger parameter class containing this submodel.

Proof. By Theorem G.6, for all sufficiently large m we may choose

$$\alpha_m := c_0 \sqrt{\frac{N}{m}}$$

with $c_0 > 0$ small enough so that $\alpha_m \leq 1/4$ and every pairwise ℓ_1-distance $\|\pi^u - \pi^v\|_1$ is strictly smaller than $\vartheta_0/4$. Hence, on the submodel $\{\eta^u\}$, the quotient metric is simply

$$d_q(\eta^u, \eta^v) = \|\pi^u - \pi^v\|_1.$$

Choose a Varshamov–Gilbert subset $V \subseteq \{-1, 1\}^s$ satisfying

$$\log |V| \geq \frac{s}{8}, \qquad d_H(u, v) \geq \frac{s}{8} \text{ for all distinct } u, v \in V.$$

For $u, v \in V$,

$$d_q(\eta^u, \eta^v) = \|\pi^u - \pi^v\|_1 = \frac{4\alpha_m}{N} d_H(u, v) \geq \frac{\alpha_m s}{2N}.$$

Since $s/N \geq 1/3$ for all $N \geq 2$,

$$d_q(\eta^u, \eta^v) \geq \frac{\alpha_m}{6} = \frac{c_0}{6} \sqrt{\frac{N}{m}} \asymp \sqrt{\frac{K_{\text{eff}}}{m}}. \tag{115}$$

We now compare the original grouped-data experiment with the oracle experiment that reveals the latent programs $Z_{1:m}$. By Theorem G.2, this oracle experiment is more informative, so it suffices to prove the lower bound there. In the oracle experiment,

$$Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} \pi^u.$$

Thus the law is simply the multinomial product measure $(\pi^u)^{\otimes m}$.

Fix $u, v \in V$. If $h = d_H(u, v)$, then the single-sample KL divergence is

$$\begin{aligned}\mathrm{KL}(\pi^u, \pi^v) &= \sum_{j:\, u_j \neq v_j} \left[\frac{1+\alpha_m}{N} \log \frac{1+\alpha_m}{1-\alpha_m} + \frac{1-\alpha_m}{N} \log \frac{1-\alpha_m}{1+\alpha_m}\right] \\ &= \frac{2\alpha_m h}{N} \log \frac{1+\alpha_m}{1-\alpha_m}. \end{aligned} \tag{116}$$

Since $\alpha_m \leq 1/4$, the inequality

$$\log \frac{1+\alpha_m}{1-\alpha_m} \leq 4\alpha_m$$

gives

$$\mathrm{KL}(\pi^u, \pi^v) \leq \frac{8\alpha_m^2 h}{N} \leq 8\alpha_m^2 \frac{s}{N} \leq 4\alpha_m^2.$$

Therefore the m-sample KL divergence in the oracle experiment satisfies

$$\mathrm{KL}\big((\pi^u)^{\otimes m}, (\pi^v)^{\otimes m}\big) \leq 4m\alpha_m^2 = 4c_0^2 N.$$

Since $\log |V| \geq s/8$ and $s \asymp N$, choosing $c_0 > 0$ sufficiently small ensures that

$$\max_{u \in V \setminus \{u^0\}} \mathrm{KL}\big((\pi^u)^{\otimes m}, (\pi^{u^0})^{\otimes m}\big) \leq \frac{1}{16} \log |V|$$

for any fixed $u^0 \in V$. Applying Theorem G.4 and the separation bound Equation (115) yields

$$\inf_{\widehat{\eta}} \sup_{u \in V} \mathbb{E}_u[d_q(\widehat{\eta}, \eta^u)] \geq c \sqrt{\frac{N}{m}} \asymp \sqrt{\frac{K_{\mathrm{eff}}}{m}}$$

for a constant $c > 0$ depending only on $\mathcal{L}$. By Theorem G.1, the same lower bound holds for the original grouped-data experiment. □

Remark G.8 (Interpretation). Theorem G.7 shows that even if the module library is known perfectly, the latent task-family law π cannot be estimated faster than $\sqrt{K_{\mathrm{eff}}/m}$. This is the irreducible grammar-complexity term in the problem.

G.4. Anchored-module lower bound

We now show that when the benchmark explicitly exposes some modules in isolation, one also recovers the $1/\sqrt{mn}$-type module-estimation difficulty.

Definition G.9 (Anchor-language condition)**.** Let $a_1, \ldots, a_s \in [r]$ be distinct symbols. We say that the admissible language $\mathcal{L}$ has an *s-anchor set* if:

1. the singleton programs $(a_1), \ldots, (a_s)$ all belong to $\mathcal{L}$;
2. no program in $\mathcal{L} \setminus \{(a_1), \ldots, (a_s)\}$ contains any of the symbols $a_1, \ldots, a_s$.

Assume Theorem G.9 with anchor set $a_1, \ldots, a_s$. Fix a grammar law $\pi^\circ \in \mathfrak{P}^\circ_{\mathcal{L}}$ and set

$$\beta_j := \pi^\circ_{(a_j)} > 0, \qquad \beta_{\max} := \max_{1 \leq j \leq s} \beta_j.$$

Fix the asymmetric baseline library Θ° from Theorem G.6, and let

$$p := d^2.$$

Choose an orthonormal basis $E_1, \ldots, E_p$ of $\mathbb{R}^{d \times d}$ under the Frobenius inner product.

For $u = (u_{j\ell}) \in \{-1, 1\}^{sp}$ and $\tau > 0$, define

$$A_{a_j}^{(u)} := A_{a_j}^\circ + \tau \sum_{\ell=1}^{p} u_{j\ell} E_\ell, \qquad j = 1, \ldots, s, \tag{117}$$

and keep all non-anchor modules fixed at their baseline values. Let η^u denote the resulting parameter with grammar law π°.

Theorem G.10 (Anchored-module minimax lower bound)**.** *Assume that $\mathcal{L}$ has an s-anchor set. Then there exists a constant $c_{\mathrm{anc}} > 0$, depending only on $(\mathcal{L}, r, d, B, \sigma, \pi^\circ)$, such that for all sufficiently large m,*

$$\mathfrak{R}_{m,n}(\{\eta^u : u \in \{-1,1\}^{sp}\}; d_q) \geq c_{\mathrm{anc}}\, \sigma\, d\sqrt{\frac{s}{mn}}. \tag{118}$$

Proof. Choose

$$\tau_{m,n} := c_1 \frac{\sigma}{\sqrt{mn}}$$

with $c_1 > 0$ sufficiently small so that (i) all perturbed anchor modules remain in $\mathbb{A}_B$, and (ii) the total library perturbation is at most $\vartheta_0/8$, where ϑ_0 is the asymmetry gap from Equation (112). Then by Theorem G.6, the quotient minimum on this submodel is attained by the identity, so for any u, v,

$$d_q(\eta^u, \eta^v) = \sum_{j=1}^{s} \big\| A_{a_j}^{(u)} - A_{a_j}^{(v)} \big\|_F.$$

If $h = d_H(u, v)$ denotes the Hamming distance in $\{-1,1\}^{sp}$, then

$$\sum_{j=1}^{s} \big\| A_{a_j}^{(u)} - A_{a_j}^{(v)} \big\|_F \geq \left(\sum_{j=1}^{s} \big\| A_{a_j}^{(u)} - A_{a_j}^{(v)} \big\|_F^2 \right)^{1/2} = 2\tau_{m,n}\sqrt{h}.$$

Choose a Varshamov–Gilbert subset $V \subseteq \{-1,1\}^{sp}$ with

$$\log |V| \geq \frac{sp}{8}, \qquad d_H(u, v) \geq \frac{sp}{8} \text{ for all } u \neq v \in V.$$

Hence for distinct $u, v \in V$,

$$d_q(\eta^u, \eta^v) \geq 2\tau_{m,n}\sqrt{\frac{sp}{8}} = c\,\tau_{m,n}\, d\sqrt{s} \asymp \sigma d\sqrt{\frac{s}{mn}}. \tag{119}$$

We now compare with the oracle experiment that reveals the latent programs $Z_{1:m}$; by Theorem G.2, it suffices to work there. In the oracle experiment, the augmented one-task law is

$$(Z_i, X_i, Y_i), \qquad Z_i \sim \pi^\circ.$$

By the anchor-language condition, perturbing an anchor module A_{a_j} affects only tasks with $Z_i = (a_j)$, because no other program contains a_j. Therefore, for $u, v \in V$, the one-task KL divergence in the oracle experiment equals

$$\begin{aligned} \mathrm{KL}(P_u^{\mathrm{or}}, P_v^{\mathrm{or}}) &= \sum_{j=1}^{s} \beta_j \cdot \frac{n}{2\sigma^2} \big\| A_{a_j}^{(u)} - A_{a_j}^{(v)} \big\|_F^2 \\ &\leq \frac{n\beta_{\max}}{2\sigma^2} \sum_{j=1}^{s} \big\| A_{a_j}^{(u)} - A_{a_j}^{(v)} \big\|_F^2 \\ &= \frac{2n\beta_{\max}\tau_{m,n}^2}{\sigma^2}\, d_H(u, v). \end{aligned} \tag{120}$$

Thus the m-task oracle KL divergence is bounded by

$$\mathrm{KL}\big((P_u^{\mathrm{or}})^{\otimes m}, (P_v^{\mathrm{or}})^{\otimes m}\big) \leq \frac{2mn\beta_{\max}\tau_{m,n}^2}{\sigma^2}\, d_H(u, v) \leq 2c_1^2\beta_{\max}\, sp.$$

Choosing c_1 sufficiently small ensures that this is at most $\frac{1}{16}\log|V|$ for all $u, v \in V$. Fano's lemma Theorem G.4, together with the separation bound Equation (119), yields

$$\inf_{\widehat{\eta}} \sup_{u \in V} \mathbb{E}_u[d_q(\widehat{\eta}, \eta^u)] \geq c'\, \sigma d\sqrt{\frac{s}{mn}}$$

for some $c' > 0$. By Theorem G.1, the same lower bound holds in the original grouped-data experiment. □

Remark G.11 (Benchmark-design meaning of anchors). Theorem G.10 isolates the statistical value of benchmark cells that expose modules in isolation. Without such anchor tasks, module-space lower bounds depend on more complicated local inverse geometry; with anchors, one recovers the classical $1/\sqrt{mn}$ parametric difficulty of estimating a shared linear operator.

G.5. Zero-observability impossibility

The next statement turns exact observational collapse into a nonvanishing minimax floor.

Theorem G.12 (Exact non-identifiability induces a positive minimax floor)**.** *Suppose there exist two parameters $\eta_0, \eta_1 \in \Xi(B, \mathcal{L})$ such that*

$$d_q(\eta_0, \eta_1) = \Delta_0 > 0 \qquad \textit{and} \qquad Q_{\eta_0}^{(n)} = Q_{\eta_1}^{(n)} \ \textit{for every } n \geq 1.$$

Then for every $m \geq 1$ and every $n \geq d + 1$,

$$\mathfrak{R}_{m,n}(\{\eta_0, \eta_1\}; d_q) \geq \frac{\Delta_0}{4}. \tag{121}$$

In particular, whenever Appendix C produces such a pair through zero contextual observability in a single-occurrence language, the structural risk cannot vanish.

Proof. If the one-task grouped laws are equal, then the m-task laws are also equal:

$$(Q_{\eta_0}^{(n)})^{\otimes m} = (Q_{\eta_1}^{(n)})^{\otimes m}.$$

Therefore the total variation distance between the two m-task laws is zero. Applying Le Cam's inequality Equation (109) gives

$$\inf_{\widehat{\eta}} \sup_{\eta \in \{\eta_0, \eta_1\}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)] \geq \frac{d_q(\eta_0, \eta_1)}{4} = \frac{\Delta_0}{4}.$$

□

G.6. Singular lower bound on the duplicated-module family

We finally prove the lower bound corresponding to the quadratic singular geometry of Appendix F. Recall the duplicated family

$$\eta_t := \eta_{\Delta_t}, \qquad \Delta_t := t\, e_1 e_1^\top, \qquad t \geq 0,$$

built around the singular center η_0 in Theorem F.14. By Theorem F.15,

$$d_q(\eta_t, \eta_0) = 2t.$$

Let $\widetilde{Q}_t^{(n)}$ denote the reduced one-task predictive law from Appendix E under η_t.

Lemma G.13 (Exact unconditional χ^2 divergence in the rank-one duplicated subfamily)**.** *Let*

$$a_t := \frac{t^2}{\sigma^2}.$$

If $a_t < 1/2$, then

$$\chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big) = \frac{1}{2}(1 - 2a_t)^{-n/2} + \frac{1}{2}(1 + 2a_t)^{-n/2} - 1. \tag{122}$$

Moreover, there exists a universal constant $C_\chi \in (0, \infty)$ such that whenever

$$a_t \leq \frac{1}{4n},$$

one has

$$\chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big) \leq C_\chi\, n^2 a_t^2 = C_\chi\, \frac{n^2 t^4}{\sigma^4}. \tag{123}$$

Proof. By Appendix F, conditional on X, the reduced one-task χ^2-divergence satisfies

$$\chi^2(q_{t,X}, q_{0,X}) = \cosh(r_{X,t}^2) - 1, \qquad r_{X,t}^2 = \frac{1}{\sigma^2}\|\Delta_t X\|_F^2.$$

Since $\Delta_t = te_1e_1^\top$,

$$\|\Delta_t X\|_F^2 = t^2 \sum_{j=1}^{n} X_{1j}^2,$$

and $\sum_{j=1}^{n} X_{1j}^2 \sim \chi_n^2$. Because the design law is the same under both parameters,

$$\chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big) = \mathbb{E}_X\big[\chi^2(q_{t,X}, q_{0,X})\big] = \mathbb{E}\big[\cosh(a_t\chi_n^2) - 1\big].$$

Using

$$\cosh(u) = \frac{e^u + e^{-u}}{2}$$

and the mgf of χ_n^2,

$$\mathbb{E}[e^{\lambda\chi_n^2}] = (1-2\lambda)^{-n/2} \qquad \text{for } \lambda < 1/2,$$

we obtain

$$\chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big) = \frac{1}{2}(1-2a_t)^{-n/2} + \frac{1}{2}(1+2a_t)^{-n/2} - 1,$$

which is Equation (122).

For the upper bound, define

$$f(a) := \frac{1}{2}(1-2a)^{-n/2} + \frac{1}{2}(1+2a)^{-n/2} - 1.$$

Then $f(0) = f'(0) = 0$, and

$$f''(a) = \frac{n(n+2)}{2}\left[(1-2a)^{-n/2-2} + (1+2a)^{-n/2-2}\right].$$

If $0 \le a \le 1/(4n)$, then

$$1 - 2a \ge 1 - \frac{1}{2n}, \qquad 1 + 2a \le 1 + \frac{1}{2n},$$

so both factors above are bounded by a universal constant. Therefore

$$\sup_{0\le a\le 1/(4n)} |f''(a)| \le C_\chi n^2$$

for some universal C_χ. Taylor's theorem with remainder gives

$$f(a_t) \le \frac{1}{2}\sup_{0\le a\le a_t} |f''(a)|\, a_t^2 \le C_\chi n^2 a_t^2,$$

which proves Equation (123). □

Theorem G.14 (Singular two-point lower bound)**.** *There exists a constant $c_{\mathrm{sing}} > 0$, depending only on σ, such that for all sufficiently large m,*

$$\mathfrak{R}_{m,n}\big(\{\eta_0, \eta_{t_{m,n}}\}; d_q\big) \ge c_{\mathrm{sing}}\,\frac{\sigma}{m^{1/4}\sqrt{n}}, \qquad t_{m,n} := c_0\,\frac{\sigma}{m^{1/4}\sqrt{n}}, \tag{124}$$

provided $c_0 > 0$ is chosen sufficiently small. Consequently,

$$\inf_{\widehat{\eta}} \sup_{0\le t\le t_{m,n}} \mathbb{E}_{\eta_t}[d_q(\widehat{\eta}, \eta_t)] \gtrsim \frac{1}{m^{1/4}\sqrt{n}}.$$

Proof. By Theorem G.2, the reduced experiment is Blackwell equivalent to the full grouped-data experiment, so it suffices to prove the lower bound for the reduced one-task laws $\widetilde{Q}_t^{(n)}$.

For $t = t_{m,n}$, we have

$$a_t = \frac{t^2}{\sigma^2} = \frac{c_0^2}{\sqrt{m}\, n}.$$

Hence $a_t \le 1/(4n)$ for all sufficiently large m. By Equation (123),

$$\chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big) \le C_\chi \frac{c_0^4}{m}.$$

For the m-task product experiment, χ^2 tensorizes:

$$1 + \chi^2\big((\widetilde{Q}_t^{(n)})^{\otimes m}, (\widetilde{Q}_0^{(n)})^{\otimes m}\big) = \Big(1 + \chi^2\big(\widetilde{Q}_t^{(n)}, \widetilde{Q}_0^{(n)}\big)\Big)^m \le \exp(C_\chi c_0^4).$$

Choose $c_0 > 0$ small enough that

$$\exp(C_\chi c_0^4) - 1 \le \frac{1}{2}.$$

Then

$$\mathrm{TV}\big((\widetilde{Q}_t^{(n)})^{\otimes m}, (\widetilde{Q}_0^{(n)})^{\otimes m}\big) \le \sqrt{\frac{1}{2}\chi^2\big((\widetilde{Q}_t^{(n)})^{\otimes m}, (\widetilde{Q}_0^{(n)})^{\otimes m}\big)} \le \frac{1}{2}.$$

Applying Le Cam's inequality Equation (109) and using Equation (102),

$$\inf_{\widehat{\eta}} \sup_{\eta \in \{\eta_0, \eta_t\}} \mathbb{E}_\eta[d_q(\widehat{\eta}, \eta)] \ge \frac{d_q(\eta_t, \eta_0)}{4}\left(1 - \frac{1}{2}\right) = \frac{2t}{8} = \frac{t}{4}.$$

Substituting $t = t_{m,n}$ proves Equation (124). □

Remark G.15 (Sharpness relative to the upper bound). The singular lower bound of Theorem G.14 matches the $\kappa = 2$ structural upper bound from Appendix F up to logarithmic factors, since the posterior rate there is $O(\delta_{m,n}^{1/2}/\sqrt{n})$ and $\delta_{m,n} \asymp m^{-1/2}$ up to logs when the model dimensions are fixed.

Remark G.16 (What Appendix G shows). The lower bounds separate three distinct sources of difficulty:

1. *grammar complexity*: even with a perfectly known module library, the latent task-family law cannot be recovered faster than $\sqrt{K_{\mathrm{eff}}/m}$;
2. *module estimation*: when the benchmark exposes modules in isolation, one inherits the classical $1/\sqrt{mn}$-type difficulty of estimating shared operators;
3. *singularity*: in overfitted duplicated-module families, structural estimation slows to $m^{-1/4}n^{-1/2}$, even though predictive learning remains much easier.

This is precisely the statistical phase structure that the main paper turns into benchmark hardness labels.

H. Experimental Protocol

This appendix specifies the benchmark generators, posterior computations, evaluation metrics, and plotting conventions used throughout Section 5. Every experiment is run in the exact reduced experiment $(X_i, \widehat{W}_i)_{i=1}^m$, which is Blackwell equivalent to the full grouped-data experiment by Appendix B. All reported quantities therefore correspond directly to the proved model class, rather than to an external approximation layer. Unless otherwise stated, all numerical calculations are performed in double precision, and all figures report medians across 32 independent replications.

H.1. Benchmark philosophy and released hardness labels

Each benchmark cell is generated from a theorem-matched stochastic task grammar and released with an explicit hardness label tuple

$$h = \big(K_{\mathrm{eff}},\ \Delta_{\mathrm{ctx}},\ \lambda_{\mathrm{lin}},\ \gamma_{\mathrm{comp}},\ \kappa,\ s_{\mathrm{anc}}\big).$$

These labels are computed from the definitions in the main text and not fit post hoc.

Effective grammar complexity. For every benchmark family,

$$K_{\mathrm{eff}} = |\mathcal{L}| - 1.$$

Contextual observability. When relevant, we compute

$$\Delta_{\mathrm{ctx}}(\eta^\star) = \min_{a \neq b} \sum_{(u,v) \in \mathcal{C}_{\mathcal{L}}} \nu_{\pi^\star}(u,v)\, \|A_v^\star (A_a^\star - A_b^\star) A_u^\star\|_F^2$$

exactly from the benchmark definition.

Linearized observability. For the regular families, $\lambda_{\mathrm{lin}}(\eta^\star)$ is computed as the smallest singular value squared of the weighted Jacobian of the program-operator map,

$$\lambda_{\mathrm{lin}}(\eta^\star) = \inf_{\|H\|_{\mathrm{lib}}=1} \|D\Phi_{\Theta^\star}(H)\|_{\pi^\star}^2,$$

using a finite matrix representation in the theorem ambient space.

Component separation. We compute

$$\gamma_{\mathrm{comp}}(\eta^\star) = \min_{z \neq z'} \|A_z^\star - A_{z'}^\star\|_F$$

by explicit enumeration of the admissible language.

Singularity order. We assign $\kappa = 1$ for the regular identifiable families and $\kappa = 2$ for the duplicated-module singular family proved in Appendix F. For the exact non-identifiable observability cell $\rho = 0$, we do not assign a finite κ; instead the cell is labeled as *non-identifiable*.

Anchor count. For the anchored families, $s_{\mathrm{anc}} = s$. For all other benchmark families, $s_{\mathrm{anc}} = 0$.

H.2. Benchmark suites

We use four theorem-native suites, plus a matched no-anchor comparison family for the anchor experiment.

H.2.1. Regular identifiable suite

For $q \in \{1, 2, 4\}$, define

$$\mathcal{L}_q^{\mathrm{reg}} = \big\{(1),\ (1,2),\ \ldots,\ (1,2,\ldots,q+1)\big\}, \qquad r = q+1, \qquad d = 2, \qquad n = 32.$$

The true grammar law is uniform on $\mathcal{L}_q^{\mathrm{reg}}$, and the module library is diagonal:

$$A_a^\star = \begin{pmatrix} 0.72 + 0.03a & 0 \\ 0 & 0.35 + 0.02a \end{pmatrix}, \qquad a = 1, \ldots, q+1.$$

This suite is used for the regular predictive and structural sanity checks shown in Appendix Figures 4 and 5. In the appendix structural figure, the left panel uses $q \in \{1, 2, 4\}$, the normalized structural comparison focuses on $q \in \{1, 2\}$, and the direct local inverse inset uses the $q = 4$ cell.

H.2.2. Observability suite

The observability phase-transition family is

$$\mathcal{L}_{\mathrm{obs}} = \{(1,3), (2,3)\}, \qquad \pi^\star = \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \qquad d = 2, \qquad m = 256.$$

The module library is

$$A_1^\star = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad A_2^\star = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \qquad A_3^\star(\rho) = \begin{pmatrix} 1 & 0 \\ 0 & \rho \end{pmatrix},$$

with

$$\rho \in \{0, 0.05, 0.10, 0.20, 0.40, 0.80\}, \qquad n \in \{4, 8, 16, 32, 64\}.$$

For this family,

$$\Delta_{\mathrm{ctx}}(\eta_\rho^\star) = \rho^2/2, \qquad \gamma_{\mathrm{comp}}(\eta_\rho^\star) = \rho.$$

The cell $\rho = 0$ is exactly non-identifiable and is treated separately in the right panel of Figure 1.

H.2.3. SINGULAR DUPLICATED-MODULE SUITE

The singular family is

$$\mathcal{L}_{\mathrm{sing}} = \{(1), (2)\}, \qquad \pi^\star = \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \qquad d = 1,$$

with

$$A_1(t) = 1 + t, \qquad A_2(t) = 1 - t, \qquad t^\star = 0.$$

We evaluate

$$n \in \{4, 8, 16, 32\}, \qquad m \in \{64, 128, 256, 512, 1024\}.$$

In this family,

$$d_q(\eta_t, \eta_0) = 2t, \qquad \kappa = 2.$$

The right panel of Figure 2 additionally evaluates the local quartic geometry at

$$t \in \{0.003, 0.0045, 0.006, 0.008, 0.011, 0.015, 0.021, 0.030, 0.042, 0.060\}$$

for each $n \in \{4, 8, 16, 32\}$.

H.2.4. ANCHORED AND MATCHED NO-ANCHOR SUITES

For $s \in \{1, 2, 4, 8\}$, define the anchored family

$$\mathcal{L}_s^{\mathrm{anc}} = \{(1), \ldots, (s)\} \cup \{(s+1), (s+1, s+2)\},$$

with $r = s + 2$, $d = 2$, fixed $n = 16$, and grammar weights

$$\pi^\star_{(j)} = \frac{\beta}{s}, \qquad j = 1, \ldots, s, \qquad \beta = \frac{1}{2},$$

while the remaining mass is split uniformly over $(s+1)$ and $(s+1, s+2)$. The module library is diagonal:

$$A_a^\star = \begin{pmatrix} 0.80 + 0.04a & 0 \\ 0 & 0.25 + 0.03a \end{pmatrix}, \qquad a = 1, \ldots, s+2.$$

To isolate the value of explicit anchor exposure, we construct a matched no-anchor family

$$\mathcal{L}_s^{\mathrm{noanc}} = \{(1, s+1), \ldots, (s, s+1)\} \cup \{(s+1), (s+1, s+2)\},$$

with the same r, the same module library, and the same grammar weights. Thus the only qualitative difference between the two suites is that in $\mathcal{L}_s^{\mathrm{anc}}$ the anchor modules appear in singleton tasks, while in $\mathcal{L}_s^{\mathrm{noanc}}$ they only appear through a shared context.

The right panel of Figure 3 uses the matched comparison at

$$s = 4, \qquad n = 16, \qquad m \in \{64, 128, 256, 512\}.$$

H.3. Data generation in the reduced experiment

For every suite and every replication:

1. sample m latent programs $Z_1, \dots, Z_m$ independently from the true grammar law $\pi^\star$;
2. for each task i, draw $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, $j = 1, \dots, n$;
3. form responses
$$y_{ij} = A^\star_{Z_i} x_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d),$$
with $\sigma = 0.15$;
4. compute the exact reduced statistic
$$\widehat{W}_i = Y_i X_i^\top (X_i X_i^\top)^{-1}.$$

All posterior computations are then performed on $(X_i, \widehat{W}_i)_{i=1}^m$.

H.4. Posterior computation

H.4.1. Regular and anchored suites

For the regular and anchor families, we restrict both the generator and fitted model to the theorem-matched diagonal subfamily. Thus each module is parameterized as

$$A_a = \operatorname{diag}(u_a, v_a).$$

The grammar weights are fixed at the true values in the anchored families and free in the regular family. In all cases, inference uses the *exact reduced likelihood*. Posterior computation proceeds as follows:

1. compute a MAP estimator using deterministic L-BFGS optimization from multiple random initializations;
2. evaluate the exact Hessian of the negative log posterior at the MAP by automatic differentiation;
3. approximate the posterior by the corresponding Gaussian Laplace approximation;
4. compute posterior quantiles from draws of this Gaussian approximation.

The regular suites use 8 optimization restarts and 192 posterior draws. For the $q = 4$ structural validation cell, we increase the computational budget to 12 restarts and 320 posterior draws. In the anchored suite we use 6 restarts and 160 posterior draws by default, increase to 10 restarts and 240 draws for $s = 8$, and to 14 restarts and 384 draws for $s = 1$, which is the visibly noisiest case.

H.4.2. Observability suite

For the observability family, posterior inference is exact on a rectangular grid over

$$(\pi, \rho) \in [\pi_{\min}, \pi_{\max}] \times [\rho_{\min}, \rho_{\max}] = [10^{-3}, 1 - 10^{-3}] \times [0, 0.9].$$

We use 161 grid points in π and 241 grid points in ρ, together with trapezoidal quadrature weights. The grammar law is parameterized as

$$\pi = (\pi, 1 - \pi),$$

where π denotes the probability of the first program $(1, 3)$. Because the reduced family is exactly two-dimensional, no approximation layer is required.

H.4.3. SINGULAR SUITE

For the duplicated-module family, posterior inference is exact on a one-dimensional grid

$$t \in [0, 0.2]$$

with 1601 grid points and trapezoidal quadrature weights. The predictive Hellinger geometry is computed exactly by Gauss–Hermite quadrature of order 80. No Laplace approximation or MCMC is used in this suite.

H.5. Evaluation metrics

Predictive posterior radius. For the regular predictive sanity check and the singular-family predictive panel, we use

$$R_{\mathrm{pred},0.9} = \inf\{r : \Pi_m(\bar{h}_X(\eta, \eta^\star) \le r) \ge 0.9\}.$$

In the regular suite, $\bar{h}_X$ is estimated on a held-out design bank using exact reduced densities and a fixed proposal bank sampled from the true predictive law. In the singular suite, $\bar{h}_X$ is computed by exact Gauss–Hermite quadrature.

Structural posterior radius. In the regular and singular suites, the structural posterior radius is

$$R_{\mathrm{str},0.9} = \inf\{r : \Pi_m(d_q(\eta, \eta^\star) \le r) \ge 0.9\}.$$

For the regular suite, d_q is evaluated by exact enumeration of the automorphism group of the language. For the duplicated family, this reduces to

$$d_q(\eta_t, \eta_0) = 2t.$$

In the observability suite, the fitted family is two-dimensional, and the restricted quotient metric becomes

$$d_{\mathrm{obs}}\big((\pi, \rho), (\pi^\star, \rho^\star)\big) = 2|\pi - \pi^\star| + |\rho - \rho^\star|,$$

which equals the general quotient distance restricted to the family with fixed A_1 and A_2. The right panel of Figure 1 therefore reports

$$R_{\mathrm{str},0.9} = \inf\{r : \Pi_m(d_{\mathrm{obs}} \le r) \ge 0.9\}.$$

Latent-program recovery. For synthetic data the true latent programs $Z_i^\star$ are known, so we report

$$M_Z = \frac{1}{m} \sum_{i=1}^{m} \Pi_m(Z_i = Z_i^\star).$$

In the observability suite this is computed exactly by integrating responsibilities over the exact posterior grid.

Anchor-module radius. In the anchor experiments, the structural target is the anchor-only library error

$$R_{\mathrm{anc},0.9} = \inf\left\{ r : \Pi_m\left(\min_{\sigma \in S_s} \sum_{j=1}^{s} \|A_j - A^\star_{\sigma(j)}\|_F \le r \right) \ge 0.9 \right\}.$$

Thus the anchor-module error is the *sum* over s anchor modules after optimal assignment. This is why the left panel of Figure 3 uses the per-anchor normalization

$$\frac{R_{\mathrm{anc},0.9}\sqrt{mn}}{d\, s^{3/2}},$$

rather than $R_{\mathrm{anc},0.9}/(d\sqrt{s/(mn)})$.

H.6. Figure-specific details

Appendix Figure 4. We evaluate the regular suite at

$$q \in \{1, 2, 4\}, \qquad m \in \{64, 128, 256, 512, 1024\}, \qquad n = 32.$$

The figure reports $R_{\mathrm{pred},0.9}$ and its normalization by

$$\delta_{m,n} = \left(\frac{K_{\mathrm{eff}} \log m + r d^2 \log(mn)}{m} \right)^{1/2}.$$

Appendix Figure 5. The structural regular figure uses the same regular suite and grid. The left panel reports $R_{\mathrm{str},0.9}$; the normalized panel focuses on the classes with the cleanest main-paper comparison ($q = 1, 2$); and the inset directly evaluates local perturbations around the $q = 4$ truth. Specifically, we draw random zero-sum grammar perturbations and random module perturbation directions, scale them by

$$\varepsilon \in \{0.003, 0.006, 0.009, 0.012, 0.018\},$$

and plot $\bar{h}_X(\eta, \eta^\star)^2$ against $d_q(\eta, \eta^\star)^2$.

Figure 1. For the observability phase transition we evaluate

$$\rho \in \{0, 0.05, 0.10, 0.20, 0.40, 0.80\}, \qquad n \in \{4, 8, 16, 32, 64\}, \qquad m = 256.$$

The left panel plots $1 - M_Z$ against

$$n\Delta_{\mathrm{ctx}} = n\rho^2/2,$$

using all positive-ρ cells. The right panel isolates the $\rho = 0$ family and overlays the corresponding structural and symmetry floors.

Figure 2. For the singular family we evaluate

$$n \in \{4, 8, 16, 32\}, \qquad m \in \{64, 128, 256, 512, 1024\}.$$

The normalized panels use

$$\frac{R_{\mathrm{pred},0.9}}{\delta_{m,n}} \qquad \text{and} \qquad \frac{R_{\mathrm{str},0.9}\sqrt{n}}{\delta_{m,n}^{1/2}},$$

with

$$\delta_{m,n} = \left(\frac{\log m + 2\log(mn)}{m}\right)^{1/2},$$

corresponding to $K_{\mathrm{eff}} = 1$, $r = 2$, and $d = 1$. The quartic-geometry panel uses the explicit (n, t)-grid listed above and a slope-1 guide on log-log axes.

Figure 3. For the anchored family we evaluate

$$s \in \{1, 2, 4, 8\}, \qquad m \in \{64, 128, 256, 512\}, \qquad n = 16.$$

The left panel uses the per-anchor normalized quantity

$$R_{\mathrm{anc},0.9}\sqrt{mn}/(d\, s^{3/2}).$$

The right panel fixes $s = 4$, $n = 16$, and compares the anchored family against the matched no-anchor family. Because the $s = 1$ regime is empirically the noisiest, the anchor figure reports interquartile bands rather than 10th–90th percentile bands.

H.7. Reproducibility and numerical conventions

Each benchmark cell is rerun with 32 independent random seeds. The plotting scripts use deterministic pseudo-random initialization schedules tied to $(s, m, \rho, n, \text{family})$ so that the experiments are exactly reproducible. GPU acceleration is used when available, but all scripts fall back to CPU execution without changing the mathematical computations. No figure in the paper relies on a training heuristic, model-selection sweep, or unreported hyperparameter search: every numerical choice used in the final plots is fixed in advance by the suite-specific protocol described above.

H.8. Criteria for successful theorem validation

The experiments are considered successful if they exhibit the following signatures:

1. **Regular predictive sanity check:** normalization by $\delta_{m,n}$ substantially stabilizes the predictive posterior radius.

2. **Regular structural sanity check:** normalization by $\delta_{m,n}$ stabilizes the structural posterior radius in the regular regime, and local perturbations satisfy an approximately linear relation between $\bar{h}_X^2$ and d_q^2.

3. **Observability:** $1 - M_Z$ approximately collapses as a function of $n\Delta_{\mathrm{ctx}}$, while the $\rho = 0$ cell exhibits a visible nonzero uncertainty floor.

4. **Singularity:** predictive radii are regular after $\delta_{m,n}$-normalization, but structural radii require the slower singular normalization $\delta_{m,n}^{1/2}/\sqrt{n}$.

5. **Anchors:** the per-anchor normalized anchor radius remains order one across the anchor-count grid, and explicit anchor cells consistently outperform matched no-anchor cells.

These signatures are exactly the empirical counterparts of the theorems in Section 3. The benchmark is therefore not only derived from the theory in construction, but validated by checking the theorem-predicted phase structure directly.

I. Additional Plots

This appendix collects supplementary theorem-validation plots that are useful for completeness but not central to the main-text narrative. In the main paper we focus on the three benchmark regimes that most directly expose our new hardness parameters—observability, singularity, and anchor exposure. The plots below instead serve as *regular-regime sanity checks*: they verify that, when the inverse map is well behaved, the posterior radii behave in the regular way predicted by the theory.

More precisely, Appendix Figure 4 validates the predictive contraction theorem in the regular identifiable family, while Appendix Figure 5 validates the corresponding regular structural behavior and directly visualizes the local inverse law

$$\bar{h}_X(\eta, \eta^\star)^2 \asymp d_q(\eta, \eta^\star)^2$$

through a local perturbation study. These figures are relegated to the appendix only because they play a supporting role: they show that the regular regime behaves as expected, whereas the main contribution of the paper lies in isolating and benchmarking the nontrivial regimes in which observability vanishes, singularity slows structural learning, or benchmark-cell design changes the effective statistical phase.

Regular predictive sanity check. The first supplementary figure reports the regular identifiable family

$$\mathcal{L}_q^{\mathrm{reg}} = \{(1), (1,2), \ldots, (1,2,\ldots,q+1)\}, \qquad q \in \{1,2,4\},$$

with $n = 32$. The left panel shows the raw predictive posterior radius $R_{\mathrm{pred},0.9}$, while the right panel normalizes by the theoretical predictive scale

$$\delta_{m,n} = \left(\frac{K_{\mathrm{eff}} \log m + r d^2 \log(mn)}{m} \right)^{1/2}.$$

The substantive point is not exact overlap across grammar classes, which would require class-independent constants, but rather that the normalization substantially stabilizes the family of curves and removes most of the m-dependence.

Regular structural sanity check and local inverse validation. The second supplementary figure uses the same regular family. Its left panel reports the structural posterior radius $R_{\mathrm{str},0.9}$. The normalized panel focuses on the lower-complexity classes $q \in \{1,2\}$, where the rate-normalized comparison is the clearest visually. The more important diagnostic is the inset, which directly probes the regular local inverse theorem by generating small perturbations around the $q = 4$ truth and plotting

$$\bar{h}_X(\eta, \eta^\star)^2 \qquad \text{against} \qquad d_q(\eta, \eta^\star)^2.$$

The approximate linearity of this relation is exactly what one expects when the inverse exponent is $\kappa = 1$. Thus the inset is the local geometric counterpart to the regular structural-contraction theorem: the posterior rate behaves regularly because the predictive geometry is locally Euclidean in the quotient parameter.

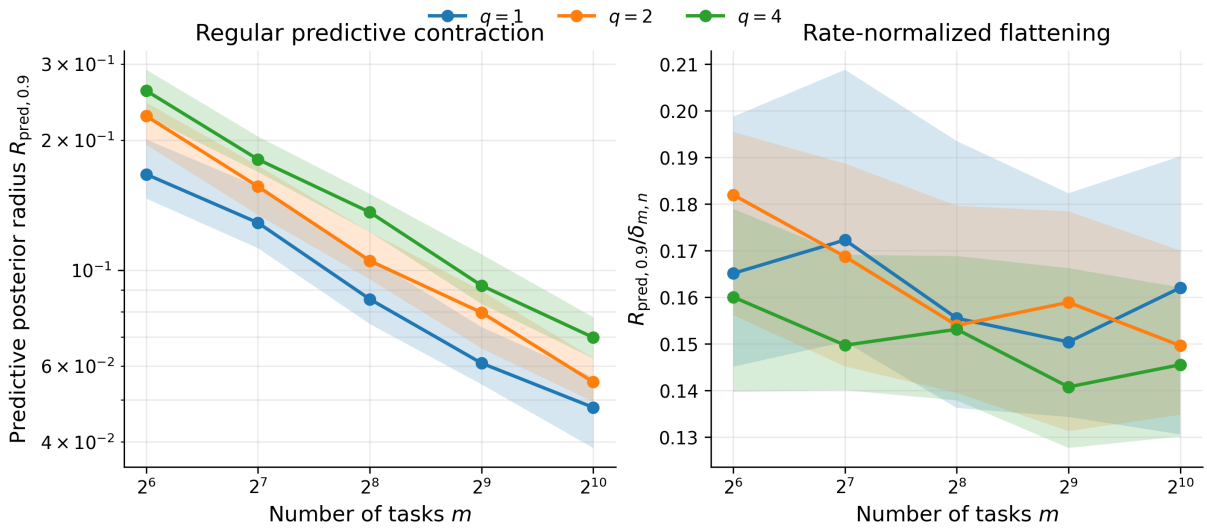

Figure 4. **Regular predictive contraction in the stochastic linear task-grammar family.** *Left:* the posterior predictive radius $R_{\mathrm{pred},0.9}$ decreases with the number of tasks m, with larger grammar complexity q yielding larger radii. *Right:* dividing by the theoretical predictive scale $\delta_{m,n}$ substantially flattens the curves, consistent with the predicted regular predictive rate up to model-class-dependent constants. Curves show medians across replications; shaded bands indicate 10th–90th percentile ranges.

How these plots complement the main text. Taken together, Appendix Figures 4 and 5 show that the regular regime behaves exactly as the theory says it should: predictive and structural uncertainty contract at the same order, and the local inverse map is first-order nondegenerate. The main text then uses this regular baseline as the reference point against which the more novel benchmark regimes are contrasted. In particular, Figure 1 shows what happens when observability collapses, Figure 2 shows what happens when the inverse map becomes singular, and Figure 3 shows how benchmark design can move the problem between qualitatively different statistical regimes.

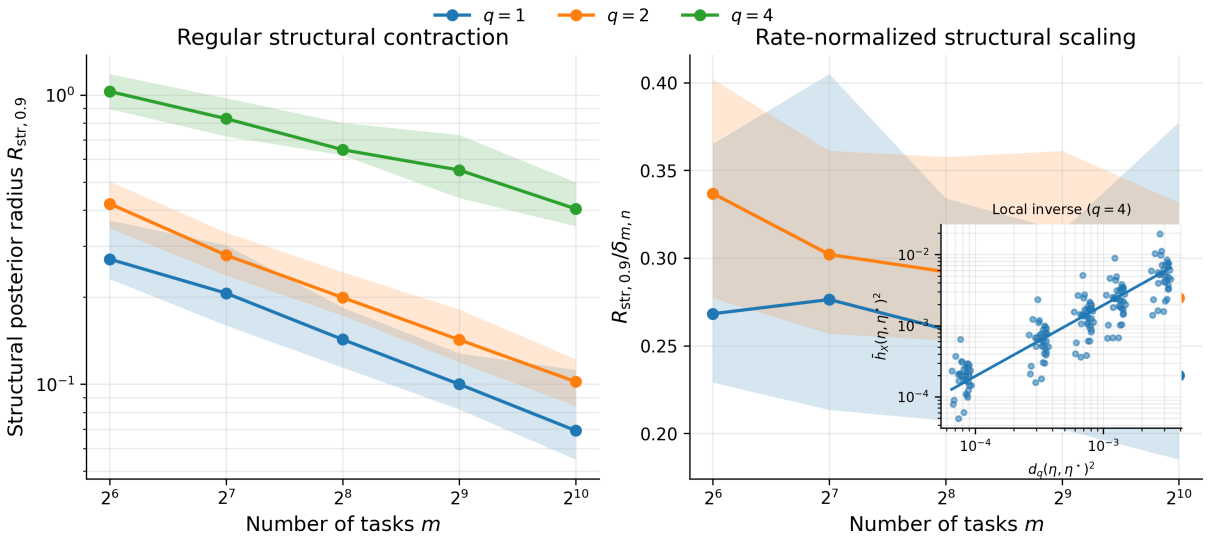

Figure 5. **Regular structural contraction and direct local inverse validation.** *Left:* the structural posterior radius $R_{\mathrm{str},0.9}$ decreases with the number of tasks m, with larger grammar classes yielding larger radii. *Right:* normalization by the regular structural scale $\delta_{m,n}$ stabilizes the lower-complexity regular classes. The inset directly validates the regular local inverse law, showing an approximately linear relationship between $\bar{h}_X(\eta,\eta^\star)^2$ and $d_q(\eta,\eta^\star)^2$ in the $q=4$ regular regime. Curves show medians across replications; shaded bands indicate 10th–90th percentile ranges.