TASK-LOCKED PREDICTIVE ADAPTATION: LEFTOVER RESIDUAL DENSITIES FROM FEW TARGET LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A trained neural operator is often reused when the unresolved forcing of a PDE changes. Users already rely on a predictive distribution for one operational functional, such as a regional mean. Recalibrating the residual as an unrestricted joint density moves that functional. Few labeled target fields are available. The leftover residual still requires a density.

We keep one shared density for the task coordinate. The labels update only the density of the remaining coefficients given the task. From the source we build a dictionary of these leftover models: scale and tail changes, coupling to the task, and both together. Sequential likelihood updates the weights. The deployed leftover density averages the sequential predictors. We call this task-locked predictive adaptation.

The adapted predictor has the same task law as the shared task density. Under i.i.d. target sampling, with the dictionary fixed from the source, the expected excess log loss relative to the source leftover density is at most $-\log(\pi_{\mathrm{ref}})/(K+1)$. Our analysis separates what the labels can buy from what the dictionary itself can represent.

We test viscous Burgers with a frozen Fourier neural operator, ten training seeds, and 32 target labels. Residuals are scored on eight Fourier coefficients. Under a joint change of forcing tail and dependence, the leftover log score improves by 0.450 nats. Under no shift the mean change is 0.0005 nats. A maximum-likelihood fit of the same dictionary satisfies its calibration criterion in every recorded fit but scores 0.0717 nats higher on held-out fields. Composed corrections improve that score by 0.050 nats over a mixture of the separate families.

1 INTRODUCTION

A distribution for one regional average is useful for a specific decision. It does not determine how errors in two regions co-occur. A surrogate reused under a change in unresolved forcing may therefore need a richer residual distribution. One route is to refit the joint density from target fields. That route moves the task distribution the application already relies on, and it spends few labels on structure the application has already specified.

We study conditional adaptation under this constraint. The supplied task marginal fixes one part of the joint law. A conditional density describes the residual variation the marginal leaves open. Updating that density leaves the marginal unchanged, for every sample and every input. The same residual predictor then answers joint queries that the marginal alone cannot.

The main design question is which conditional changes to allow. Few labels motivate using source data to choose the candidate corrections before adaptation. We use source second moments to construct residual coordinates and source response directions to organize scale, tail and dependence corrections. Target labels then fit mixture weights over this fixed dictionary. This separates the construction of candidate densities from their target-side estimation.

A fixed dictionary also constrains what unlimited data could achieve. Our central observation is that all candidates directed along one source direction share a reflection symmetry. Every convex mixture inherits it. A target with conditional variation outside that symmetry has a positive approximation error under every estimator restricted to the dictionary. The size of this error is a KL divergence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from the target to its symmetrized version. Near a symmetric reference, its size is quadratic in the part of the density change excluded by the symmetry. A Gaussian specialization separates missed mean response from covariance anisotropy.

This gives a design principle. Decide the directions in which the conditional law may change before deciding how to estimate its weights. Each added direction weakens the common symmetry. It also divides the directional prior mass among more candidates, increasing the atomic-oracle complexity penalty for a fixed directional candidate. A classical progressive-mixture bound controls that second effect for our averaging rule. The two effects differ in kind. The symmetry limits what any amount of data can express; the prior mass limits what a fixed number of labels can estimate.

Controlled Burgers experiments examine adaptation, estimation and directional coverage in the same residual coordinates. At 32 labels, the combined forcing shift yields a 0.450-nat reduction in conditional negative log-likelihood relative to the source reference. Comparisons include an unregularized fit of the same dictionary, source-informed conditional Gaussian and Student regression, and a conditional Student model aligned with the same source direction. Broader directional coverage improves three prescribed shifts, while the larger loss on the source-preferred direction makes the equal-weight average worse. These results expose the tradeoff that motivates the analysis.

Our contributions are a source-defined conditional adaptation construction that preserves a supplied task law; a quantitative convex-hull analysis linking source-direction symmetry to irreducible approximation error; and controlled comparisons that distinguish candidate coverage, source-informed estimation and prefix averaging. Full proofs and implementation definitions are deferred to the appendix.

2 A TASK MARGINAL AND A CONDITIONAL RESIDUAL LAW

The joint events of Section 1 depend on the residual between a true field and a frozen surrogate prediction. We work with that residual's eight-coefficient projection onto the sine and cosine coefficients at frequencies one to four, written $Y \in \mathbb{R}^8$. For input x, a scalar linear functional $Z = T_x(Y)$ represents the designated task. Its density $m_D(z \mid x)$ is supplied to the conditional adaptation procedure and may itself have been fitted upstream using the target sample D.

Source information S fixes an invertible coordinate map

$$\Psi_x : y \mapsto (z, u) = (T_x(y), U_x(y)), \qquad u \in \mathbb{R}^7. \tag{1}$$

The map in our implementation is linear and independent of x. Write $\Phi_x = \Psi_x^{-1}$ and define the *encoding* Jacobian $J_x(y) = |\det D\Psi_x(y)|$. For any conditional density H_D normalized in u, the adapted density is

$$q_D(y \mid x) = m_D(z \mid x) H_D(u \mid z) J_x(y), \qquad (z, u) = \Psi_x(y). \tag{2}$$

If the decoder is $y = M(z, u)$, then $J_x = |\det M|^{-1}$.

Proposition 1 (Task preservation)**.** *For every fitted sample D and input x, the pushforward of $q_D(\cdot \mid x)$ under T_x has density $m_D(\cdot \mid x)$.*

This follows from change of variables and $\int H_D(u \mid z)\, du = 1$ (Appendix A.1). It needs no sampling assumption. The identity preserves the supplied law, including any error it contains.

We use a pooled conditional module $H_D(u \mid z)$ to keep the target-side fitting problem small. Allowing $H_D(u \mid z, x)$ preserves the same factorization. Section 4 identifies the statistical cost of omitting x.

All conditional competitors use the same map and supplied marginal at each budget. Their paired log-score differences are therefore exactly conditional log-score differences:

$$\log \frac{q_D(y \mid x)}{q'_D(y \mid x)} = \log \frac{H_D(u \mid z)}{H'_D(u \mid z)}.$$

Scores are in nats per eight-coefficient residual vector. Negative NLL differences favor the first-named predictor.

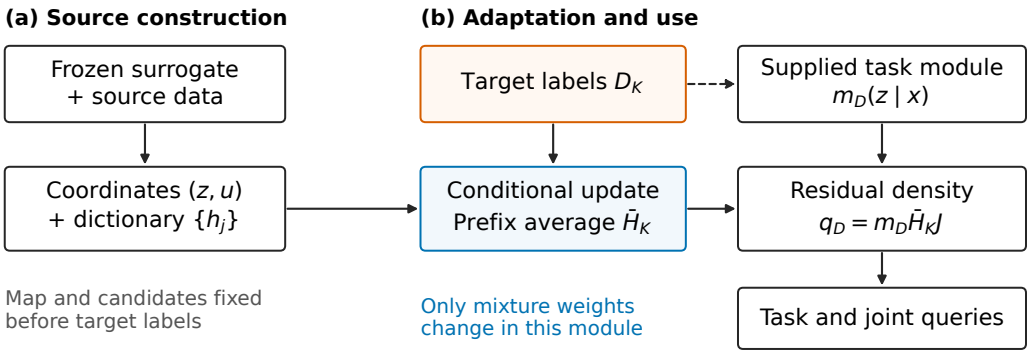


Figure 1: A supplied task distribution and an adapted conditional density form one residual predictor. (a) Source data fix the coordinates and candidate conditional densities. (b) Target labels update their mixture weights; averaging the resulting densities gives $\bar{H}_K$. The task module supplies m_D and may use the same labels upstream (dashed arrow). Combining the modules through the source map gives $q_D = m_D \bar{H}_K J_x$, preserving the supplied task marginal while supporting additional joint queries on the eight-coefficient residual.

3 Source-defined conditional adaptation

3.1 Coordinates and candidate corrections

Source second-moment geometry defines a task coordinate z and seven residual coordinates u, scaled so that the source residual covariance is the identity. A source finite-difference calculation measures how correlating each pair of forcing modes changes task–residual cross-covariance in a local linearization. It ranks all 28 pairs by the norm of that response (Appendix A.14); Appendix A.16 relates it to the full nonlinear covariance response. We select the pair with the largest response in this averaged source linearization, take its unit direction $v \in \mathbb{R}^7$, and build the single-direction dictionary along it. Appendix A.16 identifies the curvature terms omitted from this ranking. The map, directions and scalar standardization $\phi_S(z)$ are fixed before any target label is read.

The source reference h_{ref} is a likelihood-weighted mixture of 86 radial Gaussian or Student densities fitted on source calibration residuals. It retains all labelled components, including a duplicate unit Gaussian. The target dictionary contains this reference and 85 radial alternatives $h_{a,\nu}$. Here a multiplies the covariance and ν controls the tails. The grid has 17 logarithmically spaced values from 0.25 to 4 and $\nu \in \{3, 5, 10, 30, \infty\}$.

A directed correction shifts the conditional location along v and changes its spread along that direction. For a radial base h, define

$$h_{b,s}(u \mid z) = s^{-1} h\big(A_s^{-1}[u - b\phi_S(z)v]\big)\,, \qquad A_s = I + (s-1)vv^\top. \tag{3}$$

The factor s^{-1} is the inverse volume change, since $\det A_s = s > 0$. We use $b \in \{-0.8, -0.4, 0.4, 0.8\}$, with $s = 1$ and $s = \sqrt{1 - b^2}$ for each b. Applying these eight transforms to the reference gives directed candidates; applying them to each radial alternative gives composed candidates. The resulting dictionary, denoted C774 after the number of candidates it holds, has $1 + 85 + 8 + 680 = 774$ labelled densities. Appendix A.5 specifies their normalized densities and source weights.

3.2 Likelihood updating and the deployed predictor

The source-fixed prior puts mass 0.8 on the reference, 0.05 on radial candidates, 0.05 on directed candidates and 0.10 on composed candidates. Within each group, radial weights follow a fixed scale profile and the eight transforms receive equal mass. To obtain an expected log-risk guarantee relative to the source reference at each label budget, we use progressive mixing. Given $D_K = \{(z_i, u_i)\}_{i=1}^K$,

likelihood updates only the weights:

$$w_{t,j} = \frac{\pi_j \prod_{i=1}^{t} h_j(u_i \mid z_i)}{\sum_\ell \pi_\ell \prod_{i=1}^{t} h_\ell(u_i \mid z_i)}, \qquad H_t(u \mid z) = \sum_j w_{t,j} h_j(u \mid z), \quad w_{0,j} = \pi_j. \tag{4}$$

We deploy the prefix average

$$\bar{H}_K = \frac{1}{K+1} \sum_{t=0}^{K} H_t = \sum_j \bar{w}_j h_j, \qquad \bar{w}_j = \frac{1}{K+1} \sum_{t=0}^{K} w_{t,j}. \tag{5}$$

This choice admits the expected source-reference guarantee in Section 4. It stores one weight vector and requires no ensemble of fitted networks. We compare it with the final posterior H_K and the endpoint mixture $(H_0 + KH_K)/(K+1)$ in Section 5.

Sampling draws z from m_D, an index j from $\bar{w}$, and u from $h_j(\cdot \mid z)$, then applies the decoder. In the experiments the shared task module is a Student predictive fitted to the first K scalar task observations, and every conditional competitor carries it.

3.3 Expanding directional coverage

The same source response calculation supplies all 28 directions. A broader dictionary, C28, keeps the 86 undirected candidates and copies the 688 directed and composed candidates along each direction; its name counts the directions, not the candidates. It contains $86 + 28 \times 688 = 19{,}350$ labelled candidates. The total directional prior mass stays at 0.15 and is divided equally across the 28 directions. Target data do not select or remove directions. This construction lets us examine the effect of increasing coverage while holding the label budget and total directional prior mass fixed.

4 Representation and estimation limits

4.1 The error accessible to a pooled conditional model

Let $P = P_{XZU}$ be the law of a target observation. For a fixed fitted sample D, two KL chain rules give

$$\underbrace{\mathbb{E}_X \, \mathrm{KL}(P_{ZU|X} \| m_D H_D)}_{\text{joint predictive error}} = \underbrace{\mathbb{E}_X \, \mathrm{KL}(P_{Z|X} \| m_D)}_{\text{task marginal}} + \underbrace{I_P(U; X \mid Z)}_{\text{input pooling}} + \underbrace{\mathbb{E}_Z \, \mathrm{KL}(P_{U|Z} \| H_D)}_{\text{conditional approximation}}. \tag{6}$$

The identity holds in $[0, +\infty]$ after averaging over X, without an i.i.d. assumption (Appendix A.2). With the supplied marginal and coordinates fixed, fitting H_D changes only the last term. An input-dependent conditional model can also address the pooling term. Write $A(h) = \mathbb{E}_Z \, \mathrm{KL}(P_{U|Z} \| h)$ for conditional KL risk.

4.2 A shared symmetry constrains the whole dictionary

Let $G_v = \{Q \in O(7) : Qv = v\}$, including reflections. Radial densities are orthogonally invariant, and the transforms in (3) preserve invariance under G_v. Every candidate and every mixture in $\mathcal{C} = \mathrm{conv}(\mathrm{C774})$ thus satisfies $h(Qu \mid z) = h(u \mid z)$. Group averaging a target preserves its component along v and its perpendicular radius, while averaging the perpendicular orientation.

Proposition 2 (Symmetry barrier)**.** *Let $P^G_{U|Z}$ be the group-averaged target. Every $h \in \mathcal{C}$ satisfies*

$$A(h) = \underbrace{\mathbb{E}_Z \, \mathrm{KL}(P_{U|Z} \| P^G_{U|Z})}_{B_G(P)} + \mathbb{E}_Z \, \mathrm{KL}(P^G_{U|Z} \| h). \tag{7}$$

Both terms are nonnegative, including when a divergence is infinite.

The orbit-conditioning proof is in Appendix A.4. The first term is fixed by the target and the shared symmetry. More labels or a different estimator within the hull cannot reduce it. For example, every candidate assigns probability $1/2$ to $t^\top U > 0$ conditional on Z, for every unit $t \perp v$.

To quantify this restriction, let several source directions span V, and let $G = G_V$ fix V pointwise. Its orbit statistic is $\mathcal{T} = (Z, \Pi_V U, \|\Pi_{V^\perp} U\|)$.

Proposition 3 (Local symmetry cost). *For an invariant reference P_0 on (Z, U), let $dP_\epsilon = (1 + \epsilon s)dP_0$, where s is bounded, $\mathbb{E}_0 s = 0$ and $\alpha = |\epsilon|\|s\|_\infty < 1$. Put $s_G = \mathbb{E}_0[s \mid \mathcal{T}]$. Then*

$$\frac{\epsilon^2 \|s - s_G\|^2_{L^2(P_0)}}{2(1+\alpha)} \le B_G(P_\epsilon) \le \frac{\epsilon^2 \|s - s_G\|^2_{L^2(P_0)}}{2(1-\alpha)}. \tag{8}$$

The part of the density change removed by orbit averaging determines the local barrier. A change depending on Z alone contributes zero to it. Appendix A.15 proves the bound, extends it to nonsymmetric references, and separates mean displacement from covariance anisotropy in a local Gaussian model. These bounds concern the invariant class; a finite grid can add approximation error. Enlarging V relaxes the symmetry without making the dictionary universal.

The geometry becomes explicit in a local Gaussian model. Let $F \sim N(0,1)$ and $U \mid F \sim N(\epsilon F c, aI + \epsilon\Delta)$, with Δ symmetric. If Δ_G is the group average of Δ, then

$$B_G(P_\epsilon) = \epsilon^2 \left[\frac{\|\Pi_{V^\perp} c\|^2}{2a} + \frac{\|\Delta - \Delta_G\|^2_{\mathrm{F}}}{4a^2} \right] + o(\epsilon^2). \tag{9}$$

The first term measures missed task-dependent displacement. The second measures covariance changes excluded by the symmetry and can remain positive with no mean shift. Appendix A.15 gives the proof and an analytic illustration. This model clarifies which changes new directions allow; it does not assume the PDE residuals are Gaussian.

The 94-candidate Union hull has a separate restriction: positive Student tails in its transformed references cannot be canceled to reproduce a shifted Gaussian composed target. Appendix A.8 proves its strict separation from that target.

4.3 An expected guarantee for source-informed aggregation

Condition on the frozen source. Assume i.i.d. calibration and fresh test pairs, positive candidates on the target support, and a finite-risk comparison (Appendix A.3).

Theorem 1 (Classical prefix oracle). *For every probability vector $\nu \ll \pi$,*

$$\mathbb{E}_D A(\bar{H}_K) \le \sum_j \nu_j A(h_j) + \frac{\mathrm{KL}(\nu\|\pi)}{K+1}. \tag{10}$$

This is the progressive-mixture bound (Yang, 2000; Juditsky et al., 2008). The source-reference choice gives

$$\mathbb{E}_D A(\bar{H}_K) - A(h_{\mathrm{ref}}) \le \frac{-\log \pi_{\mathrm{ref}}}{K+1}, \tag{11}$$

which is 0.006762 nats at $\pi_{\mathrm{ref}} = 0.8$, $K = 32$. The expectation is over calibration samples. The oracle averages individual candidate risks; it does not compare to the optimal mixture risk. Appendix A.7 gives an example with persistent excess risk above the hull optimum. Appendix A.17 separates the gain from density averaging from the contribution of earlier predictors, explaining how the prefix and last posterior can differ.

4.4 What directional expansion changes

C28 has an exact hierarchical interpretation: it mixes 28 single-direction blocks, with posterior block weights proportional to their sequential evidence. The cumulative sequential log-loss cost relative to any one block is at most $\log 28$. Passing to deployed prefix risks requires a Jensen correction (Appendix A.17); $\log(28)/(K+1)$ alone is not their risk-difference bound.

A prescribed target $F = \phi_S(Z) \sim N(0,1)$, $U \mid F \sim N(bFw, a[I + (s^2-1)ww^\top])$ makes the representation comparison explicit. For the three specified off-axis directions, reflection gives $A(h) \ge L_{\mathrm{sym}} = \arcsin(\rho)/\pi$ throughout the C774 hull. C28 contains a matched candidate with prior $\pi_\star$, giving $\mathbb{E}A(\bar{H}^{28}_K) \le U_K = -\log \pi_\star/(K+1)$. The inequality $U_K < L_{\mathrm{sym}}$ yields sufficient budgets of 88, 81 and 78 labels (Appendix A.6). These concern that Gaussian construction. A magnitude-comparison event also detects off-axis shape changes for centered heavy-tailed elliptical targets (Appendix A.18). Section 5 separately measures adaptation on the PDE.

Control	C774 − reference	C774 − Union	C774 − mixture-MLE
No shift	$+0.0005$ $[-0.0003, +0.0013]$	-0.0001 $[-0.0003, +0.0001]$	-0.0746 $[-0.0787, -0.0705]$
Tail shift	-0.3802 $[-0.4093, -0.3511]$	$+0.0051$ $[+0.0042, +0.0060]$	-0.1021 $[-0.1128, -0.0913]$
Dependence shift	-0.0247 $[-0.0269, -0.0225]$	$+0.0011$ $[+0.0004, +0.0018]$	-0.0121 $[-0.0170, -0.0073]$
Combined shift	-0.4503 $[-0.4820, -0.4185]$	-0.0501 $[-0.0554, -0.0447]$	-0.0717 $[-0.0813, -0.0622]$
Weak dependence	-0.0037 $[-0.0051, -0.0023]$	$+0.0005$ $[+0.0002, +0.0008]$	-0.0553 $[-0.0620, -0.0487]$
Outside primary direction	$+0.0000$ $[-0.0013, +0.0013]$	-0.0001 $[-0.0004, +0.0002]$	-0.0738 $[-0.0766, -0.0710]$

Table 1: Conditional NLL differences at $K = 32$ under six Burgers forcing controls. Entries are C774 minus the named predictor, in nats per eight-coefficient residual vector; lower is better. Brackets give unadjusted 95% paired t_9 intervals over ten seed-level averages, each averaging 128 calibration draws. Union uses 94 active candidates and a redistributed prior. Mixture MLE fits the same 774 candidates without a prior penalty.

5 Controlled adaptation experiments

5.1 Protocol and comparisons

We use frozen Fourier neural operators (Li et al., 2021) on viscous Burgers and score the eight-coefficient residual defined in Section 2. The source problem has a 64-point periodic grid, viscosity 0.05, final time 0.5 and 64 internal steps. Ten source models are trained on 2,048 fields each and selected by validation MSE on 256 fields. All adaptation takes place after these models are frozen.

The controls modify eight latent forcing coefficients. A dependence shift correlates one source-prescribed pair; a tail shift multiplies all eight coefficients by a common variance-preserving Student scale. The six controls are no shift, tail, dependence, both, weaker dependence and dependence on an outside-primary pair. The combined control uses correlation 0.6 and five-degree-of-freedom tails. Appendix A.14 gives the complete shift rules and comparator definitions.

For each control and training seed, 128 independent calibration draws are paired with the same 4,096 held-out residuals. Budgets $K = 8, 16, 32, 64, 128$ use nested prefixes of each draw. We first average differences across the 128 draws within a seed, then report the mean across ten seeds with an unadjusted two-sided 95% paired t_9 interval. The seed is the unit of inference. All comparisons use the same supplied task marginal and coordinate map.

Comparators serve separate purposes. The source reference measures adaptation. Union measures the joint change in candidate composition and prior allocation. Mixture MLE, the final posterior and the endpoint mixture compare estimators on the same 774 candidates. Three further comparators are continuous rather than dictionary-based. Two of them regress the residual on the task coordinate and shrink toward source-fitted coefficients and covariance, once with a Gaussian and once with a Student density. The third fits a conditional Student along the source direction. Appendix A.14 specifies what source information each one uses.

5.2 Adapting under tail and dependence changes

Under the combined shift at $K = 32$, C774 reduces held-out conditional NLL by 0.4503 nats per residual vector relative to the source reference, with paired difference $[-0.4820, -0.4185]$. The gain increases from 0.3389 nats at eight labels to 0.5181 at 128. The no-shift difference is $+0.000526$ $[-0.000260, +0.001311]$. The outside-primary control gives $+0.000031$ $[-0.001265, +0.001327]$.

The composed dictionary scores 0.0501 nats below Union under the combined shift. Union scores 0.0051 lower under tail shift and 0.0011 lower under dependence shift. This pattern supports composed corrections for simultaneous changes. The comparison changes both the available candidates and their prior allocation, so it measures the resulting constructions together.

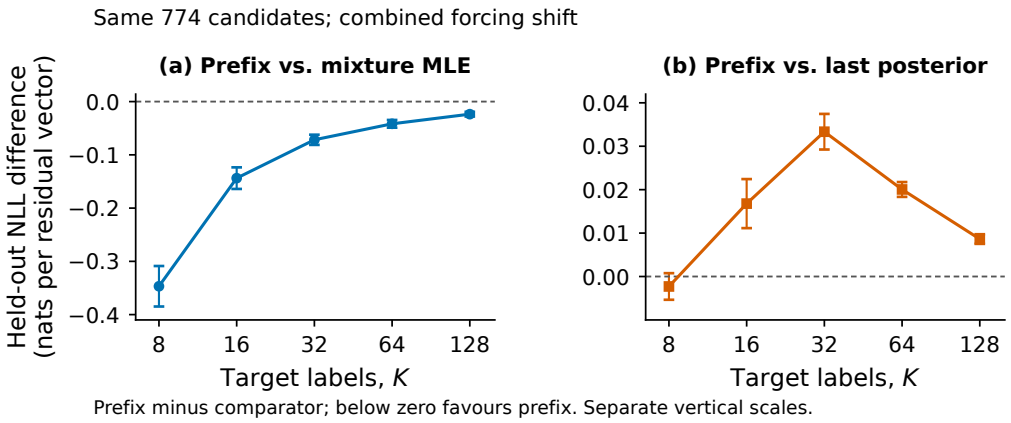


Figure 2: Weight estimation and prefix averaging have different effects within the same dictionary. Under the combined shift, the prefix average scores lower than mixture-weight maximum likelihood at every measured budget (a), but higher than the last posterior at the measured budgets $K \geq 16$ (b). Points are prefix minus comparator in total nats per residual vector; negative values favour the prefix average. The panels use separate vertical scales. Bars are unadjusted 95% paired t intervals across ten training seeds, after averaging 128 calibration draws within each seed. Lines connect measured budgets only.

5.3 Source-informed estimation and prefix averaging

Against unregularized mixture MLE on the same dictionary, the prefix average reduces combined-shift NLL by 0.0717 nats at $K = 32$, with paired interval $[-0.0813, -0.0622]$. The gap narrows from 0.3468 at eight labels to 0.0236 at 128. Every recorded fit in the two-control five-budget MLE comparison meets the all-candidate optimality-gap tolerance of 10^{-7}; the maximum recorded gap is 9.98×10^{-11}. This certifies numerical optimization of calibration NLL.

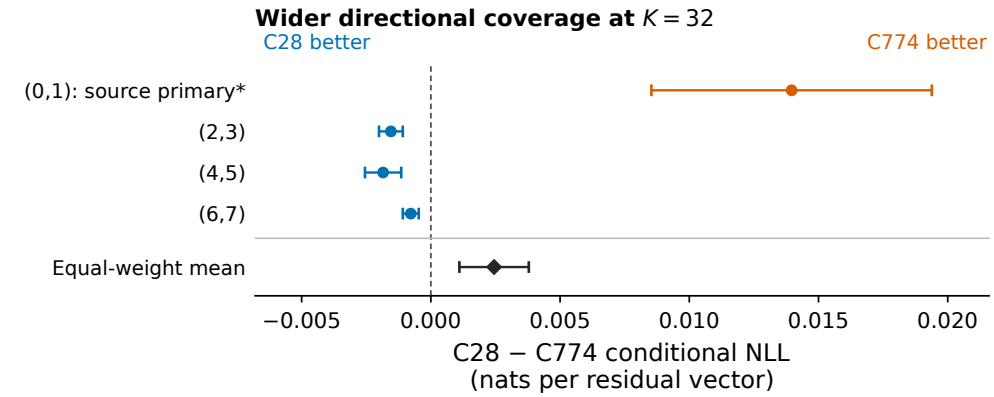
Continuous conditional alternatives also use source information. At $K = 32$ under the combined shift, C774 scores 0.1315 nats below the source-direction conditional Student fit, 1.4653 below source-informed Gaussian regression and 0.9705 below source-informed Student regression. Their paired intervals are, respectively, $[-0.1469, -0.1161]$, $[-1.5215, -1.4091]$ and $[-0.9931, -0.9478]$ (Appendix A.10). The two regressions shrink toward source-fitted coefficients and covariance; the directional model fits its location, variances and tail parameters on the target sample. These comparisons support the source-defined dictionary as a useful small-sample construction. They do not match its discrete 0.8 reference prior exactly.

Prefix averaging has a distinct role. Relative to the final posterior, it gives a $+0.0333$-nat difference at $K = 32$ and has higher NLL at the measured budgets $K \geq 16$. Relative to the endpoint mixture, the $K = 32$ difference is $+0.0353$ nats. Both comparisons retain the candidate dictionary and source prior. We use prefix averaging for its expected reference guarantee; the final and endpoint predictors provide stronger measured log scores in this regime. The exact averaging decomposition is in Appendix A.17; all five-budget results are in Appendix A.10.

5.4 The cost and benefit of directional coverage

C28 distributes the directional prior across 28 source directions. At $K = 32$, it improves on C774 for prescribed pairs $(2, 3)$, $(4, 5)$ and $(6, 7)$ by 0.001547, 0.001848 and 0.000778 nats. Pair $(0, 1)$, source-primary in nine of ten seeds, incurs a 0.013960-nat increase. The equal-weight average over four pairs is $+0.002447$ $[+0.001103, +0.003790]$, favoring C774. Under the combined control, C28 also scores higher by 0.024815 nats.

Broader coverage permits conditional changes outside the original direction, while the fixed label budget must allocate weight across a larger family. Three shifts improve and the source-preferred one does not, so coverage alone does not determine finite-sample performance. The analytic suffi-



* Source primary in 9 of 10 seeds. Bars: paired 95% intervals across 10 seeds.

Figure 3: Broader coverage improves three prescribed direction pairs, while the larger loss on $(0, 1)$ leaves the equal-weight aggregate worse. Points show C28 minus C774 conditional NLL at $K = 32$, in total nats per residual vector. Pair $(0, 1)$ is source-primary in nine of ten seeds. The diamond weights all four comparisons equally within each seed. Bars are unadjusted 95% paired t intervals across ten seeds. This measured comparison is distinct from the Gaussian analytic bound in Figure 4.

Query	C774 Brier	Radial KDE Brier	Paired difference	95% interval
Both exceed	0.012578	0.012594	−0.000016	[−0.000020, −0.000011]
Either exceeds	0.254975	0.253559	+0.001416	[+0.001196, +0.001635]

Table 2: Brier scores for two regional events under the combined shift at $K = 32$. Differences are C774 minus radial residual KDE, with unadjusted 95% paired t_9 intervals over ten seed-level averages. Brier scores are dimensionless; lower is better. Both methods share the task marginal and source-fixed event thresholds.

cient budgets concern a separate prescribed Gaussian target. Additional Gaussian comparisons in Appendix A.11 do not locate an empirical counterpart to those thresholds.

5.5 Joint queries and a second frozen architecture

We also evaluate joint regional events on the eight-coefficient residual. The two regional means use source-fixed 75th-percentile thresholds. The probabilities are obtained from the same fitted joint density. Against a radial residual KDE, the Brier difference is -0.000016 for both regions exceeding their thresholds and $+0.001416$ for either region exceeding (Table 2). The two outcomes show that conditional log-score improvement does not imply improvement for every event. Empirical event frequencies were not retained, so comparison with a constant event-frequency predictor is unavailable.

A custom multiscale convolutional surrogate provides a second frozen architecture. Its configuration is given in Appendix A.14. Across five paired seeds at $K = 32$, the difference in within-architecture NLL improvements is $+0.1505$ nats $[-0.0011, +0.3020]$. Each architecture uses its own source-whitened coordinates; the comparison concerns improvements within each coordinate system. These regional and architecture checks are follow-up measurements with unadjusted intervals.

6 Related work

Probabilistic modules and recalibration. Markov melding joins submodels through shared quantities and makes their marginal specification explicit (Goudie et al., 2019). Multivariate latent recalibration uses a post-hoc transformation to improve a conditional flow's latent calibration while retaining an explicit predictive density (Dheur & Ben Taieb, 2025). These works motivate separating a reported marginal from remaining distributional structure. Here the marginal is supplied, and

source data specify a finite conditional family whose target-side adaptation and common symmetries can be analyzed directly.

Finite-dictionary aggregation. The expected oracle bound is classical progressive mixing (Yang, 2000) and mirror averaging (Juditsky et al., 2008). Its comparison is with individual candidate risks. Convex aggregation asks a different question: estimating a density in the mixture class. Compton et al. (2026) study high-probability guarantees for model and convex aggregation without bounded density ratios. Our representation lower bound applies to the entire source-defined convex hull, while the deployed prefix estimator retains an atomic-risk upper bound.

Symmetry and approximation. Group averaging yields generalization benefits when the target respects the symmetry (Elesedy & Zaidi, 2021). Recent variational-inference analyses recover statistics under target symmetries (Marks et al., 2026; Zellinger & Vergari, 2026). Here the conditional dictionary imposes a symmetry that a target shift can violate. We quantify the resulting approximation cost and connect its local change to conditional expectations of density derivatives.

Uncertainty for frozen scientific surrogates. Linearizing a neural operator produces a function-valued Gaussian process (Magnani et al., 2025). REEF-GP fits residual uncertainty in the learned geometry-aware features of a frozen operator (Vendrell-Gallart et al., 2026). Flow-based conformal predictive distributions connect sampling to conformal prediction regions (Harris, 2026). Our predictive object is a finite residual-coefficient density with a prescribed scalar marginal; our guarantees concern preservation, conditional KL risk and dictionary approximation.

7 Conclusion

A supplied task distribution leaves conditional residual structure available for adaptation. Source-defined conditional densities use few target labels while preserving that task law. Their shared symmetries determine which changes any mixture can express, and the part of a local density perturbation that violates those symmetries determines an irreducible KL cost. The prefix guarantee controls a separate estimation question. Burgers experiments support adaptation under controlled tail and dependence shifts and measure the finite-budget effects of directional expansion. The design principle is to match conditional geometry to the changes of interest, then evaluate how the available labels estimate within that family.

References

Spencer Compton, Gábor Lugosi, Jaouad Mourtada, Jian Qian, and Nikita Zhivotovskiy. Ratio covers of convex sets and optimal mixture density estimation. *arXiv preprint arXiv:2602.16142*, 2026. Version 2.

Victor Dheur and Souhaib Ben Taieb. Multivariate latent recalibration for conditional normalizing flows. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.16636.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021.

Robert J. B. Goudie, Anne M. Presanis, David Lunn, Daniela De Angelis, and Lorenz Wernisch. Joining and splitting models with Markov melding. *Bayesian Analysis*, 14(1):81–109, 2019. doi: 10.1214/18-BA1104.

Trevor A. Harris. Flow-based conformal predictive distributions. *arXiv preprint arXiv:2602.07633*, 2026. Version 3, 7 May 2026.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008. arXiv:math/0511468.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. arXiv:2010.08895.

Emilia Magnani, Marvin Pförtner, Tobias Weber, and Philipp Hennig. Linearization turns neural operators into function-valued Gaussian processes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42421–42452, 2025.

Daniel Marks, Dario Paccagnan, and Mark van der Wilk. Symmetry guarantees statistic recovery in variational inference. *arXiv preprint arXiv:2604.18310*, 2026. Version 1.

David Pollard. A note on insufficiency and the preservation of Fisher information. *arXiv preprint arXiv:1107.3797*, 2012. Version 2.

Oriol Vendrell-Gallart, Nima Negarandeh, and Ramin Bostanabad. Geometry-aware post-hoc uncertainty quantification in operator learning. *arXiv preprint arXiv:2606.17513*, 2026. Submitted 16 June 2026.

Felix Voigtlaender. A general version of Price's theorem. *Journal of Theoretical Probability*, 34: 1474–1485, 2021. doi: 10.1007/s10959-020-01017-w. arXiv:1710.03576.

Yuhong Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000.

Lena Zellinger and Antonio Vergari. Even more guarantees for variational inference in the presence of symmetries. *arXiv preprint arXiv:2604.21407*, 2026. Version 2.

Statement on the use of generative AI

In this work, we used generative AI tools for the following tasks, for which ICLR requires disclosure: checking mathematical claims and proofs and correcting errors identified in them; assisting in the writing of proofs; providing feedback on the research methodology and the experiment design; implementing methods; and interpreting results. We additionally used generative AI tools for tasks with recommended disclosure: analysing data; running experiments; creating and editing software code; producing figures; drafting and editing the text; and formatting references. The authors have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims and artefacts.

No reviewer outside the author list has checked the theoretical statements or the proofs.

A Proofs and implementation details

This appendix is written to be read in pieces. Each subsection states what it settles and what it assumes, so a reader checking one claim does not need the others.

- **Checking the method.** Appendix A.1 proves that the constructed density keeps the supplied task marginal, and records how a change of coordinates acts on the conditional density. Appendix A.5 lists the map, the normalized densities, the priors and the source data splits. Appendix A.14 defines the comparators and the data generation.
- **Checking the analysis.** Appendix A.2 proves the three-term decomposition. Appendix A.3 proves the prefix bound and lists its assumptions self-containedly. Appendix A.4 proves that every density in the convex hull is invariant under the same symmetry group, and Appendix A.7 shows that the estimation gap need not vanish.
- **Checking what the dictionary cannot express.** Appendix A.8 proves the strict gap against the 94-candidate Union hull. Appendix A.6 proves the off-axis bound and reports the bracket on the hull optimum. Appendix A.15 quantifies the cost of the changes the shared symmetry excludes, Appendix A.16 states what the stored direction ranking actually measures, Appendix A.17 separates prior allocation from prefix averaging, and Appendix A.18 treats joint events.
- **Checking the numbers.** Appendix A.10 gives the full budget ladders. Appendix A.11 gives the ten prescribed Gaussian scenarios at all nine measured budgets. Appendix A.9 records the multi-direction prior and its cost.
- **Checking the scope.** Appendix A.12 states what each claim covers and what it does not. Appendix A.13 lists the stored records and what a re-run can settle.

A.1 Task preservation and coordinate changes

Fix D and x. Let Ψ_x be a differentiable bijection with differentiable inverse Φ_x, or an invertible linear map as used here. The product $m_D(z \mid x)H_D(u \mid z)$ is a probability density on (z, u)-space. For a Borel set $B \subseteq \mathbb{R}$,

$$\begin{aligned}\int_{\{y:T_x(y)\in B\}} q_D(y \mid x)\, dy &= \int_{B\times\mathbb{R}^7} m_D(z \mid x)H_D(u \mid z)\, dz\, du \\ &= \int_B m_D(z \mid x)\, dz.\end{aligned}$$

The change of variables uses $dy = |\det D\Phi_x(z, u)|\, dz\, du$ and $J_x(\Phi_x(z, u)) = |\det D\Phi_x(z, u)|^{-1}$. This proves Proposition 1 for every D, x, without a distributional assumption on the calibration sample.

A change of the remaining coordinates changes their density numerically. If $u' = Gu$ for an invertible 7×7 matrix G, then

$$H'_D(u' \mid z) = H_D(G^{-1}u' \mid z)|\det G|^{-1}, \qquad J'_x = |\det G|J_x.$$

Their product in (2) is unchanged. This is the sense in which the conditional law, rather than its numerical density in a particular coordinate system, is invariant to a coordinate change.

A.2 Risk decomposition under input pooling

Let $P = P_{XZU}$ and define two probability measures on the same space:

$$Q_D = P_X(dx)m_D(dz \mid x)H_D(du \mid z), \qquad R_D = P_{XZ}(dx, dz)H_D(du \mid z).$$

The KL chain rule first conditions on (X, Z):

$$\mathrm{KL}(P\|Q_D) = \mathrm{KL}(P_{XZ}\|P_X m_D) + \mathrm{KL}(P\|R_D). \tag{12}$$

The conditional law of U given (X, Z) under both comparison measures is $H_D(\cdot \mid Z)$. A second chain rule conditions on (Z, U). Under R_D, the (Z, U) marginal is $P_Z H_D$ and the conditional law of X is $P_{X|Z}$:

$$\mathrm{KL}(P\|R_D) = \mathrm{KL}(P_{ZU}\|P_Z H_D) + \mathbb{E}_{Z,U}\,\mathrm{KL}(P_{X|Z,U}\|P_{X|Z}). \tag{13}$$

The last term is $I_P(U; X \mid Z)$. Combining the two identities proves (6). Every term is a divergence of probability measures, so the identity holds in $[0, +\infty]$. If the total is finite, all three terms are finite. No signed log-ratio needs to be integrated term by term.

The averaging over X is essential. For example, let Z be constant, X uniform on $\{0, 1\}$, $U \mid X = 0 \sim \mathrm{Bern}(0.2)$, $U \mid X = 1 \sim \mathrm{Bern}(0.8)$ and $H = \mathrm{Bern}(0.4)$. The correct input-averaged total is 0.2131557543. At $X = 0$, $\mathrm{KL}(\mathrm{Bern}(0.2)\|\,\mathrm{Bern}(0.4)) = 0.0915162218$. A fixed-input decomposition that replaces the second argument by the pooled law and then adds the pooled KL does not give this value.

A.3 Expected prefix-oracle inequality

For completeness, the following sufficient conditions state all the assumptions used for the experimental application. The proof below is formulated entirely in KL divergence.

Assumption 1. *Conditional on the source information S:*

(A1) *The finite dictionary $\{h_j\}$ and prior π are fixed and S-measurable; $\pi_j \geq 0$, $\sum_j \pi_j = 1$, and $\pi_{\mathrm{ref}} > 0$.*

(A2) *$O_1, \ldots, O_N$, with $O_i = (Z_i, U_i)$ and $N = K + 1$, are i.i.d. with law P.*

(A3) *Each h_j with positive prior is a measurable conditional density, normalized in u and positive on the target support. Its likelihood is finite there.*

(A4) $\mathbb{E}|\log h_{\mathrm{ref}}(U \mid Z)| < \infty$.

(A5) *A finite M bounds all candidate densities on the target support:* $h_j(u \mid z) \le M$.

(A6) *For the comparison vector $\nu \ll \pi$,* $\sum_j \nu_j A(h_j) < \infty$ *and* $\mathbb{E}|\log h_j(U \mid Z)| < \infty$ *whenever* $\nu_j > 0$.

Translations do not change the supremum of a radial density. The finite scale and tail grid, together with $s \ge 0.6$, gives a finite uniform bound in (A5). Joint decay as $\|(z,u)\| \to \infty$ is not required. The KL argument below uses (A1)–(A3) and the finite-risk part of (A6); the remaining conditions are sufficient regularity conditions for the stated application.

Define $Q_j(dz, du) = P_Z(dz)h_j(u \mid z)\, du$, so $\mathrm{KL}(P\|Q_j) = A(h_j)$. The sequence comparison law is

$$M_\pi = \sum_j \pi_j Q_j^{\otimes N}. \tag{14}$$

It draws one candidate and uses it for the entire sequence. To compare it with $P^{\otimes N}$, introduce probability measures on an index and a sequence:

$$R_\nu(\{j\}, do_{1:N}) = \nu_j P^{\otimes N}(do_{1:N}), \qquad S_\pi(\{j\}, do_{1:N}) = \pi_j Q_j^{\otimes N}(do_{1:N}).$$

The chain rule on the index and product tensorization give

$$\mathrm{KL}(R_\nu\|S_\pi) = \mathrm{KL}(\nu\|\pi) + N\sum_j \nu_j A(h_j).$$

Marginalizing the index and applying data processing yields

$$\mathrm{KL}(P^{\otimes N}\|M_\pi) \le \mathrm{KL}(\nu\|\pi) + N\sum_j \nu_j A(h_j). \tag{15}$$

This proves the variational step directly for relative entropy, without subtracting an entropy of P.

Given the first t observations under M_π, the next observation has law $P_Z(dz)H_t(du \mid z)$. The common P_Z factors cancel in the posterior weights. Thus, relative to $(P_Z \otimes du)^{\otimes N}$, the density of M_π is $\prod_{t=0}^{K} H_t(u_{t+1} \mid z_{t+1})$. The sequential KL chain rule gives

$$\mathrm{KL}(P^{\otimes N}\|M_\pi) = \sum_{t=0}^{K} \mathbb{E}A(H_t). \tag{16}$$

The fresh $(K+1)$st observation is used only in this proof; the deployed predictor uses $O_1, \ldots, O_K$.

Convexity in the second argument of KL implies, for each calibration sample,

$$A(\bar{H}_K) \le \frac{1}{N}\sum_{t=0}^{K} A(H_t).$$

Together with (15) and (16), this proves Theorem 1. The sum includes $t = 0$. The reference choice $\nu = \delta_{\mathrm{ref}}$ proves (11). In particular, its evidence bound is an inequality: $\sum_j \pi_j \prod_i h_j(u_i \mid z_i) \ge \pi_{\mathrm{ref}} \prod_i h_{\mathrm{ref}}(u_i \mid z_i)$.

The linear comparison $\sum_j \nu_j A(h_j)$ is an average of individual risks. Convexity gives $A(\sum_j \nu_j h_j) \le \sum_j \nu_j A(h_j)$, so replacing the latter by the former would strengthen the theorem. The Bernoulli example in Appendix A.7 rules out that general replacement with a vanishing remainder for this estimator. Corollary 5.4 of Juditsky et al. (2008) states the density-estimation, best-individual-candidate version with a uniform prior; the weighted version used here follows from the index argument above.

A.4 Symmetry and the approximation barrier

Lemma 1 (Candidate invariance)**.** *Every candidate in C774, and every density in its convex hull, is invariant under $G_v = \{Q \in O(7) : Qv = v\}$.*

Proof. A radial density is a function of $\|u\|^2$, and the source reference is a mixture of such densities. For a directed or composed candidate, write $r = u - b\phi_S(z)v$. The inverse transform is $A_s^{-1} = I + (s^{-1} - 1)vv^\top$, and

$$\|A_s^{-1}r\|^2 = \|r\|^2 + (s^{-2} - 1)(v^\top r)^2.$$

If $Qv = v$, then $v^\top Qu = v^\top u$ and $\|Qu\| = \|u\|$. Both terms in the displayed quadratic form are unchanged. The log-Jacobian is $-\log s$, since $\det A_s = s$. This determinant holds for any base density; radial symmetry is used only for the quadratic-form representation. Convex combinations preserve invariance. □

Proof of Proposition 2. Fix z. The group orbit is determined by

$$T(u) = \big(v^\top u,\ \|(I - vv^\top)u\|\big).$$

At a positive perpendicular radius, the orbit is a sphere in $v^\perp$. Let σ_t be its uniform probability measure; at radius zero it is the point mass on the singleton orbit. Haar averaging $P_z = P_{U|Z=z}$ leaves the law $P_{z,T}$ of T unchanged and replaces the conditional distribution on each orbit by σ_t. Hence $P_z^G = P_{z,T}\sigma_t$. Every invariant comparison density h_z has a disintegration $H_{z,T}\sigma_t$ with the same orbit kernel.

Applying the KL chain rule to T yields

$$\mathrm{KL}(P_z\|h_z) = \mathrm{KL}(P_{z,T}\|H_{z,T}) + \int \mathrm{KL}(P_z(\cdot \mid T = t)\|\sigma_t)\, P_{z,T}(dt).$$

The integral is $\mathrm{KL}(P_z\|P_z^G)$, and the first term is $\mathrm{KL}(P_z^G\|h_z)$. Each is nonnegative and the chain rule holds in $[0, +\infty]$, including singular targets. Averaging over Z proves (7). □

Several directions. For directions spanning V, the shared group is $G_V = \{Q \in O(7) : Q|_V = I_V\}$. The orbit statistic becomes $(\Pi_V u, \|\Pi_{V^\perp} u\|)$. If $V \subseteq W$, then $G_W \subseteq G_V$, and the class of invariant comparison laws increases. The minimum KL over all invariant laws is attained by the group-averaged target. It follows that $B_{G_W} \le B_{G_V}$. At $V = \mathbb{R}^7$ this particular barrier is zero, while finite-dictionary approximation can remain positive.

The same symmetry restricts moments whenever they exist. The conditional mean lies in V. The conditional covariance has zero cross-block between V and $V^\perp$, and its block on $V^\perp$ is a scalar multiple of the identity. These statements follow by reflecting any orthogonal direction and rotating within $V^\perp$. They provide direct interpretations of what the dictionary cannot represent.

Five nonnegative contributions. Suppose the relevant joint risk and convex-hull optimum are finite. Write

$$G_{\mathcal{C}} = \inf_{h \in \mathcal{C}} \mathbb{E}_Z\, \mathrm{KL}(P_{U|Z}^G\|h), \qquad E_{\mathcal{C},K} = \mathbb{E}_D A(\bar{H}_K) - \inf_{h \in \mathcal{C}} A(h).$$

Equation (7) gives $\inf_{h\in\mathcal{C}} A(h) = B_G + G_{\mathcal{C}}$. Combining with (6) and averaging over D gives

$$\mathbb{E}_D\, \mathrm{KL}(P\|Q_D) = \mathbb{E}_D\mathbb{E}_X\, \mathrm{KL}(P_{Z|X}\|m_D) + I_P(U; X \mid Z) + B_G + G_{\mathcal{C}} + E_{\mathcal{C},K}.$$

All five terms are nonnegative. The last is excess risk above the best density in the finite hull. It is not a KL divergence between the deployed density and a fitted projection, and it need not vanish with K. With infinite optima, retain the extended-valued chain-rule identities rather than forming differences of infinities.

A.5 Frozen coordinates, densities and source data

Coordinates. Let C be the positive-definite source second-moment matrix and let $W \neq 0$ be the task vector. Choose N with orthonormal columns spanning $\ker W^\top$, and set

$$A = \frac{CW}{W^\top CW}, \quad \Sigma_\perp = (I - AW^\top)C(I - AW^\top)^\top, \quad L = \mathrm{chol}(N^\top \Sigma_\perp N), \quad R = NL.$$

For a residual coefficient vector y, encoding gives $z = W^\top y$ and $u = L^{-1}N^\top(y - Az)$; decoding gives $y = Az + Ru$. The identities $W^\top A = 1$ and $W^\top R = 0$, together with invertibility of L, make $M = [A\ R]$ invertible. Its determinant satisfies

$$|\det M| = \frac{\det L}{\|W\|}, \qquad J = |\det M|^{-1}, \qquad \log q_D = \log m_D + \log H_D - \log|\det M|.$$

To obtain the factorization, use the orthonormal basis $[W/\|W\|, N]$ on the left of M; the resulting matrix is block triangular with diagonal blocks $1/\|W\|$ and L. Table 3 reports the decoder log-determinants.

Normalized radial densities. For dimension d, covariance multiplier $a > 0$ and finite $\nu > 2$, we use

$$h^{(d)}_{a,\nu}(u) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)[(\nu-2)\pi a]^{d/2}} \left(1 + \frac{\|u\|^2}{(\nu-2)a}\right)^{-(\nu+d)/2}.$$

It is a Student density with shape $a(\nu-2)I_d/\nu$ and covariance aI_d. At $\nu = \infty$ it is $N(0, aI_d)$. The 85 radial settings combine $a = \exp(\mathrm{linspace}(\log 0.25, \log 4, 17))$ and $\nu \in \{3, 5, 10, 30, \infty\}$. The inner source family adds one unit Gaussian, retaining the duplicate as a separate label.

Reference fitting and priors. Let $r_j \propto \exp[-\frac{1}{2}(\log a_j/\log 2)^2]$, normalized over the 85 labelled radial alternatives. The inner prior places 0.8 on the added unit Gaussian and $0.2r_j$ on those alternatives. Given the source residual coordinates, inner weights are proportional to this prior times each candidate's source likelihood. The reference density is the resulting 86-component mixture. The scalar standardization $\phi_S(z) = (z - z_{\mathrm{mean}})/z_{\mathrm{std}}$ is fitted on the same source calibration sample.

The eight ordered transforms are $(-0.8, 1)$, $(-0.8, 0.6)$, $(-0.4, 1)$, $(-0.4, \sqrt{0.84})$, $(0.4, 1)$, $(0.4, \sqrt{0.84})$, $(0.8, 1)$ and $(0.8, 0.6)$. Outer prior masses are 0.8, $0.05r_j$, $0.05/8$ and $0.10r_j/8$ for the reference, radial, directed and composed candidates. Union uses masses 0.8, $0.10r_j$, $0.10/8$ and zero on the same groups. Its active set has 94 candidates.

The supplied task law. At budget K, the experimental module is the location-scale Student law

$$m_D = t_{K-1}\big(\bar z, s_z\sqrt{1 + 1/K}\big), \qquad \bar z = K^{-1}\sum_i z_i, \quad s_z^2 = (K-1)^{-1}\sum_i (z_i - \bar z)^2.$$

Its second argument is a scale, not a variance. All conditional competitors share this fitted law. Paired NLL comparisons cancel it; regional event probabilities integrate against it.

Source and target samples. Per training seed, the source calibration sample contains 2,048 fields, and a disjoint check sample contains 2,048. A design sample of 128 inputs supplies the finite-difference response directions, using step 0.025. A half-step check and an independent direction check use 128 and 256 additional draws. They do not reselect directions. Training uses 2,048 fields and validation uses 256 over 60 epochs, selecting the minimum-validation-MSE checkpoint. Target calibration comprises 16,384 fields partitioned into 128 draws of 128 rows. Each budget is a nested prefix of a draw. The target test set contains 4,096 rows. Source construction and prediction are separate from held-out scoring.

A.6 Off-axis separation on a prescribed Gaussian target

Let v, w be source-fixed unit vectors, $c = v^\top w$ with $|c| < 1$, $\delta = 1 - c^2$, and $t = (w - cv)/\sqrt{\delta}$. The map $Q = I - 2tt^\top$ reflects across the hyperplane whose normal is t. It satisfies $Qt = -t$, $Qv = v$, $Q^\top Q = I$ and $\det Q = -1$.

Let $F = \phi_S(Z) \sim N(0, 1)$ and $U \mid F \sim N(bFw, a[I + (s^2 - 1)ww^\top])$, with $a, s > 0$ and $b \neq 0$. Here F is the standardized task coordinate in the prescribed probability model. Define

$$\tau^2 = a[1 + (s^2 - 1)\delta], \qquad \gamma = \frac{b\sqrt{\delta}}{\tau}, \qquad \rho = \frac{\gamma^2}{1 + \gamma^2}.$$

Table 3: Frozen source states for the ten training seeds. The determinant is for the decoder $M = [A\ R]$; the encoding log-Jacobian has the opposite sign. Main and outside denote source-prescribed forcing pairs. The outside-primary control is represented in the saved $K = 32$ summaries for all ten seeds. Scalar standardization is fitted on source calibration data.

Seed	Val. MSE	Epoch	Last val. MSE	Map log | det |	z_{mean}	z_{std}	Pairs (main / outside)
2026091401	0.073467836	1	0.131987	-18.596146507	-0.010292353	0.246543130	$(0, 1) / (3, 6)$
2026091402	0.076968864	7	0.126202	-18.677583826	-0.009711285	0.249447644	$(0, 1) / (0, 6)$
2026091403	0.072852567	3	0.131658	-18.674292416	0.001584146	0.245967333	$(0, 1) / (0, 6)$
2026091404	0.078100204	3	0.132497	-18.657310815	-0.005722546	0.245943176	$(0, 1) / (0, 6)$
2026091405	0.075641043	8	0.126093	-18.677745816	0.000410516	0.249443064	$(0, 1) / (3, 6)$
2026091406	0.075516097	8	0.130115	-18.647661992	-0.003706263	0.246386499	$(0, 1) / (3, 6)$
2026091407	0.081463084	7	0.135813	-18.666720068	-0.010866438	0.242386167	$(0, 1) / (3, 6)$
2026091408	0.072917700	3	0.122941	-18.675884626	-0.007685409	0.249713750	$(0, 4) / (5, 6)$
2026091409	0.072499163	12	0.126355	-18.592139528	0.001416515	0.239727625	$(0, 1) / (0, 6)$
2026091410	0.072632305	8	0.134238	-18.649799625	-0.000735014	0.248388551	$(0, 1) / (0, 6)$

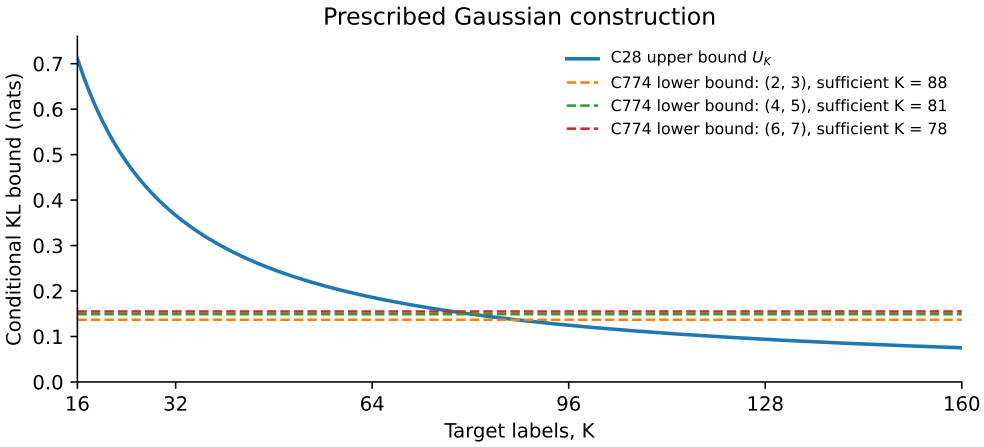


Figure 4: Analytic comparison on the prescribed Gaussian construction. The horizontal axis is the label budget K; the vertical axis is a bound on the expected conditional KL, in nats. The curve bounds the expected conditional KL of the C28 prefix predictor through a matched candidate. Each horizontal line lower-bounds every density in the C774 convex hull. Their strict inequalities yield the sufficient integer budgets listed in the legend. At $K = 32$ the upper bound exceeds all three lower bounds. No measured PDE loss appears in this figure.

Proposition 4 (Off-axis separation)**.** *Every rule $\widehat{H}_D$ whose output lies in* $\mathrm{conv}(\mathrm{C774})$ *satisfies*

$$\mathbb{E}_D A(\widehat{H}_D) \geq L_{\text{sym}} = \frac{\arcsin \rho}{\pi}.$$

If C28 contains a candidate matching the prescribed conditional Gaussian, with positive prior $\pi_\star$, then its prefix average satisfies $\mathbb{E}_D A(\bar{H}_K^{\mathrm{C28}}) \leq U_K = -\log \pi_\star/(K+1)$. The condition $U_K < L_{\text{sym}}$ separates that prefix predictor from every C774-hull-valued rule in expected risk.

Proof. A hull density is invariant under Q. Since it is absolutely continuous with respect to Lebesgue measure, the separating hyperplane has zero probability. The reflection exchanges its two sides, giving $\Pr_h(t^\top U > 0 \mid F) = 1/2$. The target projection has law $t^\top U \mid F \sim N(b\sqrt{\delta}F, \tau^2)$, so its positive-side probability is $p(F) = \Phi(\gamma F)$. Binary data processing and Pinsker's inequality yield

$$\mathrm{KL}(P_{U|F} \| h(\cdot \mid F)) \geq \mathrm{KL}(\mathrm{Bern}(p(F)) \| \mathrm{Bern}(1/2)) \geq 2(p(F) - 1/2)^2.$$

For independent standard normals E_1, E_2 independent of F, the variables $\gamma F - E_1$ and $\gamma F - E_2$ have correlation ρ. The Gaussian quadrant probability gives

$$\mathbb{E}\Phi(\gamma F)^2 = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}, \qquad \mathbb{E}\Phi(\gamma F) = \frac{1}{2}.$$

Taking expectations of the Pinsker bound proves $A(h) \geq \arcsin(\rho)/\pi$. The bound holds for every hull member, hence also sample by sample for a data-dependent member. The matching C28 candidate has zero conditional KL, and Theorem 1 with its point-mass comparison gives the upper bound. □

The numerical rows use frozen source state 2026091401. The specified target has $(a, b, s) = (2, 0.8, 0.6)$, at transform index 7 and radial index 64 (zero-based within its block). Its prior is $\pi_\star = 5.586180890750248 \times 10^{-6}$, so $-\log \pi_\star = 12.095214708088724$. The corrected source-direction calculation gives:

Pair	L_{sym}	Minimum sufficient K	U_K
$(2, 3)$	0.13681979065152935	88	0.13590128885492947
$(4, 5)$	0.14902401908293980	81	0.14750261839132590
$(6, 7)$	0.15494128354362954	78	0.15310398364669270

Each listed budget is the smallest integer satisfying the strict inequality. At $K = 32$, $U_{32} = 0.36652165782087043$ exceeds all three lower bounds. The calculation provides no separation there and concerns the prescribed Gaussian model only.

Lower and upper comparisons for the hull. The sign-event bound is valid but loose. Before Pinsker, the proof also gives the stronger expression $\mathbb{E}\,\mathrm{KL}(\mathrm{Bern}(\Phi(\gamma F))\|\,\mathrm{Bern}(1/2))$. Neither expression is asserted to equal the C774-hull infimum. For pair $(2, 3)$, the saved quadrature of an explicit feasible two-candidate mixture gives

$$L_{\mathrm{sym}} \leq \inf_{h \in \mathrm{conv(C774)}} A(h) \leq A(h_{\mathrm{feasible}}) \approx 0.338636901672.$$

The feasible mixture has mass 0.5758 on $N(-0.4Fv, 2I)$ and the remaining mass on $N(0, 2I)$. The reported upper comparison is numerical quadrature without a rigorous integration-error envelope. A minimum over single candidates is also an upper comparison for the hull infimum; it cannot replace the lower bound. The certified budget calculation uses only L_{sym} and the matched-candidate prior.

The Gaussian assumption on F is needed for the arcsine evaluation. If $F = 0$ almost surely, the sign-event lower bound is zero. This does not eliminate possible covariance-based obstructions. For a directional span V, the same proof applies with $t = \Pi_{V^\perp} w / \|\Pi_{V^\perp} w\|$ and $\delta = \|\Pi_{V^\perp} w\|^2$, provided this projection is nonzero. The reported numeric budgets use only the original single-direction construction.

A.7 A NONVANISHING GAP TO THE CONVEX-HULL ORACLE

Take $P = \mathrm{Bern}(0.5)$, $h_0 = \mathrm{Bern}(0.4)$ and $h_1 = \mathrm{Bern}(0.8)$, with positive prior on both. The target is the mixture $0.75h_0 + 0.25h_1$, so the hull optimum is zero. The mean log-likelihood ratio favoring h_1 is

$$\mathbb{E}_P \log \frac{h_1(U)}{h_0(U)} = \tfrac{1}{2}\log 2 + \tfrac{1}{2}\log(1/3) = -0.2027325541 < 0.$$

The strong law makes the posterior odds of h_1 converge to zero. Thus $H_t \to h_0$ almost surely and Cesàro averaging gives $\bar{H}_K \to h_0$. All predictive Bernoulli probabilities lie in $[0.4, 0.8]$, where the KL loss from $\mathrm{Bern}(0.5)$ is bounded and continuous. Dominated convergence then yields

$$\lim_{K \to \infty} \mathbb{E}A(\bar{H}_K) = \mathrm{KL}(\mathrm{Bern}(0.5)\|\,\mathrm{Bern}(0.4)) = 0.02041099726.$$

This is a limitation of the specified prefix estimator, even though its dictionary can represent the target exactly.

A.8 STRICT SEPARATION FROM THE 94-CANDIDATE UNION HULL

The Union family contains the source reference, 85 centered radial alternatives and eight transforms of the reference. Assume the reference gives positive mass to a finite-ν Student component. Let the target be a Gaussian composed candidate with $b \neq 0$ and $\Pr(\phi_S(Z) \neq 0) > 0$.

Every transformed reference retains a positive polynomial tail. A nonnegative mixture containing it cannot equal the conditional Gaussian target: equality almost everywhere of these continuous densities would imply equality everywhere, while along a ray the polynomial tail eventually exceeds the Gaussian tail. The same applies to finite-ν radial alternatives. Equality would therefore require zero weight on the reference, all its transforms, and all finite-ν radial alternatives. The remaining candidates are centered Gaussian radial densities. Every mixture of them has zero conditional mean, while the target mean $b\phi_S(z)v$ is nonzero on a set of positive P_Z probability. The target is not in the Union hull.

Each conditional density has unit $L^1(P_Z \otimes du)$ norm. The simplex of 94 weights is compact and the map from weights to densities is continuous in that norm. Its image is compact, hence closed. Non-membership implies a strictly positive total-variation distance to this hull. Pinsker's inequality then gives a strictly positive conditional KL infimum. This proof concerns the Union hull; the full single-direction dictionary already contains the specified composed target.

A.9 Multi-direction prior and computational cost

The 86 undirected candidates and their priors are shared between C774 and C28. Each directional candidate has one copy along each of 28 prescribed directions with one twenty-eighth of its original mass. The label count becomes 19,350. For a comparison vector placing total mass β on directional candidates, embedding it on the corresponding original-direction labels increases its prior-KL term by $\beta \log 28$. This concerns the right-hand side of the oracle inequality, not the difference of realized or expected predictor risks.

One saved combined-shift computation over five label budgets records 98 seconds for prediction, 8,875 seconds for scoring, a peak resident size of 347 MB and a 191 MB weight file. An isolated prediction replay for another setting records 32.2 seconds and reproduces the stored weights in two fresh processes. Those measurements cover their respective settings; the candidate-count ratio alone is not a runtime ratio.

A.10 Full budget ladders

Every entry below is a paired difference between two conditional densities scored on the same target fields, so the supplied distribution and the map Jacobian cancel. Within a training seed the 128 target calibration draws are averaged first; the reported entry is the mean over the ten seeds of that per-seed difference, and the bracket is an unadjusted two-sided paired t interval on nine degrees of freedom, $t_{9,0.975} = 2.2621571629$. Scores are totals in nats per residual vector and a negative entry means the first-named method scores lower.

Under the combined shift the gain of the single-direction dictionary over the source reference is 0.338912 nats at $K = 8$, 0.389874 at $K = 16$, 0.450257 at $K = 32$, 0.494799 at $K = 64$ and 0.518122 at $K = 128$. Against the same-dictionary mixture maximum-likelihood fit the held-out gap is 0.346777 nats at $K = 8$, 0.143763 at $K = 16$, 0.071741 at $K = 32$, 0.041726 at $K = 64$ and 0.023607 at $K = 128$. Against the last posterior, and with the same sign convention, the prefix average is better by 0.002294 nats at $K = 8$ and worse at every larger budget: 0.016788 nats at $K = 16$, 0.033345 at $K = 32$, 0.020046 at $K = 64$ and 0.008627 at $K = 128$. The maximum-likelihood fits satisfy the calibration optimality-gap tolerance at all 12,800 recorded draw–budget pairs, with maximum gap $9.982947801745468 \times 10^{-11}$; that certificate covers the average calibration score of the fitted weights only and is not a held-out loss certificate.

A.11 Prescribed Gaussian scenarios: the full budget ladder

The ten scenarios are the prescribed Gaussian targets of Appendix A.6 and its neighbourhood. S0 to S4 lie on the implemented composite grid with the source mixture in place, S5 to S7 place the directed component outside the source-selected direction, S8 replaces the source mixture by an exactly Gaussian one, and S9 moves the target off the grid. All ten share one frozen source state and take 128 independent repeats each, and they use the multi-direction dictionary, the single-direction dictionary and the prior used everywhere else in this paper.

Table 4: The complete budget ladder under the combined shift, at all five stored budgets. Every entry is C774 minus the row comparator, in total nats per residual vector. Each cell gives the mean followed by the lower and upper interval endpoints. Summaries use ten paired training seeds with an unadjusted two-sided paired t interval on nine degrees of freedom. The source reference is the soft prior-weighted mixture over the 86 source radial components and is not a selected component; the same-dictionary mixture maximum-likelihood fit is judged by the calibration optimality gap, whose optimum and the held-out loss are different objects and are kept distinct. The five-budget extra comparators cover the combined and no-shift controls; the other four controls are stored at $K = 32$ only.

Comparator to C774	$K = 8$	$K = 16$	$K = 32$	$K = 64$	$K = 128$
source reference	-0.338912 $[-0.369664, -0.308160]$	-0.389874 $[-0.420112, -0.359636]$	-0.450257 $[-0.482007, -0.418507]$	-0.494799 $[-0.528305, -0.461294]$	-0.518122 $[-0.552027, -0.484217]$
source-only full-dictionary reference	-0.337913 $[-0.368600, -0.307227]$	-0.388876 $[-0.419060, -0.358691]$	-0.449259 $[-0.480943, -0.417574]$	-0.493801 $[-0.527233, -0.460369]$	-0.517123 $[-0.550951, -0.483296]$
Union set	-0.034684 $[-0.039363, -0.030004]$	-0.039605 $[-0.044840, -0.034370]$	-0.050058 $[-0.055408, -0.044708]$	-0.058069 $[-0.063880, -0.052258]$	-0.065278 $[-0.072015, -0.058541]$
same-dictionary mixture MLE	-0.346777 $[-0.384808, -0.308746]$	-0.143763 $[-0.164027, -0.123500]$	-0.071741 $[-0.081308, -0.062174]$	-0.041726 $[-0.048960, -0.034491]$	-0.023607 $[-0.028028, -0.019185]$
last posterior	-0.002294 $[-0.005356, 0.000768]$	0.016788 $[0.011140, 0.022436]$	0.033345 $[0.029244, 0.037446]$	0.020046 $[0.018339, 0.021753]$	0.008627 $[0.007520, 0.009734]$
endpoint mixture	0.010696 $[0.008086, 0.013305]$	0.023402 $[0.018442, 0.028362]$	0.035326 $[0.031341, 0.039310]$	0.020686 $[0.019051, 0.022321]$	0.009018 $[0.007983, 0.010053]$
radial-only prefix (86 labels)	-0.009009 $[-0.014750, -0.003268]$	-0.024765 $[-0.031354, -0.018175]$	-0.044754 $[-0.051792, -0.037716]$	-0.059787 $[-0.066829, -0.052745]$	-0.068676 $[-0.076063, -0.061290]$
single best kernel	-0.277888 $[-0.322699, -0.233077]$	-0.034394 $[-0.048402, -0.020385]$	0.014079 $[0.008406, 0.019751]$	0.012744 $[0.010271, 0.015217]$	0.004545 $[0.003476, 0.005615]$
radial residual KDE	-0.489481 $[-0.770174, -0.208788]$	-0.158899 $[-0.216050, -0.101747]$	-0.123583 $[-0.159276, -0.087889]$	-0.101550 $[-0.115330, -0.087770]$	-0.100331 $[-0.114452, -0.086210]$
zero-location robust comparator	-0.432883 $[-0.479780, -0.385986]$	-0.204734 $[-0.231558, -0.177910]$	-0.152358 $[-0.167950, -0.136767]$	-0.143720 $[-0.156194, -0.131245]$	-0.144398 $[-0.155499, -0.133297]$
source-direction Student, conditional location	-0.964121 $[-1.062222, -0.866021]$	-0.293156 $[-0.334329, -0.251983]$	-0.131486 $[-0.146868, -0.116104]$	-0.088265 $[-0.097811, -0.078719]$	-0.074270 $[-0.081119, -0.067422]$
source-informed Gaussian regression	-2.302015 $[-2.392483, -2.211547]$	-2.243383 $[-2.317465, -2.169301]$	-1.465313 $[-1.521501, -1.409125]$	-0.923129 $[-0.966676, -0.879581]$	-0.663425 $[-0.705208, -0.621642]$
source-informed Student regression	-1.284239 $[-1.313055, -1.255423]$	-1.329262 $[-1.351560, -1.306964]$	-0.970492 $[-0.993147, -0.947837]$	-0.683529 $[-0.706171, -0.660887]$	-0.540993 $[-0.569208, -0.512778]$

The five budgets $K = 8, 16, 32, 64, 128$ were measured first. The four low budgets $K = 0, 1, 2, 4$ were added afterwards, on the same construction, to ask one question: does the measured difference between the two dictionaries change sign at the low end? A budget of zero is well defined here, because the prefix average includes $t = 0$: at $K = 0$ it is the prior mixture, and the two dictionaries are compared at their priors.

Eight scenarios keep the same sign with intervals excluding zero at all nine measured budgets. S0 and S8 have intervals covering zero at every measured budget from $K = 8$ onwards. In S8 the point estimate changes sign near zero at intermediate budgets, so unchanged signs at the endpoints do not imply an unchanged sign throughout. No scenario shows interval-resolved differences of opposite signs across the measured budgets. These measurements therefore identify no resolved empirical

Table 5: The four low budgets, added after the five already reported, on the same ten prescribed Gaussian scenarios, the same dictionaries, the same prior and one frozen source state, with 128 independent repeats per scenario. Cells are the multi-direction minus the single-direction dictionary in total nats per residual vector: the mean, and beneath it the half-width of the unadjusted 95% interval $t_{0.975,127}s/\sqrt{n}$. The sign at $K = 0$ is the sign at $K = 4$ in every row.

Scenario	$K = 0$	$K = 1$	$K = 2$	$K = 4$
S0	-0.002 $+0.001$	-0.002 $+0.001$	-0.002 $+0.001$	-0.001 $+0.001$
S1	-0.003 $+0.001$	-0.004 $+0.002$	-0.005 $+0.003$	-0.008 $+0.003$
S2	$+0.026$ $+0.001$	$+0.041$ $+0.004$	$+0.053$ $+0.005$	$+0.082$ $+0.008$
S3	$+0.018$ $+0.001$	$+0.031$ $+0.004$	$+0.050$ $+0.006$	$+0.087$ $+0.009$
S4	$+0.017$ $+0.001$	$+0.036$ $+0.005$	$+0.054$ $+0.007$	$+0.089$ $+0.010$
S5	-0.003 $+0.001$	-0.005 $+0.002$	-0.008 $+0.003$	-0.016 $+0.004$
S6	-0.006 $+0.001$	-0.012 $+0.002$	-0.018 $+0.003$	-0.028 $+0.004$
S7	-0.009 $+0.001$	-0.017 $+0.003$	-0.026 $+0.004$	-0.047 $+0.006$
S8	-0.001 $+0.001$	-0.001 $+0.001$	-0.002 $+0.001$	-0.002 $+0.001$
S9	$+0.008$ $+0.001$	$+0.012$ $+0.003$	$+0.013$ $+0.004$	$+0.021$ $+0.005$

Table 6: The five budgets measured first, on the same scenarios, for the same difference and with the same summary rule. $K = 0, 1, 2, 4$ are in Table 5. The sign is unchanged from the low-budget table in every row whose interval resolves.

Scenario	$K = 8$	$K = 16$	$K = 32$	$K = 64$	$K = 128$
S0	-0.002 $+0.002$	-0.001 $+0.002$	-0.001 $+0.001$	-0.000 $+0.000$	-0.000 $+0.000$
S1	-0.010 $+0.004$	-0.008 $+0.004$	-0.006 $+0.004$	-0.004 $+0.003$	-0.002 $+0.002$
S2	$+0.129$ $+0.010$	$+0.152$ $+0.010$	$+0.080$ $+0.008$	$+0.030$ $+0.003$	$+0.012$ $+0.001$
S3	$+0.138$ $+0.011$	$+0.141$ $+0.008$	$+0.076$ $+0.006$	$+0.030$ $+0.003$	$+0.011$ $+0.001$
S4	$+0.136$ $+0.011$	$+0.136$ $+0.008$	$+0.080$ $+0.007$	$+0.030$ $+0.003$	$+0.011$ $+0.001$
S5	-0.037 $+0.007$	-0.095 $+0.013$	-0.206 $+0.013$	-0.287 $+0.007$	-0.318 $+0.006$
S6	-0.056 $+0.007$	-0.119 $+0.011$	-0.216 $+0.013$	-0.293 $+0.008$	-0.326 $+0.006$
S7	-0.102 $+0.011$	-0.186 $+0.012$	-0.260 $+0.010$	-0.306 $+0.006$	-0.328 $+0.005$
S8	-0.001 $+0.001$	$+0.000$ $+0.001$	$+0.000$ $+0.001$	$+0.000$ $+0.001$	-0.000 $+0.000$
S9	$+0.030$ $+0.006$	$+0.040$ $+0.005$	$+0.040$ $+0.004$	$+0.031$ $+0.004$	$+0.017$ $+0.003$

crossover; they neither establish the absence of a crossover between measured budgets nor locate the analytic sufficient budgets of Proposition 4, which concern a different object.

No repeat count, direction, prior or scenario was changed after these numbers were seen.

A.12 Interpretation of the guarantees and comparisons

Task preservation holds for the supplied m_D for every D, x. It makes no calibration claim about that marginal. The probabilistic object throughout is the eight-coefficient residual projection. Pooling and symmetry are restrictions of the implemented conditional module, not restrictions required by task preservation.

The prefix oracle is an expectation over an i.i.d. calibration sample conditional on fixed source information. Its comparator is an average of individual candidate risks. It is not a per-sample or high-probability statement, nor a guarantee for the final posterior or endpoint mixture. The off-axis lower bound applies to the whole C774 hull on the prescribed Gaussian construction. Its sufficient budgets are distinct from measured PDE crossings, and it certifies no comparison at $K = 32$.

All displayed paired intervals are unadjusted. The original source-reference control family and the later same-dictionary, direction, query and architecture comparisons retain their original inferential roles. Union changes the candidate set and prior together. The continuous baselines use source information but do not match the discrete reference prior. No same-family MLE fit over C28 is reported. The architecture comparison is a difference of within-architecture gains. The two regional queries assess two specified events.

A.13 Available records and reproduction scope

The supplement contains source code, frozen source geometry and conditional state, saved prediction weights, per-setting score records, and scripts for rebuilding reported summaries. The paired summaries average calibration draws within each training seed before computing intervals across seeds. The full source-training and target-data pipeline is separate from re-aggregation of these saved results.

The records also document re-scoring of saved weights in the multi-direction matrix and weight-hash consistency. One prediction setting was replayed in two fresh processes with access restricted to source and calibration inputs; attempted reads of test residuals were rejected, and the stored weights were reproduced. This isolated replay covers that setting. It does not establish a replay of every setting or of network training. Implementation checks and the AI-assisted proof review are distinct from independent human mathematical review.

A.14 Data generation and comparator definitions

A.14.1 Forcing changes and source directions

Source initial fields have twelve independent standard-normal sine/cosine coefficients with mode amplitudes $0.35/k^{1.2}$ for $k = 1, \dots, 6$. Source forcing has eight independent standard-normal latent coefficients $g \in \mathbb{R}^8$, with amplitudes $0.3/k^{0.65}$ for $k = 1, \dots, 4$. For a prescribed pair (i, j), dependence is introduced by replacing g_j with $\rho_f g_i + \sqrt{1-\rho_f^2}\, g_j$. Tail shift then multiplies all eight forcing coordinates by the same independent factor $\sqrt{3/V}$, where $V \sim \chi^2_5$. The latter preserves their covariance. Initial-field latents remain unchanged. The controls use $(\rho_f, \text{tail}) = (0,0), (0,1), (0.6,0), (0.6,1), (0.3,0), (0.6,0)$; the last uses the source-selected outside-primary pair. The data-generation implementation is `source_data.py`.

Let $e(g)$ be the eight-coefficient residual for a source design input. The frozen surrogate does not observe unresolved forcing, so the forcing derivative is computed from the PDE solution. Central finite differences with step $\epsilon = 0.025$ give

$$J_{e,k} \approx \frac{e(g+\epsilon e_k) - e(g-\epsilon e_k)}{2\epsilon}, \qquad J_{u,k} = L^{-1}N^\top (I - AW^\top) J_{e,k}, \qquad J_{F,k} = W^\top J_{e,k}/z_{\text{std}}.$$

Here e_k in the finite-difference stencil denotes the kth coordinate vector. Over the 128 design inputs, the source pair-response vector is

$$c_{ij} = \frac{1}{128} \sum_n \left[J^{(n)}_{u,i} J^{(n)}_{F,j} + J^{(n)}_{u,j} J^{(n)}_{F,i} \right]. \tag{17}$$

Pairs are ranked by decreasing $\|c_{ij}\|$, with lexicographic tie-breaking. The main direction is its normalized vector. Among the other nondegenerate pairs, the outside-primary pair minimizes the absolute cosine with the main direction. C28 normalizes all 28 pair-response vectors. This is implemented in `remote_response.py`; it uses source design inputs only.

Why this response is relevant to conditional dependence. Fix a source linearization and write perturbations as $\delta U = J_u \xi$ and $\delta F = J_F \xi$. Replacing the covariance of ξ by $\mathrm{Cov}(\xi) + \rho_f (e_i e_j^\top + e_j e_i^\top)$ changes their cross-covariance by

$$\Delta \,\mathrm{Cov}(\delta U, \delta F) = \rho_f \left(J_{u,i} J_{F,j} + J_{u,j} J_{F,i} \right).$$

This is direct matrix multiplication. Averaging over source design inputs gives $\rho_f c_{ij}$. Thus the selected direction represents the strongest pairwise task–residual cross-covariance response in the

averaged source linearization. This interpretation is local to that linearization; it does not identify the exact covariance response of the nonlinear target PDE.

The transform grid includes both $s = 1$ and $s = \sqrt{1 - b^2}$. In the unit-variance reference model with independent F and residual noise, the latter keeps the unconditional variance along v equal to one while introducing covariance b with F. This explains the two transform types. General radial scales and target laws need not share that variance identity.

A.14.2 Estimators on the fixed dictionary

Sequential comparators. The final posterior uses $H_K = \sum_j w_{K,j} h_j$. The endpoint predictor uses $(H_0 + KH_K)/(K+1)$. Single-candidate maximum likelihood selects the candidate with largest total calibration log-likelihood among positive-prior labels, breaking ties by the smallest label. The radial-only prefix comparator uses the 86 radial labels with prior 0.8 on the added unit Gaussian and $0.2 r_j$ on the other labels. The source-only full-dictionary reference (record key `Bplus`) is the last posterior on all 774 candidates after the source calibration sample; it is distinct from the 86-component source reference.

Mixture maximum likelihood. Let $\ell_{ij} = \log h_j(u_i \mid z_i)$. Mixture MLE minimizes

$$f(w) = -\frac{1}{K} \sum_{i=1}^{K} \log \sum_{j=1}^{774} w_j e^{\ell_{ij}}, \qquad w \in \Delta_{774}.$$

No source-prior penalty is added. For a feasible w, put

$$g_j(w) = \frac{1}{K} \sum_i \frac{e^{\ell_{ij}}}{\sum_\ell w_\ell e^{\ell_{i\ell}}}, \qquad \operatorname{gap}(w) = \max_j g_j(w) - 1.$$

The identity $w^\top g(w) = 1$ and convexity give $0 \le f(w) - \inf_{w'} f(w') \le \operatorname{gap}(w)$. The implementation accepts a fit only when this all-candidate gap is at most 10^{-7} and the simplex checks hold. Rowwise log shifts stabilize the calculation. The gap concerns calibration optimization, independently of held-out generalization.

A.14.3 Source-informed Gaussian and Student regression

Write $x(z) = (1, \phi_S(z))^\top$. On source calibration observations, ordinary least squares gives a 2×7 coefficient matrix M_0. If E_0 is the source residual matrix, define

$$\widehat{\Sigma}_0 = \frac{E_0^\top E_0}{n_0 - 2}, \quad \Sigma_0 = \widehat{\Sigma}_0 + 10^{-6} \frac{\operatorname{tr} \widehat{\Sigma}_0}{7} I, \quad \Lambda_0 = 2I_2, \quad \Psi_0 = 2\Sigma_0.$$

For a target design matrix X and target coordinate matrix U, fit

$$\Lambda_K = \Lambda_0 + X^\top X, \qquad M_K = \Lambda_K^{-1}(\Lambda_0 M_0 + X^\top U),$$
$$\Psi_K = \Psi_0 + (U - XM_K)^\top (U - XM_K) + (M_K - M_0)^\top \Lambda_0 (M_K - M_0).$$

At a new z, let $c(z) = 1 + x(z)^\top \Lambda_K^{-1} x(z)$ and $\mu(z) = x(z)^\top M_K$. The Gaussian comparator has mean $\mu(z)$ and covariance $c(z)\Psi_K/(K+2)$. The Student comparator has $K + 4$ degrees of freedom, the same mean, and shape $c(z)\Psi_K/(K+4)$. Thus both have the same predictive covariance. They are seven-dimensional conditional densities fitted with source-informed regularization. Their record keys are `G_D` and `T_D`; neither is a task-marginal-only predictor. The implementation is in `family_774.py`.

A.14.4 Conditional and zero-location directional Student fits

The conditional Student comparator uses the same source direction and scalar standardization. Decompose $r = v^\top u$ and $u_\perp = (I - vv^\top)u$. The density factors into a one-dimensional Student law for $r - (\beta_0 + \beta_1 \phi_S(z))$ and a six-dimensional radial Student law for $u_\perp$. Their covariance multipliers are $a_v, a_\perp$ and their degrees of freedom are selected separately from $\{3, 5, 10, 30, \infty\}$. Log-variances lie in $[-4, 4]$.

Calibration likelihood fits location, variances and tails. The finite-degree location fit uses L-BFGS-B with two deterministic starts: least squares and a median intercept. Its settings are `ftol=1e-12`, `gtol=1e-7`, `maxiter=500` and `maxls=40`. The perpendicular variance is fitted by bounded scalar optimization. The Gaussian location limit has a closed-form least-squares fit with bounded variance. This continuous fit is locally optimized; it has no all-candidate MLE certificate. The separate zero-location robust comparator fixes $\beta_0 = \beta_1 = 0$ and fits the two variances and tail parameters. The implementations are `conditional_robust.py` and `remote_p1_score.py`.

A.14.5 Radial residual KDE

For each calibration point, let $r_i = \|u_i\|$ and $t_i = r_i^{1/3}$. A gamma-kernel mixture estimates the density of t:

$$\widehat{f}_T(t) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Gamma}(t;\ \text{shape} = \lambda t_i,\ \text{rate} = \lambda).$$

The rate grid is $\lambda_j = 10^{-5+10j/99}$, $j = 0, \dots, 99$. Calibration-only cross-validation uses $\min(10, K)$ fixed folds, averages fold NLLs equally and breaks ties by the first grid value. The implementation floors t_i and evaluated t at 10^{-12} for numerical stability. The radial vector density, away from zero, is obtained from

$$\widehat{f}_R(r) = \widehat{f}_T(r^{1/3})\frac{1}{3}r^{-2/3}, \qquad \widehat{h}(u) = \frac{\widehat{f}_R(\|u\|)}{|S^6|\|u\|^6}.$$

It is independent of z and uniform in angular direction. This is a radial residual recalibration comparator, not an unrestricted seven-dimensional KDE. It uses the radial gamma-kernel construction associated with latent recalibration (Dheur & Ben Taieb, 2025); the experiment does not reproduce that paper's normalizing-flow model. Implementations are `b1_radial.py` and `frozen_kde_cv.py`.

A.14.6 Regional events and the second architecture

The task and second regional functional average the eight-mode Fourier reconstruction on grid indices 0:8 and 32:40, respectively, with zero-based half-open indexing. Each threshold is the empirical source-calibration 75th percentile using linear quantile interpolation. The events use strict exceedance and their intersection or union. Task integration uses Gauss–Legendre quadrature in the supplied Student law. These definitions are implemented in `query_regions.py` and `remote_e5.py`.

The second backbone is a custom multiscale convolutional operator, with four channel levels $16, 32, 64, 128$, three downsampling stages, average pooling, linear upsampling, GELU activations, four full-resolution residual blocks and four residual-neck blocks; batch normalization is disabled. It contains no global Fourier mixing. Its local implementation is in `crossop/cno.py`; it is not a reproduction of the published simplified-CNO topology. The source training recipe uses AdamW with learning rate 10^{-3} and weight decay 10^{-8}, batch size 32 and 60 epochs. The learning rate halves every 15 epochs, and the checkpoint is selected by minimum validation MSE. These source training settings are distinct from the conditional adaptation procedure.

A.15 Quantifying changes excluded by the dictionary

This section studies a probability law on (Z, U) and a fixed subspace $V \subseteq \mathbb{R}^d$, with $d = 7$ in the implementation. Write Π for the orthogonal projection onto V, $\Pi_\perp = I - \Pi$, and

$$\mathcal{T} = (Z, \Pi U, \|\Pi_\perp U\|).$$

This statistic records the task, the component in V, and the remaining radius. Its fibers are the orbits of G_V. Let P^G denote group averaging in U, leaving Z fixed, and set $B_G(P) = \mathrm{KL}(P\|P^G)$. It equals the conditional symmetry term because P and P^G have the same Z marginal.

Proposition 5 (A bounded perturbation, including a nonsymmetric baseline)**.** *Let $B_G(P_0) < \infty$. Let s be bounded, $\mathbb{E}_0 s = 0$, and define $dP_\epsilon = (1 + \epsilon s)dP_0$, where $\alpha = |\epsilon|\|s\|_\infty < 1$. Put*

$$\bar{s} = \mathbb{E}_0[s \mid \mathcal{T}], \qquad \ell_0 = \log\frac{dP_0}{dP_0^G}, \qquad \sigma_s^2 = \mathbb{E}_0(s - \bar{s})^2.$$

Then

$$\frac{\epsilon^2\sigma_s^2}{2(1+\alpha)} \le B_G(P_\epsilon) - B_G(P_0) - \epsilon\mathbb{E}_0[s\ell_0] \le \frac{\epsilon^2\sigma_s^2}{2(1-\alpha)}. \tag{18}$$

If $P_0 = P_0^G$, the baseline and linear terms vanish. In that case $\bar{s}$ is the ordinary Haar average of s, and (18) *gives Proposition 3.*

Proof. Disintegrate P_0 over $\mathcal{T}$ as in the symmetry-barrier proof. Group averaging replaces its conditional orientation law by the uniform orbit kernel σ_t. Tilting changes the orbit marginal by $1+\epsilon\bar{s}(t)$, so

$$\frac{dP_\epsilon^G}{dP_0^G} = 1 + \epsilon\bar{s}(\mathcal{T}).$$

The two laws have the same orbit marginal. Because s is bounded and $B_G(P_0)$ is finite, $s\ell_0$ is integrable. Here finiteness of relative entropy implies integrability of its log-density ratio: the negative part is bounded in integral, and the positive part is finite. The remaining log-ratios are bounded since $1-\alpha \le 1+\epsilon s, 1+\epsilon\bar{s} \le 1+\alpha$. Thus no infinite quantities are subtracted in the following identity:

$$B_G(P_\epsilon) = B_G(P_0) + \epsilon\mathbb{E}_0[s\ell_0] + \mathbb{E}_0\{f(\epsilon s) - f(\epsilon\bar{s})\}, \quad f(r) = (1+r)\log(1+r) - r. \tag{19}$$

To obtain it, expand the log-density ratio into $\ell_0 + \log(1+\epsilon s) - \log(1+\epsilon\bar{s})$ and condition the last term on $\mathcal{T}$. The linear parts in f have equal expectations.

Taylor's formula at $\epsilon\bar{s}$ gives

$$f(\epsilon s) - f(\epsilon\bar{s}) = \epsilon f'(\epsilon\bar{s})(s-\bar{s}) + \epsilon^2(s-\bar{s})^2\int_0^1 \frac{1-t}{1+\epsilon\bar{s}+t\epsilon(s-\bar{s})}\,dt.$$

The first term has expectation zero by conditional expectation. The denominator lies in $[1-\alpha, 1+\alpha]$. Integrating $1-t$ gives the two bounds. If P_0 is invariant, its conditional orientation is uniform, so conditional averaging is Haar averaging and $\ell_0 = 0$. □

The bounds quantify the approximation cost of the entire invariant class. Their upper endpoint does not upper-bound the best risk in a finite dictionary. For a nonsymmetric baseline, the linear term in (18) may have either sign. A change can reduce an existing symmetry mismatch. Whitening a covariance matrix alone does not make a distribution invariant.

A.15.1 From a physical perturbation to a conditional density change

Let ξ collect the source input, unresolved forcing and any simulator randomness. A frozen measurable map produces $O = (Z, U) = T(\xi)$. This map includes the surrogate and the fixed coordinate transform. Consider an exact bounded tilt of the source law,

$$\frac{d\mathsf{R}_\epsilon}{d\mathsf{R}_0}(\xi) = 1 + \epsilon\psi(\xi), \qquad \mathbb{E}_0\psi = 0, \quad |\epsilon|\|\psi\|_\infty < 1.$$

Define three conditional expectations under the source law:

$$s_O = \mathbb{E}_0[\psi \mid Z, U], \qquad s_\mathcal{T} = \mathbb{E}_0[\psi \mid \mathcal{T}], \qquad s_Z = \mathbb{E}_0[\psi \mid Z].$$

For every bounded test function of O, pushing forward the tilted law shows $dP_\epsilon/dP_0 = 1 + \epsilon s_O$. The analogous ratio for the task marginal is $1 + \epsilon s_Z$. Consequently

$$\frac{d}{d\epsilon}\log\frac{dP_{\epsilon,U|Z}}{dP_{0,U|Z}}\bigg|_{\epsilon=0} = s_O - s_Z. \tag{20}$$

The change in the supplied marginal and the conditional change are separated at the level of the density derivative.

Successive conditional expectations are orthogonal projections in $L^2(\mathsf{R}_0)$. The nested sigma-fields generated by Z, $\mathcal{T}$, O and ξ therefore give the exact identity

$$\|\psi\|_2^2 = \|s_Z\|_2^2 + \|s_\mathcal{T} - s_Z\|_2^2 + \|s_O - s_\mathcal{T}\|_2^2 + \|\psi - s_O\|_2^2. \tag{21}$$

For example, the cross product of $s_O - s_\mathcal{T}$ with any $\mathcal{T}$-measurable square-integrable variable has zero expectation; the other cross products vanish by the same argument. The four terms describe the

task response, the remaining invariant response, the orientation response excluded by the dictionary, and the response lost by observing only (Z, U). These squared score norms partition Fisher information about the specified scalar perturbation at the source law; they are not Shannon information or a claim about recovering a full physical field.

Applying Proposition 5 with $s = s_O$ makes its quadratic coefficient $\|s_O - s_{\mathcal{T}}\|_2^2/2$. This links a perturbation of the physical input law to the approximation barrier after the frozen observation map. The ingredients are conditional expectation and the usual score-projection principle (Pollard, 2012). The bounded-tilt statement is exact under its stated assumptions. For a general differentiable statistical path, score projection still describes its first derivative under suitable regularity; quadratic-mean differentiability alone is not invoked as a sufficient condition for an expansion of forward KL. The Gaussian calculation below supplies its own domination argument.

A.15.2 Gaussian forcing paths through a nonlinear observation map

The exact tilt is useful for a finite perturbation bound. A likelihood-ratio condition also covers Gaussian covariance changes followed by an arbitrary fixed measurable observation map.

Proposition 6 (A regular source path through a frozen map)**.** *Let $R_\epsilon = d\mathsf{R}_\epsilon/d\mathsf{R}_0$ and suppose, for some $2 < p \le 3$, that*

$$\|R_\epsilon - 1 - \epsilon\psi\|_{L^p(\mathsf{R}_0)} = o(|\epsilon|), \qquad \psi \in L^p(\mathsf{R}_0), \quad \mathbb{E}_0\psi = 0.$$

Let $O = T(\xi) = (Z, U)$ be any fixed measurable map, and define $s_O = \mathbb{E}_0[\psi \mid O]$ and $s_{\mathcal{T}} = \mathbb{E}_0[\psi \mid \mathcal{T}]$. If its source pushforward P_0 is G-invariant, then

$$B_G(P_\epsilon) = \frac{\epsilon^2}{2}\|s_O - s_{\mathcal{T}}\|_2^2 + o(\epsilon^2). \tag{22}$$

For a nonsymmetric reference, the same quadratic expression holds after subtracting $B_G(P_0) + \mathbb{E}_0[(r_\epsilon - 1)\ell_0]$, provided $\ell_0 = \log(dP_0/dP_0^G) \in L^{p/(p-1)}(P_0)$, where $r_\epsilon = \mathbb{E}_0[R_\epsilon \mid O]$.

Proof. Conditional expectation is a contraction in L^p, so

$$r_\epsilon = 1 + \epsilon s_O + o_{L^p}(\epsilon), \qquad \bar{r}_\epsilon = \mathbb{E}_0[R_\epsilon \mid \mathcal{T}] = 1 + \epsilon s_{\mathcal{T}} + o_{L^p}(\epsilon).$$

Set $\varphi(x) = x\log x - x + 1$ for $x \ge 0$, with $0\log 0 = 0$. For $2 < p \le 3$ there is a finite constant C_p such that

$$|\varphi(1+y) - y^2/2| \le C_p|y|^p, \qquad y \ge -1.$$

Near zero this follows from the cubic Taylor remainder. On the rest of a bounded interval it follows by continuity, including $y = -1$. As $y \to \infty$, $\varphi(1+y) = O(y\log y)$ and the ratio of the left side to $|y|^p$ is bounded. Thus

$$\mathbb{E}_0\varphi(r_\epsilon) = \tfrac{1}{2}\epsilon^2\|s_O\|_2^2 + o(\epsilon^2), \qquad \mathbb{E}_0\varphi(\bar{r}_\epsilon) = \tfrac{1}{2}\epsilon^2\|s_{\mathcal{T}}\|_2^2 + o(\epsilon^2).$$

The exact orbit identity is

$$B_G(P_\epsilon) = B_G(P_0) + \mathbb{E}_0[(r_\epsilon - 1)\ell_0] + \mathbb{E}_0[\varphi(r_\epsilon) - \varphi(\bar{r}_\epsilon)].$$

For the nonsymmetric case, Hölder's inequality and the stated moment of ℓ_0 make the baseline term finite; the L^p bounds make the two entropy increments finite. For an invariant reference $\ell_0 = 0$. Orthogonality of conditional expectation gives $\|s_O\|_2^2 - \|s_{\mathcal{T}}\|_2^2 = \|s_O - s_{\mathcal{T}}\|_2^2$, proving both claims. □

For $g_\epsilon \sim N(0, \Sigma_0 + \epsilon\Delta)$ with Σ_0 positive definite and Δ symmetric, its likelihood ratio and first two derivatives are a quadratic-exponential factor times polynomials in g. On a sufficiently small interval around zero, their pth powers have a common integrable Gaussian envelope for any fixed finite p. Taylor's theorem in L^p yields the required condition, with

$$\psi(g) = \tfrac{1}{2}\{g^\top\Sigma_0^{-1}\Delta\Sigma_0^{-1}g - \operatorname{tr}(\Sigma_0^{-1}\Delta)\}.$$

No differentiability of the frozen map T is needed for this score-projection result. It applies to nonlinear residual maps as well as thresholded observations. The invariance assumption is on the source pushforward, not merely on the Gaussian forcing. If that pushforward is nonsymmetric, the

additional baseline term must be retained; it generally contains second-order changes as well as a linear response.

As a finite check, take independent standard Gaussians at the source, perturb their correlation to ϵ, and observe only their signs (Z, U). The source pushforward is uniform on four states and invariant under flipping U. Here $s_O = (2/\pi)ZU$, $s_{\mathcal{T}} = 0$, and $B_G(P_\epsilon) = 2\epsilon^2/\pi^2 + o(\epsilon^2)$. The exact sign probabilities give this coefficient directly. This example illustrates score transport through a nonsmooth observation; it is not an additional experiment on the PDE.

A.15.3 Mean and covariance changes in a Gaussian reference model

Corollary 1 (Local mean and covariance contributions)**.** *Let $F \sim N(0,1)$ and, for sufficiently small $|\epsilon|$, let*

$$U \mid F \sim N(\epsilon F c,\ aI_d + \epsilon\Delta), \qquad a > 0, \quad \Delta = \Delta^\top.$$

Write $m = d - \dim V > 0$ and

$$\Delta_G = \Pi\Delta\Pi + \frac{\operatorname{tr}(\Pi_\perp \Delta)}{m}\Pi_\perp.$$

For the group fixing V pointwise,

$$B_{G_V}(P_\epsilon) = \epsilon^2 \left\{ \frac{\|\Pi_\perp c\|^2}{2a} + \frac{\|\Delta - \Delta_G\|_{\mathrm{F}}^2}{4a^2} \right\} + o(\epsilon^2). \tag{23}$$

Proof. The density derivative at zero, divided by the reference density, is

$$s(F,U) = \frac{Fc^\top U}{a} + \frac{U^\top \Delta U - a \operatorname{tr}\Delta}{2a^2}.$$

At zero, F and $U \sim N(0, aI)$ are independent. Haar averaging sends c to Πc and Δ to Δ_G. Reflection removes the cross-blocks and averaging within $V^\perp$ replaces that block by its trace divided by m. Hence

$$s - s_G = \frac{F(\Pi_\perp c)^\top U}{a} + \frac{U^\top(\Delta - \Delta_G)U}{2a^2}.$$

The trace of $\Delta - \Delta_G$ is zero. The two terms are orthogonal in L^2 because F is independent and centered. Gaussian second and fourth moments give

$$\mathbb{E}_0(s - s_G)^2 = \frac{\|\Pi_\perp c\|^2}{a} + \frac{\|\Delta - \Delta_G\|_{\mathrm{F}}^2}{2a^2}.$$

The Gaussian score is unbounded, so the bounded-tilt proposition does not apply directly. A separate differentiation argument gives the required expansion. On a sufficiently small closed interval around zero, the joint Gaussian law of (F, U) and all its group transforms have covariance eigenvalues bounded above and away from zero. The first two density derivatives are bounded by a polynomial times a common integrable Gaussian envelope. Their log densities are bounded in absolute value by a quadratic polynomial; the log derivatives of the group mixture are polynomially bounded as well, being weighted averages of component log derivatives. These bounds justify differentiating the entropy difference

$$B_G(P_\epsilon) = \int p_\epsilon \log p_\epsilon - \int p_\epsilon^G \log p_\epsilon^G$$

twice. Group invariance of Lebesgue measure and $\log p_0$ cancels the terms containing $p_0'' - (p_0^G)''$. The first derivative vanishes and the second is $\mathbb{E}_0 s^2 - \mathbb{E}_0 s_G^2 = \mathbb{E}_0(s - s_G)^2$. Taylor's theorem proves (23). □

The mean term measures the part of the task-dependent displacement outside V. The covariance term includes cross-covariance with $V^\perp$ and anisotropy inside it. It can be positive with $c = 0$. These are two distinct mechanisms by which a target violates the dictionary symmetry. Equation (23) is local to the stated Gaussian reference; it is not a fitted law for the Burgers scores.

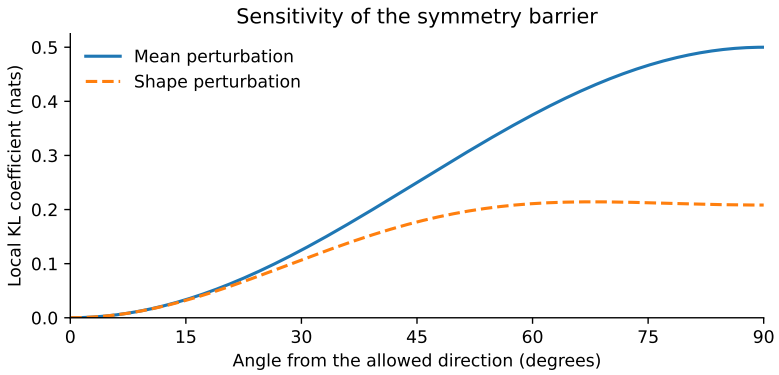


Figure 5: Analytic local sensitivity to a missed direction. In dimension seven with $a = 1$ and $V = \mathrm{span}(v)$, let θ be the angle from v to a unit vector w. The mean curve uses $c = w, \Delta = 0$; the covariance curve uses $c = 0, \Delta = ww^\top$. They show the coefficient of ϵ^2 in (23), namely $\sin^2\theta/2$ and $[1 - \cos^4\theta - \sin^4\theta/6]/4$. Both changes are allowed along v. Off-axis covariance anisotropy can create a barrier even without a conditional mean shift. These are analytic coefficients for the invariant class, with no measured PDE quantities.

A source-subspace criterion under the mean-only model. For specified response vectors $c_1, \ldots, c_L$ with weights $\beta_\ell \geq 0$, $\sum_\ell \beta_\ell = 1$, let $C = \sum_\ell \beta_\ell c_\ell c_\ell^\top$. The average mean contribution is

$$\frac{\mathrm{tr}\, C - \mathrm{tr}(\Pi C)}{2a}.$$

Among rank-r projections, it is minimized by the top r eigenvectors of C, with minimum $\sum_{i>r} \lambda_i(C)/(2a)$. To see this, write $C = \sum_i \lambda_i e_i e_i^\top$ and $\mathrm{tr}(\Pi C) = \sum_i \lambda_i \|\Pi e_i\|^2$. The coefficients belong to $[0, 1]$ and sum to r, so their weighted sum is at most the sum of the r largest eigenvalues; the stated projection attains it. This classical spectral calculation gives a design criterion for the local mean-only model. A covariance perturbation adds the second term of (23), and the implemented single-pair ranking is not changed to this subspace rule.

A sufficient budget for expanding the family. Suppose a source-fixed expanded dictionary contains a candidate h_+ with prior $\pi_+ > 0$ and risk $r_+ = A_{P_\epsilon}(h_+) < \infty$. Let L_ϵ be a proved lower bound on $B_G(P_\epsilon)$. Under the prefix-oracle assumptions, its prefix rule separates in expected risk from every rule confined to the original invariant hull whenever

$$r_+ + \frac{-\log \pi_+}{K+1} < L_\epsilon. \tag{24}$$

This follows by placing the atomic upper bound below the invariant-class lower bound. For a symmetric bounded-tilt reference, $L_\epsilon = \epsilon^2\sigma_s^2/[2(1+\alpha)]$ is valid. For a prescribed target with a matched fixed candidate ($r_+ = 0$), a sufficient condition is

$$K + 1 > \frac{2(1+\alpha)\log(1/\pi_+)}{\epsilon^2\sigma_s^2}, \qquad \sigma_s^2 > 0.$$

This condition combines symmetry-breaking signal and prior mass. It is a sufficient comparison, not a necessary sample size or a predicted empirical crossover. Candidate matching is an assumption, not a consequence of expanding a finite dictionary.

A.16 The source-response calculation and nonlinear curvature

Let $g_\theta \sim N(0, \Sigma_0 + \theta\Delta_{ij})$, where $i \neq j$, $\Delta_{ij} = e_i e_j^\top + e_j e_i^\top$, and the covariance stays positive definite. All coordinate maps and standardizations are held fixed. Assume the functions $U(g), F(g)$ are twice continuously differentiable, with sufficient Gaussian integrability to differentiate the displayed moments. Polynomial growth of the functions and the derivatives in the product calculation

is a sufficient condition. Other independent source randomness may be averaged over under the same integrability conditions.

Price's covariance derivative identity (Voigtlaender, 2021) states

$$\frac{d}{d\theta}\mathbb{E}[f(g_\theta)]\bigg|_0 = \mathbb{E}_0[\partial_{ij} f(g_0)].$$

There is no factor $1/2$ for this off-diagonal symmetric perturbation. Apply it to UF, U and F, and differentiate $\mathrm{Cov}(U,F) = \mathbb{E}[UF] - \mathbb{E}[U]\mathbb{E}[F]$. The product rule gives

$$\begin{aligned}\frac{d}{d\theta}\,\mathrm{Cov}(U,F)\bigg|_0 &= \mathbb{E}_0[\partial_i U\,\partial_j F + \partial_j U\,\partial_i F] \\ &\quad + \mathrm{Cov}_0(U,\partial_{ij}F) + \mathrm{Cov}_0(\partial_{ij}U,F).\end{aligned} \tag{25}$$

The stored source response averages finite-difference estimates of the first line. It is the exact response of the corresponding local linearizations. The second line contains the nonlinear curvature contribution. This makes explicit which part of the full covariance response the dictionary construction uses.

A simple example distinguishes stencil accuracy from this modeling approximation. Let (g_1, g_2) be centered standard Gaussians with correlation θ, and set

$$F = g_1, \qquad U = (g_1^2 - 1)g_2/\sqrt{2}.$$

At $\theta = 0$, F and U each have variance one and zero covariance. The Jacobian-product expectation in (25) is zero. Gaussian moments give

$$\mathrm{Cov}_\theta(U,F) = (3\theta - \theta)/\sqrt{2} = \sqrt{2}\,\theta.$$

Its derivative is $\sqrt{2}$, entirely contributed by $\mathrm{Cov}_0(\partial_{12}U, F)$. Even exact first derivatives do not remove this curvature term. The example concerns a smooth Gaussian-input system; it does not measure curvature error in the Burgers experiment.

An equivalent score expression. The Gaussian source-law score is

$$\psi(g) = \tfrac{1}{2}\{g^\top \Sigma_0^{-1}\Delta_{ij}\Sigma_0^{-1} g - \mathrm{tr}(\Sigma_0^{-1}\Delta_{ij})\}.$$

Differentiating the Gaussian density under the integral gives $\partial_\theta \mathbb{E}f = \mathbb{E}_0[f\psi]$ and $\mathbb{E}_0\psi = 0$. It follows that the full response is

$$\frac{d}{d\theta}\,\mathrm{Cov}(U,F)\bigg|_0 = \mathbb{E}_0[(U - \mathbb{E}_0 U)(F - \mathbb{E}_0 F)\psi]. \tag{26}$$

For $\Sigma_0 = I$, $\psi = g_i g_j$. This is a source-only alternative for evaluating the exact moment response if the paired latent forcing and residual observations are available. The experiments use the finite-difference construction already specified in the method.

A.17 Prior mass, sequential evidence and prefix averaging

A.17.1 Risk improvement from averaging and the contribution of earlier predictors

Fix a calibration sample and suppose all the risks below are finite. Let $N = K + 1$ and

$$J_K = \mathbb{E}_{Z,U}\left[\log \bar{H}_K(U \mid Z) - \frac{1}{N}\sum_{t=0}^{K} \log H_t(U \mid Z)\right].$$

The pointwise integrand is nonnegative by concavity of the logarithm. Subtracting the finite KL risks gives the exact identities

$$A(\bar{H}_K) = \frac{1}{N}\sum_{t=0}^{K} A(H_t) - J_K, \tag{27}$$

$$A(\bar{H}_K) - A(H_K) = \frac{1}{N}\sum_{t=0}^{K} [A(H_t) - A(H_K)] - J_K. \tag{28}$$

The first term in (28) compares earlier predictors with the endpoint. It is positive when those earlier predictors have higher average risk, but it has no fixed sign in general. The second term is the gain from density averaging. The identity explains how a prefix estimator with an expected guarantee can score higher than the last posterior. The experimental comparisons measure the left side; they do not separately estimate the two right-hand terms.

A.17.2 C28 IS A MIXTURE OVER SOURCE-DIRECTION PREDICTORS

For each of the $L = 28$ prescribed directions, form a conceptual C774 block with that direction and the same radial reference and prior. Let

$$\mathcal{E}_{\ell,t} = \sum_j \pi_j \prod_{i=1}^{t} h_{\ell j}(u_i \mid z_i), \qquad \mathcal{E}_{\ell,0} = 1.$$

The C28 prior is exactly the uniform mixture of these block priors. Collapsing the 28 duplicate copies of each shared radial candidate gives its implemented 19,350-label representation. Thus

$$\mathcal{E}_{28,t} = \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{E}_{\ell,t}, \qquad H_{28,t} = \sum_{\ell=1}^{L} \frac{\mathcal{E}_{\ell,t}}{\sum_m \mathcal{E}_{m,t}} H_{\ell,t}. \tag{29}$$

The second equality follows either by expanding the posterior, or by taking the ratio of the next-step evidence to the current evidence. For any sequence and any block ℓ, telescoping then yields

$$\sum_{t=0}^{N-1} \log \frac{H_{\ell,t}(u_{t+1} \mid z_{t+1})}{H_{28,t}(u_{t+1} \mid z_{t+1})} = \log \frac{\mathcal{E}_{\ell,N}}{\mathcal{E}_{28,N}} \le \log L. \tag{30}$$

This pathwise statement compares cumulative sequential log losses. It needs no sampling assumption. For i.i.d. observations independent of the frozen source, averaging over data gives

$$\frac{1}{N}\sum_{t=0}^{N-1} \mathbb{E}[A(H_{28,t}) - A(H_{\ell,t})] \le \frac{\log L}{N}.$$

At $N = K+1$, the extra observation is a proof device; neither deployed prefix uses it. Applying (27) to the two prefixes gives, when the terms are finite,

$$\mathbb{E}A(\bar{H}_{28,K}) - \mathbb{E}A(\bar{H}_{\ell,K}) \le \frac{\log L}{K+1} + \mathbb{E}J_{\ell,K} - \mathbb{E}J_{28,K}. \tag{31}$$

The Jensen correction is essential. The value $\log(28)/33$ alone is not a bound on the risk difference between the deployed prefixes. Equation (29) identifies an exact mechanism of the implemented expansion: evidence must be allocated across direction blocks as well as within them.

A.17.3 NEAR-OPTIMAL PRIOR MASS AND IDENTICAL CANDIDATES

Optimizing the right side of the atomic oracle over ν gives

$$\mathbb{E}A(\bar{H}_K) \le -\frac{1}{N}\log \sum_j \pi_j \exp\{-NA(h_j)\}, \qquad N = K+1. \tag{32}$$

For finite risks, put $Z_A = \sum_j \pi_j e^{-NA(h_j)}$ and $\nu_j^* = \pi_j e^{-NA(h_j)}/Z_A$. The oracle objective is

$$-\frac{\log Z_A}{N} + \frac{\mathrm{KL}(\nu\|\nu^*)}{N}.$$

It is minimized at $\nu = \nu^*$, proving the formula. Infinite-risk candidates contribute zero and can be omitted as long as one finite-risk comparison remains. If $A_* = \min_j A(h_j)$ and $S_\eta = \{j : A(h_j) \le A_* + \eta\}$, then

$$\mathbb{E}A(\bar{H}_K) \le A_* + \eta - \frac{\log \pi(S_\eta)}{N}.$$

This is still an atomic comparison. It shows why useful prior mass is more informative than candidate count alone. Splitting one candidate into identical copies while preserving their total prior changes neither the evidence nor any posterior predictive: the copies have identical likelihood factors and their weights add back to the original one.

Persistent gaps are possible even inside a two-candidate hull. Let $h_0 \neq h_1$ be mutually absolutely continuous with finite divergences $D_{01} = \mathrm{KL}(h_0 \| h_1) > 0$ and $D_{10} = \mathrm{KL}(h_1 \| h_0) > 0$. Let $P_\eta = (1-\eta)h_0 + \eta h_1$, $0 < \eta < 1$, with positive prior on both candidates. The expected log-likelihood ratio is

$$\mathbb{E}_{P_\eta} \log(h_1/h_0) = -(1-\eta)D_{01} + \eta D_{10}.$$

For $\eta < D_{01}/(D_{01} + D_{10})$, the strong law forces posterior mass on h_1 to zero, so both the last posterior and its Cesaro prefix average converge to h_0. The hull optimum is zero while the limiting risk is $\mathrm{KL}(P_\eta \| h_0) > 0$. Finiteness of the two mutual divergences implies integrability of the log-likelihood ratio and finite risks of both candidates under P_η. Convexity bounds every mixture risk by the maximum of those two risks. Lower semicontinuity and the bound $A(q) \leq A(h_0) - \log w_0$ as $w_0 \to 1$ show risk convergence; the uniform finite bound also gives convergence in expectation. The equality case at the threshold has no strict drift and is not covered. This argument generalizes the asymmetric Bernoulli example without replacing the deployed estimator by a different mixture-fitting method.

A.18 Consequences for joint events and approximate symmetry

The following events explain what the symmetry restriction can mean for probabilities. They are analytic queries on the prescribed Gaussian construction, distinct from the two measured regional events.

A task-dependent sign query. In the off-axis construction, let $T = t^\top U$ and $\gamma = b\sqrt{1-c^2}/\tau > 0$. Every invariant conditional density predicts $\Pr(T > 0 \mid F) = 1/2$. The target probability is $p(F) = \Phi(\gamma F)$. Under squared probability loss, the excess conditional Brier risk over the true probability is

$$\mathbb{E}_F[p(F) - 1/2]^2 = \frac{\arcsin(\rho)}{2\pi} = \frac{L_{\mathrm{sym}}}{2}, \qquad \rho = \frac{\gamma^2}{1+\gamma^2}.$$

This follows from $\mathbb{E}\Phi(\gamma F) = 1/2$ and the Gaussian quadrant calculation used to prove L_{sym}.

A joint event when the supplied task marginal is correct. For this statement additionally supply the true standard-normal law of F. Let $E = \{F > 0, T > 0\}$. Every invariant conditional predictor then assigns it probability $1/4$. Under the target, the correlation of F and T is $\gamma/\sqrt{1+\gamma^2}$, so

$$P(E) = \frac{1}{4} + \frac{\arctan\gamma}{2\pi}.$$

The unconditional excess Brier risk is exactly $[\arctan(\gamma)/(2\pi)]^2$. Indeed, for an event with true probability p and prediction q, $\mathbb{E}(q - \mathbf{1}_E)^2 = p(1-p) + (q-p)^2$. A negative b can be treated by reversing the sign of F in the event. These formulas do not apply to a misspecified supplied task marginal without accounting for its error. They make no claim about improvement on every event.

A shape-sensitive event for heavy-tailed elliptical targets. The sign query detects a conditional location shift. A different bounded event detects shape anisotropy without any moment assumptions. Let $\dim V^\perp \geq 2$, let w be a unit direction outside V, and set

$$\delta = \|\Pi_\perp w\|^2 > 0, \qquad t = \Pi_\perp w/\sqrt{\delta}.$$

Choose a unit $r \in V^\perp$ orthogonal to w. Consider, conditionally on each z, the centered elliptical law

$$U = \sqrt{a}\,[I + (s-1)ww^\top]R_z\Omega, \qquad a > 0, \quad s > 0,$$

where Ω is uniform on the unit sphere in $\mathbb{R}^d$, $R_z > 0$ almost surely, and R_z is independent of Ω given z. Its radial distribution can vary with z and can be heavy-tailed. The orthogonal map exchanging t and r and fixing their orthogonal complement belongs to G_V. Every invariant density consequently assigns probability $1/2$ to

$$E_{\mathrm{shape}} = \{|t^\top U| > |r^\top U|\}.$$

The boundary is a union of two hyperplanes and is null for a Lebesgue density. Under the elliptical target,

$$p_{\mathrm{shape}} = \frac{2}{\pi} \arctan \sqrt{1 + (s^2 - 1)\delta}, \qquad A(h) \geq \mathrm{KL}\big(\mathrm{Bern}(p_{\mathrm{shape}}) \| \, \mathrm{Bern}(1/2)\big) \geq 2(p_{\mathrm{shape}} - 1/2)^2. \tag{33}$$

To prove the probability formula, write $A_s = I + (s-1)ww^\top$. The row vectors $A_s t$ and $A_s r = r$ are orthogonal, with lengths $\sqrt{1 + (s^2-1)\delta}$ and one. The common radial factor cancels in their magnitude comparison. Two orthogonal coordinates of a uniform sphere have the same ratio as two independent standard Gaussians, whose ratio is Cauchy. Integrating its density over the appropriate interval gives the displayed arctangent. Binary data processing proves the first lower bound and Pinsker proves the second.

This obstruction is independent of radial scale and tails and is positive when $s \neq 1$. It remains valid even if the target has no second moments, in which case s describes shape rather than covariance. It gives a separate mechanism from the mean-based off-axis construction. No matched atom or empirical comparison for this centered elliptical target is asserted for C28.

Approximate reflection symmetry. Let Q be an orthogonal reflection with $Qt = -t$. For each conditional density h_z, define $h_z^{\mathrm{sym}} = (h_z + Q_{\#} h_z)/2$ and use $\mathrm{TV}(P, Q) = \sup_E |P(E) - Q(E)|$. Suppose all candidate densities satisfy

$$\mathrm{TV}(h_j(\cdot \mid z), h_j^{\mathrm{sym}}(\cdot \mid z)) \leq \delta(z).$$

Convexity gives the same inequality for every mixture, including a data-dependent one. Lebesgue densities give zero mass to $t^\top U = 0$, so each symmetrized density assigns probability $1/2$ to $t^\top U > 0$. Thus any such mixture has sign probability within $\delta(z)$ of $1/2$. If the target sign probability is $p(z)$, binary data processing and Pinsker give

$$A(h) \geq 2\mathbb{E}_Z \left(\left| p(Z) - \tfrac{1}{2} \right| - \delta(Z) \right)_+^2 . \tag{34}$$

The exact-symmetry bound is recovered at $\delta = 0$. The tolerance here is total variation from the symmetrized law, which equals half the total variation from the reflected law. A few pointwise log-density checks do not establish such a global tolerance.