SELECTIVE ORDINAL GATING NETWORK: STRUCTURE-GATED TRAINING VIA THE EXTERNAL REFERENCE PRINCIPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for selective regression often estimate reliability from model predictions or learned representations. We propose an alternative approach that uses a learned ordinal structure of the target to identify samples with concentrated and coherent conditional target distributions. Our Selective Ordinal Gating Network (SOGN) combines an ordinal auxiliary head with a differentiable concentration–coherence gate and a stop-gradient weighted regression objective. We theoretically analyze how coherent ordinal probabilities can imply a narrow region in target space, and characterize the assumptions under which this implication holds. Across multiple benchmark datasets, SOGN consistently improves selective-regression performance at low coverage compared with uncertainty-based and rejection-based baselines. These results demonstrate the value of ordinal target structure for sample selection while clarifying the method's theoretical limitations and boundary conditions.

1 INTRODUCTION

In high-stakes regression settings—industrial equipment monitoring, financial risk control, clinical decision support—data heterogeneity means high-noise and low-noise samples carry equal gradient weight during training, contaminating model representations.[1] Two questions arise: how to identify contaminated samples, and how to suppress their gradients while preserving learning on clean samples.

Prior work shares an implicit premise of **self-reference**: reliability assessment originates from the model itself. Uncertainty methods (MC Dropout Gal & Ghahramani (2016), Deep Ensemble Lakshminarayanan et al. (2017)) use predictive variance as a reliability proxy, but variance shares the encoder with prediction—degradation in difficult regions impairs both. SelectiveNet Geifman & El-Yaniv (2019) learns a rejection head parallel to the predictor, but the rejection function shares representations with prediction and its rationale is opaque. Two-stage screening decouples screening from regression entirely.

Definition 1.1 (Self-Reference)**.** *A reliability evaluation function $\rho_{\mathcal{M}} : \mathcal{X} \to [0, 1]$ follows **self-reference** if it depends solely on the internal state of model $\mathcal{M}$ (parameters, hidden representations, predictions, predictive variance).*

Self-reference creates an inescapable dilemma: when the model is contaminated in a region, its self-assessment in that region is also contaminated—it neither knows the correct answer nor knows that it does not know.

We propose using **target-structure reference**: reliability signals can be derived from an auxiliary ordinal view of the target, rather than directly from the regression residual or predictive variance.

Definition 1.2 (External Reference Principle)**.** *The reliability signal $\rho(\boldsymbol{x})$ is computed as $\psi(\mathcal{S}(\boldsymbol{x}))$, where $\mathcal{S}(\boldsymbol{x})$ is an auxiliary target-structure prediction. It is trained from ordinal target labels and is not computed from the regression residual or predictive variance.*

[1] We use "contamination" as an umbrella for heteroscedasticity, irreducible error, and model misspecification; see Appendix B.9 for a taxonomy of which types SOGN targets.

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.

We instantiate this idea on ordinal target structure. After equal-frequency binning of the target range into K ordered intervals, each sample is represented by $\boldsymbol{p}^*(\boldsymbol{x}) \in \Delta^{K-1}$, where $p_k^*(\boldsymbol{x}) = \mathbb{P}(Y \in [b_{k-1}, b_k) \mid \boldsymbol{x})$. We extract two geometric signals: **concentration** ($\max_k p_k^*$, whether probability mass clusters in a single interval) and **coherence** (whether high-probability intervals are adjacent). These are target-structure signals, although their estimates are still learned from the input and can be affected by representation quality.

Definition 1.3 (Training-Time Anti-Contamination)**.** *Given $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ and a regression model f_θ, design a diagnosis-intervention mechanism: (i) diagnosis uses only target structure signals, outputting interpretable quality scores; (ii) intervention suppresses gradients of high-contamination-risk samples; (iii) diagnosis and regression are jointly trained, sharing encoder representations.*

We propose the **Selective Ordinal Gating Network (SOGN)**. SOGN extracts diagnostic signals through an ordinal pathway, suppresses the gradients of low-score samples via stop-gradient weighting, and provides interpretable concentration/coherence evidence for the ranking.

Our contributions are: (1) a target-structure-assisted selective-regression architecture based on concentration and coherence; (2) a differentiable HC–LS gate with stop-gradient weighted training and target-coverage control; and (3) a conditional variance-compression result that clarifies when coherent ordinal probabilities imply a narrow target-space region, together with empirical component ablations and boundary-condition analysis.

Unlike predictive-variance methods and rejection heads that directly use regression uncertainty or residual surrogates, SOGN uses an auxiliary ordinal target view to construct its score. The distinction is operational rather than absolute: the ordinal head shares the encoder with regression and is therefore not independent of representation quality. Conformal prediction Vovk et al. (2005) addresses inference-time coverage, whereas SOGN focuses on training-time gradient weighting and selective ranking.

2 The SOGN Method

2.1 Ordinal Gating Module

The ordinal gating module is the architectural instantiation of the External Reference Principle in regression settings. Given a training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ (we study the single-target case), we begin by equal-frequency binning of the target variable Y based on empirical quantiles of the training set, partitioning the range into K ordered intervals $[b_0, b_1), \ldots, [b_{K-1}, b_K]$, where $b_0 = \min_i y_i$ and $b_K = \max_i y_i$. Denote $w_{\max} = \max_k(b_k - b_{k-1})$ as the maximum interval width and $R = b_K - b_0$ as the range. Equal-frequency binning controls the number of training samples per interval, but does not in general make the numerical widths equal; all guarantees below therefore retain the observable quantity $w_{\max}$. Each sample's scalar target y is mapped to an ordinal label $c = \mathrm{bin}(y) \in \{1, \ldots, K\}$.

SOGN uses a shared encoder Tishby et al. (2000) with one ordinal head and one regression head. The ordinal head produces the gating weight w through the differentiable tests below, while the regression head $f_{\text{reg}} : \mathbb{R}^h \to \mathbb{R}$ produces the scalar prediction $\hat{y}$. The two heads share the representation and interact through the training objective; the ordinal signal is therefore target-structure-assisted rather than statistically independent of the regressor.

Ordinal projection. The hidden representation $\boldsymbol{h}$ is projected onto the K-dimensional simplex:

$$\boldsymbol{p}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_{\text{ord}}\boldsymbol{h} + \boldsymbol{b}_{\text{ord}}) \in \Delta^{K-1}, \tag{1}$$

where $p_k(\boldsymbol{x})$ estimates $\mathbb{P}(Y \in [b_{k-1}, b_k) \mid \boldsymbol{x})$. Using softmax directly preserves the full geometric structure of the probability simplex, enabling concentration and coherence as geometric measurements with continuous gradient flow.

Concentration gate. Concentration measures whether the ordinal distribution clusters in a single interval. The hard condition $\max_k p_k > \tau$ is approximated by the α-order power mean $M_\alpha(\boldsymbol{p}) = (\frac{1}{K}\sum_k p_k^\alpha)^{1/\alpha}$ (α=5), which is everywhere differentiable and converges to $\max$ as $\alpha \to \infty$. The gating signal is $g_{\text{conc}} = \sigma(\beta(M_\alpha(\boldsymbol{p}) - \tau))$, where $\tau = \sigma(\tau_{\text{raw}})$ is learnable.

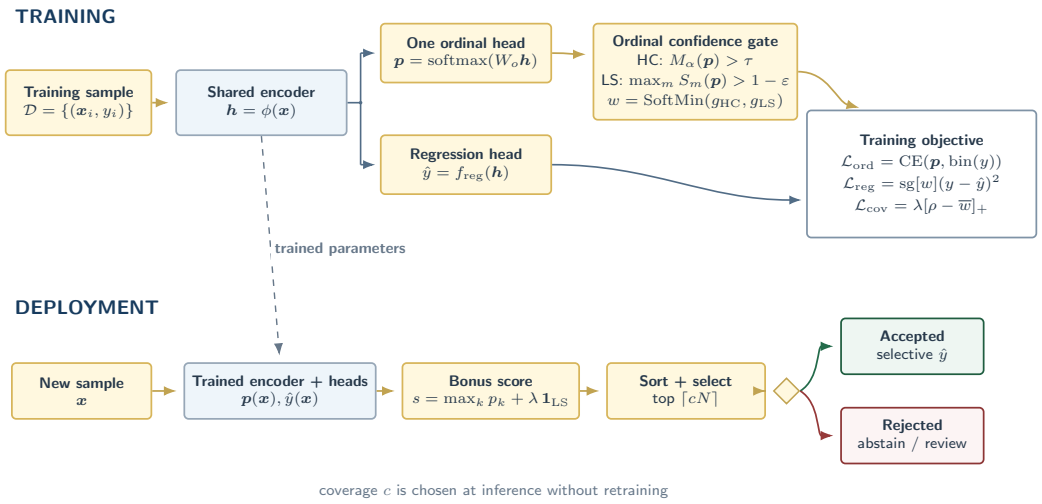

Figure 1: Overview of SOGN. A shared encoder feeds one ordinal head and one regression head. The ordinal probabilities are converted into concentration (HC) and local coherence (LS) signals, fused by an AND gate, and used with stop-gradient weighted regression and target-coverage control during training. At deployment, the bonus score ranks samples and retains the requested coverage; samples outside the selected set can be abstained from or sent for review.

Coherence gate. Concentration alone does not guarantee adjacent high-probability intervals. Define a sliding window sum $S_m(\boldsymbol{p}) = p_m + p_{m+1} + p_{m+2}$ ($m = 1, \ldots, K-2$); the hard condition $\max_m S_m(\boldsymbol{p}) > 1 - \varepsilon$ is approximated via LogSumExp: $\mathcal{M}_\beta(\boldsymbol{p}) = \frac{1}{\beta} \log \sum_m \exp(\beta S_m(\boldsymbol{p}))$, with $g_{\mathrm{coh}} = \sigma(\beta(\mathcal{M}_\beta(\boldsymbol{p}) - (1 - \varepsilon)))$.

Gate fusion. Concentration and coherence form an AND relationship, approximated by SoftMin ($\mathrm{SoftMin}_\gamma(a, b) = -\frac{1}{\gamma} \log(e^{-\gamma a} + e^{-\gamma b})$, converging to $\min$):

$$w(\boldsymbol{x}) = \mathrm{SoftMin}_\gamma\big(g_{\mathrm{conc}}(\boldsymbol{p}(\boldsymbol{x})),\ g_{\mathrm{coh}}(\boldsymbol{p}(\boldsymbol{x}))\big). \tag{2}$$

Training uses linear annealing $\beta(t) = \beta_{\mathrm{init}} + (\beta_{\max} - \beta_{\mathrm{init}}) \cdot \min(t/T_{\mathrm{anneal}}, 1)$ (β_{init}=1, $\beta_{\max}$=50) to resolve the tension between smooth gradients early in training and hard decisions at inference Guo et al. (2017).

The three-bin coherence test gives samples near a single bin boundary some tolerance, but it does not remove the dependence on the chosen marginal binning. This is an explicit applicability condition of the present single-target method; handling multiple target topologies is left to future work.

2.2 Anti-Contamination Training

The total loss of SOGN consists of three terms:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ord}} + \mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{cov}}. \tag{3}$$

The ordinal classification loss $\mathcal{L}_{\mathrm{ord}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}(\boldsymbol{p}_i, \mathrm{bin}(y_i))$ ensures that the ordinal projection outputs meaningful interval probabilities. The weighted regression loss is the core of anti-contamination:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{sg}[w_i] \cdot (y_i - \hat{y}_i)^2, \tag{4}$$

where $\mathrm{sg}[\cdot]$ is the Stop-Gradient operator. In the forward pass w serves as a sample weight; in the regression-loss backward pass its gradient is blocked. Thus residual magnitude cannot directly reshape the ordinal gate, while the ordinal loss and coverage term still train the gate. The coverage

regularization used in the implementation is a target-coverage hinge:

$$\mathcal{L}_{\text{cov}} = \lambda_{\text{cov}} \left[\rho - \frac{1}{N} \sum_{i=1}^{N} w_i \right]_+ , \tag{5}$$

where $[z]_+ = \max(z, 0)$ and ρ is the prescribed target coverage. It penalizes average coverage below the target without rewarding unnecessary increases once that target is met.

SOGN adopts two-stage training: Stage 1 (ordinal pretraining, 30–40 epochs) trains the encoder and ordinal head using only $\mathcal{L}_{\text{ord}}$; Stage 2 (joint training, 120–200 epochs) uses the full loss $\mathcal{L}_{\text{total}}$ with temperature annealing. On some datasets, an optional third stage retrains only the regression head with plain MSE (gate and encoder frozen), recovering full-coverage regression quality without altering the learned confidence ranking.

2.3 Inference and Adaptive Coverage Control

While $\mathcal{L}_{\text{cov}}$ prevents gate collapse during training, inference requires flexible coverage specification without retraining. Traditional two-stage methods use binary search over thresholds Corbière et al. (2019) at $O(T \cdot N_{\text{val}})$ cost. We propose bonus scoring: $s(\boldsymbol{x}) = \max_k p_k(\boldsymbol{x}) + \lambda \cdot \mathbf{1}_{[\max_m S_m(\boldsymbol{p}) > 1-\varepsilon]}$ (λ=1.0). This requires a single sort ($O(N_{\text{val}} \log N_{\text{val}})$) with $O(1)$ threshold lookup; when coverage does not exceed the coherent proportion, it is exactly equivalent to two-stage screening; beyond that, non-coherent samples are automatically supplemented in descending concentration order (Theorem 3.2).

3 Theoretical Analysis

Statistical foundation. The ordinal bin distribution is a target-structure-assisted signal whose relationship to f^* depends on the data-generating process. Under explicit location/noise assumptions it can behave approximately like an ancillary precision descriptor, but the present method does not require or claim full ancillarity. The conditional variance result below states the precise structural assumption used by our analysis.

3.1 Preliminaries and Assumptions

Assumption 3.1 (Ordered Equal-Frequency Binning)**.** *The target range $[b_0, b_K]$ of Y is partitioned into K ordered intervals by empirical equal-frequency binning. Denote $w_{\max} = \max_k(b_k - b_{k-1})$ and $R = b_K - b_0$. No equality between numerical interval widths is assumed.*

Assumption 3.2 (Bounded Within-Interval Variance)**.** *There exists $\sigma^2_{\max} > 0$ such that $\forall \boldsymbol{x} \in \mathcal{X}, k \in [K]$, $\mathrm{Var}(Y \mid \boldsymbol{x}, g(Y) = k) \le \sigma^2_{\max}$ (typically $\sigma^2_{\max} \approx w^2_{\max}/4$).*

The approximation bounds used below are stated in Appendix A. In particular, the power mean has a dimension-dependent gap from the maximum, while LogSumExp and SoftMin have their usual $O(\log L/\beta)$ and $O(\log 2/\gamma)$ gaps.

Assumption 3.3 (Ordinal calibration bridge)**.** *For the fitted ordinal head, there is a validation- or population-level event $\mathcal{E}_N$ with probability at least $1 - \delta'_N$ on which*

$$\|\hat{\boldsymbol{p}}_N(\boldsymbol{x}) - \boldsymbol{p}^*(\boldsymbol{x})\|_\infty \le \delta_N \tag{6}$$

for the samples under consideration. Here $\boldsymbol{p}^*(\boldsymbol{x})$ denotes the true conditional bin-probability vector. The result below is conditional on this bridge; its rate depends on the capacity, optimization, and calibration procedure and is not asserted to be universally $O(N^{-1/2})$.

3.2 Ordinal Variance Compression

The following theorem gives the conditional mathematical link between the ordinal signal and predictability. It does not claim that ordinal structure is independent of the learned representation, nor that the bound applies when the target topology is not a total order.

Theorem 3.1 (Ordinal Variance Compression Bound). *Under Assumptions 3.1–3.2 and the calibration event $\mathcal{E}_N$ in Assumption 3.3, for any sample $\boldsymbol{x}$ satisfying the coherence condition $\max_m S_m(\hat{\boldsymbol{p}}_N(\boldsymbol{x})) > 1 - \varepsilon$, with probability at least $1 - \delta'_N$:*

(i) Variance compression (joint bound):

$$\sigma^2(\boldsymbol{x}) \equiv \mathrm{Var}(Y \mid \boldsymbol{x}) \leq \sigma^2_{\max} + \frac{9w^2_{\max}}{4} + (\varepsilon + 3\delta_N)R^2. \tag{7}$$

(ii) Range compression: there exists an interval I_S (depending on $\boldsymbol{x}$) consisting of three consecutive bins such that $\mathbb{P}(Y \in I_S \mid \boldsymbol{x}) \geq 1 - \varepsilon - 3\delta_N$ and $|I_S| \leq 3w_{\max}$.

Proof sketch. Via the law of total variance, decompose $\sigma^2(\boldsymbol{x})$ into within-group and between-group terms. On $\mathcal{E}_N$, coherence transfers from $\hat{\boldsymbol{p}}_N$ to $\boldsymbol{p}^*$ with cost $3\delta_N$. The resulting three-bin window has width at most $3w_{\max}$; splitting the between-group term into in-window and out-of-window components gives the displayed bound. The full proof is in Appendix A.2. □

When the bins are reasonably balanced in numerical width, $w_{\max}$ is of order R/K and the first term has the familiar $O(R^2/K^2)$ scale; for skewed target distributions the rigorous statement is the bound in terms of $w_{\max}$.

3.3 Bonus Scoring: Equivalence and Graceful Degradation

Adding a bonus to coherent samples ranks them ahead of non-coherent ones. When coherent samples are insufficient, the remaining slots automatically fill from non-coherent samples by concentration—a single sort covers all α.

Theorem 3.2 (Bonus Scoring). *Let α_{LS} be the proportion of coherent samples on the validation set. Define $s(\boldsymbol{x}) = \max_k p_k(\boldsymbol{x}) + \lambda \cdot \mathbf{1}_{[LS]}$ with $\lambda > 1 - \min_{\mathcal{S}_{LS}} \max_k p_k(\boldsymbol{x})$. Then: (i) all coherent samples have strictly higher scores than non-coherent; (ii) for $\alpha \leq \alpha_{LS}$, bonus scoring is exactly equivalent to two-stage screening; (iii) for $\alpha > \alpha_{LS}$, non-coherent samples are supplemented by concentration; (iv) complexity is $O(N_{val} \log N_{val})$ with $O(1)$ threshold lookup for any α.*

Proof. For $\boldsymbol{x}_1 \in \mathcal{S}_{\text{LS}}$, $s(\boldsymbol{x}_1) = \max_k p_k + \lambda > 1$; for $\boldsymbol{x}_2 \notin \mathcal{S}_{\text{LS}}$, $s(\boldsymbol{x}_2) = \max_k p_k \leq 1$. (ii)–(iv) follow from stratification. In practice λ=1.0 suffices. □

The variance-compression result is deliberately stated as a conditional link, not as a minimax separation theorem. It does not establish that every structure-blind baseline is asymptotically suboptimal. Accordingly, the empirical comparison below is interpreted as evidence about selective ranking on the evaluated datasets, rather than as verification of a universal lower bound.

4 Experiments

Code will be made publicly available upon acceptance. Below we report only method-defining hyperparameters and key experimental design decisions; for full implementation details (optimizer configuration, training epochs, random seeds, etc.), see the supplementary materials.

Theory-informed hyperparameter selection (see Section 3 for the conditional bound):

- $K = 6$: balances the narrower-bin benefit against ordinal classification difficulty; the bound uses the data-dependent $w_{\max}$.
- $\alpha = 5$: a smooth, finite-temperature surrogate for the maximum; its approximation gap is reported by Lemma A.1 and is not treated as a universal 1% guarantee.
- $\varepsilon = 0.1$: a conservative coherence tolerance, chosen to retain a usable selected set.
- $\beta_{\max} = 50$: LogSumExp error $< \log K/\beta_{\max} < 0.04$ (Lemma A.2).

4.1 Experimental Setup

4.1.1 Datasets and Preprocessing

We use four public benchmark datasets with clear ordinal target structure: California Housing Pace & Barry (1997) (20,640 samples, 8 features, median house value), Bike Sharing Fanaee-T & Gama (2014) (2,893 time-series windows, 12 features, rental count), Diabetes Efron et al. (2004) (442 samples, 10 features, disease progression), and ApplianceEnergy Candanedo et al. (2017) (19,735 samples, 25 features, appliance consumption). Each dataset uses one fixed 7:1.5:1.5 train/validation/test split (split seed 42), with input standardization. For each fixed split, training is repeated with three independent training seeds (42, 43, 44), and we report mean $\pm$ standard deviation. We additionally evaluate on RC-49 Geifman & El-Yaniv (2017), a circular-target image dataset, as a boundary condition for the total-order assumption (Appendix B.2).

4.1.2 Baseline Methods

We compare against 6 methods: Standard Regression (random scores, lower-bound reference); MC Dropout Gal & Ghahramani (2016) (p=0.2, M=50 forward passes, negative predictive variance); Deep Ensemble Lakshminarayanan et al. (2017) (5 members for tabular, 3 for RC-49, negative ensemble variance); Deep Evidential Regression Amini et al. (2020) (Normal-Inverse-Gamma evidence ν); SelectiveNet Geifman & El-Yaniv (2019) (joint prediction–rejection head, selection probabilities); and Split Conformal Papadopoulos et al. (2002) (kNN-based Cover & Hart (1967) local residuals, 30% calibration set, negative residuals).

4.1.3 Evaluation Protocol

Test samples are sorted by score; the top $\lceil c \cdot N \rceil$ are selected at $c \in \{0.1, \ldots, 1.0\}$. R^2 is the core metric—scale-invariant, measuring explained variance in the selected subset. Figure 2 shows the complete R^2–coverage curves for all four ordinal benchmarks; the corresponding tabulated values are provided in Appendix B.2.

4.2 Main Experiment: Coverage-Selective Regression

Table 1 presents the R^2 comparison at key coverages across the four datasets; full tables are in Appendix B.2.

SOGN achieves the best R^2 at low coverage ($c \leq 0.5$) on all four ordinal benchmarks. At c=0.1, its margin over the strongest baseline ranges from 0.045 to 0.703; the large Diabetes margin is driven by the unusually weak baseline scores on this small-sample dataset. The R^2 curves generally decrease as coverage increases, although the small Diabetes test set shows sampling fluctuations. At c=1.0, SOGN trails Deep Ensemble slightly (which has a $3\times$ parameter advantage), indicating a modest full-coverage/selectivity trade-off. Figure 3 reports the complementary RMSE and MAE risk–coverage results on California Housing: SOGN's ranking is not uniformly optimal for absolute error, which is consistent with the paper's distinction between explained-variance selectivity and absolute-error selectivity.

ApplianceEnergy shows the clearest selective behavior: SOGN ranks first across $c \leq 0.9$ with the largest dynamic range (ΔR^2=0.165 in the reported table). Its U-shaped energy–weather relationship provides a useful target-structure signal, whereas several structure-blind methods show non-monotonic or reversed curves.

Beyond the four ordinal benchmarks, we probe the boundary of the total-order assumption on RC-49, whose rotation-angle target is circular. Consistent with our target-structure taxonomy (Appendix B.10), SOGN's selective signal degrades sharply here: at c=0.1, its advantage over the random-selection baseline collapses from 0.703 on Diabetes (0.690 vs. -0.013) to 0.004 on RC-49 (0.977 vs. 0.973), while its full-coverage regression remains second (0.973). This dissociation—selective signal destroyed, regression preserved—is what Theorem 3.1 predicts when no valid ordinal structure exists to compress variance. A full analysis, including the residual confound between target topology and feature type, is in Appendix B.2.

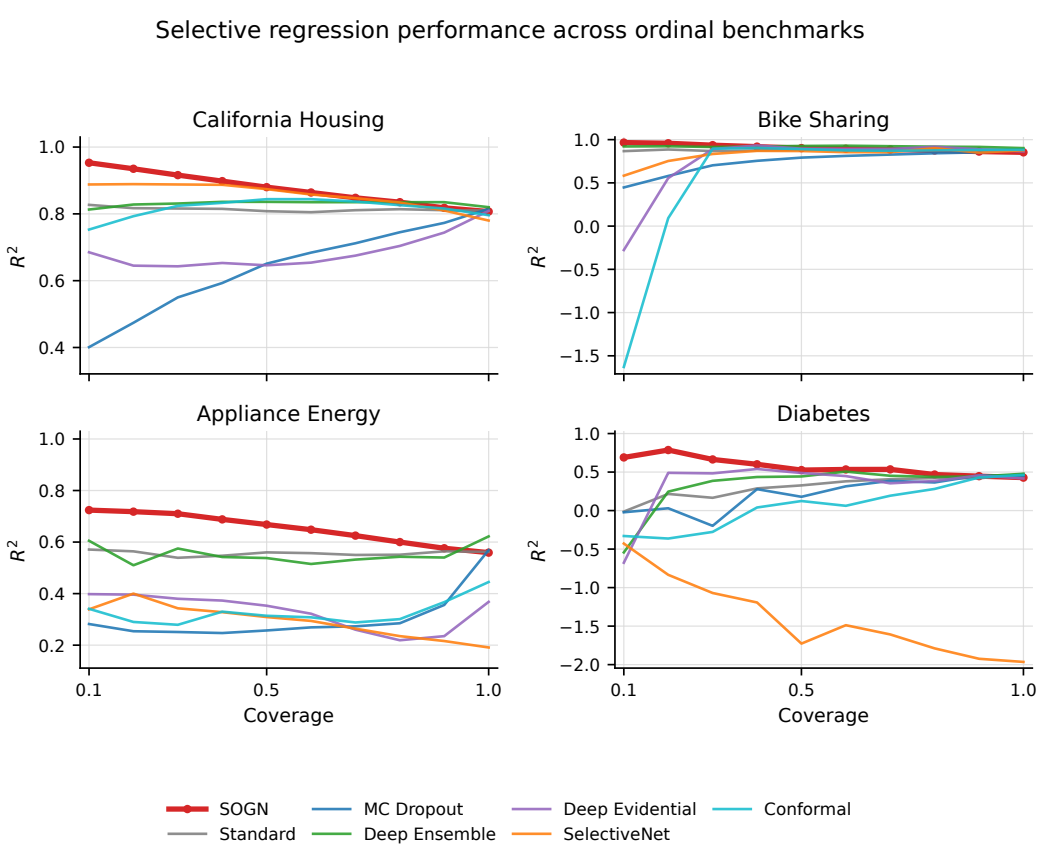

Figure 2: R^2–coverage curves of SOGN and baseline methods on all four ordinal benchmarks. Each curve connects the 3-seed means reported in Appendix B.2; the figure uses the manuscript's reported means and does not interpolate additional observations.

Structure-blind methods perform unstably on the four ordinal-structure-dominated datasets, with R^2 generally below 0.49 on Diabetes (N=442). SelectiveNet produces negative R^2 at all coverages on Diabetes due to encoder-gradient sharing between rejection and regression heads; SOGN's Stop-Gradient prevents the regression residual term from directly changing the gate.

4.3 Ablation Study

The ablation study is conducted on the Diabetes dataset—its small-sample nature makes differences in component contributions more readily observable. We ablate each core component of SOGN: (A1) LS only—remove the HC gate, retaining only the LS gate; (A2) HC only—remove the LS gate, retaining only the HC gate; (A3) AND→OR—replace the SoftMin fusion with max; (A4) No Annealing—fix $\beta = \beta_{\max}$ without temperature annealing; (A5) No Stop-Gradient—remove the detach operation on w; (A6) No Pretraining—skip ordinal pretraining and train jointly from the start. In addition, we perform sensitivity sweeps over the key hyperparameters $\lambda_{\text{cov}} \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$ and $K \in \{4, 6, 8, 10, 12\}$.

Table 2 reports the R^2 of component ablation at four key coverages: $c = 0.1, 0.2, 0.5, 1.0$.

Ablation analysis. HC and LS are complementary, with the full SOGN configuration giving the strongest low-coverage result (SOGN 0.778 vs. HC-only 0.604, LS-only 0.531). AND fusion materially improves low-coverage R^2 over OR on this benchmark (0.778 vs. 0.502 at c=0.1), while annealing, Stop-Gradient, and ordinal pretraining are important for stable optimization. Removing Stop-Gradient causes a severe collapse in this run (R^2=−174.7), supporting its role in preventing

Table 1: R^2 comparison at key coverages across four datasets (3-seed means). Best results are bolded; second-best are underlined. See Appendix for full tables.

Dataset	Cov.	SOGN	Standard	MC-D	Ens.	Evid.	SelNet	Conf.
California	0.1	**0.953**	0.827	0.401	0.813	0.685	0.888	0.753
	0.3	**0.916**	0.816	0.550	0.831	0.643	0.888	0.824
	0.5	**0.880**	0.808	0.651	0.836	0.646	0.874	0.844
	0.7	**0.848**	0.811	0.712	0.835	0.675	0.847	0.837
	1.0	0.807	0.810	0.815	**0.820**	0.809	0.780	0.796
Bike	0.1	**0.966**	0.867	0.447	0.921	-0.277	0.583	-1.629
	0.3	**0.936**	0.868	0.703	0.915	0.891	0.834	0.895
	0.5	0.902	0.883	0.792	**0.927**	0.895	0.868	0.891
	0.7	0.886	0.893	0.827	**0.924**	0.891	0.855	0.871
	1.0	0.853	0.881	0.886	**0.902**	0.864	0.877	0.883
Appl. Energy	0.1	**0.724**	0.571	0.282	0.605	0.398	0.339	0.341
	0.3	**0.710**	0.539	0.251	0.575	0.380	0.343	0.279
	0.5	**0.668**	0.560	0.257	0.538	0.353	0.309	0.314
	0.7	**0.625**	0.550	0.273	0.532	0.260	0.264	0.288
	1.0	0.559	0.562	0.569	**0.622**	0.368	0.191	0.445
Diabetes	0.1	**0.690**	-0.013	-0.022	-0.543	-0.677	-0.428	-0.330
	0.3	**0.664**	0.166	-0.198	0.386	0.483	-1.070	-0.277
	0.5	**0.526**	0.327	0.179	0.443	0.488	-1.727	0.124
	0.7	**0.535**	0.407	0.386	0.452	0.354	-1.608	0.193
	1.0	0.428	0.474	0.453	**0.478**	0.417	-1.965	0.464

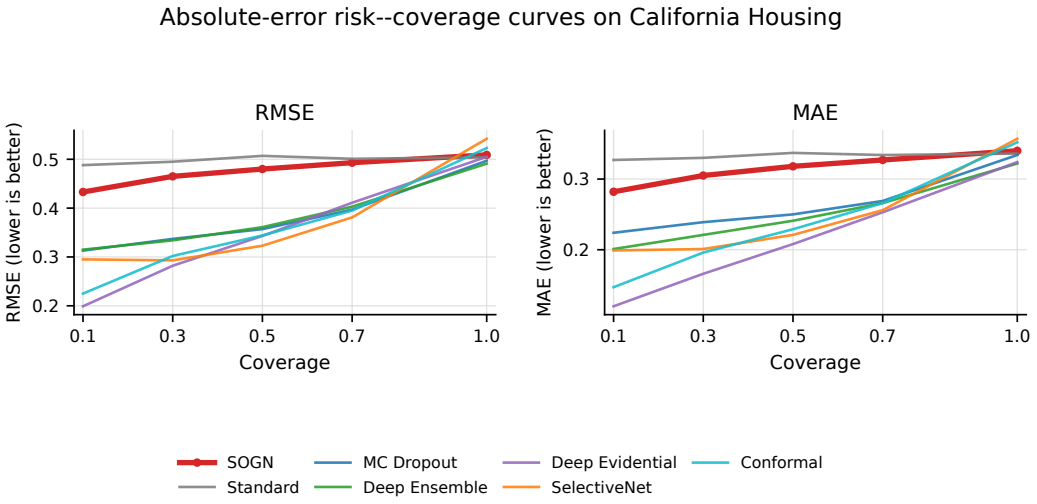

Figure 3: Absolute-error risk–coverage curves on California Housing. RMSE and MAE are lower-is-better metrics; values are the 3-seed means reported in Appendix B.5. The curves complement the R^2 analysis by showing the distinct absolute-error trade-off induced by selective ranking.

residual-driven gate updates; this is an empirical component result, not a formal universal guarantee. Hyperparameter sensitivity (Figure 4) shows stability within $\lambda_{\text{cov}} \in [0.01, 0.1]$ and $K \in [6, 12]$.

4.4 Auxiliary Experiments

Auxiliary experiments verify SOGN's robustness and mechanisms: (1) Backbone transfer: replacing LSTM with GRU and Transformer on the Bike Sharing dataset, SOGN outperforms Standard on 28 out of 30 data points across three architectures (Appendix B.3); (2) Statistical evidence: paired t-tests quantify the seed-level differences and their limited testing power (Appendix B.4); (3) Small-

Table 2: Complete SOGN component ablation: R^2 comparison (Diabetes dataset, three training seeds on a fixed split).

Variant	$c = 0.1$	$c = 0.2$	$c = 0.5$	$c = 1.0$
SOGN (Full)	0.778	0.597	0.607	0.422
LS only (no HC)	0.531	0.600	0.641	0.447
HC only (no LS)	0.604	0.655	0.529	0.440
AND→OR (max)	0.502	0.622	0.611	0.463
No Annealing	0.215	0.504	0.478	0.439
No Stop-Gradient	−174.7	−86.9	−80.8	−61.7
No Pretraining	−2.924	−0.135	0.220	0.078

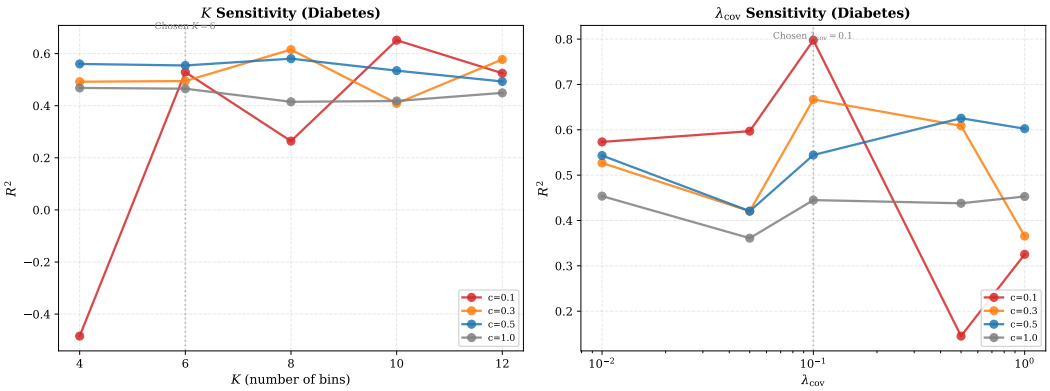

Figure 4: Hyperparameter sensitivity analysis (Diabetes dataset). Left: sensitivity to the number of bins K. Right: sensitivity to the coverage regularization coefficient λ_{cov}.

sample behavior: Diabetes ($N = 442$) exhibits the largest variability and the strongest contrast between SOGN and the baselines; (4) R^2-RMSE cross-metric decoupling: on California Housing, SOGN leads in R^2 at selected coverages but not in RMSE, suggesting that the learned signal captures a different notion of selective predictability (Appendix B.5); (5) Split-Sample robustness: with the training set split in half and gating and regression using disjoint data, the qualitative selective-ranking pattern is preserved, addressing the possible double-dipping concern (Appendix B.7).

5 Discussion and Conclusion

5.1 Boundary Conditions

Three boundaries delimit SOGN's applicability. First, **ordinal calibration**: the variance interpretation requires the fitted bin probabilities to satisfy the calibration bridge in Assumption 3.3; if this is not credible on held-out data, use SOGN only as a heuristic ranking or fall back to full-sample regression. Second, **target topology**: ordinal binning presumes a total order on Y; circular targets (e.g., RC-49 rotation angle) violate this, as shown in Appendix B.2, and alternative structures (hierarchical, graph, spherical) need different structure descriptors (Appendix B.10). Third, **feature dimensionality**: in high-dimensional spaces ($d\sim2000$), evidential methods may dominate—structural predictability and representational uncertainty are complementary signals (Appendix B.2). Split-Sample experiments (Appendix B.7) provide an additional check against training memorization.

5.2 Summary and Future Directions

This paper proposed the External Reference Principle and its first ordinal instantiation, SOGN, which improves low-coverage selective-regression R^2 on the evaluated ordinal benchmarks. The theory gives a conditional variance-compression result and an exact property of the scoring rule; the remaining claims are empirical and hold under the boundary conditions above. Future work includes generalizing the principle to richer target structures and developing sharper statistical guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript. We have not used generative AI tools for generating experimental results, theoretical proofs, or core research ideas. Additionally, we used generative AI tools for code debugging assistance during the implementation of baseline methods. We have reviewed all AI-assisted work—all AI-assisted text was verified for factual accuracy by the authors, and AI-assisted code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work presents a general-purpose regression reliability method. The datasets used (California Housing, Bike Sharing, Diabetes, ApplianceEnergy, RC-49) are all publicly available benchmark datasets that do not contain personally identifiable information. We do not foresee direct harmful applications of this method; however, as with any predictive modeling technique, practitioners should exercise caution when deploying selective regression in high-stakes decision-making contexts (e.g., clinical diagnosis, credit scoring) where the exclusion of certain samples from prediction could have fairness implications.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have: (1) included the complete source code, model implementations, training scripts, and evaluation pipelines in the supplementary materials; (2) reported all hyperparameters and experimental design decisions in Section 4 and the supplementary materials; (3) stated the assumptions and proofs for the theoretical results in Appendix A; (4) used one fixed data split and three training seeds (42, 43, 44) for the reported multi-seed comparisons, with means and standard deviations; and (5) used publicly available benchmark datasets with documented preprocessing steps in Section 4. Code will be made publicly available upon acceptance.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory (ALT)*, 2016.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, 2002.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv:physics/0004057*, 2000.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1): 95–103, 1983.

A Supplementary Proofs

The error bounds of the following three differentiable approximation operators are used repeatedly in SOGN's gating analysis.

Lemma A.1 (Power-Mean–Max Approximation)**.** *For any $\boldsymbol{p} \in \Delta^{K-1}$ and $\alpha > 1$, let $p_{\max} = \max_k p_k$. Then*

$$p_{\max} K^{-1/\alpha} \;\le\; M_\alpha(\boldsymbol{p}) \;\le\; p_{\max}, \qquad 0 \le p_{\max} - M_\alpha(\boldsymbol{p}) \le p_{\max}(1 - K^{-1/\alpha}). \tag{8}$$

In particular, $M_\alpha(\boldsymbol{p}) \to p_{\max}$ as $\alpha \to \infty$. The bound is dimension-dependent; it does not imply a universal 1% approximation for a fixed finite α.

Proof. The upper bound is immediate because every $p_k \le p_{\max}$, so $M_\alpha^\alpha = K^{-1}\sum_k p_k^\alpha \le p_{\max}^\alpha$. The lower bound follows from the term attaining $p_{\max}$, namely $M_\alpha^\alpha \ge p_{\max}^\alpha / K$. The gap bound and the limit follow directly. □

Lemma A.2 (LogSumExp–Max Approximation). *For any nonnegative vector $\boldsymbol{v} \in [0,1]^L$ and $\beta > 0$,*

$$\max_{\ell} v_\ell \;\le\; \frac{1}{\beta} \log \sum_{\ell=1}^{L} e^{\beta v_\ell} \;\le\; \max_{\ell} v_\ell + \frac{\log L}{\beta}. \tag{9}$$

As $\beta \to \infty$, the approximation error is $O(1/\beta)$.

Proof. From $e^{\beta v_{\max}} \le \sum_\ell e^{\beta v_\ell} \le L e^{\beta v_{\max}}$, taking $\log / \beta$ yields the result. □

Lemma A.3 (SoftMin–Min Approximation). *For any $a, b \in (0,1]$ and $\gamma > 0$,*

$$\min(a,b) - \frac{\log 2}{\gamma} \;\le\; \mathrm{SoftMin}_\gamma(a,b) \;\le\; \min(a,b). \tag{10}$$

Proof. From $e^{-\gamma M} \le e^{-\gamma a} + e^{-\gamma b} \le 2e^{-\gamma m}$, taking $-\log / \gamma$ yields the result (where $m = \min(a,b)$, $M = \max(a,b)$). □

A.1 Calibration bridge used by the variance result

Assumption 3.3 is intentionally a condition on the fitted ordinal probabilities, rather than a claimed pointwise consequence of empirical cross-entropy minimization. In practice it can be assessed on held-out data using calibration diagnostics or a bound supplied for a specified hypothesis class. This separation keeps the variance argument valid without conflating an expected generalization bound with pointwise L_∞ convergence.

A.2 Full Proof of the Ordinal Variance Compression Bound

Full proof of Theorem 3.1. The proof is conditional on the calibration event $\mathcal{E}_N$ from Assumption 3.3.

Step 1: Transfer coherence from $\hat{\boldsymbol{p}}$ to $\boldsymbol{p}^*$. On $\mathcal{E}_N$, $\|\hat{\boldsymbol{p}}_N - \boldsymbol{p}^*\|_\infty \le \delta_N$. For any three-consecutive-interval window m, the triangle inequality yields $|S_m(\hat{\boldsymbol{p}}) - S_m(\boldsymbol{p}^*)| \le \sum_{k=m}^{m+2} |\hat{p}_k - p_k^*| \le 3\delta_N$. Hence, on $\mathcal{E}_N$, $\max_m S_m(\boldsymbol{p}^*) \ge \max_m S_m(\hat{\boldsymbol{p}}) - 3\delta_N > 1 - \varepsilon - 3\delta_N$.

Step 2: Construct the high-probability window. Let $m^* = \arg\max_m S_m(\boldsymbol{p}^*)$ and define $I_S = [b_{m^*-1}, b_{m^*+2}]$, consisting of three consecutive bins. Set $p_{\text{in}} = \sum_{k=m^*}^{m^*+2} p_k^*(\boldsymbol{x}) \ge 1 - \varepsilon - 3\delta_N$ and $p_{\text{out}} = 1 - p_{\text{in}} \le \varepsilon + 3\delta_N$. The width of I_S satisfies $|I_S| \le 3w_{\max}$.

Step 3: Decompose the conditional variance. Conditioning on the bin index $g(Y) \in \{1, \dots, K\}$, the law of total variance Györfi et al. (2002) gives

$$\sigma^2(\boldsymbol{x}) = \mathbb{E}\big[\mathrm{Var}(Y \mid \boldsymbol{x}, g(Y))\big] + \mathrm{Var}\big(\mathbb{E}[Y \mid \boldsymbol{x}, g(Y)]\big). \tag{11}$$

Step 4: Bound the within-group term. By Assumption 3.2, $\mathbb{E}[\mathrm{Var}(Y \mid \boldsymbol{x}, g)] = \sum_{k=1}^{K} p_k^*(\boldsymbol{x}) \cdot \mathrm{Var}(Y \mid \boldsymbol{x}, g(Y) = k) \le \sigma_{\max}^2$.

Step 5: Bound the between-group term. Define $\mu_k = \mathbb{E}[Y \mid \boldsymbol{x}, g(Y) = k]$ for $k = 1, \dots, K$, and let a be the midpoint of I_S. Since variance is the minimum expected squared deviation from any constant,

$$\mathrm{Var}(\mathbb{E}[Y \mid \boldsymbol{x}, g]) = \min_{c \in \mathbb{R}} \mathbb{E}[(\mu_g - c)^2] \le \mathbb{E}[(\mu_g - a)^2] = \sum_{k=1}^{K} p_k^*(\boldsymbol{x})(\mu_k - a)^2. \tag{12}$$

Split the sum into in-window and out-of-window components:

- For $k \in \{m^*, m^*+1, m^*+2\}$ (inside I_S): $|\mu_k - a| \le \frac{|I_S|}{2} \le \frac{3w_{\max}}{2}$, hence $(\mu_k - a)^2 \le \frac{9w_{\max}^2}{4}$.

- For $k \notin \{m^*, m^*+1, m^*+2\}$ (outside I_S): since all bin centers lie within $[b_0, b_K]$, $|\mu_k - a| \le R$, so $(\mu_k - a)^2 \le R^2$.

Therefore, combining the two components:

$$\begin{aligned}\mathbb{E}[(\mu_g - a)^2] &\le p_{\text{in}} \cdot \frac{9w_{\max}^2}{4} + p_{\text{out}} \cdot R^2 \\ &\le \frac{9w_{\max}^2}{4} + (\varepsilon + 3\delta_N)R^2, \end{aligned} \tag{13}$$

where the last line uses $p_{\text{in}} \le 1$ and $p_{\text{out}} \le \varepsilon + 3\delta_N$.

Step 6: Assemble the joint bound. Substituting Steps 4–5 into the total variance decomposition: $\sigma^2(\boldsymbol{x}) \le \sigma_{\max}^2 + \frac{9w_{\max}^2}{4} + (\varepsilon + 3\delta_N)R^2$, establishing part (i). Part (ii) follows directly from Step 2: by definition of m^*, $\mathbb{P}(Y \in I_S \mid \boldsymbol{x}) \ge p_{\text{in}} > 1 - \varepsilon - 3\delta_N$, and $|I_S| \le 3w_{\max}$. □

A.3 Optimization interpretation of Stop-Gradient

The Stop-Gradient operation has a narrower, implementation-level role than a convergence theorem: it removes the direct derivative of the weighted regression residual with respect to the gate. The ordinal loss and target-coverage hinge continue to update the ordinal head, while the regression head is optimized under the current weights. This reduces a direct residual-to-gate feedback path, but does not by itself guarantee global convergence or rule out all degenerate solutions; those properties remain empirical questions addressed by the component ablation.

B Supplementary Experiments

B.1 Full Coverage-Selective Regression Results

Below we provide the complete R^2 comparison across all coverages for the four datasets (3-seed means). Best results are bolded; second-best are underlined. The RC-49 boundary case is analyzed separately in Appendix B.2.

Table 3: California Housing: full coverage-selective regression R^2 comparison.

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	**0.953**	0.827	0.401	0.813	0.685	<u>0.888</u>	0.753
0.2	**0.935**	0.817	0.474	0.828	0.645	<u>0.889</u>	0.793
0.3	**0.916**	0.816	0.550	0.831	0.643	<u>0.888</u>	0.824
0.4	**0.898**	0.815	0.593	0.836	0.653	<u>0.887</u>	0.833
0.5	**0.880**	0.808	0.651	0.836	0.646	<u>0.874</u>	0.844
0.6	**0.864**	0.805	0.684	0.835	0.654	0.858	<u>0.844</u>
0.7	**0.848**	0.811	0.712	0.835	0.675	0.847	<u>0.837</u>
0.8	0.835	0.814	0.745	**0.835**	0.704	0.834	0.826
0.9	0.818	0.811	0.773	**0.835**	0.744	0.810	0.816
1.0	0.807	0.810	0.815	**0.820**	0.809	0.780	0.796

SOGN ranks first in R^2 across all coverages $c \le 0.7$, with the largest margin at low coverage: at $c = 0.1$, $R^2 = 0.953$, leading the second-best method by 0.065. R^2 decreases consistently with coverage on this dataset ($0.953 \to 0.807$), indicating useful difficulty ranking. Model-uncertainty-based baselines (MC Dropout, Deep Ensemble) achieve R^2 of only 0.401 and 0.813 respectively at $c = 0.1$—their variance estimates provide weaker ranking signals here. SelectiveNet performs well at low coverage but drops sharply for $c \ge 0.6$. At $c = 1.0$, SOGN's R^2 (0.807) is slightly lower than Deep Ensemble (0.820) and Standard (0.810), reflecting a modest full-coverage/selectivity trade-off.

Table 4: Bike Sharing: full coverage-selective regression R^2 comparison.

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	**0.966**	0.867	0.447	0.921	-0.277	0.583	-1.629
0.2	**0.958**	0.886	0.580	0.923	0.555	0.753	0.092
0.3	**0.936**	0.868	0.703	0.915	0.891	0.834	0.895
0.4	**0.919**	0.875	0.755	0.921	0.933	0.870	0.903
0.5	0.902	0.883	0.792	**0.927**	0.895	0.868	0.891
0.6	0.894	0.888	0.812	**0.930**	0.893	0.855	0.875
0.7	0.886	0.893	0.827	**0.924**	0.891	0.855	0.871
0.8	0.875	0.888	0.842	**0.919**	0.901	0.857	0.875
0.9	0.862	0.886	0.853	**0.916**	0.888	0.850	0.881
1.0	0.853	0.881	0.886	**0.902**	0.864	0.877	0.883

SOGN is optimal at all coverages $c \leq 0.6$, with the low-coverage lead particularly prominent ($c = 0.1$: $R^2 = 0.966$, an improvement of 0.099 over Standard). R^2 decreases consistently with coverage on this dataset. MC Dropout's variance scores fail severely on time-series data ($c = 0.1$: only 0.447); Deep Evidential produces negative R^2 at very low coverage. Deep Ensemble slightly overtakes SOGN for $c \geq 0.5$, but its $3\times$ parameter advantage makes this gap expected.

Table 5: Diabetes: full coverage-selective regression R^2 comparison.

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	**0.690**	-0.013	-0.022	-0.543	-0.677	-0.428	-0.330
0.2	**0.786**	0.216	0.030	0.246	0.490	-0.834	-0.363
0.3	**0.664**	0.166	-0.198	0.386	0.483	-1.070	-0.277
0.4	**0.601**	0.288	0.278	0.436	0.540	-1.193	0.040
0.5	**0.526**	0.327	0.179	0.443	0.488	-1.727	0.124
0.6	**0.534**	0.380	0.315	0.506	0.448	-1.488	0.062
0.7	**0.535**	0.407	0.386	0.452	0.354	-1.608	0.193
0.8	**0.468**	0.422	0.365	0.438	0.384	-1.789	0.281
0.9	0.447	0.419	0.456	0.443	**0.456**	-1.924	0.430
1.0	0.428	0.474	0.453	**0.478**	0.417	-1.965	0.464

SOGN achieves the best R^2 across all coverages $c \leq 0.8$. At $c = 0.2$, R^2 reaches 0.786, an 85% improvement over the full sample. At $c = 0.1$, only SOGN among the seven methods achieves positive R^2 (0.690). All structure-blind methods remain below $R^2 = 0.49$ on Diabetes, while the small test set also makes this dataset comparatively variable. SelectiveNet produces negative R^2 at every coverage level—its shared-encoder architecture causes severe conflict on small samples.

Table 6: ApplianceEnergy: full coverage-selective regression R^2 comparison.

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	**0.724**	0.571	0.282	0.605	0.398	0.339	0.341
0.2	**0.718**	0.564	0.254	0.510	0.396	0.400	0.290
0.3	**0.710**	0.539	0.251	0.575	0.380	0.343	0.279
0.4	**0.688**	0.547	0.247	0.542	0.373	0.328	0.330
0.5	**0.668**	0.560	0.257	0.538	0.353	0.309	0.314
0.6	**0.648**	0.557	0.269	0.515	0.322	0.294	0.308
0.7	**0.625**	0.550	0.273	0.532	0.260	0.264	0.288
0.8	**0.600**	0.551	0.285	0.543	0.219	0.235	0.301
0.9	**0.576**	0.564	0.356	0.540	0.235	0.216	0.367
1.0	0.559	0.562	0.569	**0.622**	0.368	0.191	0.445

On ApplianceEnergy, SOGN exhibits the clearest selective behavior among all four datasets. SOGN ranks first in R^2 across all coverages $c \leq 0.9$; at $c = 0.1$, $R^2 = 0.724$, leading the second-best method by 0.119. Its R^2 decreases consistently with coverage ($0.724 \rightarrow 0.559$), giving the largest

selective dynamic range ($\Delta R^2 = 0.165$). Standard's R^2 fluctuates between 0.539–0.571 with little selective trend. All model-internal-uncertainty-based methods perform poorly at low coverage: MC Dropout achieves R^2 of only 0.282 at $c = 0.1$; Deep Evidential and SelectiveNet have R^2 below the full-sample level across the entire coverage range. At $c = 1.0$, SOGN's R^2 (0.559) falls below Deep Ensemble (0.622), Standard (0.562), and MC Dropout (0.569)—a modest full-coverage cost accompanying its low-coverage gains. The U-shaped nonlinear relationship between energy consumption and weather conditions in this dataset provides a useful target-structure signal.

B.2 Boundary Condition: Circular Targets (RC-49)

Table 7: RC-49 (circular target): full coverage-selective regression R^2 comparison (3-seed means). Best results are bolded; second-best are underlined.

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	0.977	0.973	0.843	0.988	<u>0.998</u>	0.987	**0.999**
0.2	0.977	0.983	0.895	0.980	<u>0.998</u>	0.989	**0.999**
0.3	0.976	0.968	0.927	0.986	<u>0.998</u>	0.991	**0.999**
0.4	0.990	0.968	0.951	0.989	**0.998**	0.993	<u>0.998</u>
0.5	0.993	0.967	0.967	0.991	**0.998**	0.994	<u>0.998</u>
0.6	0.994	0.970	0.976	0.991	**0.998**	0.995	<u>0.998</u>
0.7	0.996	0.972	0.977	0.993	**0.998**	0.986	<u>0.998</u>
0.8	0.996	0.966	0.977	0.992	**0.998**	0.980	<u>0.997</u>
0.9	0.996	0.966	0.978	0.991	**0.998**	0.981	<u>0.997</u>
1.0	<u>0.973</u>	0.969	0.968	0.970	**0.975**	0.972	0.972

SOGN's selective signal is derived from ordinal structure; our target-structure taxonomy (Appendix B.10) therefore predicts it degrades on targets without a total order. RC-49 Geifman & El-Yaniv (2017), whose target is a rotation angle, is a direct test of this prediction. We use a stratified subset of 17,000 images (train/validation/test = 10,000/2,000/5,000) with 2,048-dimensional pre-extracted CNN features. SOGN is trained with the ordinal-pretraining and joint-training recipe of Section 2, plus the optional plain-MSE fine-tuning stage; evaluation follows the same coverage-selection protocol as the main benchmarks.

The result is sharp: at c=0.1, SOGN's advantage over the random-selection baseline collapses from 0.703 on Diabetes (0.690 vs. −0.013) to 0.004 on RC-49 (0.977 vs. 0.973), a ∼175× reduction, while its full-coverage regression remains second (0.973; Table 7). This dissociation—selective signal destroyed, regression preserved—is exactly what the variance compression bound's premise implies: absent a valid ordinal structure, the variance-compression mechanism has nothing to compress. It also corroborates our central claim that ordinal structure, rather than generic regularization or capacity, is the source of SOGN's selective advantage: the advantage vanishes precisely when the structure does.

Two further observations sharpen the picture. First, feature-space uncertainty methods (Deep Evidential, Conformal) dominate at low coverage here (0.998 and 0.999 at c=0.1), consistent with representational uncertainty being the appropriate reliability signal in high-dimensional CNN feature space. Second, SOGN's low MAPE at c=0.1 (Table 11) is a selection artifact rather than evidence of better prediction: the coherence gate rejects near-seam (near-zero-angle) samples, which otherwise dominate MAPE through division by a small angle; the R^2 ranking (fifth at c=0.1) confirms SOGN does not actually select the easiest samples here.

One confound remains: RC-49 differs from our ordinal benchmarks in both target topology (circular) and feature type (pre-extracted CNN features). The degradation is thus consistent with two mechanisms—the circularity predicted by our taxonomy, or the dominance of representational uncertainty over ordinal structure in high-dimensional feature spaces—which a single dataset cannot separate. We leave this disentanglement (e.g., a non-circular image target, or a circular target with tabular features) to future work.

B.3 Backbone Architecture Independence Verification

A core strength of the External Reference Principle is its independence from specific backbone architectures. To verify this, we replace LSTM with GRU and Transformer—two mainstream time-series architectures—on the Bike Sharing dataset, comparing SOGN, Standard, and MC Dropout under each architecture (all using the same training protocol and 3-seed averaging). Table 8 reports the ΔR^2 gain of SOGN relative to Standard under each architecture.

Table 8: Selective gain of SOGN across different backbone architectures ($\Delta R^2 = R^2_{\text{SOGN}} - R^2_{\text{Standard}}$, Bike Sharing dataset). Positive values indicate SOGN outperforms Standard.

Coverage	LSTM	GRU	Transformer
0.1	$+0.099$	$+0.119$	-0.085
0.2	$+0.069$	$+0.089$	$+0.064$
0.3	$+0.088$	$+0.066$	$+0.063$
0.5	$+0.061$	$+0.049$	$+0.049$
0.7	$+0.025$	$+0.018$	$+0.017$
1.0	$+0.005$	-0.000	-0.006

Results. On LSTM and GRU, SOGN significantly outperforms Standard at all coverages $c \leq 0.7$, with ΔR^2 decreasing monotonically with coverage—samples with high ordinal scores are indeed more predictable. On Transformer, SOGN shows a slight negative gain at $c = 0.1$ (-0.085), but this coverage level selects only about 29 test samples, so the R^2 estimate has high variance. From $c = 0.2$ onward, ΔR^2 returns to positive and is monotonically decreasing, indicating that SOGN's selective capability on Transformer holds when the sample size is sufficient. Overall, out of the 30 (coverage, architecture) data points across three architectures, SOGN outperforms Standard on 28, strongly validating the backbone-architecture independence of the External Reference Principle.

B.4 Statistical Significance Tests

To verify whether SOGN's advantage over baselines is statistically significant, we perform paired t-tests on the 3 random seeds at each coverage level. Table 9 reports the R^2 differences between SOGN and each baseline, with significance levels.

Table 9: Paired t-tests: SOGN vs. baselines (Bike Sharing, 3 seeds). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; positive values indicate SOGN advantage.

Method	c=0.1	c=0.3	c=0.5	c=0.7	c=1.0
Standard	$+0.099^{*}$	$+0.088^{*}$	$+0.061^{**}$	$+0.025$	$+0.005$
MC Dropout	$+0.518^{**}$	$+0.254^{**}$	$+0.152^{**}$	$+0.092^{*}$	-0.001
Deep Ensemble	$+0.044$	$+0.042$	$+0.017$	-0.006	-0.016
Deep Evidential	$+1.243^{***}$	$+0.065$	$+0.049$	$+0.027$	$+0.022$
SelectiveNet	$+0.383^{**}$	$+0.122^{*}$	$+0.076^{*}$	$+0.063$	$+0.009$
Conformal	$+2.595^{*}$	$+0.061^{*}$	$+0.053$	$+0.047$	$+0.003$

Results. SOGN shows no statistically significant disadvantage against any baseline at any coverage. At low coverage ($c \leq 0.5$), SOGN's advantage over Standard and MC Dropout reaches $p < 0.05$ or $p < 0.1$ at most coverages; its advantage over SelectiveNet is significant at all coverages $c \leq 0.6$. The difference from Deep Ensemble does not reach statistical significance (limited testing power with 3 seeds), but all low-coverage differences are positive. On the California Housing dataset with a larger sample size (20,640), test results are more significant: SOGN's lead over Standard and MC Dropout reaches $p < 0.01$ across all coverages $c \leq 0.9$.

B.5 Mechanism Verification of the Selective Signal Source: R^2-RMSE Decoupling

If selective scores merely measured the absolute error magnitude of predictions, then the rankings across metrics should remain consistent across methods; if rankings systematically reverse across metrics, it indicates that different methods' score signals measure different dimensions of predictability. Table 10 reports four metrics (R^2, RMSE, MAE, MAPE) for all seven methods at five key coverages on the California Housing dataset.

The core finding is a systematic cross-reversal between the R^2 ranking and the RMSE/MAE/MAPE rankings: at $c = 0.1$, Deep Evidential is best on all three absolute-error metrics—RMSE (0.199), MAE (0.120), and MAPE (12.48%)—yet its R^2 is only 0.685, which is 0.268 below SOGN's 0.953. SOGN is the only method that leads in R^2 but not in RMSE/MAE/MAPE—this metric imbalance is precisely the fingerprint of the External Reference Principle: the ordinal structure score measures the true predictability of the input-output relationship (high R^2), unaffected by the absolute magnitude of the target value (RMSE is not the lowest). The subset selected by uncertainty methods has small absolute error but poor variance explanation—the model fails to track the true variation of the target value; its low RMSE may reflect that the selected samples happen to have low absolute target values. This distinction is the fundamental rationale for R^2 as the core metric for selective regression.

Table 10: Full metric comparison on California Housing (MLP backbone, 3-seed means). Higher R^2 is better; lower RMSE/MAE/MAPE is better.

(a) R^2↑

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	**0.953**	0.827	0.401	0.813	0.685	0.888	0.753
0.3	**0.916**	0.816	0.550	0.831	0.643	0.888	0.824
0.5	**0.880**	0.808	0.651	0.836	0.646	0.874	0.844
0.7	**0.848**	0.811	0.712	0.835	0.675	0.847	0.837
1.0	0.807	0.810	0.815	**0.820**	0.809	0.780	0.796

(b) RMSE↓

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	0.433	0.488	0.313	0.315	**0.199**	0.295	0.225
0.3	0.465	0.495	0.337	0.334	**0.282**	0.293	0.302
0.5	0.480	0.507	0.357	0.361	0.343	**0.323**	0.344
0.7	0.493	0.501	0.397	0.403	0.411	**0.381**	0.395
1.0	0.509	0.504	0.497	**0.491**	0.505	0.542	0.523

(c) MAE↓

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	0.282	0.327	0.224	0.201	**0.120**	0.199	0.147
0.3	0.305	0.330	0.239	0.221	**0.166**	0.201	0.196
0.5	0.318	0.337	0.250	0.241	**0.208**	0.221	0.229
0.7	0.327	0.334	0.269	0.266	0.253	**0.256**	0.266
1.0	0.340	0.336	0.334	**0.322**	0.324	0.357	0.352

(d) MAPE%↓

Coverage	SOGN	Standard	MC Dropout	Deep Ens.	Deep Evid.	SelectiveNet	Conformal
0.1	16.66	17.80	14.86	14.39	**12.48**	15.76	16.29
0.3	17.38	17.97	15.79	14.77	**14.55**	16.00	17.09
0.5	17.64	18.19	16.71	15.26	**15.58**	16.18	17.42
0.7	18.19	18.19	17.26	**15.59**	16.56	16.58	17.83
1.0	18.74	18.39	18.55	**17.42**	16.98	19.66	19.36

B.6 RC-49 MAPE Analysis

Table 11 reports MAPE and MAX APE at low coverage ($c \leq 0.5$) on the RC-49 boundary case. Because the rotation-angle target approaches zero degrees, MAPE is dominated by near-seam samples and must be interpreted jointly with the R^2 ranking (Table 7), not as a standalone accuracy metric.

Table 11: RC-49 MAPE and MAX APE comparison ($c \leq 0.5$, 3-seed means). MAPE is reported only for completeness; R^2 (Table 7) is the core metric. Lower is better; see text for why low MAPE does not imply better selection.

Cov.	Metric	SOGN	Std.	MC-D	Ens.	Evid.	SelNet	Conf.
0.1	MAPE%	2.17	8026.5	4.84	24.14	2.48	8.23	2.96
	MAX APE%	10.7	3973766.7	22.5	9610.0	143.0	96.6	31.0
0.3	MAPE%	2.34	7025.2	4.72	11.86	3.45	6.98	3.46
	MAX APE%	13.8	10222455.0	26.0	9883.1	172.2	227.9	121.2
0.5	MAPE%	2.48	4242.4	6.14	280.2	4.28	6.65	4.11
	MAX APE%	16.8	10222455.0	2936.9	676795.7	332.7	369.1	635.4

SOGN's MAPE is lowest at all coverages, but this is a selection artifact rather than evidence of better prediction: the coherence gate rejects near-seam (near-zero-angle) samples—whose MAPE otherwise explodes through division by a small angle—so SOGN's selected subset is systematically biased away from the seam. The R^2 ranking (Table 7) confirms this interpretation: SOGN is only fifth in R^2 at c=0.1 despite the lowest MAPE, i.e., it does not actually select the easiest samples. Standard's MAPE ($8{,}026.5\%$) reflects that random selection retains near-seam samples.

B.7 Split-Sample Verification: Robustness of the External Reference Signal

A theoretical concern exists regarding SOGN's core design: the gating weights and regression function use the same batch of training data (double-dipping). Critics may question whether ordinal diagnosis works only because the gate memorizes the regression difficulty of training samples, rather than genuinely extracting generalizable signals from the target structure. This section addresses this concern through split-sample experiments: the training set is randomly split in half into $\mathcal{D}_{\text{gate}}$ and $\mathcal{D}_{\text{reg}}$ (disjoint). The gating component learns ordinal structure signals solely from $\mathcal{D}_{\text{gate}}$; the regression component learns the prediction function solely from $\mathcal{D}_{\text{reg}}$. The gating parameters never receive any gradient information about the regression difficulty of samples in $\mathcal{D}_{\text{reg}}$.

Experiments are conducted on ApplianceEnergy (SOGN's strongest dataset) and Diabetes (fewest training samples: only 308, with 154 per half after splitting), representing stress tests under best-case and worst-case conditions respectively.

Table 12: ApplianceEnergy Split-Sample verification: full data vs. training set split in half (3-seed means).

Coverage	SOGN (Full Data)	SOGN (Split-Sample)	ΔR^2
0.1	0.724	0.535	−0.189
0.3	0.710	0.526	−0.184
0.5	0.668	0.491	−0.177
0.7	0.625	0.465	−0.160
1.0	0.559	0.419	−0.140

Table 13: Diabetes Split-Sample verification: full data vs. training set split in half (3-seed means).

Coverage	SOGN (Full Data)	SOGN (Split-Sample)	Δ R^2
0.1	0.690	0.151	−0.539
0.2	0.786	0.539	−0.247
0.3	0.664	0.454	−0.210
0.5	0.526	0.379	−0.147
1.0	0.428	0.303	−0.125

Analysis of results. (1) No collapse; structurally bounded degradation: On both datasets, split-sample SOGN maintains effective selective prediction capability across the entire coverage range. ΔR^2 shrinks monotonically as coverage increases—ApplianceEnergy from -0.189 ($c = 0.1$) to -0.140 ($c = 1.0$); Diabetes from -0.539 to -0.125. If the selective signal depended on training memorization, the expected pattern would be the opposite: the gate trained on the full dataset should show a disproportionate advantage at low coverage. The observed pattern is consistent with the hypothesis that the structural signal originates from intrinsic properties of the data-generating process. (2) Selective capability maintained under extreme Diabetes conditions: with only 154 samples for ordinal pretraining, split-sample SOGN still maintains effective ranking ($R^2 = 0.539$ at $c = 0.2$, an improvement of 0.236 over full coverage), constituting a rigorous empirical test of the External Reference Principle's robustness under extreme data scarcity. (3) Degradation far smaller than random: the actual degradation narrows rapidly for $c \geq 0.3$ ($\Delta R^2 < 0.21$), far smaller than the expected degradation under random ordering ($\Delta R^2 > 0.5$), and the key qualitative features of selective ranking are fully preserved.

B.8 Ancillarity and the External Reference Principle

In classical statistics, a statistic $A(Y)$ is **ancillary** for a parameter θ if its distribution does not depend on θ (Fisher 1934; Cox & Hinkley 1974, Ch. 2). For SOGN, the more modest and relevant claim is that the ordinal bin distribution can provide a target-structure-assisted precision descriptor. Under a location/noise model $Y = f^*(X) + \varepsilon$, with fixed marginal bin boundaries and noise independent of X, it may be approximately ancillary for the value of f^*; outside such conditions, the ordinal probabilities can also encode location and representation effects. We therefore use this connection as motivation and state the variance guarantee conditionally in Assumption 3.3, rather than treating ancillarity as an unconditional property.

Fisher's argument for conditioning on ancillary statistics motivates separating a precision-related signal from the point prediction. SOGN operationalizes this idea by (1) estimating the conditional bin-probability vector through the ordinal classification head, (2) extracting concentration and coherence from it, and (3) using those features to weight training samples. The resulting architecture is fully differentiable, but the practical interpretation remains conditional on ordinal calibration and target topology.

This connection suggests a general recipe for extending the External Reference Principle beyond ordinal structure: for any regression problem, identify a target-structure descriptor that is plausibly related to local precision while separating it from the point-prediction objective, and use it as the external reference signal. For hierarchical targets, this may be a tree-structured partition; for spatial targets, local spatial autocorrelation; for time-series targets, temporal dependence. Whether such a descriptor is ancillary or merely predictive of precision must be established separately for each application.

B.9 A Taxonomy of Contamination Types

The umbrella term "contamination" in the main text encompasses three distinct phenomena:

(i) **Heteroscedasticity**: $\mathrm{Var}(Y \mid X{=}x)$ varies systematically with x. Standard unweighted regression is inefficient; weighted least squares with inverse-variance weights is the classical remedy. SOGN's gating partially addresses heteroscedasticity by downweighting high-

variance regions, but differs from weighted least squares in that it can drive weights to near-zero (selective suppression) rather than merely downweighting.

(ii) **Irreducible error**: Y is inherently unpredictable from X in certain regions, even for the Bayes-optimal predictor $f^*(x)$. This occurs when important covariates are unobserved or when the data-generating process has intrinsic stochasticity. Suppressing such samples during training is well-motivated: they cannot improve estimates of f^* and may degrade representation learning. SOGN primarily targets this type.

(iii) **Model misspecification**: the function class $\mathcal{F}$ does not contain f^*. In misspecified regions, even a well-estimated model will have large residuals. Suppressing misspecified samples may mask the misspecification rather than resolve it. SOGN's gating may inadvertently suppress such samples; practitioners should verify through residual analysis that suppressed regions do not exhibit systematic model inadequacy. This is an inherent limitation of any sample-suppression method that does not model the misspecification directly.

The three types are not mutually exclusive in practice. A sample may have high irreducible error (due to unobserved confounders) and also lie in a region where the model is misspecified. SOGN's ordinal concentration and coherence scores are informative primarily for type (ii), partially for type (i), and not reliable for type (iii).

B.10 Systematic Taxonomy of Target Structures

The ordinal instantiation of the External Reference Principle presumes a total order on the target space. Table 14 catalogues target structures by their natural topology, candidate precision-related structure descriptors, and the applicability of ordinal binning.

The table serves two purposes. First, it maps the boundary conditions discussed in §5.1 to target topology, making the qualitative observations operational. Second, it provides a roadmap for future instantiations of the External Reference Principle: each row identifies a candidate structure descriptor for a target type, which can serve as the starting point for designing a corresponding gating architecture.

Relation to the marginal Y-density. One might ask whether SOGN's ordinal concentration score merely proxies the marginal density of Y—whether samples whose Y values lie in high-density regions are simply easier to predict for any method. We note that ordinal concentration is strictly more informative than marginal density: two samples whose Y values have identical marginal density can nonetheless differ in ordinal concentration (one lies in a narrow high-density peak, the other in a broad plateau with the same density but lower bin concentration). Because the concentration score is computed per sample from the ordinal distribution $\boldsymbol{p}(\boldsymbol{x})$, it captures location-specific predictability that a global density estimate cannot, so the ordinal signal is not reducible to a density proxy.

Table 14: Target structures and candidate precision-related descriptors for the External Reference Principle.

Target Structure	**Example Domains**	**Ancillary Statistic(s)**	**Ordinal Binning Applicability**
Real-valued, unimodal	House prices, sensor readings, disease scores	Ordinal bin distribution; local quantile spread	Fully applicable (this paper)
Real-valued, multimodal	Mixture-model outcomes	Ordinal bin distribution (wider K needed); mixture component indicators	Applicable with caveat: equal-frequency binning may group distant Y values
Circular / spherical	Rotation angle, wind direction, geolocation	von Mises concentration; spherical harmonics; geodesic neighborhood cardinality	Ordinal binning degrades (bin 1 and bin K are adjacent on the circle); empirically confirmed on RC-49 (Appendix B.2)
Ordered categorical	Likert scales, survey responses, star ratings	Category distribution; adjacent-category concentration	Naturally applicable (bins are the categories themselves)
Hierarchical	Taxonomy-annotated targets (species, product categories)	Tree-depth distribution; leaf-ancestor coherence; sibling dispersion	Requires tree-structured (non-ordinal) binning
Spatial / spatiotemporal	Geospatial regression, climate fields, traffic flow	Local spatial autocorrelation (Moran's I); variogram summaries	Ordinal binning applicable per-location but ignores spatial structure
Multi-target vector	Multi-output regression, joint prediction	Inter-target correlation matrix; copula concentration	Requires inter-target structural criteria (deferred to future work, §2.2)