PREDICT BY POINTING: A TASK-AGNOSTIC OUTPUT HEAD FOR TABULAR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models learn in-context: given a labeled support set, they predict unlabeled queries in a single forward pass. Simple point losses extend naturally to vector-valued targets but are often less competitive than distributional regression objectives, whose heads are typically scalar-specific. Consequently, there is no common prediction head spanning support-adaptive classification, strong scalar regression, and native multi-output regression. In this work, we introduce **P**rediction **O**ver **In**-context **T**argets (**POINT**), a task-agnostic output head that instead treats the labeled support set itself as the prediction space. POINT assigns probability directly to the support set; classification probabilities are obtained by summing mass over examples of the same class, while regression predictions are obtained by taking the expectation of continuous support labels. The output space therefore adapts to the context rather than being fixed by the task. To handle this new space, POINT is trained with **POINT-CE**, a support-distribution cross-entropy objective to handle both classification and regression simultaneously. We further show that, under support coverage, minimizing POINT-CE in expectation recovers conditional class probabilities for classification and the conditional mean for regression. This connects the same support-level objective to the quantities targeted by categorical likelihoods and distributional regression. Across tabular in-context-learning experiments, POINT achieves competitive performance on both classification and regression while retaining the same support-based prediction principle. Experiment results suggest that predicting by pointing to labels already present in context can provide a common prediction head for tabular models.

1 INTRODUCTION

Tabular foundation models (TFMs) are emerging as a powerful paradigm for general-purpose tabular prediction (Hollmann et al., 2025; Qu et al., 2026). By pretraining across diverse prediction tasks, they learn to infer from labeled examples in-context, often without task-specific parameter updates (Müller et al., 2022), and offer the potential for models that transfer predictive knowledge across datasets and task types. This is possible as, at a statistical level, the underlying operation is ultimately similar regardless of the target type (e.g., a category, a real number, a vector, etc.): to infer a conditional distribution from context. Nevertheless, existing TFMs implement this operation through task-specific output parameterizations. Classification typically uses categorical logits and cross-entropy, whereas regression may use a bar distribution (Hollmann et al., 2025) or quantiles (Qu et al., 2026). Though effective, these separate heads prevent classification and regression from directly sharing knowledge at the prediction layer during joint pretraining, potentially limiting cross-task transfer. Moreover, strong distributional regression heads are typically formulated for a single scalar target, so multi-output regression requires target-wise prediction or additional output machinery.

A narrower question is thus whether one architecture can be optimized for all tasks in tandem. More concretely, we ask ourselves: *can classification and regression share a single parameterized prediction head and a common distributional training interface?* To this end, we introduce **P**rediction **O**ver **In**-context **T**argets (**POINT**), to make the labeled support set itself the output space. Unlike prior approaches, POINT scores query–support pairs and normalizes them into a distribution over support rows, rather than over predetermined class logits, regression bins, or quantile levels. Pushing this distribution through the observed labels yields class probabilities by summing mass within

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.

each class, or a discrete distribution over continuous targets whose expectation gives the regression prediction. The learned scoring and normalization mechanisms remain unchanged across task types.

Because POINT uses the same learned scorer and support-probability simplex across task types, classification and regression update a common prediction head during pretraining. This creates an opportunity for cross-task transfer: classification can teach discrimination among support examples, while regression can teach graded allocation of probability mass. The same construction extends naturally to vector-valued regression. When support labels are vectors, the same support distribution jointly aggregates all target dimensions in a single forward pass. More broadly, a shared support-indexed output space could provide a common interface for distribution-level distillation between tabular foundation models, extending a knowledge-transfer mechanism widely used in language models (Hinton et al., 2015; Sanh et al., 2019; Xu et al., 2024).

Yet a core difficulty remains: supervision. A regression target typically matches no observed support label, whereas a categorical target may match many. To handle this, we introduce a novel objective, **POINT-CE**, which constructs a target distribution over support examples. For classification, probability mass is distributed over support rows with the correct class. For regression, a label-space reference distribution is tilted subject to a mean-preserving constraint, so that its weighted support labels recover the observed target exactly. Cross-entropy against these as targets trains the *same support-probability simplex* for both task types, providing a unified approach for optimization.

Our approach connects two established lines of work. Kernel regression, pointer networks, and matching networks already demonstrate prediction through context-dependent weighting or selection (Nadaraya, 1964; Watson, 1964; Vinyals et al., 2015; 2016). Meanwhile, histogram-loss methods show that regression can benefit from cross-entropy against smoothed target distributions, with evidence linking these gains to optimization (Imani & White, 2018; Imani et al., 2026). POINT brings these ideas together by using observed support labels as the output space and moment-preserving distributions for supervision, connecting the same support-level objective to classification likelihoods and quantile-based regression. Empirically, POINT is competitive across classification and regression and extends to multi-output prediction. Our contributions are thus threefold:

- We introduce POINT and POINT-CE, a shared support-level cross-entropy objective that expresses categorical and continuous prediction through a common distribution over labeled support examples, without task-specific output coordinates in the learned head.
- We establish conditional Fisher consistency under support-coverage conditions, show that excess POINT-CE risk controls classification log loss and regression mean error, and connect the induced predictive distribution to integrated pinball risk while accounting for smoothing, discretization, and out-of-hull targets.
- We evaluate component-level classification and regression models under documented pretraining and inference budgets, and isolate the roles of retrieval geometry, probability supervision, robust point losses, and teacher smoothing through systematic ablations.

2 Related work

Tabular foundation models. Prior-data fitted networks (PFNs) amortize inference under a distribution of synthetic tasks (Müller et al., 2022; Nagler, 2023). TabPFN and TabICL apply this approach to tabular prediction (Hollmann et al., 2025; Qu et al., 2025). TabICLv2 uses separate classification and regression models, with 999-quantile pinball training for the latter (Qu et al., 2026). Joint models already exist: LimiX supports multiple structured-data tasks, and TabH2O jointly trains classification and regression with a dual-head architecture (Zhang et al., 2025; Pfeiffer et al., 2026). Our distinction is a support-indexed prediction head and a common moment-preserving CE core, not the first model to share a backbone across tasks. More closely, TabPFN-3 (Grinsztajn et al., 2026; Jäger et al., 2026) replaces its parametric classification head with an attention-based retrieval decoder that averages in-context one-hot labels, but supervises it through class-aggregated cross-entropy and retains a separate bar-distribution head for regression.

Support-based prediction. Kernel regression aggregates observed targets (Nadaraya, 1964; Watson, 1964); Matching Networks learn attention-weighted support-label votes (Vinyals et al., 2016). Prototypical Networks instead aggregate representations into class prototypes (Snell et al., 2017).

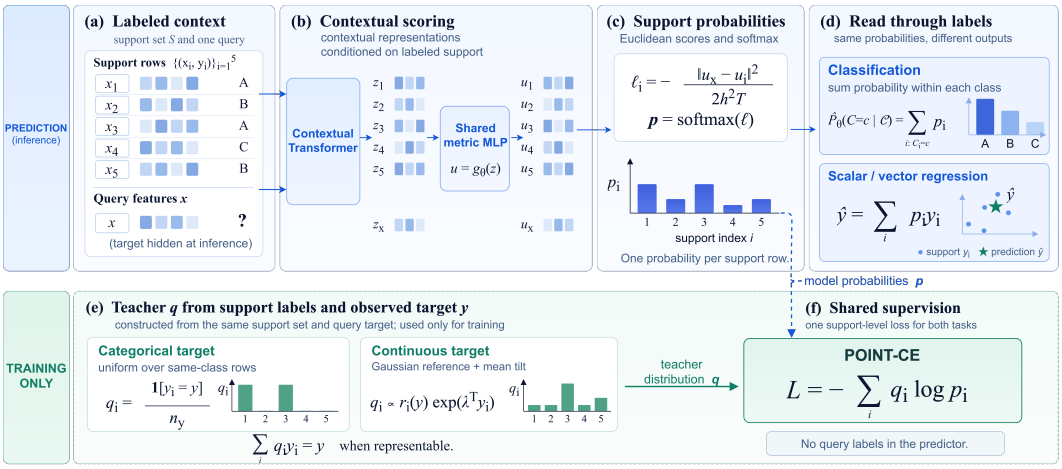


Figure 1: POINT uses the labeled support set as its output space. (a) The labeled context contains support rows and an unlabeled query. (b) A contextual encoder and shared metric head score query–support pairs. (c) Softmax produces a probability distribution over support rows. (d) Reading these probabilities through the observed labels gives classification or scalar/vector regression predictions. (e) During training, the observed query target defines a teacher distribution over support rows. (f) POINT-CE applies the same support-level cross-entropy across tasks. The query target is used only to construct the teacher and is never provided to the predictor.

TABR learns retrieval-augmented tabular representations and reports gains from using shared key projections with squared Euclidean distance rather than a vanilla query–key dot product (Gorishniy et al., 2024). We follow that geometric choice, but keep *actual labels* as output atoms rather than decode a retrieved latent value into a task-specific prediction.

Distributional supervision. Histogram losses convert regression targets into probability targets and can improve optimization (Imani & White, 2018). Categorical distributional reinforcement learning uses barycentric projection onto a prescribed grid (Bellemare et al., 2017). Our support atoms are instead irregular, episode-dependent observed labels. The Gaussian moment constraint is an information projection (Csiszár, 1975), not a new divergence. Proper scoring-rule theory relates probabilistic forecasts, log scores, and CRPS (Gneiting & Raftery, 2007); quantile regression provides a complementary parameterization (Koenker & Bassett, 1978). Section 4 combines these connections into support-specific calibration and approximation statements.

3 A SHARED SUPPORT-DISTRIBUTION HEAD

Our major contribution is a TFM paradigm that always predicts a probability distribution over labeled support rows; classification and regression differ only in how this distribution is supervised and read out. We detail the internal mechanics and the essential components that enable unified learning across different task types, with a summary in Figure 1.

3.1 ONE PROBABILITY SIMPLEX, DIFFERENT TARGET REPRESENTATIONS

Consider a sample (or *an episode*) for a tabular prediction task. Each episode consists of a labeled context $\mathcal{C}$, which is a tuple $(x, \mathcal{S}, \tau)$ that contains a support set $\mathcal{S}$, a task description τ and a query x. $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{S}$ consists of multiple labeled tabular rows (*support rows*) on which the prediction on x is conditioned. During training, the target y for task τ on query x is available. Categorical targets (e.g. classification tasks) are represented as one-hot vectors, whereas continuous targets (such as for regression) are represented directly as support-standardized scalars or vectors.

We retain the TABICL context-conditioned encoder backbone and replace only the output head. This encoder produces contextual embeddings of the support set and query $z_1, \ldots, z_S, z_x \in \mathbb{R}^d$. These

are further passed through a shared metric MLP g_θ to produce

$$u_x = g_\theta(z_x), \quad u_i = g_\theta(z_i), \qquad \ell_i = -\frac{\|u_x - u_i\|_2^2}{2h^2 T}, \qquad p_i = \mathrm{softmax}_i(\ell), \tag{1}$$

where h is a bandwidth hyper-parameter. The probabilities p_i define a distribution over support rows. Pushing this distribution through their observed targets induces the predictive measure $F_p(\cdot \mid \mathcal{C}) = \sum_{i=1}^{S} p_i \, \delta_{y_i}$, where δ_{y_i} denotes a Dirac point mass at support target y_i.

For classification, let C be the class-valued target, C_i the class of support row i, and c a class. The same support distribution then yields the classification and regression readouts

$$\widehat{P}_\theta(C = c \mid \mathcal{C}) = \textstyle\sum_{i:C_i=c} p_i, \qquad \widehat{y}_\theta = \mathbb{E}_{F_p}[Y] = \textstyle\sum_{i=1}^{S} p_i y_i. \tag{2}$$

For classification, $\widehat{y}_\theta$ is a vector of class probabilities, since each y_i is one-hot; equivalently, its probability for class c is the total mass assigned to support rows of class c. For scalar regression, $\widehat{y}_\theta$ is the weighted average of the support targets, and the same expression extends directly to vector-valued regression. Thus, the same support probabilities produce classification, scalar regression, and multi-output regression predictions without replacing the learned prediction head.

Why Euclidean scoring? Expanding Equation 1 gives $u_x^\top u_i/(h^2T) - \|u_i\|^2/(2h^2T)$ plus a query-only constant. The support-norm penalty distinguishes distance from unrestricted dot-product ranking. TABR's ablations support shared projections and L_2 distance as a useful tabular default, not as a universally optimal metric (Gorishniy et al., 2024).

3.2 MOMENT-PRESERVING PROBABILITY TARGETS

To supervise the same support-probability head across different target types, we construct a teacher distribution over support rows whose weighted targets match the observed query target. Standard classification and regression losses alone are not sufficient, as shown in our ablation (Table 3). For a target y and a reference distribution $r(y)$ over the support rows, we define

$$q_S(y) = \arg\min_{q \in \Delta^{S-1}} \mathrm{KL}(q \| r(y)) \quad \text{subject to} \quad \textstyle\sum_i q_i y_i = y. \tag{3}$$

where y denotes a one-hot vector for classification, and i indexes the S labeled support examples. For scalar targets outside the support range, we clip the teacher target to the nearest endpoint. We choose q so that aggregating the support targets under q recovers the observed target, while deviating as little as possible from $r(y)$. This teacher distribution determines which support rows receive probability mass and thus shapes the aggregation learned by POINT. We use uniform and Gaussian references by default, and explore references distilled from classical predictors.

Categorical targets use a uniform reference over the support rows. For classification, let C_i denote the class of support row i and let c be the query class. If class c is present in the support set, the moment constraint forces all mass onto rows of that class, and the KL minimizer is

$$q_{S,i}(c) = \frac{\mathbf{1}[C_i = c]}{n_c}, \qquad n_c = \textstyle\sum_{j=1}^{S} \mathbf{1}[C_j = c].$$

For **continuous targets** $y_i, y \in \mathbb{R}^m$, we want the teacher to favor support targets close to the observed target. We therefore use the Gaussian reference $r_i(y) \propto \exp(-\|y_i - y\|_2^2/(2\sigma^2))$. On a finite and irregular support set, this reference need not aggregate exactly to y. Applying Equation 3 gives

$$q_i(y; \lambda) = \frac{\exp[-\|y_i - y\|_2^2/(2\sigma^2) + \lambda^\top y_i]}{\sum_j \exp[-\|y_j - y\|_2^2/(2\sigma^2) + \lambda^\top y_j]}, \qquad \textstyle\sum_i q_i y_i = y. \tag{4}$$

The tilt $\lambda \in \mathbb{R}^m$ is chosen to match the teacher barycenter with the target. For scalar regression, λ can be found by bisection; for vector-valued regression, it is obtained by solving the corresponding low-dimensional convex problem. Thus, the Gaussian term determines locality, while the tilt corrects the finite-support bias so that aggregation under the teacher recovers the observed target.

A shared interface for model distillation. The reference distribution used to construct the POINT teacher can also be supplied by another predictor. Many classical methods, including k-NN, CART, Random Forests, Extra Trees, and positive-kernel methods, admit predictions of the form $\widehat{y}^{\mathrm{cl}}(x) = \sum_i w_i^{\mathrm{cl}}(x) y_i$, where $w_i^{\mathrm{cl}}(x) \geq 0$ and $\sum_i w_i^{\mathrm{cl}}(x) = 1$ (Breiman, 2001; Geurts et al., 2006; Lin &

Jeon, 2006). We can therefore directly use these support weights as the reference distribution, $r_i(x) = w_i^{\mathrm{cl}}(x)$, replacing the uniform categorical or Gaussian regression reference.

This provides a form of knowledge distillation (Bucila et al., 2006; Hinton et al., 2015) that transfers *intermediate support relevance*, allowing POINT to absorb the locality or partition structure of stronger classical predictors. This is particularly useful for future tabular foundation models. Current TFMs use task-specific output parameterizations, such as class logits, regression bins, or quantiles, whose predictive distributions are difficult to align across tasks and models. By applying the same POINT output head, different TFMs can instead express predictions over the same support-indexed output space, enabling direct distribution-level distillation while retaining a unified output head.

3.3 A COMMON LOSS AND PRACTICAL AUXILIARY TERMS

To enable learning using a common model, we introduce POINT-CE, an ordinary cross-entropy applied to distributions over the support simplex. Given the target support distribution $q_S(y)$ and the predicted support distribution $p(\mathcal{C})$, we define

$$\mathcal{L}_{\text{POINT-CE}}(\mathcal{C}, y) = -\sum_{i=1}^{S} q_{S,i}(y) \log p_i(\mathcal{C}). \tag{5}$$

The teacher distribution is required only during training. Importantly, Equation 5 has exactly the same form for classification and regression: the task type changes how $q_S(y)$ is constructed, but not the learned scoring head, the support simplex, or the cross-entropy objective.

In practice, we optionally combine POINT-CE with a task-specific point-prediction loss,

$$\mathcal{L} = \mathcal{L}_{\text{POINT-CE}} + \gamma_\tau \mathcal{L}_{\text{aux}}^{(\tau)}, \text{ where } \mathcal{L}_{\text{aux}}^{(\tau)} = \begin{cases} -\log \widehat{P}(y \mid \mathcal{C}), & \tau = \text{class}, \\ \mathrm{SL}_\beta(\widehat{y} - y), & \tau = \text{reg}. \end{cases} \tag{6}$$

where $\widehat{P}(c \mid \mathcal{C}) = \sum_{i:y_i=c} p_i(\mathcal{C})$, $\widehat{y} = \sum_i p_i(\mathcal{C}) y_i$, γ_τ controls the strength of the task-specific auxiliary loss relative to the shared POINT-CE objective, and SL_β denotes the Smooth L1 (Huber) loss with transition parameter β. Thus, POINT-CE provides the shared distributional supervision, while the auxiliary terms act only as task-specific optimization aids. Our theoretical results below concern the common POINT-CE objective unless stated otherwise.

4 WHY ONE SUPPORT LOSS CAN SERVE BOTH TASKS

We first analyze what POINT-CE learns, separately from how the required representation is obtained. Fix a context $\mathcal{C}$ with support targets $y_1, \ldots, y_S$ and true conditional query distribution $P^*(\cdot \mid \mathcal{C})$. The teacher $q_S(Y)$ can be determined fully by the support set and query target. Unless stated, all logarithms are natural. Full proofs and additional results are given in Appendix A.

4.1 A COMMON LOSS WITH TASK-APPROPRIATE POPULATION TARGETS

Theorem 4.1 (Conditional Fisher consistency and excess risk). *Let* $\pi_i = \mathbb{E}[q_{S,i}(Y) \mid \mathcal{C}]$ *and* $\mathcal{R}_{\text{PCE}}(p) = \mathbb{E}[-\sum_i q_{S,i}(Y) \log p_i \mid \mathcal{C}]$. *Over the support simplex,*

$$\mathcal{R}_{\text{PCE}}(p) - \mathcal{R}_{\text{PCE}}(\pi) = \mathrm{KL}(\pi \| p) =: \varepsilon(p). \tag{7}$$

If the teacher preserves the target, $\sum_i q_{S,i}(Y) y_i = Y$, *then*

$$\textstyle\sum_i \pi_i y_i = \mathbb{E}[Y \mid \mathcal{C}], \qquad \|\sum_i (p_i - \pi_i) y_i\|_2^2 \le \frac{D^2}{2} \varepsilon(p), \tag{8}$$

where $D = \max_{i,j} \|y_i - y_j\|_2$.

For classification, let C_i *denote the class of support row* i *and define* $P_p(c) = \sum_{i:C_i=c} p_i$. *If every possible query class occurs in support, then*

$$\mathrm{KL}(P^*(C \mid \mathcal{C}) \| P_p) \le \varepsilon(p). \tag{9}$$

Therefore, excess POINT-CE risk controls the excess classification log loss and the excess squared-error risk of the regression mean. For $\widehat{y} = \sum_i p_i y_i$, a regression target is infeasible when it lies

outside the convex hull of the support targets, since no convex combination of the support targets can equal it. Projecting such a target onto the support hull contributes an explicit bias:

$$\|\widehat{y} - \mathbb{E}[Y \mid \mathcal{C}]\|_2 \leq D\sqrt{\varepsilon(p)/2} + \|\mathbb{E}[\Pi_{\mathcal{H}_S}(Y) - Y \mid \mathcal{C}]\|_2 , \tag{10}$$

where $\Pi_{\mathcal{H}_S}$ denotes projection onto the convex hull of the support targets. For scalar regression, this projection simply clips Y to the support range $[\min_i y_i, \max_i y_i]$[1].

For a present class c, its uniform teacher also gives the exact decomposition

$$\mathcal{L}_{\mathrm{PCE}}(p, c) = -\log P_p(c) + \log n_c + \mathrm{KL}(U_c \| p(\cdot \mid c)), \tag{11}$$

where n_c is the number of support rows in class c and U_c is uniform over those rows. The final term governs how probability is distributed among support rows within the correct class, beyond matching the class-level probability alone.

4.2 Finite-support approximation and distributional regression

For scalar support targets y_i, the Gaussian reference is $r_i(y) \propto \exp[-(y_i - y)^2/(2\sigma^2)]$. Its minimum-KL projection subject to $\sum_i q_i y_i = y$ has the form $q_i(\lambda) \propto r_i(y) e^{\lambda y_i}$. For an interior target, the tilt is unique because the derivative of its weighted mean is $\mathrm{Var}_{q(\lambda)}(y_i) > 0$.

For scalar regression, define the finite-support resolution

$$\rho_S(y) = \min_{q \in \Delta^{S-1}, \sum_i q_i y_i = y} \sum_i q_i (y_i - y)^2.$$

The Gaussian information projection satisfies

$$\sum_i q_{\mathcal{S},i}(y)(y_i - y)^2 \leq \rho_S(y) + 2\sigma^2 \log S \leq \tfrac{\Delta_S^2}{4} + 2\sigma^2 \log S, \tag{12}$$

where Δ_S is the largest gap between consecutive distinct support targets. If y exactly matches a support target, then $\rho_S(y) = 0$. The Gaussian teacher remains the minimum-KL projection over all support rows, balancing locality with entropy relative to its Gaussian reference. Related smoothed distributional regression objectives have been studied in Imani & White (2018); Imani et al. (2026).

Theorem 4.2 (Distribution and integrated-pinball transfer). *Let $F_p = \sum_i p_i \delta_{y_i}$, let F^* be the conditional law of an integrable Y, set $D = y_{\max} - y_{\min}$ and $\epsilon_S = \mathbb{E}[\sum_i q_{S,i}(Y)|y_i - Y| \mid \mathcal{C}]$. Then*

$$W_1(F_p, F^*) \leq D\sqrt{\varepsilon(p)/2} + \epsilon_S, \tag{13}$$

where W_1 is the Wasserstein-1 distance. Define $\rho_t(r) = \max\{tr, (t-1)r\}$ and the conditional integrated pinball risk $\mathcal{R}_{\mathrm{IP}}(F) = \mathbb{E}[\int_0^1 \rho_t(Y - Q_F(t))\, dt \mid \mathcal{C}]$. Directly in terms of pinball risk,

$$0 \leq \mathcal{R}_{\mathrm{IP}}(F_p) - \mathcal{R}_{\mathrm{IP}}(F^*) \leq W_1(F_p, F^*) \leq D\sqrt{\varepsilon(p)/2} + \epsilon_S. \tag{14}$$

For the Gaussian teacher, with largest adjacent-target gap Δ_S,

$$\epsilon_S \leq \sqrt{\Delta_S^2/4 + 2\sigma^2 \log S} + \mathbb{E}\left[|Y - \mathrm{clip}(Y, y_{\min}, y_{\max})| \mid \mathcal{C}\right]. \tag{15}$$

Appendix A.5 gives tighter scalar versions that retain the local support geometry and cumulative probability mass. The pinball step uses quantile optimality and the one-Lipschitz dependence of $\rho_t(y - q)$ on q (Koenker & Bassett, 1978); it does not require another scoring rule. The bound separates prediction learning, support-target resolution, smoothing, and tail coverage. Distributional consistency requires these terms to vanish, not just mean preservation. For K levels $t_k = k/(K+1)$ and targets in the support interval, Appendix A.6 adds an explicit quadrature term $4D/(K+1)$ to Equation 14. This connects to the finite multi-quantile objective of TabICLv2 (Qu et al., 2026).

We provide a theoretical bridge between POINT-CE and the distributional regression objectives used by modern TFMs. Although POINT does not directly optimize quantile losses, small POINT-CE excess risk together with sufficient support coverage controls the induced predictive distribution and, consequently, its integrated pinball risk. Thus POINT-CE can target the same distributional quantities sought by multi-quantile regression without requiring a fixed set of quantile outputs.

[1] Out-of-range targets are rare on our evaluated regression benchmarks: 89 of 667,877 query targets (0.0133%) fall outside their split-specific support range.

4.3 WHY CONTEXTUAL REPRESENTATIONS CAN OUTPERFORM FIXED KERNELS

Ordinary k-NN and fixed-kernel smoothers choose weights from covariates, then aggregate labels. In POINT, the Transformer can use *labeled context to construct the geometry itself*:

$$K_{\theta,\mathcal{S}}(x, x_i) = \exp\big(-\|u_\theta(x; \mathcal{S}) - u_\theta(x_i; \mathcal{S})\|^2 / 2h^2 T\big)\,. \tag{16}$$

Although POINT ultimately predicts by weighting support labels, those weights can depend nonlinearly on the labeled context. Prior work shows that pretrained Transformers can learn such context-dependent representations (Kim & Suzuki, 2024; Nishikawa et al., 2025). To make this distinction concrete, we study a simple regression family with one unknown relevant input coordinate, where contextual representations can identify that coordinate but label-independent weighting fails. The following theorem formalizes this separation for the Euclidean POINT readout.

Theorem 4.3 (A continuous-target separation from label-independent weighting). *Draw J uniformly from $\{1, \dots, d\}$ independently for each episode. Draw $X_1, \dots, X_n, X$ independently and uniformly from $[-1, 1]^d$, and set $Y_i = X_{iJ}$ and $Y = X_J$, with $1 \le n < d$. Any predictor $\widehat{f} = \sum_i w_i(X_{1:n}, X) Y_i$ whose weights do not depend on the current support labels satisfies*

$$\mathbb{E}[(\widehat{f} - Y)^2] \ge (d - n)/3d. \tag{17}$$

The weights may be signed and may use independently pretrained parameters. Conversely, for every $\lambda > 0$ and $\zeta > 0$, a sufficiently expressive masked Transformer followed by a tied Euclidean POINT *head with logits $-\lambda(u_x - u_i)^2$ can achieve*

$$\mathbb{E}[(\widehat{f}_\theta - Y)^2] \le \tfrac{8}{(n+1)(n+2)} + \tfrac{\log n}{\lambda} + \zeta. \tag{18}$$

The support labels identify the relevant coordinate, allowing the representation to focus on it and suppress irrelevant dimensions. The final readout still averages observed support labels. The separation therefore comes from adapting the representation to the labeled support set, instead of the weighted-sum readout itself. The lower bound covers fixed-metric k-NN, Gaussian kernel regression, and fixed-kernel ridge regression.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Our models are initialized using a TABICLV2-style encoder and an identity-initialized $512 \to 1024 \to 512$ metric MLP with tied Euclidean scoring (Qu et al., 2026). We pre-train using joint-training budgets containing 64K / 8K (small) and 320K / 40K (large) classification / regression episodes sampled from synthetic priors (Qu et al., 2026), respectively. Each joint model supplies one final checkpoint for both tasks. Appendix C.1 gives the architecture and optimizer. We test models on classification across 57 real datasets, reporting mean accuracy and log loss, and regression using Airfoil, Concrete, and QSAR fish toxicity for development (30 splits each) and ten further TABARENA regression datasets (132 splits) for broader evaluation (Erickson et al., 2025), reporting dataset root mean square error (RMSE) and macro dataset-normalized RMSE; classical estimators are fitted and tuned on support data only.

5.2 SHARED VERSUS TASK-SPECIFIC OUTPUT HEADS

We compare the shared output head used by POINT with three shared-encoder, task-specific controls: a logit classifier paired with a 999-quantile pinball regressor (*Logit + Quantile*) (Qu et al., 2026), a hard 100-bin regressor (*Logit + Bar*) (Müller et al., 2022; Hollmann et al., 2025), and a Gaussian-smoothed 100-bin regressor (*Logit + Histogram*) (Imani & White, 2018). A POINT variant with separate metric MLPs for classification and regression isolates the effect of sharing the head.

Table 1 shows a consistent improvement across small and large pretraining budgets. The shared POINT head gives the best classification accuracy and log loss at both budgets, improves regression over the separate-metric variant, and remains competitive with the strongest task-specific regression head after longer training.

Table 1: Shared and task-specific output heads under small and large pretraining budgets. Classification accuracy and log-loss are averaged over 57 datasets. Regression entries are mean RMSE ± split SD over 30 splits. GM is the geometric mean of the three RMSE means. Bold marks the best displayed mean within each budget.

Budget	Head	Classification		Regression RMSE			
		Acc. (%) ↑	Log loss ↓	Airfoil ↓	Concrete ↓	QSAR ↓	GM ↓
Small	POINT (shared)	**82.04**	**0.451**	$\mathbf{2.161}_{\pm 0.127}$	$\mathbf{5.806}_{\pm 0.382}$	$\mathbf{0.886}_{\pm 0.049}$	**2.232**
	POINT (separate)	81.77	0.471	$2.577_{\pm 0.085}$	$6.191_{\pm 0.279}$	$0.903_{\pm 0.010}$	2.433
	Logit + Quantile	81.94	0.493	$3.025_{\pm 0.154}$	$6.135_{\pm 0.386}$	$0.897_{\pm 0.048}$	2.553
	Logit + Bar	81.49	0.538	$3.840_{\pm 0.146}$	$8.006_{\pm 0.386}$	$0.934_{\pm 0.049}$	3.062
	Logit + Histogram	81.31	0.519	$2.910_{\pm 0.167}$	$6.121_{\pm 0.339}$	$0.891_{\pm 0.045}$	2.513
Large	POINT (shared)	**83.01**	**0.428**	$\mathbf{1.436}_{\pm 0.083}$	$4.949_{\pm 0.319}$	$\mathbf{0.878}_{\pm 0.053}$	**1.841**
	POINT (separate)	82.91	0.430	$1.577_{\pm 0.115}$	$5.028_{\pm 0.323}$	$0.884_{\pm 0.051}$	1.914
	Logit + Quantile	82.88	0.431	$1.455_{\pm 0.116}$	$\mathbf{4.858}_{\pm 0.328}$	$0.883_{\pm 0.048}$	**1.841**
	Logit + Bar	82.34	0.464	$2.124_{\pm 0.165}$	$5.651_{\pm 0.412}$	$0.879_{\pm 0.047}$	2.193
	Logit + Histogram	82.80	0.464	$1.611_{\pm 0.078}$	$5.001_{\pm 0.334}$	$0.886_{\pm 0.048}$	1.925

Table 2: TABARENA regression on ten datasets and 132 splits. Entries are RMSE, except macro nRMSE; lower is better. POINT uses query-local contexts. Values are rounded for display; **bold** marks the best result in each row.

Dataset	POINT (Ours)		Task-specific heads			Traditional ML				
	Base	+CE	Quant.	Hist.	Bar	ET	kNN	CART	KRR	NW
Fiat 500	761.0	761.2	758.4	763.2	764.9	**746.4**	825.1	839.8	767.8	911.4
Food Delivery	8.258	8.264	8.410	8.260	10.349	**7.940**	9.390	8.434	9.950	9.380
QSAR-TID-11	**0.896**	0.898	0.991	1.010	1.242	1.058	0.944	1.058	1.104	1.213
Diamonds	608.8	607.4	1269.5	956.4	1050.1	**545.8**	1742.1	679.3	729.7	2019.5
Insurance	4660	4641	4673	**4597**	4744	4678	5597	4910	4870	5785
Houses	0.234	**0.234**	0.238	0.244	0.258	0.240	0.288	0.286	0.441	0.399
Miami housing	90244	**89862**	101983	101883	119682	92726	114367	132824	124643	158572
Protein	**3.389**	3.391	3.640	3.906	4.300	3.460	3.840	4.544	5.020	5.260
Superconductivity	9.471	9.475	11.650	11.345	14.281	**9.400**	10.740	12.102	15.740	21.440
Wine quality	0.625	0.623	0.641	0.659	0.664	**0.614**	0.631	0.767	0.727	0.743
Macro nRMSE	0.4631	**0.4628**	0.5042	0.5022	0.5666	0.4696	0.5466	0.5488	0.6036	0.6777

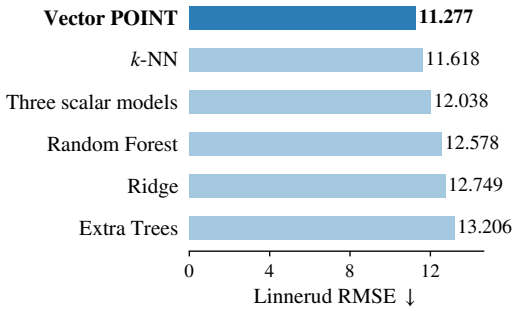
5.3 BROADER REGRESSION AND TEACHER TRANSFER

Table 2 evaluates the larger-budget checkpoints. For long tables, POINT uses FAISS to retrieve 512 support rows per query before encoding, without query labels (Douze et al., 2026). Traditional baselines include Extra Trees (Geurts et al., 2006), k-NN, CART, kernel ridge regression (KRR), and RBF Nadaraya–Watson regression (RBF-NW) (Nadaraya, 1964; Watson, 1964).

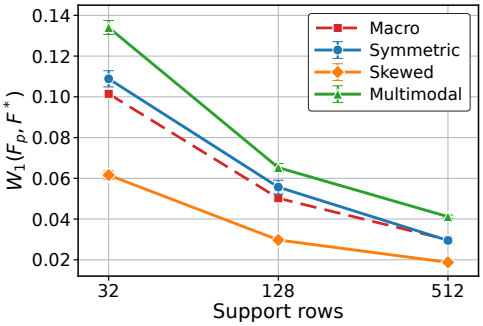
CE-teacher transfer. The + *CE teacher* variant replaces 10% regression episodes with a 75% Gaussian / 25% Extra-Trees teacher blend. Extra-Trees support weights are projected onto the same mean-preserving constraint as the Gaussian teacher, so both define distributions over the same support rows before being mixed in the POINT-CE loss. This tests whether POINT can absorb the tree model's partition structure through its shared support-distribution interface. Macro nRMSE improves slightly from 0.4631 to 0.4628, with mixed dataset-level changes.

5.4 MULTI-OUTPUT REGRESSION

We also verify the effects of one support distribution predicting a vector without separate readouts per dimension. On the Linnerud dataset, a small POINT model with a uniform-reference vector teacher reaches RMSE 11.277, versus 12.038 for three independent scalar POINT models with pinball supervision and 11.618 for k-NN (Figure 2a). It predicts all three targets in one pass.



(a) Three-target Linnerud RMSE.



(b) Larger supports reduce W_1.

Figure 2: Left: multi-output regression on Linnerud, where one support distribution predicts all three targets in a single pass. Right: larger support sets improve CE-only distributional recovery. Error bars in the right panel show 95% batch-mean intervals.

5.5 Matched objective ablations

Table 3 ablates the task-specific components of POINT-CE under the small pretraining budget, enabling neither term, the classification term only, the regression term only, or both. The classification-only variant already improves all reported metrics over removing POINT-CE, while the regression-only variant has mixed effects. Combining both terms gives the best result on every metric.

Table 3: Matched joint-objective ablation. Classification accuracy and log-loss are averaged over 57 datasets. GM is geometric-mean RMSE. Best is **bolded**.

Objective	Classification		Regression RMSE			
	Acc. (%) ↑	Log loss ↓	Airfoil ↓	Concrete ↓	QSAR ↓	GM ↓
No POINT-CE	80.65	0.507	$2.515_{\pm 0.124}$	$6.899_{\pm 0.422}$	$0.901_{\pm 0.055}$	2.500
+ Classification CE	81.33	0.476	$2.225_{\pm 0.104}$	$6.569_{\pm 0.408}$	$0.890_{\pm 0.051}$	2.352
+ Regression CE	81.28	0.488	$3.089_{\pm 0.143}$	$6.264_{\pm 0.456}$	$0.899_{\pm 0.055}$	2.591
Full POINT-CE	**82.04**	**0.451**	$\mathbf{2.161}_{\pm 0.127}$	$\mathbf{5.806}_{\pm 0.382}$	$\mathbf{0.886}_{\pm 0.049}$	**2.232**

5.6 Finite-support distributional prediction

Our theory predicts better distributional recovery as the support becomes denser. We test this with one 12-block checkpoint trained for 16K episodes using only Gaussian POINT-CE ($\sigma = 0.05$), then evaluate it with 32, 128, and 512 nested support rows on three conditional families. Macro W_1 drops from 0.1015 to 0.0298, and excess integrated pinball risk from 0.00770 to 0.00062, with W_1 improving in every family (Table 4; Figure 2b).

Table 4: CE-only distributional prediction across three conditional families. Lower is better.

Rows	32	128	512
W_1	0.101	0.050	**0.030**
Pinball	0.105	0.099	**0.098**
Excess pinball	0.0077	0.0018	**0.0006**
RMSE	0.358	0.348	**0.346**
Outside hull (%)	6.3	1.6	**0.4**

6 Conclusion

POINT provides a common prediction mechanism for categorical and continuous targets by treating the labeled support set itself as the output space. Moment-preserving supervision connects classification and regression through the same support-level cross-entropy objective, while the finite-support analysis links this objective to distributional regression. Across our experiments, one shared prediction head performs competitively on both tasks and extends naturally to vector-valued outputs. These results suggest that support-indexed prediction is a useful direction for building more unified tabular foundation models.

REPRODUCIBILITY STATEMENT

The method, all theorem assumptions, and complete proofs appear in Sections 3–4 and Appendix A. Appendix B specifies target construction, masks, and stable numerical operations. Appendix C distinguishes experimental lineages and inference budgets. Following the ICLR 2027 author guidelines[2] we will share our code with the reviewers and area chairs through a link to an anonymous repository, which we will post as a comment once the discussion forums open.

ETHICS STATEMENT

Support-based predictions depend directly on the labels supplied at inference and can reproduce biases, mislabeled examples, or sensitive information in that context. Attention weights are not causal explanations. These development experiments do not validate medical, financial, or other consequential use. Dataset terms, support-label privacy, subgroup performance, and the appropriateness of any deployment require separate review. No new human-subject study or raw-data release is claimed.

AI USE STATEMENT

An AI assistant was used to help restructure the manuscript, draft and revise prose and mathematical arguments, assemble tables from experiment reports, research references, generate LaTeX and numerical sanity checks, and run a subset of the reported experiments. The authors reviewed the AI-assisted experimental outputs and manuscript content and take responsibility for the final results and claims.

REFERENCES

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017. URL http://proceedings.mlr.press/v70/bellemare17a.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos (eds.), *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pp. 535–541. ACM, 2006. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 12(2):346–361, 2026. doi: 10.1109/TBDATA.2025.3618474. URL https://doi.org/10.1109/TBDATA.2025.3618474.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1697e3fb412da11dc9488249f9e7bbc9-Abstract-Datasets_and_Benchmarks_Track.html.

[2] https://iclr.cc/Conferences/2027/AuthorGuidelines

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. Tabr: Tabular deep learning meets nearest neighbors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rhgIgTSSxW.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kürşat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://doi.org/10.48550/arXiv.2605.13986.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network, 2015. URL http://arxiv.org/abs/1503.02531.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/S41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2162–2171. PMLR, 2018. URL http://proceedings.mlr.press/v80/imani18a.html.

Ehsan Imani, Kai Luedemann, Sam Scholnick-Hughes, Esraa Elelimy, and Martha White. Investigating the histogram loss in regression. *Journal of Machine Learning Research*, 27(66):1–54, 2026. URL http://jmlr.org/papers/v27/24-0260.html.

Benjamin Jäger, Nick Erickson, Léo Grinsztajn, Felix Birkel, Klemens Flöge, Oscar Key, Kürşat Kaya, Jonas Kübler, Adèle Frankel, Tobias Schröder, Anurag Garg, Jan Hendrik Metzen, David Salinas, Simon Bing, Kristina Collins, Tuana Çelik, Vahid Balazadeh, Lydia Sidhoum, Tomás Pereda, Brendan Roof, Andrej Tschalzev, Siyuan Guo, Philipp Singer, Lennart Purucker, Jake Robertson, Marie Salmon, Philipp Jund, Jerry Chen, Diana Kriuchkova, Arthur Cahu, Eliott Kalfon, Adrian Hayler, Georg Grab, Vitor Monteiro, Lilly Wehrhahn, Dominik Safaric, Clara Cornu, Alan Arazi, Rylee Grace, Simone Alessi, Mihir Manium, Bernhard Schölkopf, Yann LeCun, Madelon Hulsebos, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3.5: Technical report, 2026. URL https://arxiv.org/abs/2609.17895.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24527–24561. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/kim24af.html.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*,

6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5. URL https://doi.org/10.1016/S0893-6080(05)80131-5.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. doi: 10.1198/016214505000001230.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020. URL https://doi.org/10.1137/1109020.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25660–25676. PMLR, 2023. URL https://proceedings.mlr.press/v202/nagler23a.html.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46554–46585. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/nishikawa25a.html.

Pascal Pfeiffer, Dmitry Gordeev, Mathias Müller, Laura Fink, Joan Salvà Soler, Mark Landry, Branden Murray, Marcos V. Conde, and Sri Satish Ambati. Tabh2o: A unified foundation model for tabular prediction, 2026. URL https://doi.org/10.48550/arXiv.2605.18383.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026. URL https://arxiv.org/abs/2602.11139.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/45e9827fcd83cfb39285669e13b2a30d-Abstract-Conference.html.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2692–2700, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581572X. URL http://www.jstor.org/stable/25049340.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, Ningbo Dai, Renzhe Xu, Shuyang Li, Tianyang Zhang, Yue He, Yuanrui Wang, Yunjia Zhang, Zijing Xu, Dongzhe Li, Fang Gao, Hao Zou, Jiandong Liu, Jiashuo Liu, Jiawei Xu, Kaijie Cheng, Kehan Li, Linjun Zhou, Qing Li, Shaohua Fan, Xiaoyu Lin, Xinyan Han, Xuanyue Li, Yan Lu, Yuan Xue, Yuanyuan Jiang, Zimu Wang, Zhenlei Wang, and Peng Cui. Limix: Unleashing structured-data modeling capability for generalist intelligence, 2025. URL https://doi.org/10.48550/arXiv.2509.03505.

A PROOFS AND MATHEMATICAL SCOPE

Except in the explicitly episode-averaged representation comparisons below, all statements condition on a fixed context $\mathcal{C}$. The support atoms and target transformation are then fixed, but the query label remains random. We allow simplex-boundary optima and use $0 \log 0 = 0$. Finite-logit softmax heads approximate boundary optima as limits. If $p_i = 0$ where $\pi_i > 0$, the corresponding risk and KL are infinite. We write PCE as a mathematical subscript for POINT-CE. In the scalar proofs, a_i denotes the scalar support value y_i.

A.1 CONDITIONAL FISHER CONSISTENCY AND EXCESS PREDICTION RISK

Proof of Theorem 4.1. Linearity of conditional expectation gives

$$\mathcal{R}_{\mathrm{PCE}}(p) = -\sum_i \mathbb{E}[q_{S,i}(Y) \mid \mathcal{C}] \log p_i = -\sum_i \pi_i \log p_i \tag{19}$$

$$= H(\pi) + \mathrm{KL}(\pi \| p). \tag{20}$$

Gibbs' inequality proves Eq. (7), including uniqueness of the minimizing probability vector. For a teacher whose mean is $\Pi_{\mathcal{H}_S}(Y)$, finite summation gives

$$\sum_i \pi_i y_i = \mathbb{E}\left[\sum_i q_{S,i}(Y) y_i \middle| \mathcal{C}\right] = \mathbb{E}[\Pi_{\mathcal{H}_S}(Y) \mid \mathcal{C}]. \tag{21}$$

When no projection is needed, this is $\mathbb{E}[Y \mid \mathcal{C}]$. Let $t = \mathrm{TV}(p, \pi) = \frac{1}{2}\sum_i |p_i - \pi_i|$. If $t = 0$, the barycenters agree. Otherwise write the positive and negative parts of $p - \pi$ as tr^+ and tr^-, with probability vectors r^+, r^-. Then

$$\sum_i (p_i - \pi_i) y_i = t\left(\sum_i r_i^+ y_i - \sum_i r_i^- y_i\right).$$

Both sums lie in the same convex hull, whose Euclidean diameter is D. The norm is at most Dt. Pinsker's inequality for natural logarithms gives $t^2 \le \mathrm{KL}(\pi\|p)/2$, proving Eq. (8).

For classification, class coverage and the one-hot moment constraint imply $\sum_{i:y_i=e_c} q_{S,i}(Y) = \mathbf{1}[Y = e_c]$. Consequently $\sum_{i:y_i=e_c} \pi_i = P^*(c \mid \mathcal{C})$. Applying the log-sum inequality within each class yields

$$\sum_c P^*(c \mid \mathcal{C}) \log \frac{P^*(c \mid \mathcal{C})}{\sum_{i:y_i=e_c} p_i} \le \sum_i \pi_i \log \frac{\pi_i}{p_i}.$$

This proves Eq. (9). Adding and subtracting $\sum_i \pi_i y_i$ and applying the triangle inequality proves Eq. (10).

Squared error and classification decisions. If Y has finite second moment and needs no projection, define the Bayes regression function

$$\mu^*(\mathcal{C}) = \mathbb{E}[Y \mid \mathcal{C}].$$

Since $\widehat{y} = \sum_i p_i y_i$, the orthogonality identity

$$\mathbb{E}[\|Y - \widehat{y}\|_2^2 \mid \mathcal{C}] - \mathbb{E}[\|Y - \mu^*(\mathcal{C})\|_2^2 \mid \mathcal{C}] = \|\widehat{y} - \mu^*(\mathcal{C})\|_2^2$$

turns the weighted-mean bound into a squared-error excess-risk bound. Inverse target scaling is required before interpreting it in raw data units. For $c^* \in \arg\max_c P^*(c)$ and $\widehat{c} \in \arg\max_c P_p(c)$,

$$P^*(c^*) - P^*(\widehat{c}) \le \|P^* - P_p\|_1 \le \sqrt{2\mathrm{KL}(P^*\|P_p)} \le \sqrt{2\varepsilon(p)}.$$

Ties may use a fixed convention. If a possible query class is absent from support, its predicted mass is necessarily zero and the class-probability consistency claim does not apply. For regression, moment recovery is for the projected target unless the conditional target law is supported inside the support hull. In particular, finite random support does not automatically provide coverage even when all samples follow the same law.

A.2 Classification row-level versus group-level supervision

Proposition A.1 (Exact decomposition of the categorical teacher loss). *For a present target class y, let $G_y = \{i : y_i = y\}$, $n_y = |G_y|$, U_y be uniform on G_y, and $p(i \mid y) = p_i/P_p(y)$. Then Eq. (11) holds.*

Proof. Write $p_i = P_p(y)r_i$ on G_y. Since $\sum_{i\in G_y} r_i = 1$,

$$-\frac{1}{n_y}\sum_{i\in G_y}\log p_i = -\log P_p(y) - \frac{1}{n_y}\sum_{i\in G_y}\log r_i \tag{22}$$

$$= -\log P_p(y) + \log n_y + \mathrm{KL}(U_y\|r). \tag{23}$$

At the population POINT-CE optimum, $p_i = P^*(y_i \mid \mathcal{C})/n_{y_i}$, so class aggregation gives P^*. Class NLL has the same optimal class masses. Any nonnegative combination with a positive POINT-CE coefficient retains those unrestricted class probabilities.

The $\log n_y$ term is model-independent; the within-class KL is not. It discourages concentrating a class's mass on only one support example. A label-permutation-equivariant encoder makes the head equivariant to a bijection of class identifiers, but one-hot output aggregation alone does not guarantee equivariance of an arbitrary encoder.

A.3 Gaussian information projection

Proposition A.2 (Existence, uniqueness, and locality). *Let $a_{\min} < a_{\max}$, $\sigma > 0$, $a_{\min} < y < a_{\max}$, and $r_i = \exp[-(a_i - y)^2/(2\sigma^2)]/Z_0$. The constrained problem*

$$\min_{q_i\ge 0,\ \sum_i q_i = 1} \mathrm{KL}(q\|r) \quad \textit{subject to} \quad \sum_i q_i a_i = y \tag{24}$$

has a unique solution $q_i = r_i e^{\alpha a_i}/\sum_j r_j e^{\alpha a_j}$ and a unique finite α. It satisfies Eq. (27). At an endpoint, its limiting solution is uniform over matching endpoint rows.

Proof. The feasible set is nonempty, closed, and compact. The reference has positive entries, and strict convexity of KL gives uniqueness. The interior moment admits a positive feasible allocation. A minimizer can not have a zero entry: mixing it with a strictly positive feasible allocation has negative infinite directional derivative in that coordinate. Lagrange multipliers therefore give

$$\log q_i - \log r_i + 1 - \alpha a_i + \nu = 0, \qquad q_i = \frac{r_i e^{\alpha a_i}}{\sum_j r_j e^{\alpha a_j}}.$$

For $Z(\alpha) = \sum_i r_i e^{\alpha a_i}$, its mean $\mu(\alpha) = (\log Z)'$ obeys

$$\mu'(\alpha) = (\log Z)'' = \sum_i q_i a_i^2 - \left(\sum_i q_i a_i\right)^2 = \mathrm{Var}_{q(\alpha)}(a) > 0. \tag{25}$$

Moreover $\mu(\alpha)$ tends to the two support endpoints as $\alpha \to \mp\infty$. Continuity and strict monotonicity yield one finite solution for each interior target. At an endpoint, feasible mass lies only on rows with that label; their equal reference weights yield a uniform optimum. If all labels coincide, that uniform distribution is the only relevant limiting construction.

For any feasible q,

$$\mathrm{KL}(q\|r) = \frac{\sum_i q_i(a_i - y)^2}{2\sigma^2} - H(q) + \log Z_0.$$

Define the finite-support resolution at y by

$$\rho_S(y) := \min_{\substack{q\in\Delta_S \\ \sum_i q_i a_i = y}} \sum_i q_i(a_i - y)^2. \tag{26}$$

Let q^0 attain this minimum. Since the Gaussian information projection q^σ minimizes $\mathrm{KL}(q\|r)$ over the same feasible set,

$$\mathrm{KL}(q^\sigma\|r) \le \mathrm{KL}(q^0\|r).$$

Using the preceding decomposition and cancelling $\log Z_0$ gives

$$\sum_i q_i^\sigma (a_i - y)^2 \le \rho_S(y) + 2\sigma^2 \big[H(q^\sigma) - H(q^0)\big] \le \rho_S(y) + 2\sigma^2 \log S.$$

For scalar targets, if Δ_S denotes the largest gap between consecutive distinct support values, then

$$\rho_S(y) \le \frac{\Delta_S^2}{4},$$

and hence

$$\sum_i q_i^\sigma (a_i - y)^2 \le \frac{\Delta_S^2}{4} + 2\sigma^2 \log S. \tag{27}$$

If y coincides with a support value, then $\rho_S(y) = 0$. Duplicate support rows do not affect the bound.

Zero-width limit and reference measure. On fixed scalar support, minimizing variance at a fixed mean is equivalent to minimizing $\sum_i q_i a_i^2$. As $\sigma \downarrow 0$, the Gaussian information projection increasingly prioritizes this minimum-spread feasible distribution, since the variance term is scaled by $1/(2\sigma^2)$ while entropy remains bounded. Hence any accumulation point minimizes the variance subject to the mean constraint. For scalar support, this minimum is attained using the closest support values needed to represent the target, or an exact match when available. Rows with identical target values share their total mass uniformly. For any finite $\sigma > 0$, however, the Gaussian projection is generally soft and assigns positive mass to more than these limiting support values. The optimization is a minimum relative-entropy projection toward a local Gaussian reference, rather than maximum Shannon entropy under the mean constraint alone, consistent with the information-projection viewpoint (Csiszár, 1975).

A.4 Distributional error and a direct pinball bound

Proof of Theorem 4.2. Sample $Y \sim F^*$ and then $I \mid Y \sim q_S(Y)$. The marginal law of a_I is $F_\pi = \sum_i \pi_i \delta_{a_i}$, so this coupling proves

$$W_1(F_\pi, F^*) \le \mathbb{E}[|a_I - Y| \mid \mathcal{C}] = \epsilon_S.$$

A coupling between p and π matches their common mass and moves the remaining $\mathrm{TV}(p, \pi)$ mass at most D. Therefore

$$W_1(F_p, F_\pi) \le D \cdot \mathrm{TV}(p, \pi) \le D\sqrt{\varepsilon(p)/2}.$$

The triangle inequality proves Eq. (13). For an in-range target, Cauchy–Schwarz and Proposition A.2 give

$$\sum_i q_i |a_i - y| \le \sqrt{\sum_i q_i (a_i - y)^2} \le \sqrt{\Delta_S^2/4 + 2\sigma^2 \log S}.$$

With $\bar{Y} = \mathrm{clip}(Y, a_{\min}, a_{\max})$, apply $|a_i - Y| \le |a_i - \bar{Y}| + |\bar{Y} - Y|$ and take expectations. This proves Eq. (15). For a singleton support interval, set $\Delta_S = 0$; its approximation cost is exactly $\mathbb{E}|a_1 - Y|$.

For the pinball result, put $Q_F(t) = \inf\{z : F(z) \ge t\}$ for $t \in (0, 1)$. The subgradient of $q \mapsto \mathbb{E}[\rho_t(Y - q) \mid \mathcal{C}]$ is the interval $[F^*(q^-) - t, F^*(q) - t]$, where $F^*(q^-)$ denotes the left limit of the CDF at q. It contains zero at every valid t-quantile, including $Q_{F^*}(t)$; hence each excess quantile's population risk is nonnegative. Also,

$$|\rho_t(y - q) - \rho_t(y - q')| \le \max\{t, 1 - t\}|q - q'| \le |q - q'|.$$

Integrating this inequality and using the one-dimensional transport identity gives

$$0 \le \mathcal{R}_{\mathrm{IP}}(F_p) - \mathcal{R}_{\mathrm{IP}}(F^*) \tag{28}$$

$$\le \int_0^1 |Q_{F_p}(t) - Q_{F^*}(t)|\, dt = W_1(F_p, F^*). \tag{29}$$

This proves Eq. (14) directly from pinball loss. Integrability justifies the expectations and the transport identity.

A.5 SHARPER SCALAR TRANSFER BOUNDS

The bounds in Theorem 4.2 use only the largest support diameter D. For scalar targets we can retain the ordered support geometry instead. The resulting expressions are more involved, but are never weaker than the diameter–Pinsker relaxation.

Cumulative-mass Wasserstein bound. Let $a_1 < \cdots < a_M$ be the distinct scalar support targets and $d_k = a_{k+1} - a_k$. Define

$$P_k = \sum_{i:y_i \le a_k} p_i, \qquad \Pi_k = \sum_{i:y_i \le a_k} \pi_i, \qquad \pi_i = \mathbb{E}[q_{S,i}(Y) \mid \mathcal{C}].$$

For $v \in [0,1]$ and $e \ge 0$, let

$$b(v,e) = \max\{|u - v| : u \in [0,1],\ \mathrm{kl}(u\|v) \le e\}, \tag{30}$$

where kl is Bernoulli KL, and define

$$B_1(p,e) = \sum_{k=1}^{M-1} d_k b(P_k, e), \qquad B_2(p,e) = \frac{1}{2}\sum_{k=1}^{M-1} d_k b(P_k,e)^2. \tag{31}$$

Proposition A.3 (Target-geometry transfer). *For $e \ge \mathrm{KL}(\pi\|p)$,*

$$W_1(F_p, F_\pi) = \sum_{k=1}^{M-1} d_k |P_k - \Pi_k| \le B_1(p,e), \tag{32}$$

$$\frac{1}{2}\int_{\mathbb{R}} (F_p(z) - F_\pi(z))^2\, dz = \frac{1}{2}\sum_{k=1}^{M-1} d_k (P_k - \Pi_k)^2 \le B_2(p,e). \tag{33}$$

Moreover,

$$B_1(p,e) \le D\sqrt{e/2}, \qquad B_2(p,e) \le De/4. \tag{34}$$

Consequently,

$$W_1(F_p, F^*) \le B_1(p, \varepsilon(p)) + \epsilon_S \le D\sqrt{\varepsilon(p)/2} + \epsilon_S. \tag{35}$$

Proof. For each support gap, apply the log-sum inequality to the two groups $\{i : y_i \le a_k\}$ and $\{i : y_i > a_k\}$:

$$\mathrm{kl}(\Pi_k\|P_k) \le \mathrm{KL}(\pi\|p) \le e.$$

Hence $|\Pi_k - P_k| \le b(P_k, e)$. In one dimension, $W_1(F_p, F_\pi) = \int |F_p - F_\pi|$, and the two CDFs are constant on each interval (a_k, a_{k+1}), giving Eq. (32). The same interval decomposition gives Eq. (33). Bernoulli Pinsker yields $b(v,e) \le \sqrt{e/2}$; summing the support gaps, whose total is D, gives Eq. (34). Finally, $W_1(F_\pi, F^*) \le \epsilon_S$ by the teacher coupling used in the proof of Theorem 4.2; the triangle inequality gives Eq. (35).

The exact first equality in Eq. (32) is sharper still when π is available, because it uses the realized cumulative mass error rather than a common KL radius at every support gap.

A sharper integrated-pinball bound. For the pinball normalization used in this paper,

$$\mathcal{R}_{\mathrm{IP}}(F) - \mathcal{R}_{\mathrm{IP}}(F^*) = \frac{1}{2}\int_{\mathbb{R}} (F(z) - F^*(z))^2\, dz. \tag{36}$$

This follows by writing $\rho_t(Y - q) = t(Y - q) + (q - Y)_+$, integrating the excess risk first over q and then over t, and applying Tonelli's theorem. It also shows that the integrated pinball excess is controlled by CDF error directly, not only through a quantile-Lipschitz relaxation.

Combining Eq. (36) with Proposition A.3 gives

$$\begin{aligned} 0 \le \mathcal{R}_{\mathrm{IP}}(F_p) - \mathcal{R}_{\mathrm{IP}}(F^*) \le \min\Big\{ &\tfrac{1}{2}\big[B_1(p,\varepsilon(p)) + \epsilon_S\big], \\ &\big(\sqrt{B_2(p,\varepsilon(p))} + \sqrt{\epsilon_S/2}\big)^2 \Big\}. \end{aligned} \tag{37}$$

Indeed, $|F_\pi - F^*| \le 1$ implies $\|F_\pi - F^*\|_2^2 \le \|F_\pi - F^*\|_1 = W_1(F_\pi, F^*) \le \epsilon_S$, and the second line follows from the L_2 triangle inequality. In the idealized case $F_\pi = F^*$, Eq. (37) reduces to

$$\mathcal{R}_{\mathrm{IP}}(F_p) - \mathcal{R}_{\mathrm{IP}}(F^*) \le B_2(p, \varepsilon(p)) \le D\varepsilon(p)/4,$$

which is linear, rather than square-root, in the POINT-CE excess risk.

Together with Proposition A.5, the same refinement gives the finite-grid corollary

$$0 \le \mathcal{R}_K(F_p) - \mathcal{R}_K(F^*) \le B_{\mathrm{IP}} + \frac{4D}{K+1}, \tag{38}$$

where B_{IP} denotes the right-hand side of Eq. (37) before the final period.

Local teacher spread instead of worst-case $\log S$**.** Let

$$\bar{Y} = \mathrm{clip}(Y, y_{\min}, y_{\max}), \qquad \tau_S = \mathbb{E}[|Y - \bar{Y}| \mid \mathcal{C}].$$

For each $\bar{Y}$, let $q_S^0(\bar{Y})$ be a maximum-entropy minimizer of $\rho_S(\bar{Y})$, and define

$$h_S(Y) = H(q_S(Y)) - H(q_S^0(\bar{Y})).$$

The same KL comparison used in Proposition A.2 gives

$$\sum_i q_{S,i}(Y)(y_i - \bar{Y})^2 \le \rho_S(\bar{Y}) + 2\sigma^2 h_S(Y), \qquad 0 \le h_S(Y) \le \log S.$$

Therefore

$$\epsilon_S \le \tau_S + \sqrt{\mathbb{E}[\rho_S(\bar{Y}) + 2\sigma^2 h_S(Y) \mid \mathcal{C}]} \le \tau_S + \sqrt{\Delta_S^2/4 + 2\sigma^2 \log S}. \tag{39}$$

Thus the main-text smoothing term is a worst-case relaxation of a more local quantity. If $a \le y \le b$ are consecutive distinct support targets, then

$$\rho_S(y) = (y - a)(b - y), \tag{40}$$

with $\rho_S(y) = 0$ at an exact support match.

A distribution-dependent bound for the regression mean. The same KL direction also admits a sharper scalar mean bound. Put $\mu_p = \sum_i p_i y_i$, $\mu_\pi = \sum_i \pi_i y_i$, and for $e \ge \mathrm{KL}(\pi\|p)$ define

$$B_\mu(p, e) = \max_{s \in \{-1,1\}} \inf_{t>0} \frac{e + \log \sum_i p_i \exp\{st(y_i - \mu_p)\}}{t}. \tag{41}$$

Proposition A.4 (Mean transfer in the correct KL direction). *For scalar support targets,*

$$|\mu_\pi - \mu_p| \le B_\mu(p, e) \le D\sqrt{e/2}. \tag{42}$$

If $V_p = \sum_i p_i(y_i - \mu_p)^2$*, then also*

$$B_\mu(p, e) \le \sqrt{2V_p e} + \frac{D}{3}e. \tag{43}$$

When the teacher preserves the projected target,

$$|\mu_p - \mathbb{E}[Y \mid \mathcal{C}]| \le B_\mu(p, \varepsilon(p)) + |\mathbb{E}[\Pi_{\mathcal{H}_S}(Y) - Y \mid \mathcal{C}]|. \tag{44}$$

Proof. For $s \in \{-1, 1\}$ and $t > 0$, define

$$r_i = \frac{p_i \exp\{st(y_i - \mu_p)\}}{\sum_j p_j \exp\{st(y_j - \mu_p)\}}.$$

Nonnegativity of $\mathrm{KL}(\pi\|r)$ gives

$$s(\mu_\pi - \mu_p) \le \frac{e + \log \sum_i p_i \exp\{st(y_i - \mu_p)\}}{t}.$$

Taking the infimum over t and both signs proves the first inequality. Hoeffding's log-moment bound on an interval of width D gives the second. For Eq. (43), expand the exponential under p and use $\mathbb{E}_p|Y - \mu_p|^m \le D^{m-2}V_p$ for $m \ge 2$, yielding

$$\log \mathbb{E}_p e^{st(Y - \mu_p)} \le \frac{V_p t^2}{2(1 - Dt/3)}, \qquad 0 < t < 3/D.$$

Optimizing gives the stated Bernstein form. Finally, $\mu_\pi = \mathbb{E}[\Pi_{\mathcal{H}_S}(Y) \mid \mathcal{C}]$, so the triangle inequality gives Eq. (44).

The variance in Eq. (43) is under p, not π. This direction matters because the excess risk is $\mathrm{KL}(\pi\|p)$. For example, with support values $(0, 1)$, $\pi = (1, 0)$ and $p = (0.9, 0.1)$, the mean difference is 0.1 while $\mathrm{KL}(\pi\|p) = -\log 0.9$ and $V_\pi = 0$; replacing V_p by V_π would give a false bound.

What the logged teacher KL measures. For a fixed context, write $q_Y = q_S(Y)$. Then

$$\mathbb{E}[\mathrm{KL}(q_Y\|p) \mid \mathcal{C}] = \underbrace{\mathrm{KL}(\pi\|p)}_{\varepsilon(p)} + \underbrace{\mathbb{E}[\mathrm{KL}(q_Y\|\pi) \mid \mathcal{C}]}_{\text{teacher variability}}. \tag{45}$$

This follows by inserting $\log(q_{Y,i}/p_i) = \log(q_{Y,i}/\pi_i) + \log(\pi_i/p_i)$ and averaging over Y. The second term can remain positive even when $p = \pi$. Therefore a logged query-specific teacher KL is an upper bound on the conditional POINT-CE excess risk only after taking the corresponding conditional expectation; it is not itself the population excess risk. This distinction is important when assessing whether the bounds above are numerically informative.

Fixed-embedding rank does not imply a universal error floor. The affine-rank observation in Appendix A.11 conditions on fixed support embeddings while varying the query embedding. It does not imply that every learned contextual Euclidean head has a positive approximation floor whenever $S - 1$ exceeds the embedding width. For example, for any single strictly positive probability vector p and any $\lambda > 0$, one scalar dimension is enough to realize it by taking $u_x = 0$ and

$$u_i = \sqrt{-\log p_i/\lambda},$$

for which logits $-\lambda(u_x - u_i)^2 = \log p_i$ and softmax recovers p exactly. This construction does not show that one finite network can realize every context simultaneously; it only shows that the fixed-embedding rank statement by itself cannot establish a universal positive lower bound on $\varepsilon(p)$.

Means and finite support. For a finite-first-moment law,

$$\mathbb{E}_F[Y] = \int_0^1 Q_F(t)\,dt. \tag{46}$$

For F_p, the mean is exactly $\sum_i p_i a_i$, the POINT regression readout. Theorem 4.2 also controls the full predictive law in Wasserstein distance. The bound separates the effects of finite-support resolution, Gaussian smoothing, support-range coverage, and POINT-CE learning error. As these terms vanish $\Delta_S \to 0$, $\sigma^2 \log S \to 0$, vanishing expected tail distance, and $D\sqrt{\varepsilon(p)} \to 0$, F_p converges to F^* in W_1.

A.6 A FINITE-GRID CONNECTION TO MULTI-QUANTILE TRAINING

The integrated objective is a continuum idealization. The next proposition handles the uniform interior grid used by a K-quantile objective and makes the additional approximation explicit.

Proposition A.5 (Finite quantile grid). *Suppose both Y and the forecasts are supported in $[a, b]$, with $D = b - a$. Set $t_k = k/(K+1)$ and*

$$\mathcal{R}_K(F) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\rho_{t_k}(Y - Q_F(t_k)) \mid \mathcal{C}].$$

Then

$$|\mathcal{R}_K(F) - \mathcal{R}_{\mathrm{IP}}(F)| \le \frac{2D}{K+1}, \tag{47}$$

$$0 \le \mathcal{R}_K(F_p) - \mathcal{R}_K(F^*) \le D\sqrt{\varepsilon(p)/2} + \epsilon_S + \frac{4D}{K+1}. \tag{48}$$

Moreover,

$$\left|\frac{1}{K}\sum_{k=1}^{K} Q_F(t_k) - \mathbb{E}_F[Y]\right| \le \frac{D}{K+1}.$$

Proof. For a distribution F, define

$$g_F(t) = \mathbb{E}[\rho_t(Y - Q_F(t)) \mid \mathcal{C}].$$

Then

$$\mathcal{R}_{\mathrm{IP}}(F) = \int_0^1 g_F(t)\,dt, \qquad \mathcal{R}_K(F) = \frac{1}{K}\sum_{k=1}^{K} g_F(t_k),$$

where $t_k = k/(K+1)$.

We first bound the variation of g_F. For $s < t$, insert the intermediate term $\rho_t(Y - Q_F(s))$ to obtain

$$\begin{aligned}|g_F(t) - g_F(s)| \le \mathbb{E}[|\rho_t(Y - Q_F(t)) - \rho_t(Y - Q_F(s))||\mathcal{C}] \\ + \mathbb{E}[|\rho_t(Y - Q_F(s)) - \rho_s(Y - Q_F(s))||\mathcal{C}]\,.\end{aligned}$$

The first term is at most

$$|Q_F(t) - Q_F(s)|$$

because pinball loss is one-Lipschitz in its prediction. For the second term,

$$|\rho_t(r) - \rho_s(r)| = |t - s|\,|r|,$$

and since both Y and the forecast quantiles lie in $[a, b]$, $|Y - Q_F(s)| \le D$. Hence

$$|g_F(t) - g_F(s)| \le |Q_F(t) - Q_F(s)| + D|t - s|.$$

Because Q_F is nondecreasing and takes values in an interval of width D, its total variation is at most D. Therefore the total variation of g_F on $[0, 1]$ is at most

$$\mathrm{TV}(g_F) \le D + D = 2D.$$

Let

$$\mu_K = \frac{1}{K}\sum_{k=1}^{K} \delta_{k/(K+1)}$$

be the empirical measure on the quantile grid. Its discrepancy from the uniform distribution on $[0, 1]$ is at most $1/(K+1)$. The one-dimensional integration-by-parts bound for bounded-variation functions therefore gives

$$\begin{aligned}|\mathcal{R}_K(F) - \mathcal{R}_{\mathrm{IP}}(F)| &= \left|\int_0^1 g_F(t)\,d\mu_K(t) - \int_0^1 g_F(t)\,dt\right| \\ &\le \frac{\mathrm{TV}(g_F)}{K+1} \\ &\le \frac{2D}{K+1}.\end{aligned}$$

This proves Eq. (47).

At each grid level t_k, the true quantile $Q_{F^*}(t_k)$ minimizes population pinball risk. Hence

$$\mathcal{R}_K(F_p) - \mathcal{R}_K(F^*) \ge 0.$$

For the upper bound, add and subtract the two integrated risks:

$$\begin{aligned}\mathcal{R}_K(F_p) - \mathcal{R}_K(F^*) = \big[\mathcal{R}_K(F_p) - \mathcal{R}_{\mathrm{IP}}(F_p)\big] \\ + \big[\mathcal{R}_{\mathrm{IP}}(F_p) - \mathcal{R}_{\mathrm{IP}}(F^*)\big] \\ + \big[\mathcal{R}_{\mathrm{IP}}(F^*) - \mathcal{R}_K(F^*)\big].\end{aligned}$$

Applying the preceding quadrature bound to F_p and F^* and Theorem 4.2 to the middle term gives

$$\mathcal{R}_K(F_p) - \mathcal{R}_K(F^*) \le D\sqrt{\frac{\varepsilon(p)}{2}} + \epsilon_S + \frac{4D}{K+1}.$$

Finally,

$$\mathbb{E}_F[Y] = \int_0^1 Q_F(t)\,dt.$$

Since Q_F is nondecreasing with total variation at most D, the same quadrature argument gives

$$\left|\frac{1}{K}\sum_{k=1}^{K} Q_F(t_k) - \mathbb{E}_F[Y]\right| \le \frac{D}{K+1}.$$

The bounded-support assumption is needed for these constants; unbounded forecasts require separate tail or quadrature control. TabICLv2's $K = 999$ grid is one instance (Qu et al., 2026). The quantiles in this proposition come from valid monotone distributions. It does not identify arbitrary crossing outputs of a quantile network with such a distribution, nor prove that finite-grid and integrated training have identical minimizers over a restricted network class.

Histogram likelihood. For a fixed finite bin I_k with mass b_k, the piecewise-constant density $f_b(y) = b_{k(y)}/|I_{k(y)}|$ satisfies

$$-\log f_b(y) = -\log b_{k(y)} + \log |I_{k(y)}|. \tag{49}$$

More generally, fixed normalized within-bin shapes g_k give $-\log f(y) = -\log b_{k(y)} - \log g_{k(y)}(y)$. Only the mass term is learned through b_k. POINT's atomic forecast is not a density until a smoothing kernel or binning rule is supplied. This is a relationship between learning interfaces, not equality of their continuous likelihoods.

A.7 Contextual representations versus label-independent kernels

The separation is between *weights fixed with respect to the new episode's labels* and *weights that adapt through labeled context*. It is not between all Transformers and all kernel algorithms. The latter may themselves use supervised feature learning.

Kim & Suzuki (2024) analyze an MLP followed by linear attention and establish nonlinear feature-learning properties in a mean-field, two-timescale regime. These are pretraining representation results, not guarantees for our finite model. Nishikawa et al. (2025) prove that gradient-pretrained nonlinear Transformers can extract prompt-specific features on a single-index task family and outperform fixed-basis algorithms under their assumptions. We borrow the *mechanism and comparison class*, not their training-rate theorem for a different head. Attention–kernel connections also support intrinsic-geometry analyses (Shen et al., 2026). Finally, the empirical shared-key Euclidean preference in TabR motivates our scorer, not a universal optimality claim (Gorishniy et al., 2024). The two separations proved below are explicit calculations for POINT-compatible readouts.

Lemma A.6 (A lower bound for covariate-only label weighting). *Let $A \in \mathbb{R}^{n\times d}$ collect support covariates as rows, let the independent query satisfy $\mathbb{E}[XX^\top] = \nu I_d$, where $\nu = \mathbb{E}[X_j^2]$. Let J be uniform on $[d]$, independent of A, X. Let $Y_{1:n} = Ae_J$ and $Y = X_J$. For every $w = w(A, X)$,*

$$\mathbb{E}[(w^\top Y_{1:n} - Y)^2] \geq \nu(1 - n/d).$$

The weights need not be positive or sum to one. Independently pretrained parameters may be conditioned upon throughout.

Proof. Conditional on A, X, average over J:

$$\mathbb{E}_J[(w^\top Ae_J - X_J)^2 \mid A, X] = \frac{1}{d}\|A^\top w - X\|_2^2.$$

Let Π_A be orthogonal projection onto the row space of A. For every $w(A, X)$,

$$\|A^\top w - X\|_2^2 \geq \|(I - \Pi_A)X\|_2^2.$$

Since X is independent of A, the conditional expectation of the right-hand side is $\nu \operatorname{tr}(I - \Pi_A) = \nu(d - \operatorname{rank} A) \geq \nu(d - n)$. Divide by d and average over A.

The crucial step averages over J while keeping w fixed. It fails when the weights depend on Ae_J, precisely the information exploited by the contextual representation. This class includes any fixed-metric k-NN weighting, Nadaraya–Watson weighting, and $w = (K(A, A) + \gamma I)^{-1}k(A, X)$ for fixed-kernel ridge regression. The kernel and γ may depend on independent pretraining or current covariates, but not on current labels. Label-based metric fitting, label-based choice of k or γ, and arbitrary nonlinear processing of labels lie outside the lower bound.

Proof of Theorem 4.3. The lower bound follows from Lemma A.6 with $\nu = 1/3$.

Step 1: identify the relevant coordinate from support. Define observable statistics

$$s_j = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_{ij})^2, \qquad j \in [d].$$

We have $s_J = 0$. For $j \neq J$, independent continuously distributed coordinates give $s_j > 0$ almost surely. Therefore $J = \arg\min_j s_j$ is identifiable from labeled support alone. This is a noiseless-coordinate property; it is not asserted for arbitrary noisy regressions.

Step 2: use the same scalar Euclidean map for support and query. With $\psi_{\mathcal{S}}(z) = z_J$, define

$$p_i = \frac{\exp[-\lambda(X_J - X_{iJ})^2]}{\sum_l \exp[-\lambda(X_J - X_{lJ})^2]}, \qquad \widehat{f}_0 = \sum_i p_i Y_i.$$

These are POINT probabilities with $\lambda = 1/(2h^2T)$ and a one-dimensional tied metric. Let $c_i = (X_J - X_{iJ})^2$ and $c_{\min} = \min_i c_i$. The Gibbs weights minimize $\sum_i p_i c_i - \lambda^{-1} H(p)$ over the simplex. Comparing with a point mass on a nearest support gives

$$\sum_i p_i c_i \le c_{\min} + \lambda^{-1} H(p) \le c_{\min} + \frac{\log n}{\lambda}.$$

Because $Y_i = X_{iJ}$, Jensen's inequality yields

$$(\widehat{f}_0 - Y)^2 = \left[\sum_i p_i (X_{iJ} - X_J)\right]^2 \le \sum_i p_i c_i \le c_{\min} + \frac{\log n}{\lambda}.$$

Step 3: bound nearest-label resolution. For any $z \in [-1, 1]$ and $0 \le r \le 2$, a uniform support coordinate lies within distance r of z with probability at least $r/2$. Thus $R = \min_i |X_{iJ} - X_J|$ satisfies

$$\Pr(R > r \mid X_J) \le (1 - r/2)^n, \qquad \mathbb{E}R^2 \le \int_0^2 2r(1 - r/2)^n \, dr = \frac{8}{(n+1)(n+2)}.$$

Combining the last two inequalities proves the desired upper bound for the ideal feature selector, without the ζ term. The estimate includes queries outside the observed support interval; the nonzero nearest-label term accounts for that coverage limitation.

Step 4: realize the selector with a masked Transformer. To avoid a discontinuous equality test, for $\gamma > 0$ set

$$g_j^\gamma = \frac{e^{-\gamma s_j}}{\sum_l e^{-\gamma s_l}}, \qquad \psi_{\mathcal{S}}^\gamma(z) = \sum_j g_j^\gamma z_j.$$

Almost surely $g^\gamma \to e_J$. The resulting Euclidean-weighted prediction converges to $\widehat{f}_0$ for fixed finite λ. All predictions and targets lie in $[-1, 1]$, so dominated convergence controls their squared risks.

For each fixed n, d, γ, a row-wise feed-forward network can approximate $(Y_i - X_{ij})^2$ on the compact input domain. A masked attention layer with zero query/key logits computes their uniform average over labeled support tokens and broadcasts s to every token. Residual channels retain each token's covariates; a row-wise network then approximates the continuous map $(s, z) \mapsto \sum_j g_j^\gamma(s) z_j$. The same map is used for support and query. Only the first statistics computation reads support labels, and query tokens are excluded from that computation. Continuous functions on these finite-dimensional compact domains can be approximated by sufficiently wide feed-forward networks with biases and nonpolynomial activations (Leshno et al., 1993); for fixed λ, the final softmax-weighted prediction is continuous in their outputs. First choose γ, then choose network approximation errors sufficiently small, to make the total excess squared risk at most any specified $\zeta > 0$.

This establishes a Transformer implementation with fixed test-time parameters, support-only adaptation, and the native convex-label readout. Its feature-layer size and precision may depend on d, n, λ, ζ.

Interpretation of the separation. In the regime $1 \ll n \ll d$, the covariate-only lower bound stays close to $1/3$, whereas the contextual upper bound can approach zero as n grows and the softmax resolution increases. This is not a contradiction to consistency of k-NN when $n \to \infty$ at fixed d; the lower bound becomes uninformative once $n \ge d$. The construction also does not confer an advantage over a kernel method given the correct coordinate, or one that learns it from the same labeled support. It proves that *the support-weighted output form need not inherit the limitation of a fixed input-only geometry*. The particular support-residual selector is a sufficient example, not a proposed new production module.

A.8 A FINITE-ALPHABET COUNTERPART WITH A FEATURE-IDENTIFICATION MARGIN

The continuous construction uses exact noiseless coordinate identification. A binary-valued example supplies a finite statistical margin and makes context-size dependence explicit.

Proposition A.7 (Logarithmic-context separation). *Draw $X_1, \dots, X_n, X$ independently and uniformly from $\{-1, 1\}^d$, draw an independent uniform $J \in [d]$, and let $Y_i = X_{iJ}$, $Y = X_J$. Every covariate-only linear smoother has squared risk at least $1 - n/d$. A masked Transformer with sufficiently wide ReLU feed-forward layers and a tied scalar Euclidean readout with logits $-\lambda(u_x - u_i)^2$ can attain*

$$\mathbb{E}[(\widehat{f} - Y)^2] \le 4d\, 2^{-n} + 4n^2 e^{-8\lambda}. \tag{50}$$

Thus for a desired error $\delta > 0$, $n \ge \log_2(8d/\delta)$ and $\lambda \ge \frac{1}{8}\log(8n^2/\delta)$ suffice for the upper bound to be at most δ.

Proof. The lower bound uses Lemma A.6 with $\nu = 1$. Choose the smallest coordinate $\widehat{J}$ for which $X_{ij} = Y_i$ for all support rows. The true J always qualifies. Each incorrect coordinate qualifies with probability 2^{-n}, so $\Pr(\widehat{J} \ne J) \le (d-1)2^{-n}$. Independently of the selection rule, absence of a support with $X_{iJ} = X_J$ has probability 2^{-n}. Except on a union of probability at most $d2^{-n}$, the selected coordinate is correct and at least one support has the query's sign.

Use $\psi_{\mathcal{S}}(z) = z_{\widehat{J}}$. On this good event, correct-label scores are zero and wrong-label scores are -4λ. If $n_+ \ge 1$ supports have the correct label and n_- have the opposite label, wrong-label mass is

$$m_- = \frac{n_- e^{-4\lambda}}{n_+ + n_- e^{-4\lambda}} \le n e^{-4\lambda}.$$

The regression error is $2m_-$, giving squared error at most $4n^2 e^{-8\lambda}$. On the exceptional event the squared error is at most four. This proves Eq. (50).

Here an explicit finite-margin implementation is available. Uniform masked support attention computes $c_j = n^{-1}\sum_i Y_i X_{ij}$. Products on the finite binary input set are realizable by row-wise ReLU networks. Since $c_j \in \{-1, -1 + 2/n, \dots, 1\}$, the gate

$$b_j = \mathrm{ReLU}\Big(1 + \frac{n}{2}(c_j - 1)\Big)$$

is exactly the indicator that all support labels match coordinate j. A second layer forms $e_j = \mathrm{ReLU}(b_j - \sum_{l<j} b_l)$, which selects the first matching coordinate. A row-wise network computes $\sum_j e_j z_j$ and applies the same map to every support and query. All operations are support-permutation invariant, and query labels are never supplied.

On the good event, the teacher for target $Y \in \{-1, 1\}$ is uniform over correct-label supports both for categorical supervision and for the scalar Gaussian endpoint rule. The constructed head satisfies

$$\mathrm{KL}(q(Y)\|p) = \log\left(1 + \frac{n_-}{n_+} e^{-4\lambda}\right) \le n e^{-4\lambda}.$$

Thus the representation construction can also approximate the common teacher, rather than merely produce a suitable mean through arbitrary cancelling weights. These are two-valued targets; the preceding theorem is the continuous-target result.

A.9 FROM REPRESENTATION APPROXIMATION TO POINT-CE RISK

Proposition A.8 (Logit approximation transfers to prediction guarantees). *Fix a context and let $p^0 = \mathrm{softmax}(\ell^0)$ and $p^\theta = \mathrm{softmax}(\ell^\theta)$ on the same valid supports. If some scalar c satisfies $\max_i |\ell^\theta_i - \ell^0_i - c| \le \eta$, then*

$$\varepsilon(p^\theta) \le \varepsilon(p^0) + 2\eta. \tag{51}$$

Consequently all preceding conditional bounds apply with $\varepsilon(p^0) + 2\eta$ in place of $\varepsilon(p^\theta)$.

Proof. Softmax ignores a common shift, so take $c = 0$. The log-sum-exp function is one-Lipschitz in sup norm. Hence $|\log p^\theta_i - \log p^0_i| \le 2\eta$ for every valid atom. Therefore

$$\mathrm{KL}(\pi\|p^\theta) - \mathrm{KL}(\pi\|p^0) = \sum_i \pi_i \log \frac{p^0_i}{p^\theta_i} \le 2\eta.$$

Use Eq. (7).

For an explicit Euclidean consequence, suppose comparator embeddings have norm at most B and every support and query embedding is approximated to Euclidean error r. For logits $-\lambda\|u_x - u_i\|^2$ with the same λ, their sup-norm error is at most

$$\eta \le \lambda(8Br + 4r^2).$$

Indeed, each difference vector has norm at most $2B$, its perturbation has norm at most $2r$, and the difference of squared norms is bounded by $2(2B)(2r) + (2r)^2$. This quantifies how inaccurate representations can weaken the loss guarantee; it does not show that pretraining achieves a specified r.

A.10 Practical auxiliaries are not automatically mean-consistent

The population consistency theorem concerns POINT-CE. The implemented classification auxiliary shares its Bayes optimum. A squared-error regression auxiliary also shares the mean optimum when projection is unnecessary. A Smooth L1 auxiliary need not.

Explicit counterexample. Let support values be $\{0, 1\}$ and $P^*(Y = 1) = 0.2$. Both targets are endpoints, so the Gaussian teacher is one-hot for any positive width. Write $t = p(Y = 1)$. The POINT-CE optimum and conditional mean are $t^* = 0.2$. The conditional objective with the experimental coefficients is

$$J(t) = 0.8\mathrm{SL}_{0.1}(t) + 0.2\mathrm{SL}_{0.1}(t - 1) + 0.3[-0.8\log(1 - t) - 0.2\log t].$$

At $t = 0.2$, the CE derivative is zero but the Smooth L1 derivative is $0.8 - 0.2 = 0.6$. Hence $J'(0.2) \neq 0$: the composite need not recover the conditional mean. This rules out a distribution-free equivalence between the entire practical loss and full-distribution quantile prediction.

For deterministic in-hull targets the teacher barycenter equals the target and minimizes Smooth L1, so the optima agree. They also agree whenever the conditional mean minimizes the auxiliary, including symmetric conditional laws with appropriate integrability and no clipping bias. Smooth L1 is an empirical optimization choice, not a universal consistency theorem for the composite.

Finite-logit numerical supervision. For a normalized detached teacher, differentiation of log-softmax gives $\partial\mathcal{L}_{\mathrm{PCE}}/\partial\ell_i = p_i - q_i$. A floored-probability loss $-q_i \log \max(p_i, \epsilon)$ is constant in p_i below the floor and changes this derivative; adding ϵ also changes it. Stable evaluation uses log-softmax and excludes masked terms before multiplication. These numerical identities do not assert an empirical generalization gain. The Smooth L1 derivative bound concerns its scalar prediction argument, not the norm of all network-parameter gradients.

A.11 Shared training and finite-dimensional representability

For a task mixture with weights w_τ, the common-core objective is

$$\sum_\tau w_\tau \mathbb{E}_{\mathcal{C},Y|\tau}\left[-\sum_i q^{(\tau)}_{S,i}(Y)\log p_{\theta,i}(\mathcal{C})\right].$$

Every task updates the same scorer through a gradient of the form $p - q$. This enables shared learning but does not force finite-capacity optima to improve on all tasks. If a regression risk R_r has L-Lipschitz gradient, write $g_r = \nabla R_r$, $g_c = \nabla R_c$, and $\theta^+ = \theta - \eta(g_r + \omega g_c)$. The descent lemma gives

$$R_r(\theta^+) \le R_r(\theta) - \eta\|g_r\|^2 - \eta\omega\langle g_r, g_c\rangle + \frac{L\eta^2}{2}\|g_r + \omega g_c\|^2.$$

The gradient inner product describes the first-order interaction; curvature and stochastic effects remain. This is not a theorem for all clipped Muon updates or nonsmooth network states.

For fixed support embeddings, the Euclidean head's log odds are

$$\log\frac{p_i}{p_S} = \frac{u_x^\top(u_i - u_S)}{h^2T} - \frac{\|u_i\|^2 - \|u_S\|^2}{2h^2T}.$$

As a function of u_x these lie in an affine subspace of dimension at most the embedding width. A fixed-width head is therefore not an unrestricted parameterization of all support distributions when $S - 1$ exceeds that width. Contextual learning can make task-relevant targets representable, as in Theorem 4.3, but does not abolish approximation or optimization error. Our consistency result is over the simplex, our separation is constructive, and the cited training theorems retain their own assumptions. None asserts unconditional superiority over kernel methods, guaranteed positive cross-task transfer, or benchmark parity at a fixed pretraining budget.

B Implementation, target construction, and extensions

B.1 Numerically stable Gaussian target construction

Given scalar support labels a_i, retain only finite, valid labels and compute their extrema. First apply the same support-fitted normalization and $[-10, 10]$ clipping used by the model, then let $\bar{y} = \mathrm{clip}(y, a_{\min}, a_{\max})$ for the teacher only. The point-loss target remains the normalized query target; it is not additionally clipped to the support extrema.

For interior targets, use centered coordinates $d_i = a_i - \bar{y}$ and logits

$$b_i = -d_i^2/(2\sigma^2), \qquad q_i(\alpha) = \mathrm{softmax}_i(b_i + \alpha d_i).$$

Centering changes the logits by a common constant relative to Eq. (4), so the probabilities are unchanged. Solve $\sum_i q_i(\alpha) d_i = 0$ by first bracketing the root and then bisection. The monotone mean makes this a scalar root-finding problem. Exponentials must use log-sum-exp normalization; simply normalizing the unshifted Gaussian can underflow or introduce a biased mean. Root bracketing must fail explicitly rather than silently return an unconstrained teacher. The production implementation's tolerance and iteration cap should be released with the checkpoint. The mathematical reference implementation in the source package uses float64 and tests mean error; it is not claimed to be the production GPU code.

At a minimum/maximum label target, use a uniform distribution over the matching endpoint rows. If all valid support labels coincide, use the uniform distribution over that one value's rows. Exact duplicate values are handled as equal atoms; no tolerance-based merging of nearby but different regression labels is assumed. An episode with no valid support must be rejected by a stated data policy. An invalid query target contributes neither point nor distributional loss and must not contaminate normalization statistics.

B.2 Beyond scalar regression and single-label classification

For fully observed vector targets $v_i \in \mathbb{R}^k$, Eq. (3) can be defined on their convex hull. With an interior feasible target and a positive Gaussian reference, its dual uses a vector multiplier:

$$q_i \propto r_i \exp(\alpha^\top v_i), \qquad \sum_i q_i v_i = v.$$

The Hessian is the covariance matrix of the support atoms. It may be singular when the atoms lie in a lower-dimensional affine subspace; uniqueness of the probability solution does not require a unique ambient-space multiplier. A practical solver should work in the affine span. This is an extension of the mathematical interface, not an implemented multivariate performance claim.

Multilabel targets can be represented by binary vectors, but a single convex combination cannot represent every possible marginal vector unless it lies in the observed-label convex hull. Coordinate-wise heads or masks change the joint-distribution interpretation. For partially observed outcomes, each coordinate needs its own valid-support set or an explicit joint missingness model. An arbitrary sequence, structured object, survival outcome, or causal estimand is not automatically covered merely by writing a vector encoding. Each requires a meaningful representation, feasible teacher construction, and task-appropriate evaluation. These limits define the intended sense of a *universal prediction interface*, rather than an unrestricted claim of an ultimate loss for all tasks.

B.3 Nearest-neighbor interpretation

The head is a learned kernel regressor/classifier with task-conditioned embeddings. At fixed non-tied distances, sending temperature to zero concentrates on nearest support rows; increasing it toward

infinity yields uniform label averaging. These limits do not imply superiority over a tuned raw-feature kNN at a finite learned temperature. Include raw-feature and frozen-embedding kNN controls, selected using support-only validation, in the final empirical suite.

C Additional experimental results

C.1 Training and evaluation protocols

The joint models use the same dense encoder and synthetic graph-SCM task family. Table 5 lists the configuration for the shared-metric model. Separate-metric models replace its metric MLP with two task-specific copies; the encoder remains shared. The output-head comparisons instead replace both the prediction parameterization and its associated loss. All task-input adapters are kept explicit: sharing an output head does not remove the need to encode categorical and continuous support labels differently.

Table 5: Joint-model training configuration. Episode budgets refer to generated tables, not individual rows. The classification task multiplier is applied to the complete classification objective.

Component	Setting
Encoder	3 column, 3 row, and 12 dense ICL blocks; width 512
Metric MLP	$512 \to 1024 \to 512$, GELU, identity-function initialization
Retrieval	Tied Euclidean distance; learned positive bandwidth; $T = 2$
Classification loss	$L_{\text{NLL}} + 0.3L_{\text{POINT-CE}}$
Regression loss	$L_{\text{Smooth L1, 0.1}} + 0.3L_{\text{POINT-CE}}$
Scalar teacher	Gaussian reference, $\sigma = 0.05$, exact mean tilt
Optimizer	Muon; bfloat16 mixed precision; gradient clipping at 10
Learning rate	Peak 8×10^{-4}; 1% warmup; cosine decay to 10^{-7}
Task batches	64 classification episodes or 8 regression episodes
Smaller budget	64,000 classification + 8,000 regression episodes
Larger budget	320,000 classification + 40,000 regression episodes
Parameters	28,587,039 shared; 29,637,152 with separate metric heads

Dataset groups. The 57-dataset classification suite, the three regression development datasets, and the ten-dataset TabArena regression suite are different evaluation sets. They are not interchangeable with a complete TabArena benchmark run. The development datasets are Airfoil Self-Noise, Concrete Compressive Strength, and QSAR Fish Toxicity, each evaluated on 30 fixed splits. The broader suite has 30 splits each for Fiat 500 and Healthcare Insurance Expenses and nine splits for each remaining dataset. Results are first summarized within datasets; macro results give each dataset equal weight rather than weighting by its number of splits. Several design and serving choices were informed by these evaluations, so subsequent comparisons on the same datasets are development evidence rather than a fresh confirmation set.

We use 57 real OpenML classification datasets reconstructed from Supplementary Table 6 of the TabPFNv2 development suite used by Qu et al. (2026). For each dataset, we use two fixed stratified 80/20 splits, giving 114 dataset–split evaluations in total. We retain at most 10,000 rows per dataset and cap the training set at 2,048 rows. For OpenML datasets with multiple declared targets, we use the first declared binary classification target. Reported classification accuracy and log loss are unweighted averages over the 114 evaluations.

Prediction and preprocessing. Query labels are excluded from the encoder and from support-fitted preprocessing. Scalar regression targets are standardized using support statistics and clipped to $[-10, 10]$ for numerical stability. Separately, in the teacher construction, out-of-range scalar targets are clipped to the support-label interval; the point loss still uses the observed normalized target. The neural comparisons use a single inference view unless explicitly labeled otherwise. The regression development results are shown in Tables 1.

Classical baselines. The classical regressors operate on preprocessed support features, not POINT embeddings. Their hyperparameters are selected by three-fold support-only cross-validation. The

Table 6: Classification evaluation suite. We use the 57 real OpenML classification datasets recoverable from the TabPFNv2 development suite reported by Qu et al. (2026).

Dataset	OpenML ID	Dataset	OpenML ID
adaprior	1037	Midwest_Survey_nominal	42532
allbp	40707	mw1	1071
analcatdata_asbestos	459	national-longitudinal-survey-binary	43892
analcatdata_cyyoung8092	465	olindda_outliers	42793
ar3	1060	page-blocks	30
ar6	1064	penguins	42585
baseball	185	regime_alimentaire	42172
birds	40588	ringnorm	1496
blogger	1463	rl	44160
breast-tissue	1465	rmftsa_sleepdata	679
cjs	473	scene	312
clean1	40665	Smartphone-Based_Recognition	4153
cleveland-nominal	40711	squash-stored	340
compas-two-years	42192	synthetic_control	377
confidence	468	teachingAssistant	1115
corral	40669	thyroid-allbp	40474
delta_ailerons	803	thyroid-ann	40497
dermatology	35	Titanic	40704
eye_movements	44130	tokyo1	40705
GAMETES_Epistasis	40646	Tour-and-Travels-Churn	45545
grub-damage	338	user-knowledge	1508
heart-c	49	volcanoes-a1	1527
heart-h	51	vote	56
heart-statlog	53	waveform-5000	60
heloc	45026	wholesale-customers	1511
hill-valley	1479	xd6	40693
ionosphere	59	JapaneseVowels	375
jungle_chess_2pcs_endgame_panther_lion	40997	led24	40677
mc1	1056		

reported exact kernel-ridge implementation uses a fixed-seed subset of at most 750 support rows for tractability; it should not be read as an unrestricted full-support kernel-ridge result. CART and Random Forest select minimum leaf size from $\{1, 2, 4, 8\}$, and Random Forest uses 200 trees for its final fit.

C.2 Multi-output teacher construction

This experiment tests teacher construction separately from learned prediction. For each target dimension $m \in \{2, 4, 8\}$, we generate 30 episodes with 64 support vectors and query targets strictly inside their convex hull. The uniform-reference exponential tilt, simplex least squares, and nonnegative LASSO followed by a simplex refit all enforce nonnegative weights summing to one. LASSO selects an active set before the refit: an L_1 penalty on the normalized nonnegative simplex would otherwise be constant.

The exponential tilt is accurate and fast in this small interior-hull construction. The learned Linnerud and synthetic results in Figure 2a provide the separate model-level test. Those small models use the uniform-reference vector teacher; the experiments do not yet establish the benefit of a Gaussian vector reference or accurate recovery of a full joint target distribution.

C.3 Theory-aligned synthetic experiments

The hidden-coordinate experiment varies whether the model can use support labels to identify the coordinate that determines the target. Table 8 includes an explicit label-based feature selector as a positive control. The label-contextual model identifies the coordinate on all clean test episodes and has approximately the same error as the selector. This supports context-dependent geometry, not a universal advantage over label-adaptive kernel methods.

Table 7: Constructing vector teachers. Error measures reconstruction of the known training target, not generalization to a hidden query target. Time is per teacher solve; lower is better.

Teacher	m	Vector RMSE	Time (ms)
Exponential tilt	2	4.35×10^{-12}	1.50
	4	3.75×10^{-12}	0.28
	8	4.73×10^{-12}	0.36
Simplex least squares	2	3.41×10^{-8}	140.81
	4	9.99×10^{-8}	265.47
	8	1.10×10^{-7}	495.07
Nonnegative LASSO + refit	2	2.97×10^{-8}	9.09
	4	1.02×10^{-7}	34.57
	8	1.09×10^{-7}	163.48

Table 8: Hidden-coordinate prediction. Entries are reported mean query MSE on the synthetic evaluation. Lower is better.

Predictor	MSE ↓
Label-contextual Transformer + POINT	**0.0631**
Support-label feature selector + support-CV kernel	0.0637
Fixed RBF kernel ridge	0.4774
Target-blind contextual Transformer + POINT	0.8128
Noncontextual representation + POINT	0.8152
Raw-feature 4-NN	0.8461
RBF Nadaraya–Watson	1.0197

D SCOPE AND LIMITATIONS

POINT shares the learned support-probability head, but not every part of the prediction pipeline. Categorical and continuous targets still require different input representations. Our theoretical results concern conditional population risk with a student-independent teacher. They do not characterize optimization rates or behavior under arbitrary distribution shift, and fixed Gaussian smoothing introduces finite-support approximation error. The practical Smooth L1 auxiliary can also change the population optimum of the regression objective.

The experiments have corresponding limits. Broader benchmark evaluation, matched retrieval controls, and larger multi-output studies would be needed to establish how well the findings extend beyond the settings studied here.