CDF-FIRST: CONDITIONAL DENSITY ESTIMATION WITH AN EXPLICIT SMOOTHING BANDWIDTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Decisions that depend on a tail probability or a quantile need the conditional distribution, not a conditional mean. Estimators of it differ in where they obtain normalization: a closed-form family, a Jacobian determinant, a simplex over cells, or a partition function. We present CDF-First, which takes it from a boundary condition: each factor of an autoregressive CDF is a bounded monotone network pinned to 0 and 1 at the ends of a declared domain, and the density is read from an interval probability of half-width δ. At every parameter value, and with no smoothness assumption on the target, the model is an exactly normalized conditional density, and each factor equals the model's own convolved with a uniform kernel of half-width δ; the objective is proper for that smoothed conditional, not the unsmoothed one. The smoothing is therefore a stated parameter, and we test whether it is the one applied. On four synthetic targets with known ground truth it bounds the realized smoothing from above, within 5%, in 253 of the 258 fits for which the smoothing is identifiable, which realize a median of 84% of it; raising δ over two decades improves accuracy on three targets without measurably degrading the fourth. Accuracy is competitive rather than superior: lowest error on three of the four synthetic targets, and the lowest mean held-out negative log-likelihood on four of seven UCI benchmarks, tied on two and second on one.

1 INTRODUCTION

A predicted mean with an error bar determines neither a tail probability, a quantile, nor a risk functional, yet many learning systems act on exactly these: a risk-sensitive agent chooses actions from the lower quantiles of its return distribution (Dabney et al., 2018b), simulation-based inference reads credible regions from a conditional density over simulator parameters (McNamara et al., 2026; Brehmer & Cranmer, 2020; Cranmer et al., 2021), and a probabilistic forecaster is judged by the coverage of its intervals (Dai et al., 2026; Anh & Phong, 2026; Yalavarthi et al., 2025; Chu & Chen, 2026). Conditional density estimation targets the distribution $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ of a response $\mathbf{y} \in \mathbb{R}^{d_y}$ given a covariate $\mathbf{x} \in \mathbb{R}^{d_x}$, scored by proper scoring rules (Gneiting & Raftery, 2007). Neural estimators differ chiefly in how they enforce $\int \hat{p}(\mathbf{y} \mid \mathbf{x})\, \mathrm{d}\mathbf{y} = 1$: a closed-form output family (Bishop, 1994), the Jacobian determinant of an invertible map (Papamakarios et al., 2021; Durkan et al., 2019), a simplex over cells (Chen et al., 2022), or an estimated partition function. Each is well developed, and our starting point is not a deficiency in any of them.

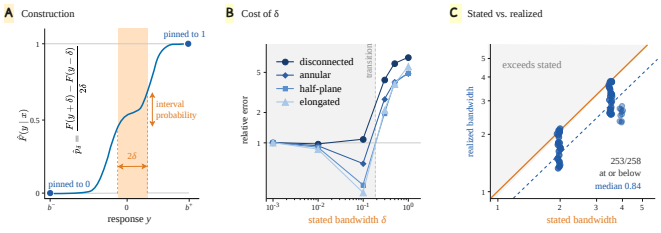


Figure 1: **(A)** The construction (schematic): a bounded monotone CDF, pinned to 0 and 1, with the density read from an interval of half-width δ. **(B)** Relative error against known ground truth as δ varies: below the transition (shaded) a larger stated bandwidth costs nothing; above it, error rises several-fold. **(C)** Realized against stated bandwidth for the fits in which smoothing is identifiable: 253 of 258 lie at or below the diagonal within 5%; dotted, the median ratio 0.84 (Section 4.3).

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.

Instead, they share a quantity they rarely expose. An estimator fitted to a finite sample returns a density at some effective resolution, set jointly by its noise, its regularization and the point at which training stopped. Classical kernel estimators make that resolution a single explicit bandwidth chosen by cross-validation (Zámečník et al., 2024; Tenreiro, 2022); for the neural estimators above it is a joint consequence of several settings that no reported number summarizes. A practitioner can therefore neither read off the scale at which such an estimator reports, nor fix it in advance. The consequence is not only practical: a sharper estimate places more mass near the observations and scores better on held-out likelihood without having learned more, so two estimators separated by a fraction of a nat cannot be told apart on fit rather than sharpness unless both report their resolution; the gap between them is a Kullback–Leibler divergence that depends on resolution alone (Appendix C). We do not claim this invalidates published comparisons, and Section 4.4 finds no evidence of it on the targets we study. The observation is classical; the quantity is nonetheless unavailable for the estimators above.

This paper asks two questions. *Can a neural conditional density estimator be built so that its smoothing is a stated parameter*; and if so, *does that value describe the smoothing the fitted model applies?* The second does not follow from the first: training acts on whatever density the estimator returns, so a model may compensate for a smoothing it is told to apply, and a parameter the fit silently undoes is a label, not a control. Neither is settled by reporting one hyperparameter, since several jointly fix the resolution, nor by driving the smoothing to zero, which an estimator that differences a CDF cannot do: the two values subtracted differ by an amount proportional to the scale, and that difference is lost to cancellation first.

We propose *CDF-First*, which obtains normalization from a boundary condition. Each factor of an autoregressive conditional CDF is a bounded monotone network, pinned to 0 and 1 by its own values at the ends of a declared domain, so no partition function, Jacobian determinant or cell simplex is involved. The density is read from the probability of an interval of half-width δ, divided by its width, rather than from a derivative of the learned CDF (Figure 1A), so a single quantity sets the scale at which the estimator reports, and it appears in the objective rather than emerging from training.

The construction works because both steps are exact. Pinning a strictly increasing function to 0 and 1 makes the model a normalized conditional density at every parameter value, not merely at a converged optimum (Proposition 3.2). Reading the density from an interval probability is not an approximation to a derivative: each conditional factor is exactly the model's own factor convolved with a uniform kernel of half-width δ, with no smoothness assumption on the target (Proposition 3.2). The smoothing is governed by δ and a learned noise scale the model also reports, and its cost is stated rather than hidden: the objective is proper for the smoothed conditional and not the unsmoothed one (Proposition 3.3). It also answers the second question in the direction one should expect: because training acts on the smoothed density, a fitted model may sharpen its own law to offset part of it, so the stated bandwidth should bound the realized one from above rather than equal it (Remark 3.2). We test that prediction against known ground truth. The realized bandwidth stays below the stated one (Figure 1C), and accuracy is unaffected until the stated bandwidth becomes the dominant source of error (Figure 1B).

Contributions.

- A conditional density estimator normalized by a boundary condition, exactly normalized at every parameter value (Section 3).
- Two properties proved without smoothness assumptions: the returned density is exactly the model's own convolved with a uniform kernel of half-width δ, and the objective is proper for that smoothed conditional and not the unsmoothed one, which we state as a limitation (Propositions 3.2–3.3, verified in Section 4.2).
- A measurement, against known ground truth, of whether the stated bandwidth is the applied one: it bounds the realized smoothing from above, within 5%, in 253 of 258 identifiable fits, which realize a median of 84% of it (Section 4.3).
- The same measurement on five baselines, none of whose resolutions — nor ours — is recoverable at their operating points (Section 4.4). Accuracy is competitive, not superior: lowest error on three of four synthetic targets and the lowest held-out negative log-likelihood on four of seven UCI benchmarks, tied on two and second on one, at one to two orders of magnitude more parameters than several baselines (Sections 4.5, 4.6).

2 Related Works

Conditional density estimators differ less in architecture than in where normalization comes from. Kernel and least-squares estimation needs no parametric family but degrades quickly in d_x+d_y (Hall et al., 2004; Sugiyama et al., 2010); uniquely, its resolution is an explicit cross-validated bandwidth, which we retain over a learned conditional map. Mixtures take it from a closed-form output family (Bishop, 1994), autoregressive variants keeping fixed conditionals (Larochelle & Murray, 2011; Germain et al., 2015), at the cost of a component budget; binned models from a simplex over cells (Oord et al., 2016; Li et al., 2021), hierarchical smoothing reducing spikiness (Chen et al., 2022), at the cost of a resolution fixed by cell width; flows from a Jacobian determinant (Rezende & Mohamed, 2015; Dinh et al., 2016; Papamakarios et al., 2017), where spline couplings (Durkan et al., 2019), smooth flows (Köhler et al., 2021), continuous-time forms (Papamakarios et al., 2021) and flow matching (Lipman et al., 2022; Kerrigan et al., 2024) differ in making that determinant cheap and relaxing bijectivity costs exact likelihoods (Cornish et al., 2020). Energy-based conditionals estimate the constant (Strauss & Oliva, 2021), arbitrary-conditioning flows trading the other way (Li et al., 2020); we assume it fixed. Our factorization is a Knothe–Rosenblatt rearrangement (Rosenblatt, 1952; Bogachev et al., 2005), also learned by monotone classes (Baptista et al., 2024) and invertible networks (Kruse et al., 2021), both baselines, which enforce monotonicity through an integrand where we use endpoints. Quantile (Meinshausen, 2006; Dabney et al., 2018a) and copula (Elidan, 2010) methods model other functionals, conformal prediction giving coverage without a density (Angelopoulos & Bates, 2023); ours is a proper scoring rule for a smoothed conditional (Gneiting & Raftery, 2007).

Nearest is Chilinski & Silva (2020), differentiating an unconditional univariate CDF: we condition and factorize by masking, bound by endpoint values (Proposition 3.2, at every parameter value), and difference rather than differentiate (Proposition 3.2 an identity, the step size a reported bandwidth). Quasi-Bayesian predictives (Ghalebikesabi et al., 2023; Huk et al., 2024) assume more. Ours commits to compact support. Across all of them resolution recurs in disguise — dequantization noise (Theis et al., 2016; Ho et al., 2019), tail bounds, cell width, weight decay, early stopping — each reported alone, none a number for the total, which we supply and test (Section 4.3).

3 The CDF-First Framework

3.1 Problem definition

Let $(\mathbf{X}, \mathbf{Y})$ be random elements with $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathbf{Y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, and let $\mathcal{D}_N = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ be i.i.d. observations; nothing is latent. The estimand is the conditional distribution $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$, through its density.

Assumption 1 (Absolute continuity). *For $P_{\mathbf{X}}$-almost every $\mathbf{x}$, $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{d_y}$.*

Assumption 1 licenses $p(\mathbf{y} \mid \mathbf{x}) := \mathrm{d}P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}/\mathrm{dLeb}_{d_y}$, and is the only assumption this paper makes about the data. It is violated by responses with atoms, including finitely rounded measurements. Log-densities are reported in nats with respect to Lebesgue measure on the *raw, unstandardized* target space; since $\log p$ is not invariant under reparameterization of $\mathbf{y}$, every method we compare against is re-evaluated in this convention (Section 4). Writing $\mathbf{y}_{<i} = (y_1, \ldots, y_{i-1})$,

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{d_y} p_i(y_i \mid \mathbf{x}, \mathbf{y}_{<i}), \qquad p_i = \partial F_i / \partial y_i, \tag{1}$$

where $F_i(y_i \mid \mathbf{x}, \mathbf{y}_{<i}) = \mathbb{P}(Y_i \le y_i \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_{<i} = \mathbf{y}_{<i})$. The map $y_i \mapsto F_i$ is the Knothe–Rosenblatt rearrangement (Rosenblatt, 1952): an identity available to every method of Section 3. What distinguishes them is which side of it carries the parameters, and from where the constraint $\int \hat{p}\,\mathrm{d}\mathbf{y} = 1$ is obtained.

3.2 Overview

CDF-First parametrizes F_i rather than p_i. A bounded monotone network is pinned to 0 and 1 by its own values at the two ends of a declared compact domain, which supplies normalization from a

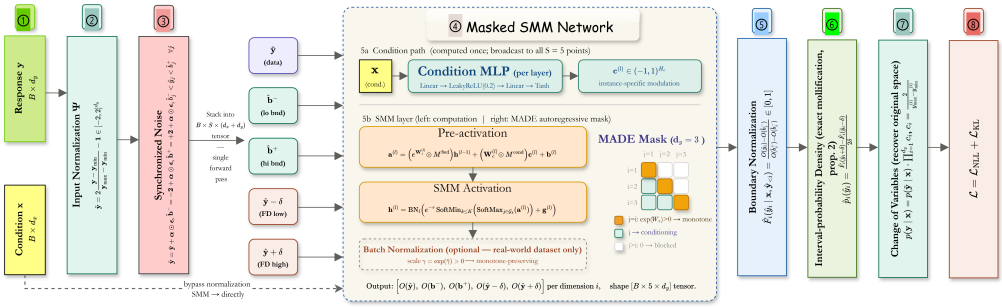


Figure 2: The CDF-First construction, stages ① to ⑧. The response is scaled by Ψ to $[-1,1]^{d_y}$ ② and, with both endpoints of the declared domain $\mathcal{B} = [-2,2]^{d_y}$, shifted by one shared noise draw equation 2 ③. The resulting $S = 5$ stencil points equation 3 are stacked into a single $B \times S \times (d_x + d_y)$ tensor and passed through the masked monotone network in one forward pass ④; the conditioning path is computed once, from the first slot, and broadcast (5a), and each layer applies equation 4–equation 5 under the mask that enforces the factorization equation 1 (5b). The two boundary slots convert the unbounded monotone output O_i into a CDF by equation 6 ⑤, the two offset slots convert it into the mollified density by equation 7 ⑥, the Jacobian of Ψ returns it to raw units ⑦, and the objective is equation 8 plus equation 9 ⑧.

boundary condition, and the density is recovered from the probability of an interval of half-width δ rather than from a derivative (Figure 2). Two properties follow from the construction alone. They hold at every parameter value, for every $\mathbf{x}$, and require nothing of the target beyond Assumption 1: the model is an exactly normalized conditional density (Proposition 3.2), and each conditional factor it returns is exactly the corresponding factor of its own law convolved with $\mathcal{U}[-\delta,\delta]$ (Proposition 3.2). The second is what makes δ a bandwidth rather than a discretizations error.

[Stated bandwidth] With $\kappa_{\delta,\alpha_i} = \mathcal{U}[-\delta,\delta] * \mathcal{N}(0,\alpha_i^2)$, the estimand of the fitted model is $p_i^{\delta,\alpha} = p_i * \kappa_{\delta,\alpha_i}$, whose standard deviation $(\delta^2/3 + \alpha_i^2)^{1/2}$ we call the stated bandwidth. Both δ and α_i are reported by the model.

Domain and scaling. With $\mathbf{y}_{\min}, \mathbf{y}_{\max}$ taken from the training split alone, $\Psi(\mathbf{y}) = 2(\mathbf{y} - \mathbf{y}_{\min}) \oslash (\mathbf{y}_{\max} - \mathbf{y}_{\min}) - \mathbf{1}$ maps the training range to $[-1,1]^{d_y}$, and the model declares the domain $\mathcal{B} = [-2,2]^{d_y}$ in these coordinates: the training range extended by half its width on each side. All values of δ and $\boldsymbol{\alpha}$ are quoted in these coordinates.

Stochastic smoothing and the stencil. With $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ drawn once per batch and $\boldsymbol{\alpha} \in \mathbb{R}_+^{d_y}$ learnable, a single draw shifts the response and both endpoints by the same amount,

$$\hat{\mathbf{y}} = \Psi(\mathbf{y}) + \boldsymbol{\alpha} \odot \boldsymbol{\epsilon}, \qquad \hat{\mathbf{b}}^{\pm} = \pm 2 \cdot \mathbf{1} + \boldsymbol{\alpha} \odot \boldsymbol{\epsilon}, \tag{2}$$

which leaves the endpoint gap unchanged; at inference $\boldsymbol{\epsilon} = \mathbf{0}$. The network is evaluated at $S = 5$ points, stacked along a slot axis,

$$\big[\,\hat{\mathbf{y}},\ \hat{\mathbf{b}}^{-},\ \hat{\mathbf{b}}^{+},\ \hat{\mathbf{y}} - \delta,\ \hat{\mathbf{y}} + \delta\,\big]. \tag{3}$$

The conditioning path — the covariate together with $\mathbf{y}_{<i}$ — is computed once, from slot 0, and broadcast, while only y_i varies across slots. The conditioning arguments are therefore held exactly fixed, and the two evaluations in equation 7 are two points of one and the same conditional.

Masked monotone network. With $\mathbf{h}^{(0)} = \hat{\mathbf{y}}$ and layer-wise covariate embeddings $\mathbf{c}^{(l)}$, each of the L layers computes

$$\mathbf{a}^{(l)} = \big(e^{\mathbf{W}_z^{(l)}} \odot M^{\mathrm{fwd}}\big)\mathbf{h}^{(l-1)} + \big(\mathbf{W}_c^{(l)} \odot M^{\mathrm{cond}}\big)\mathbf{c}^{(l)} + \mathbf{b}^{(l)}, \tag{4}$$

$$\mathbf{h}^{(l)} = \mathrm{BN}_l\Big(e^{-\tau}\,\mathrm{SoftMin}_{k\le K}\big(\mathrm{SoftMax}_{j\in\mathcal{G}_k}(\mathbf{a}^{(l)})\big) + \mathbf{g}^{(l)}\Big), \tag{5}$$

where the $H = KG$ pre-activations are partitioned into K groups of size G, $\mathrm{SoftMax}_j(a) = \log\sum_j e^{a_j}$ and $\mathrm{SoftMin}_k(b) = -\log\sum_k e^{-b_k}$ are the smooth extrema of Igel (2023), and τ is a

fixed temperature. The masks $M^{\text{fwd}}_{rc} = \mathbf{1}[\pi(r) \geq \pi(c)]$ and $M^{\text{cond}}_{rc} = \mathbf{1}[\pi(r) > \pi(c)]$, with π assigning a response index to every row and column, enforce $\partial F_i / \partial \hat{y}_j = 0$ for $j > i$, which is what makes equation 1 hold rather than merely be asserted. Three choices make O_i strictly increasing in $\hat{y}_i$: positive response weights $e^{\mathbf{W}_z}$, an additive rather than gated covariate, and a positive batch-norm scale $e^{\hat{\mathbf{s}}}$ with detached statistics shared across slots. Appendix E states what each rules out.

Normalization by boundary condition. Let O_i denote the i-th output of the last layer, monotone in $\hat{y}_i$ but unbounded. Using the two boundary slots of equation 3,

$$\hat{F}_i(\hat{y}_i) = \frac{O_i(\hat{y}_i) - O_i(\hat{b}_i^-)}{O_i(\hat{b}_i^+) - O_i(\hat{b}_i^-)}, \tag{6}$$

with $\hat{F}_i$ set to 0 below $\hat{b}_i^-$ and 1 above $\hat{b}_i^+$.

[Exact normalization] Under equation 2–equation 6, for every parameter value, every $\mathbf{x}$ and every i, $\hat{y}_i \mapsto \hat{F}_i(\hat{y}_i)$ is strictly increasing on $[\hat{b}_i^-, \hat{b}_i^+]$ with $\hat{F}_i(\hat{b}_i^-) = 0$ and $\hat{F}_i(\hat{b}_i^+) = 1$. Hence $\hat{p}_i \geq 0$ with $\int \hat{p}_i \, \mathrm{d}y_i = 1$, and the induced $\hat{p}$ requires no estimated normalizing constant, no Jacobian determinant and no bin simplex.

Density recovery. Using the two offset slots,

$$\hat{p}_i^{\delta}(\hat{y}_i) = \frac{\hat{F}_i(\hat{y}_i + \delta) - \hat{F}_i(\hat{y}_i - \delta)}{2\delta}. \tag{7}$$

[Exact mollification] Let $\hat{P}_i$ be the law on $[\hat{b}_i^-, \hat{b}_i^+]$ with CDF $\hat{F}_i$ and $U \sim \mathcal{U}[-1, 1]$ independent of it. For every $\delta > 0$, $\hat{p}_i^{\delta}$ is exactly the density of $\hat{Y}_i + \delta U$; equivalently $\hat{p}_i^{\delta} = \hat{p}_i * \mathcal{U}[-\delta, \delta]$, with no Taylor remainder and no smoothness requirement on $\hat{F}_i$.

Proofs are in Appendix D; neither proposition places any condition on the target, and we verify both numerically rather than assume them (Section 4.2). The training signal is therefore a probability in $[0, 1]$ rather than an unbounded point density. Forward-mode automatic differentiation is a legitimate alternative, computing a *different* object, the point-wise derivative of the learned CDF, at $O(d_y)$ passes against the stencil's $O(1)$; we treat it as an ablation, not a claimed saving.

[Scope for $d_y > 1$, and an upper bound] Each factor is exactly mollified, and their product is a normalized density, but it is not the joint law convolved with a kernel on $\mathbb{R}^{d_y}$, since $\mathbf{y}_{<i}$ is not perturbed; statements about mollification are per factor. Moreover, Proposition 3.2 relates $\hat{p}_i$ to $\hat{p}_i^{\delta}$, not $\hat{p}_i^{\delta}$ to p_i: because equation 8 acts on the mollified density, a fitted model may sharpen $\hat{F}_i$ to offset part of the smoothing, so the stated bandwidth should bound the realized one from above rather than equal it. We measure the realized bandwidth against known ground truth in Section 4.3.

[No consistency claim] We do not claim consistency: that would require F_i^* to lie in the closure of the class equation 4–equation 6, control of the empirical process over it, and that the empirical minimizer be found; we assert none of the three. The parameterization is also not identifiable, though the induced law is identified by $\hat{F}_i$ point-wise.

3.3 Training details

Objective. In the units of Section 3.1, with $c_i = 2/(y^{(i)}_{\max} - y^{(i)}_{\min})$ the diagonal Jacobian of Ψ,

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{d_y} \Big[\log \hat{p}_i^{\delta}\big(\hat{y}_i^{(n)} \mid \mathbf{x}^{(n)}, \hat{\mathbf{y}}_{<i}^{(n)}\big) + \log c_i \Big]. \tag{8}$$

Since $\hat{y}_i \pm \delta$ are constants with respect to the parameters, back-propagation requires only first-order reverse passes, and gradients are unbiased for the noise-averaged objective because $\hat{F}_i \in [0, 1]$ is uniformly bounded.

[Properness] The population objective is minimized uniquely, up to Lebesgue-null sets, at $\hat{p}^{\delta} = p^{\delta,\alpha}$. It is therefore a strictly proper scoring rule for the mollified conditional of Definition 3.2, and it is improper for p itself. The proof uses Gibbs' inequality and requires no smoothness of the target.

We state the second half as a limitation: CDF-First estimates a mollified conditional at a stated scale, not p. We deliberately give no bound relating $p^{\delta,\alpha}$ to p: one is available under smoothness of $p(\cdot \mid \mathbf{x})$, but every target we study has jump discontinuities, so it would apply to no experiment here. The distance is instead measured against known ground truth in Section 4.3.

The noise scale must not collapse, since $\alpha_i \to 0$ makes equation 8 unbounded above on a finite sample, so we add

$$\mathcal{L}_{\mathrm{KL}} = \frac{\lambda}{d_y} \sum_{i=1}^{d_y} \tfrac{1}{2}\left(\alpha_i^2 - 1 - \log \alpha_i^2\right), \tag{9}$$

minimized at $\alpha_i = 1$; a larger λ therefore pulls $\boldsymbol{\alpha}$ *up* toward unity, and the value it settles at balances equation 9 against equation 8, which favors small α_i. Because equation 9 depends on $\boldsymbol{\alpha}$ alone it shifts the stated bandwidth but not the conditional shape at fixed $\boldsymbol{\alpha}$; λ is ablated in Section 4.

Optimization, data and hardware. The objective is $\mathcal{L}_{\mathrm{NLL}} + \mathcal{L}_{\mathrm{KL}}$, optimized with Adam (Kingma & Welling, 2019), with checkpoints selected by held-out negative log-likelihood on a validation split disjoint from training and test. Synthetic targets use $N = 2{,}000$ draws; the UCI benchmarks use a 10-fold sliding-window partition with chunks fixed across folds, covariates standardized on the training split alone and $\mathbf{y}_{\min}, \mathbf{y}_{\max}$ taken from training only. The full protocol and hardware are in Appendix E.1. Algorithm 1 in Appendix B states the procedure step by step.

4 Experiments

Section 3 leaves four obligations, and this section is organized around them. Do Propositions 3.2 and 3.2 hold in the implementation, or only on paper (Section 4.2)? Is the stated bandwidth the one a fitted model applies, which Remark 3.2 predicts it need not be (Section 4.3)? Can the same question be put to estimators that expose no bandwidth at all (Section 4.4)? And does making the smoothing explicit cost accuracy (Sections 4.5 to 4.7)?

4.1 Setup

Synthetic tasks. We design four 2D tasks with $x \sim \mathcal{U}(-1, 1)$ and known $p(\mathbf{y} \mid x)$: *Disconnected*, *annular*, *half-plane*, and *elongated*. Each task uses 2,000 training samples; error is measured against the ground-truth density on a 100^2 grid over $[-8, 8]^2$ at $x \in \{-0.75, -0.25, 0.25, 0.75\}$, as the grid sum of squared deviations (SSE). Full specifications are in Appendix E.2.

Real-world tasks. Seven UCI datasets (Table 6, Appendix E) spanning $d_y \in \{1, 2, 3, 4\}$. Input/output splits follow the standard partitioning (Chen et al., 2022) (detailed in Appendix E.2). Performance: test NLL over ten 70/30 splits (mean±std, lower is better). Every hyperparameter for both settings is listed in Table 7; no per-dataset tuning was performed.

4.2 Propositions 3.2 and 3.2, verified

Both propositions concern the construction, not the data, so we check them on six fitted checkpoints of the disconnected target (three seeds at $\delta = 10^{-2}$, one at each of $10^{-3}, 10^{-1}, 3\times10^{-1}$), at two conditioning values in both precisions: 24 measurements (Figure 3). In every one, $|\hat{F}_i(\hat{b}_i^-)|$, $|\hat{F}_i(\hat{b}_i^+) - 1|$ and $\max_{j>i}|\partial\hat{F}_i/\partial\hat{y}_j|$ are identically zero, so the boundary condition and the triangular structure hold exactly.

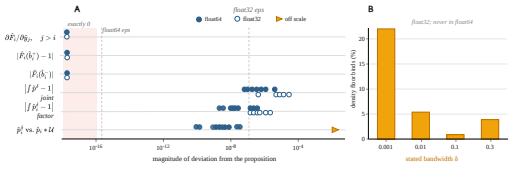


Figure 3: Deviation from each identity; one marker per measurement, filled for double and open for single precision. Quantities that are zero appear in the left lane, and dashed lines mark the two machine epsilons. Single-precision mollification values lie beyond the axis (triangle) and are not interpretable.

In double precision $\hat{F}_i$ is strictly increasing (smallest increment 5.3×10^{-10}), the density integrates to one within 4.4×10^{-7} per factor and 5.0×10^{-6} jointly, and matches $\hat{p}_i * \mathcal{U}[-\delta, \delta]$, obtained by quadrature from the auto-differentiated

$\hat{F}_i$, to relative error $\leq 4.5 \times 10^{-8}$. The residual falls with the bandwidth, to 3.5×10^{-10} at $\delta = 3 \times 10^{-1}$: the cancellation of Section E.1, since equation 7 subtracts values differing by order $2\delta\hat{p}_i$.

Single precision is the binding constraint. Increments of order 10^{-9} fall below the resolution of a quantity in $[0, 1]$, so $\hat{F}_i$ is non-decreasing rather than strictly increasing, and the density floor binds on 22% of grid evaluations at $\delta = 10^{-3}$ and 0.6% at $\delta = 10^{-1}$ — never in double precision — returning the floor rather than equation 7. The mollification check, being relative to an interval probability lost to the same cancellation, carries no information there and is not reported. Almost all affected points lie where the target places no mass; the denominator floor never binds.

4.3 Is the stated bandwidth the one the model applies?

Remark 3.2 predicts the stated bandwidth bounds the realized smoothing from above rather than equaling it. We sweep δ over $\{10^{-3}, \ldots, 1\}$ on the four targets with five seeds, in two arms — $\boldsymbol{\alpha}$ learned, and frozen at 10^{-4} so δ is the only smoothing term — recover from each fit the kernel width that best reproduces it from the known truth, and compare against the width the same measurement returns for CDF-First's own kernel applied to the truth. The claim is refuted if the realized width exceeds the stated one.

Below $\delta = 10^{-1}$ no fit is identified: smoothing explains only 0.35 to 0.58 of the squared error, so no width can be read. Above it, 253 of 258 identified fits lie at or below the stated value within 5% (Figure 4), the median realized fraction being 0.84 frozen and 0.83 learned — agreement to two decimals between arms that divide the smoothing differently. Accuracy locates the same boundary: below it a larger δ is free and improves three of four targets by 25–42%; above it, error rises 3.1 to 4.9-fold.

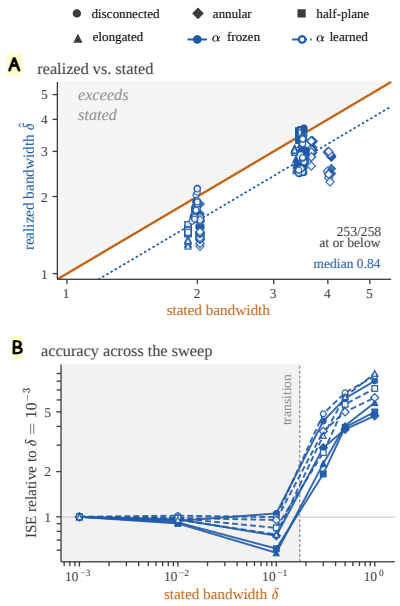


Figure 4: One marker per identified fit ($R \geq 0.8$; none occurs below $\delta = 10^{-1}$, and $\delta = 1$ is excluded because the fitted width lies beyond the searched range). Shading marks the region the prediction forbids; dotted, the median ratio. Filled: α frozen; open: α learned.

4.4 Can a baseline's bandwidth be read from its output?

Section 1 observed that a neural estimator's resolution is not reported; can it be recovered anyway? Applied unchanged to five baselines spanning the routes of Section 2 at default configurations, the measurement of Section 4.3 identifies none of 400 fits: largest value 0.80, per-method medians 0.35 to 0.49 (Figure 5A). Every fitted density is wrong in shape rather than blurred, so no bandwidth can be read from it; the same holds for CDF-First at its operating point (median 0.52). On these targets we therefore find no evidence that accuracy differences between these estimators are resolution differences. What distinguishes the stated bandwidth is not a different resolution but a movable one: raising δ carries CDF-First across the threshold, where the smoothing becomes measurable and bounded.

A second pattern, reported without a mechanism (Figure 5B): the two Jacobian-based baselines incur 2.1 and 2.9 times CDF-First's error on the two void-containing targets, against 0.9 and 1.2 on the two without; the mixture's contrast is smaller, and the two without a Jacobian show almost none. Comparing a method with itself across target types is less exposed to the parameter gap than an absolute comparison, but this rests on five methods at unmatched capacity.

4.5 Accuracy on the synthetic targets

Making the smoothing explicit is worth little if it costs accuracy, so we compare squared error against the known truth on the four targets (Table 1). CDF-First attains the lowest error on three of

Table 1: Squared error against the known truth, mean over four conditioning values; lower is better).

	void		no void	
method	disc.	ann.	half	elong.
MDN	2.13	1.50	2.32	1.04
DDN	1.27	1.23	1.93	0.91
MAF	3.37	3.62	2.52	1.25
RNF	3.38	1.69	3.40	1.18
NSF	3.47	2.53	3.41	1.18
HINT	1.74	1.71	1.74	**0.58**
ATM	2.97	3.56	3.22	2.23
NL-CDF	1.80	1.52	2.61	1.27
CDF-First	**0.89**	**0.71**	**1.29**	0.75

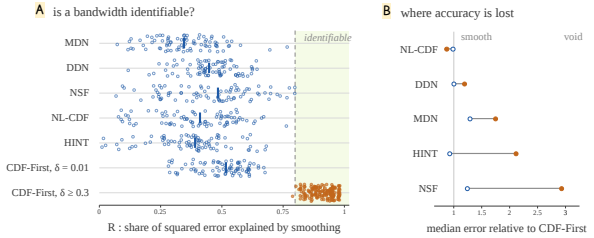


Figure 5: **(A)** One marker per fit, 400 baseline and 240 CDF-First; ticks are medians. R is the share of squared error smoothing explains; a bandwidth is readable only in the shaded region, which no baseline reaches, nor CDF-First at its operating point. **(B)** Median error relative to CDF-First on the two void-containing targets (filled) against the two without (open).

them, by 34 to 74% over the next method, and is second on the elongated Gaussian, 28% behind the hierarchical invertible network. Read by normalization route, the closest competitors on the two targets whose support contains a void are the two estimators that normalize without a Jacobian determinant, while the four that use one incur 2.1 to 4.5 times our error there against 1.1 to 2.1 times on the two targets without. Appendix F.2 shows the fitted densities against the truth for every method and target.

4.6 UCI BENCHMARKS

The stated bandwidth is the contribution; this subsection asks what it costs. On seven UCI benchmarks with d_y from one to four, CDF-First has the lowest mean held-out negative log-likelihood on four, ties on two, and trails the adaptive transport map on the seventh (Table 2), so making the smoothing explicit does not cost predictive accuracy here. We claim no more: the comparison is at default configurations, with parameter counts differing by up to two orders of magnitude, and the shortfall of 0.08 nats is smaller than the fold-to-fold dispersion of either method. The ordering is

Table 2: Held-out negative log-likelihood, in nats on the raw target ($\times 10^{-1}$; lower is better; mean and standard deviation over ten random seeds). **Bold**: lowest per dataset.

method	Fish	Concrete	Energy	Parkinsons	Temp.	Air	Skillcraft
MDN	0.92 0.03	0.76 0.09	−0.99 0.15	1.26 0.05	1.08 0.03	1.26 0.04	2.27 1.14
DDN	0.92 0.04	0.59 0.03	−1.66 0.18	0.36 0.04	0.74 0.04	0.93 0.05	1.66 0.07
MAF	1.10 0.09	1.10 0.09	−0.13 0.29	1.54 0.08	1.25 0.04	1.14 0.08	4.06 0.28
NSF	1.57 0.19	1.32 0.29	9.23 1.63	1.72 0.13	1.79 0.16	1.50 0.17	6.72 0.64
RNF	1.24 0.08	1.24 0.09	1.89 0.20	2.30 0.07	1.84 0.10	1.70 0.11	4.49 0.19
RealNVP	0.95 0.14	0.37 0.14	−1.18 0.17	0.89 0.04	**0.44** 0.04	0.76 0.04	1.99 0.59
HINT	0.88 0.10	0.42 0.11	−1.05 0.73	0.98 0.04	0.56 0.05	**0.74** 0.04	4.26 0.95
ATM	1.18 0.33	0.36 0.07	−1.87 0.28	1.17 0.04	0.87 0.035	1.09 0.03	**1.48** 0.10
NL-CDF	0.87 0.04	0.44 0.10	−2.54 0.17	0.57 0.03	0.77 0.06	1.44 0.04	3.41 0.08
CDF-First	**0.81** 0.04	**0.24** 0.10	**-3.1** 0.22	**0.32** 0.11	**0.44** 0.10	**0.74** 0.03	1.56 0.36

Table 3: Bandwidth sweep, elongated target.

δ	SSE
10^{-1}	0.0745
10^{-2}	0.0756
10^{-3}	0.0799
3×10^{-5}	0.0771
10^{-5}	0.0798
5×10^{-6}	0.0886
10^{-6}	0.1130
5×10^{-7}	0.1340

nonetheless worth recording: CDF-First is lowest or tied on all five datasets with $d_y \leq 2$, tied on the one with $d_y = 3$, and second on the one with $d_y = 4$. Two features of the construction would produce it, and our experiments do not separate them. By Remark 3.2 the returned joint is a product of per-coordinate mollifications, so smoothing accumulates across coordinates while a transport map applies none; and the masked blocks of equation 4 grow as d_y^2, so capacity per conditional falls. The transport map is itself triangular, so the factorization alone does not distinguish them. A δ sweep would separate the two: if the first dominates, a smaller bandwidth should close the gap. One dataset at $d_y = 4$ cannot settle it, and we claim nothing beyond. Appendix F.3 reports calibration on the two univariate datasets, where CDF-First is lowest on one and third of five on the other.

4.7 Ablation

Each component of Section 3 should earn its place. On the annular target we replace one at a time and report squared error against the known truth, averaged over three conditioning values (Table 4). Removing the noise injection of equation 2 raises error by 21%, so the stochastic smoothing contributes accuracy and not only a bandwidth term. Replacing the smooth min-max of equation 5 by its hard counterpart raises it four-fold, and by a monotone multilayer perceptron 4.4-fold; the hard operators route gradient through the active branch alone, and neither variant recovers the annulus. These are single runs on one target, so we report the ordering and not the margins.

Table 4: Ablation on the annular target: squared error against the known truth, mean over three conditioning values. Single run per variant.

variant	component removed	SSE	× full
CDF-First	—	**0.129**	1.0
no noise	equation 2	0.157	1.2
hard min-max	equation 5	0.517	4.0
monotone MLP	equation 5	0.566	4.4

4.8 Choice of the stated bandwidth

Accuracy is insensitive to δ over four decades: squared error varies by 7% between 10^{-1} and 10^{-5} (Table 3), and degrades only below that, reaching 1.8 times the best value at $\delta = 5 \times 10^{-7}$. The onset is where the interval probability $2\delta\hat{p}_i$ approaches the single-precision resolution of $\hat{F}_i \in [0, 1]$: at $\delta = 5 \times 10^{-6}$ it is 3×10^{-7} against a machine epsilon of 1.2×10^{-7}. This is the cancellation of Section E.1, and the same effect that binds the density floor in Section 4.2. We use $\delta = 10^{-2}$ throughout, two decades inside the flat region and four above the onset.

4.9 Sample efficiency

Averaged over the four synthetic targets, CDF-First has the lowest error at every training-set size, by 12 to 31% over the best baseline (Figure 6). The margin is widest where data is scarcest: at $N = 200$ it is lowest on all four targets individually, against two of four at $N = 2000$, and the best baseline is not the same method at every size — the binned model at $N \in \{200, 500, 2000\}$ and the conditional CDF estimator at $N = 1000$ — so the comparison is against an envelope rather than against one competitor. Its error grows by a factor of 1.95 as N falls from 2000 to 200, within the range of 1.5 to 3.7 spanned by the baselines, so the advantage comes from the level rather than from a flatter curve. These are single runs at default configurations and are not tested. Appendix F.1 gives the per-target breakdown and the degradation factors.

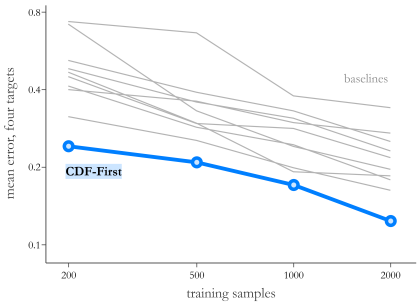


Figure 6: Mean error, four targets; grey, the nine baselines.

5 Conclusion

Normalization can come from a boundary condition rather than from a family, a determinant or a partition function, making an estimator's smoothing a quantity it reports. We present CDF-First, which takes it from a boundary condition. A monotone network pinned to the ends of a declared domain is exactly normalized at every parameter value, and reading its density from an interval probability rather than a derivative returns an exact mollification of its own law at a stated half-width. Three costs are explicit: the objective is proper for the smoothed conditional, not the target; the construction assumes compact support; and for $d_y > 1$ the mollification is per coordinate, not joint. Against known ground truth, the stated bandwidth bounded the realized smoothing in 253 of 258 identifiable fits, at a median of 84%, with accuracy competitive rather than superior in single runs. We discuss the limitations of this work in Appendix G.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing of the manuscript: grammar, phrasing, and copy-editing of text written by the authors. We have not used generative AI tools for research ideation, method design, experimental design, implementation, data analysis, or the production of any result reported in this paper, and the remaining disclosure categories do not apply to this work. We have reviewed all AI-assisted text: every sentence was checked by an author against our own results, and any phrasing that altered, weakened, or overstated a claim was corrected or removed.

ETHICS STATEMENT

This work studies conditional density estimation on synthetic targets and on seven public UCI regression benchmarks. No human subjects were involved, no new dataset is released, and the benchmarks we use are long-standing public resources that contain no personally identifying information; we use them under their existing terms and cite their sources. The method is a general estimator with no application-specific deployment in this paper, and we see no foreseeable harm arising from it beyond those common to density estimation as a whole. We note one property with practical consequences, which Section 3 and Appendix G state in full: the estimator assigns zero density outside a declared domain, so a response falling outside the range seen in training receives arbitrarily poor likelihood rather than a degraded one. A practitioner relying on the model for a tail probability or a coverage guarantee should check that condition before acting on its output. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

The construction is specified in Section 3, with each equation labelled and referenced from the training procedure of Appendix B; the assumptions are stated before use and the propositions are proved in full in Appendix D, which also records the numerical safeguards that the proofs require. The synthetic targets are given as sampling equations and ground-truth densities in Appendix E.2, so they can be regenerated exactly; the UCI benchmarks are public, and the column assignment, the scaling of covariates and responses, and the ten-fold sliding-window partition are described in the same appendix. Every hyperparameter, for both settings, is listed in Appendix E.3, and the hardware and software environment in Appendix E.3. Propositions 3.2 and 3.2 are verified numerically on fitted checkpoints rather than assumed, by a script included in the supplementary material whose diagnostics are reported in Section 4.2.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Nguyen Thi Ngoc Anh and Nguyen The Phong. Probabilistic electric load forecasting via adaptive online learning and multi-objective optimization: A na-tp forecasting approach. *Applied Soft Computing*, pp. 114858, 2026.

Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, 24(6):2063–2108, 2024.

Christopher M. Bishop. Mixture density networks. WorkingPaper 4288, Aston University, 1994.

Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335, 2005.

Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in neural information processing systems*, 33:442–453, 2020.

Bing Chen, Mazharul Islam, Jisuo Gao, and Lin Wang. Deconvolutional density network: Modeling free-form conditional distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6183–6192, Jun. 2022. doi: 10.1609/aaai.v36i6.20567. URL `https://ojs.aaai.org/index.php/AAAI/article/view/20567`.

Pawel Chilinski and Ricardo Silva. Neural likelihoods via cumulative distribution functions. In *Conference on uncertainty in artificial intelligence*, pp. 420–429. PMLR, 2020.

Betty XY Chu and Cathy WS Chen. Robust prediction intervals for time series forecasting: A bootstrap and bayesian approach. *Journal of Forecasting*, 2026.

Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pp. 2133–2143. PMLR, 2020.

Miles Cranmer, Daniel Tamayo, Hanno Rein, Peter Battaglia, Samuel Hadden, Philip J Armitage, Shirley Ho, and David N Spergel. A bayesian neural network predicts the dissolution of compact planetary systems. *Proceedings of the National Academy of Sciences*, 118(40):e2026053118, 2021.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018b.

Xilin Dai, Zhijian Xu, Wanxu Caii, and Qiang Xu. From samples to scenarios: A new paradigm for probabilistic forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 140721–140750, 2026.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *CoRR*, abs/1605.08803, 2016.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Gal Elidan. Copula bayesian networks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 881–889. JMLR.org, 2015.

Sahra Ghalebikesabi, Chris C Holmes, Edwin Fong, and Brieuc Lehmann. Quasi-bayesian non-parametric density estimation via autoregressive predictive updates. In *Uncertainty in Artificial Intelligence*, pp. 658–668. PMLR, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730, 2019.

David Huk, Yuanhe Zhang, Mark Steel, and Ritabrata Dutta. Quasi-bayes meets vines. *Advances in Neural Information Processing Systems*, 37:40359–40392, 2024.

C. Igel. Smooth min-max monotonic networks. *ICML*, 2023.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. *Advances in Neural Information Processing Systems*, 37:93602–93642, 2024.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, November 2019. ISSN 1935-8237. doi: 10.1561/2200000056.

Jakob Kruse, Gianluca Detommaso, Ullrich Köthe, and Robert Scheichl. HINT: Hierarchical invertible neural transport for density estimation and Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8191–8199, 2021.

Jonas Köhler, Andreas Krämer, and Frank Noé. Smooth normalizing flows, 2021.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 29–37, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Rui Li, Brian J. Reich, and Howard D. Bondell. Deep distribution regression. *Computational Statistics & Data Analysis*, 159:107203, 2021. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2021.107203.

Yang Li, Shoaib Akbar, and Junier Oliva. Acflow: Flow models for arbitrary conditional likelihoods. In *International Conference on Machine Learning*, pp. 5831–5841. PMLR, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Declan McNamara, Yicun Duan, and Jeffrey Regier. Neural posterior estimation with latent basis expansions. In *International Conference on Learning Representations*, volume 2026, pp. 202–244, 2026.

Nicolai Meinshausen. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, December 2006. ISSN 1532-4435.

Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4797–4805, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1530–1538. JMLR.org, 2015.

Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

Ryan Strauss and Junier B Oliva. Arbitrary conditional distributions with energy. *Advances in Neural Information Processing Systems*, 34:752–763, 2021.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pp. 781–788, 2010.

Carlos Tenreiro. Kernel density estimation for circular data: a fourier series-based plug-in approach for bandwidth selection. *Journal of Nonparametric Statistics*, 34(2):377–406, 2022.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016.

Vijaya Krishna Yalavarthi, Randolf Scholz, Stefan Born, and Lars Schmidt-Thieme. Probabilistic forecasting of irregularly sampled time series with missing values via conditional normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21877–21885, 2025.

Stanislav Zámečník, Ivana Horová, Stanislav Katina, and Kamila Hasilová. An adaptive method for bandwidth selection in circular kernel density estimation: S. zámečník et al. *Computational Statistics*, 39(4):1709–1728, 2024.

A NOTATION

A summary of the notation and abbreviations is given in Table 5.

Table 5: Symbols used in Section 3, with the equation that introduces each. S is the number of stencil slots and K the number of SMM groups; the two are distinct. $\hat{p}_i$ is the model's own density and $\hat{p}_i^{\delta}$ the mollified density it returns.

symbol	meaning	symbol	meaning
d_x, d_y	covariate, response dimension	$S = 5$	stencil slots, equation 3
N	sample size	B	batch size
$\mathbf{x}, \mathbf{y}$	covariate, response	L	layers
Ψ	scaling to $[-1, 1]^{d_y}$	K, G	SMM groups, group size
$\mathcal{B} = [-2, 2]^{d_y}$	declared domain	$H = KG$	units per layer
$\mathbf{b}^{\pm}, \hat{\mathbf{b}}^{\pm}$	domain endpoints, perturbed	τ	SMM temperature, equation 5
δ	stated bandwidth, equation 7	$\mathbf{g}^{(l)}$	SMM bias, equation 5
$\boldsymbol{\alpha}$	learned noise scale, equation 2	$e^{\hat{\mathbf{s}}}$	BN scale
$\boldsymbol{\epsilon}$	noise draw, equation 2	$\mathbf{W}_z, \mathbf{W}_c$	response, covariate weights
λ	weight of equation 9	$M^{\mathrm{fwd}}, M^{\mathrm{cond}}$	masks, equation 4
$F_i, \hat{F}_i$	true, modeled conditional CDF	π	response ordering
$p_i, \hat{p}_i$	true, modeled density	O_i	unnormalized output, equation 6
$\hat{p}_i^{\delta}$	returned density, equation 7	c_i	Jacobian of Ψ, equation 8
$p_i^{\delta,\alpha}$	estimand, Definition 3.2	$\mathcal{G}_k$	k-th SMM group

B TRAINING ALGORITHM

Four details are load-bearing. The covariate is never perturbed, so line 6 draws noise for the response alone; the same draw shifts $\hat{\mathbf{y}}$ and both endpoints, which leaves the endpoint gap at 4 and keeps $\hat{b}_i^- < \hat{y}_i < \hat{b}_i^+$. The offsets $\hat{y}_i \pm \delta$ on line 7 are constants with respect to $(\Theta, \boldsymbol{\alpha})$, so no gradient flows through them, and back-propagation needs only first-order reverse passes. Computing the conditioning path once from slot 0 (line 8) holds $\mathbf{y}_{<i}$ fixed across the stencil, which is what makes the two evaluations on line 15 two points of the same conditional. Validation is run with $\boldsymbol{\epsilon} = \mathbf{0}$, as is inference. One step costs $O(BSLH^2)$ time and $O(BSLH)$ memory with $H = KG$ and $S = 5$, independently of d_y.

C RESOLUTION AND THE LIKELIHOOD

Section 1 observes that an estimator whose resolution is not reported cannot be compared on fit alone. This appendix states that observation exactly. It is a property of the likelihood, not a claim about any particular comparison, and nothing in the paper depends on it.

[Resolution enters the likelihood] Let κ_h be a smoothing kernel and write $p^h = p * \kappa_h$ for the density an estimator reporting at resolution h targets. For $Y \sim p$,

$$\mathbb{E}\big[\log p^h(Y)\big] \;=\; \mathbb{E}\big[\log p(Y)\big] - \mathrm{KL}\big(p \,\|\, p^h\big), \tag{10}$$

by Gibbs' inequality, with no assumption on p beyond Assumption 1. Two estimators that recover their targets equally well but report at resolutions $h_1 \neq h_2$ therefore differ in expected held-out log-likelihood by $\mathrm{KL}(p\|p^{h_2}) - \mathrm{KL}(p\|p^{h_1})$, which depends on resolution alone. The penalty is of fourth order in the bandwidth: if $p(\cdot \mid \mathbf{x}) \in C^2$ and κ_h has variance h^2,

$$\mathrm{KL}\big(p \,\|\, p^h\big) = \frac{h^4}{8}\,\mathbb{E}\left[\left(\frac{p''}{p}\right)^2\right] + O(h^6), \tag{11}$$

so a finer resolution buys likelihood without buying fit.

Proof. Equation equation 10 is the definition of the Kullback–Leibler divergence rearranged, and is finite whenever $p \ll p^h$, which convolution with a kernel of full support guarantees. For equation 11,

Algorithm 1 CDF-First: training

Require: training split $\mathcal{D}$, validation split $\mathcal{V}$; learning rate η, batch size B, epochs E, evaluation interval T
Require: stated bandwidth δ, KL weight λ, temperature τ, layers L, groups K, group size G, ordering π
Ensure: parameters $(\Theta^\star, \boldsymbol{\alpha}^\star)$ selected by validation likelihood

$\mathbf{y}_{\min}, \mathbf{y}_{\max} \leftarrow$ range of $\mathcal{D}$ only; set Ψ and $c_i = 2/(y^{(i)}_{\max} - y^{(i)}_{\min})$
initialise Θ; $\boldsymbol{\alpha} \leftarrow e^{-2}\mathbf{1}$; $\mathcal{L}^\star \leftarrow \infty$
for $e = 1$ **to** E **do**
  for each minibatch $\{(\mathbf{x}^{(b)}, \mathbf{y}^{(b)})\}_{b=1}^{B}$ **do**
    $\tilde{\mathbf{y}} \leftarrow \Psi(\mathbf{y})$, mapping the training range to $[-1, 1]^{d_y}$
    draw one $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ for the batch; $\mathbf{x}$ is not perturbed
    $\hat{\mathbf{y}} \leftarrow \tilde{\mathbf{y}} + \boldsymbol{\alpha} \odot \boldsymbol{\epsilon}$, $\hat{\mathbf{b}}^{\pm} \leftarrow \pm 2 \cdot \mathbf{1} + \boldsymbol{\alpha} \odot \boldsymbol{\epsilon}$ {equation 2}
    stack the $S = 5$ slots $[\hat{\mathbf{y}}, \hat{\mathbf{b}}^-, \hat{\mathbf{b}}^+, \hat{\mathbf{y}} - \delta\mathbf{1}, \hat{\mathbf{y}} + \delta\mathbf{1}]$ {equation 3}
    compute the conditioning path $\mathbf{c}^{(1:L)}$ once from slot 0 and broadcast it to all slots
    for $l = 1$ **to** L **do**
      $\mathbf{a}^{(l)} \leftarrow (e^{\mathbf{W}_z^{(l)}} \odot M^{\mathrm{fwd}})\mathbf{h}^{(l-1)} + (\mathbf{W}_c^{(l)} \odot M^{\mathrm{cond}})\mathbf{c}^{(l)} + \mathbf{b}^{(l)}$ {equation 4}
      $\mathbf{h}^{(l)} \leftarrow \mathrm{BN}_l(e^{-\tau}\,\mathrm{SoftMin}_{k \le K}(\mathrm{SoftMax}_{j \in \mathcal{G}_k}(\mathbf{a}^{(l)})) + \mathbf{g}^{(l)})$ {equation 5}
    end for
    $O_i \leftarrow h_i^{(L)}$ at each slot
    $\hat{F}_i \leftarrow \big(O_i(\hat{y}_i) - O_i(\hat{b}_i^-)\big)/\big(O_i(\hat{b}_i^+) - O_i(\hat{b}_i^-)\big)$, clipped to $[0, 1]$ {equation 6}
    $\hat{p}_i^\delta \leftarrow \big(\hat{F}_i(\hat{y}_i + \delta) - \hat{F}_i(\hat{y}_i - \delta)\big)/(2\delta)$, floored at 10^{-8} {equation 7}
    $\mathcal{L} \leftarrow -\frac{1}{B}\sum_b \sum_i [\log \hat{p}_i^\delta + \log c_i] + \mathcal{L}_{\mathrm{KL}}$ {equation 8,equation 9}
    update $(\Theta, \boldsymbol{\alpha})$ by one Adam step with rate η
  end for
  if $e \bmod T = 0$ **then**
    evaluate $\mathcal{L}_{\mathrm{NLL}}$ on $\mathcal{V}$ with $\boldsymbol{\epsilon} = \mathbf{0}$
    if it improves on $\mathcal{L}^\star$ **then**
      $\mathcal{L}^\star \leftarrow$ it; $(\Theta^\star, \boldsymbol{\alpha}^\star) \leftarrow (\Theta, \boldsymbol{\alpha})$
    end if
  end if
end for
return $(\Theta^\star, \boldsymbol{\alpha}^\star)$

a second-order expansion gives $p^h = p + \frac{h^2}{2}p'' + O(h^4)$, and $\mathrm{KL}(p \| p + \varepsilon) = \int \varepsilon^2/(2p) + O(\varepsilon^3)$ with $\varepsilon = \frac{h^2}{2}p''$ yields the stated constant. Note that the first-order term vanishes identically, since $\mathbb{E}_p[p''/p] = \int p'' = 0$ for any density with vanishing tails; the leading behavior is therefore quartic rather than quadratic in h.

Two qualifications. Equation equation 11 assumes smoothness that none of the synthetic targets of Section 5 possesses, so it describes the mechanism rather than our experiments; equation 10 requires none. And the mechanism does not imply that any reported comparison is confounded: Section 4.4 finds the resolution of the estimators we compare to be unidentifiable at their operating points, and therefore finds no evidence that the differences between them are differences in resolution.

D PROOFS

Throughout, fix $\mathbf{x}$ and a coordinate i, and suppress both where no ambiguity arises. Write $\hat{b}_i^{\pm}$ for the perturbed endpoints of equation 2, which satisfy $\hat{b}_i^+ - \hat{b}_i^- = 4$ for every draw of $\boldsymbol{\epsilon}$ because the same shift is applied to both.

D.1 A MONOTONICITY LEMMA

Let $a \in \mathbb{R}^m$. Then $\mathrm{SoftMax}(a) = \log \sum_j e^{a_j}$ is strictly increasing in every argument, and for $b \in \mathbb{R}^K$, $\mathrm{SoftMin}(b) = -\log \sum_k e^{-b_k}$ is strictly increasing in every argument.

Proof. $\partial\,\mathrm{SoftMax}(a)/\partial a_j = e^{a_j}/\sum_{j'} e^{a_{j'}} > 0$ and $\partial\,\mathrm{SoftMin}(b)/\partial b_k = e^{-b_k}/\sum_{k'} e^{-b_{k'}} > 0$, both finite and strictly positive for all finite arguments.

D.2 PROPOSITION 3.2: EXACT NORMALIZATION

Proof. We first show, by induction on the layer index l, that every component of $\mathbf{h}^{(l)}$ belonging to response block i is strictly increasing in $\hat{y}_i$ and does not depend on $\hat{y}_j$ for $j > i$.

Base. $\mathbf{h}^{(0)} = \hat{\mathbf{y}}$, so component i equals $\hat{y}_i$: strictly increasing in $\hat{y}_i$ and independent of every $\hat{y}_j$ with $j \neq i$.

Step. Let r index a pre-activation in block i. By equation 4,

$$a_r^{(l)} = \sum_c \big[e^{\mathbf{W}_z^{(l)}}\big]_{rc} M_{rc}^{\mathrm{fwd}}\, h_c^{(l-1)} + \sum_{c'} \big[\mathbf{W}_c^{(l)}\big]_{rc'} M_{rc'}^{\mathrm{cond}}\, c_{c'}^{(l)} + b_r^{(l)}.$$

The second sum and $b_r^{(l)}$ do not involve $\hat{\mathbf{y}}$ and are therefore constant in $\hat{y}_i$. In the first sum, $M_{rc}^{\mathrm{fwd}} = \mathbf{1}[\pi(r) \geq \pi(c)]$ vanishes whenever $\pi(c) > \pi(r) = i$, so only blocks c with $\pi(c) \leq i$ contribute. By the induction hypothesis, a contributing term with $\pi(c) < i$ is constant in $\hat{y}_i$, and a term with $\pi(c) = i$ is strictly increasing in $\hat{y}_i$; its coefficient $[e^{\mathbf{W}_z}]_{rc}$ is strictly positive. Hence $a_r^{(l)}$ is strictly increasing in $\hat{y}_i$, and is independent of $\hat{y}_j$ for $j > i$ because no such block contributes at any layer.

Applying Lemma D.1 twice, $\mathrm{SoftMin}_{k \leq K}(\mathrm{SoftMax}_{j \in \mathcal{G}_k}(\mathbf{a}^{(l)}))$ is strictly increasing in each $a_r^{(l)}$ of block i; multiplication by $e^{-\tau} > 0$ and the addition of $\mathbf{g}^{(l)}$ preserve this. Finally BN_l applies $u \mapsto e^{\hat{s}}(u - \mu)/\sqrt{\sigma^2 + \varepsilon} + t$ with $e^{\hat{s}} > 0$ and with μ, σ^2 detached and shared across stencil slots, so it is a strictly increasing affine map with coefficients that do not depend on $\hat{y}_i$. This closes the induction, and with $l = L$ gives that O_i is strictly increasing in $\hat{y}_i$ and satisfies $\partial O_i/\partial \hat{y}_j = 0$ for $j > i$.

Since $\hat{b}_i^- < \hat{b}_i^+$, strict monotonicity gives $O_i(\hat{b}_i^+) - O_i(\hat{b}_i^-) > 0$, so equation 6 is well defined and is a strictly increasing affine rescaling of O_i. Evaluating it at the endpoints gives $\hat{F}_i(\hat{b}_i^-) = 0$ and $\hat{F}_i(\hat{b}_i^+) = 1$. Extended by 0 below $\hat{b}_i^-$ and 1 above $\hat{b}_i^+$, $\hat{F}_i$ is non-decreasing on $\mathbb{R}$ with limits 0 and 1, hence the CDF of a probability measure $\hat{P}_i$ supported in $[\hat{b}_i^-, \hat{b}_i^+]$; its density $\hat{p}_i$ with respect to Lebesgue measure, where it exists, is non-negative and satisfies $\int \hat{p}_i \,\mathrm{d}y_i = \hat{F}_i(\hat{b}_i^+) - \hat{F}_i(\hat{b}_i^-) = 1$.

For the joint density, the masking result $\partial O_i/\partial \hat{y}_j = 0$ for $j > i$ means $\hat{F}_i$ depends on $\hat{\mathbf{y}}$ only through $(\hat{y}_i, \hat{\mathbf{y}}_{<i})$, so equation 1 applies to the model as stated. Integrating iteratively, for each fixed $\hat{\mathbf{y}}_{<i}$ the inner integral is 1, whence $\int \prod_i \hat{p}_i \,\mathrm{d}\mathbf{y} = 1$ by Tonelli's theorem, the integrand being non-negative and measurable. No quantity in this argument was estimated, and no property of p was used.

D.3 PROPOSITION 3.2: EXACT MOLLIFICATION

Proof. Let $\hat{Y}_i \sim \hat{P}_i$ with CDF $\hat{F}_i$ as above, let $U \sim \mathcal{U}[-1, 1]$ be independent of $\hat{Y}_i$, and fix $\delta > 0$. Write G for the CDF of $Z = \hat{Y}_i + \delta U$. Conditioning on U and using Fubini's theorem, which applies because the integrand is bounded by 1,

$$G(z) = \mathbb{E}\big[\hat{F}_i(z - \delta U)\big] = \frac{1}{2}\int_{-1}^{1} \hat{F}_i(z - \delta u)\,\mathrm{d}u = \frac{1}{2\delta}\int_{z-\delta}^{z+\delta} \hat{F}_i(t)\,\mathrm{d}t,$$

the last step by the substitution $t = z - \delta u$. The integrand is bounded and monotone, hence Riemann integrable, so G is the integral of a locally integrable function and is therefore locally absolutely continuous. Differentiating with respect to z using the Lebesgue differentiation theorem,

$$G'(z) = \frac{\hat{F}_i(z + \delta) - \hat{F}_i(z - \delta)}{2\delta} = \hat{p}_i^{\delta}(z),$$

at every z at which $\hat{F}_i$ is continuous at both $z+\delta$ and $z-\delta$, hence at all but countably many z, since a monotone function has at most countably many discontinuities. Thus $\hat{p}_i^\delta$ is a version of the density of Z, which is the first claim; equivalently $\hat{p}_i^\delta = \hat{p}_i * \mathcal{U}[-\delta,\delta]$ in the sense of measures, an identity that does not require $\hat{P}_i$ to admit a density. Non-negativity follows from monotonicity of $\hat{F}_i$, and

$$\int_{\mathbb{R}} \hat{p}_i^\delta(z)\,\mathrm{d}z = \lim_{z\to\infty} G(z) - \lim_{z\to-\infty} G(z) = 1.$$

No Taylor expansion was used and no differentiability of $\hat{F}_i$ was assumed, so the identity is exact for every $\delta > 0$ and every parameter value.

The extension of $\hat{F}_i$ by 0 and 1 outside $[\hat{b}_i^-, \hat{b}_i^+]$ is what makes $\hat{P}_i$ a probability measure and is therefore required by the argument. Evaluating equation 6 outside the domain instead, as an unclipped network would, leaves the values in equation 7 unchanged at every point more than δ inside the domain, but breaks the normalization identity within δ of the endpoints. The released implementation applies the extension.

D.4 Proposition 3.3: properness

Proof. Fix $\mathbf{x}$ and write $p^{\delta,\alpha}$ for the density of Definition 3.2, which exists by Assumption 1 and the fact that convolution with κ_{δ,α_i} preserves absolute continuity. For any candidate density q with $\int q\,\mathrm{d}\mathbf{y} = 1$,

$$\mathcal{R}(q \mid \mathbf{x}) - \mathcal{R}(p^{\delta,\alpha} \mid \mathbf{x}) = \int p^{\delta,\alpha} \log \frac{p^{\delta,\alpha}}{q}\,\mathrm{d}\mathbf{y} = \mathrm{KL}\big(p^{\delta,\alpha} \,\|\, q\big) \;\geq\; 0,$$

by Gibbs' inequality, with equality if and only if $q = p^{\delta,\alpha}$ almost everywhere. Taking the expectation over $\mathbf{X}$ preserves the inequality. The objective is therefore strictly proper for $p^{\delta,\alpha}$.

It is improper for p whenever the mollification is non-degenerate. If p has finite second moment then, U and ε being independent of $\mathbf{Y}$ and centered,

$$\mathrm{Var}\big(p_i^{\delta,\alpha}\big) = \mathrm{Var}(p_i) + \frac{\delta^2}{3} + \alpha_i^2 \;>\; \mathrm{Var}(p_i)$$

for any $\delta > 0$ or $\alpha_i > 0$, so $p^{\delta,\alpha} \neq p$ and the unique minimiser of $\mathcal{R}$ is not p. Neither part of the argument uses differentiability of p; only Assumption 1 is required.

A bound on $\|p^{\delta,\alpha} - p\|_\infty$ of order $\delta^2 + \alpha_{\max}^2$ follows from a second-order Taylor expansion when $p(\cdot \mid \mathbf{x}) \in C^2$. Every synthetic target used in this paper has a jump discontinuity, so that hypothesis fails for all of them and we do not state the bound. The distance between the estimand and the target is measured directly in Section 4.3.

[Unbiased gradients] The gradient of the empirical objective equation 8 evaluated at a single draw of $\boldsymbol{\epsilon}$ is unbiased for the gradient of its expectation over $\boldsymbol{\epsilon}$.

Proof. $\hat{F}_i$ takes values in $[0,1]$ and $\delta > 0$ is fixed, so $\hat{p}_i^\delta \in [0, 1/(2\delta)]$ and $\log \hat{p}_i^\delta$ is bounded above; the density floor used in the implementation bounds it below. The integrand and its parameter-gradient are therefore dominated by an integrable function uniformly in $\boldsymbol{\epsilon}$, and differentiation under the expectation is justified by dominated convergence. The stencil offsets $\hat{y}_i \pm \delta$ are constants with respect to the parameters, so no gradient flows through them.

E Implementation details

E.1 What the three monotonicity choices rule out

Section 3.3 states that positive response weights, an additive covariate and a positive batch-norm scale make O_i strictly increasing. Each excludes a specific failure. Replacing $e^{\mathbf{W}_z}$ by an unconstrained weight allows a negative coefficient on the diagonal block, after which $\hat{F}_i$ decreases on part

of the domain and equation 6 returns values outside $[0, 1]$. Letting the covariate enter multiplicatively, as a gate of unrestricted sign, preserves monotonicity only where the gate is positive, so the model would be a valid CDF for some $\mathbf{x}$ and not others. Using the usual learnable batch-norm scale permits a negative scale, which reverses the ordering at that layer; detaching the statistics and sharing them across the stencil axis additionally ensures that the same affine map is applied to all S slots, without which $\hat{F}_i(\hat{b}_i^{\pm})$ would not be computed by the same function as $\hat{F}_i(\hat{y}_i)$ and Proposition 3.2 would fail.

E.2 Protocol

Synthetic targets. Four two-dimensional conditionals with known ground truth and $x \sim \mathcal{U}(-1, 1)$, chosen so that two have a region of zero density inside their support and two do not, and so that each presents a different obstacle: disconnected support, a cut through the mode, an elongated support with hard edges, and an annular support (Figure 7).

- **Disconnected.** Two square components whose separation narrows with x. With $\lambda \sim \text{Bernoulli}(0.5)$, $a^{(1)}, a^{(2)} \overset{\text{iid}}{\sim} \mathcal{U}(-5 + x,\ -1 + x)$ and $b^{(1)}, b^{(2)} \overset{\text{iid}}{\sim} \mathcal{U}(1 - x,\ 5 - x)$,

$$y^{(i)} = \lambda a^{(i)} + (1 - \lambda) b^{(i)}, \qquad i = 1, 2.$$

  The conditional is uniform on $[-5 + x, -1 + x]^2 \cup [1 - x, 5 - x]^2$, each component of area 16, so the density is $1/32$ on the support and zero between the two squares.

- **Half-plane.** An isotropic Gaussian cut through its mode and rotated with x. With $a \sim \mathcal{N}(0, 2^2)$ and $b \sim \mathcal{N}(0, 2^2)$ independent,

$$y^{(1)} = |a| \cos(x\pi) - b \sin(x\pi), \qquad y^{(2)} = |a| \sin(x\pi) + b \cos(x\pi),$$

  giving density $(4\pi)^{-1} e^{-\|\mathbf{y}\|^2/8}$ on the half-plane in which the rotated first coordinate is positive, and zero on the other. The cut passes through the mode, so the density jumps from its maximum to zero across a line.

- **Elongated.** A Gaussian in one rotated coordinate and a uniform in the other, so the support is a long strip with hard ends. With $a \sim \mathcal{N}(0, 1)$, $b \sim \mathcal{U}(-6, 6)$ and $c = (x - 0.75)/2$,

$$y^{(1)} = a \cos(c\pi) - b \sin(c\pi), \qquad y^{(2)} = a \sin(c\pi) + b \cos(c\pi).$$

  The orientation rotates with x; the density is discontinuous where the uniform factor ends, at $b = \pm 6$.

- **Annular.** A ring whose eccentricity varies with x. With $d \sim \mathcal{U}(0, 2)$ and $\theta \sim \mathcal{U}(0, 2\pi)$,

$$y^{(1)} = (4 + 2x + d) \cos\theta, \qquad y^{(2)} = (4 - 2x + d) \sin\theta.$$

  The semi-axes are $4 + 2x$ to $6 + 2x$ and $4 - 2x$ to $6 - 2x$, so the ring is elongated along $y^{(2)}$ at $x = -1$, circular at $x = 0$ and elongated along $y^{(1)}$ at $x = 1$. Radius and angle are uniform, so the induced density on the plane is not: it carries the Jacobian of the polar map and is higher on the inner edge. The interior is empty.

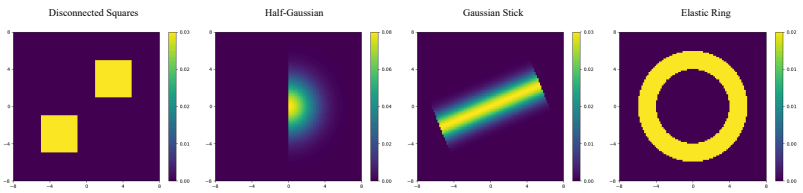


Figure 7: The four synthetic targets. Left to right: disconnected, half-plane, elongated and annular, each at three values of x.

UCI benchmarks. Seven regression datasets, summarized in Table 6. For each, d_x columns are taken as the covariate $\mathbf{x}$ and the remaining d_y as the response $\mathbf{y}$; the assignment is fixed per dataset and is recorded in the released loader. Covariates are standardized to zero mean and unit variance, and the response is mapped to $[-1, 1]$ by Ψ (Section 3); both transforms are fitted on the training split of each fold and applied unchanged to validation and test, so no test statistic enters training. Log-densities are reported in nats on the raw response, obtained by adding $\sum_i \log c_i$ with $c_i = 2/(y^{(i)}_{\max} - y^{(i)}_{\min})$.

Folds follow the sliding-window partition of Section 3.3: the data is shuffled once and cut into ten chunks that are then fixed, and fold k tests on chunk k, validates on chunk $k + 1$ together with the first half of chunk $k + 2$, and trains on the remainder — a $75/15/10$ split in which every chunk serves as test exactly once.

Table 6: UCI benchmarks. Sample counts are after removing rows with missing values.

dataset	samples	d_x	d_y	domain
Fish	908	6	1	toxicology
Concrete	1,030	8	1	civil engineering
Energy	768	8	2	building
Parkinsons	5,875	16	2	medical
Temperature	7,588	21	2	environmental
Air	8,891	8	3	environmental
Skillcraft	3,338	15	4	gaming

E.3 HYPERPARAMETERS

All seven UCI datasets use the same configuration; no per-dataset tuning was performed.

Table 7: Configuration, as released.

	synthetic	UCI
layers L	3	3
groups $K \times$ size G	32×32	32×32
feature widths	$4, 4, 1$	$8, 8, 2$
covariate widths	$16, 16, 4$ times d_x	$16, 16, 1$ times d_x
batch normalization	on	on
stated bandwidth δ	10^{-2}	10^{-2}
swept over	$\{10^{-3}, 10^{-2}, 10^{-1}, 0.3, 0.5, 1\}$	—
KL weight λ	10^{-2}	5×10^{-3}
initial $\log \alpha$	-2	-5
frozen-α arm	$\alpha = 10^{-4}$	—
optimiser	Adam, lr 3×10^{-4}	Adam, lr 3×10^{-4}
batch size	256	256
epochs	20,000	20,000
model selection	validation NLL every 250	early stopping, patience 500

Baseline configurations. Nine baselines appear in the UCI comparison and five of them on the synthetic targets, chosen to span the normalization routes of Section 2. Parameter counts, where measured, are given for the synthetic configuration at $d_y = 2$.

- *Closed-form family.* MDN: 20 Gaussian components with full covariance (2,276 parameters).
- *Cell simplex.* DDN: 256 bins per dimension, 16 latent codes, $\beta = 0.02$, permutation order 1 (11,490).
- *Jacobian determinant.* MAF: 5 autoregressive transforms with linear spline couplings. NSF: 5 rational-quadratic spline transforms (10,972). RNF: 5 radial transforms with elliptical basis functions. RealNVP: 6 affine coupling layers. HINT: 3 blocks of recursive affine coupling with max hierarchy depth 2 (17,928).
- *Transport map.* ATM: 3 rd-degree modified Hermite basis, adaptive greedy selection, at most 60 terms per map component.

- *CDF bound.* NL-CDF: x-transform MLP depth 3, width 128; conditional CDF MLP depth 2, width 256; positive transform square (15,074).

Every baseline uses its authors' open-source implementation at the configuration above, trained with Adam on the same hardware and the same data splits as CDF-First. We did not match capacity or tuning budget: the counts range from 2,276 to 17,928 against 388,065 for CDF-First, so the comparisons in Section 4 are between methods at their default settings rather than at equal cost, and we report them on that basis.

NL-CDF for conditional multivariate tasks. The NL-CDF framework of Chilinski & Silva (2020) admits covariates by concatenating $\mathbf{x}$ to the input, and multivariate responses either by an autoregressive factorization or by a Gaussian copula. We use the autoregressive variant, since it factorizes the joint exactly as equation 1 does and so isolates the differences that concern us. Each conditional factor $F_i(y_i \mid \mathbf{x}, \mathbf{y}_{<i})$ is modeled by a network whose paths from y_i to the output carry non-negative weights, which is how monotonicity in y_i is enforced. Optimizer, learning rate, batch size and early-stopping rule are those used for CDF-First, and the folds are identical.

Two differences are load-bearing. Boundedness comes from a sigmoidal output rather than from endpoint values on a declared domain, so the model is a valid CDF by construction but its normalization is not pinned at every parameter value in the sense of Proposition 3.2. And the density is obtained by differentiating the learned CDF rather than from an interval probability, so there is no bandwidth to state and Proposition 3.2 has no counterpart: the returned density is the derivative of whatever CDF was fitted, at whatever resolution the fit happens to have. Section 4.4 measures that resolution and finds it not identifiable, as for the other baselines.

Experimental environment All experiments ran on a single NVIDIA A800 80 GB PCIe accelerator (81,920 MiB of device memory, of which about 79.2 GiB is available to the process; compute capability 8.0), hosted on a machine with two Intel Xeon Platinum 8358P processors at 2.60 GHz, 64 cores in total across two sockets, and 1.0 TB of system memory. The software stack was PyTorch 2.5.1 with CUDA 12.1 and cuDNN 9.1.0, under NVIDIA driver 535.154.05, which supports CUDA up to 12.2. No experiment used more than one GPU, and every number reported in Section 4 and in this appendix was produced in this single-device configuration, so that runs are comparable across methods and across sample sizes.

F Extended Experiments

F.1 Sample Efficiency

Protocol. We repeat the synthetic comparison of Section 4.5 at four training-set sizes, $N \in \{200, 500, 1000, 2000\}$, holding everything else fixed: the same four targets, the same four conditioning values, the same 400^2 evaluation grid, and the same squared-error metric against the known ground truth. Ten estimators are compared: the nine baselines of Section 4.1 together with CDF-First, each at its default configuration and each trained once per cell. No hyperparameter is retuned as N changes, so the comparison measures how each method's default setting degrades with data rather than how well it could be adapted to a small sample.

Aggregate behavior. Averaged over the four targets, CDF-First attains the lowest error at every sample size (Figure 8). Its error is 0.241 at $N = 200$, 0.209 at $N = 500$, 0.171 at $N = 1000$ and 0.124 at $N = 2000$, against a best baseline of 0.313, 0.253, 0.192 and 0.162 respectively — a margin of 30, 21, 12 and 31 percent. The identity of that best baseline changes with N: the binned model at $N \in \{200, 500, 2000\}$ and the conditional CDF estimator at $N = 1000$, so the comparison is against an envelope rather than against a single competitor.

Per-target behavior. The aggregate conceals a pattern that runs the other way. Counting targets individually, CDF-First is lowest on all four at $N = 200$, on three of four at $N = 500$, on one of four at $N = 1000$, and on two of four at $N = 2000$. It is lowest on the annular target at every size; on the elongated target it is lowest only at $N = 200$, placing second or third elsewhere, and on the half-plane and disconnected targets it alternates with RealNVP, HINT and NL-CDF as N

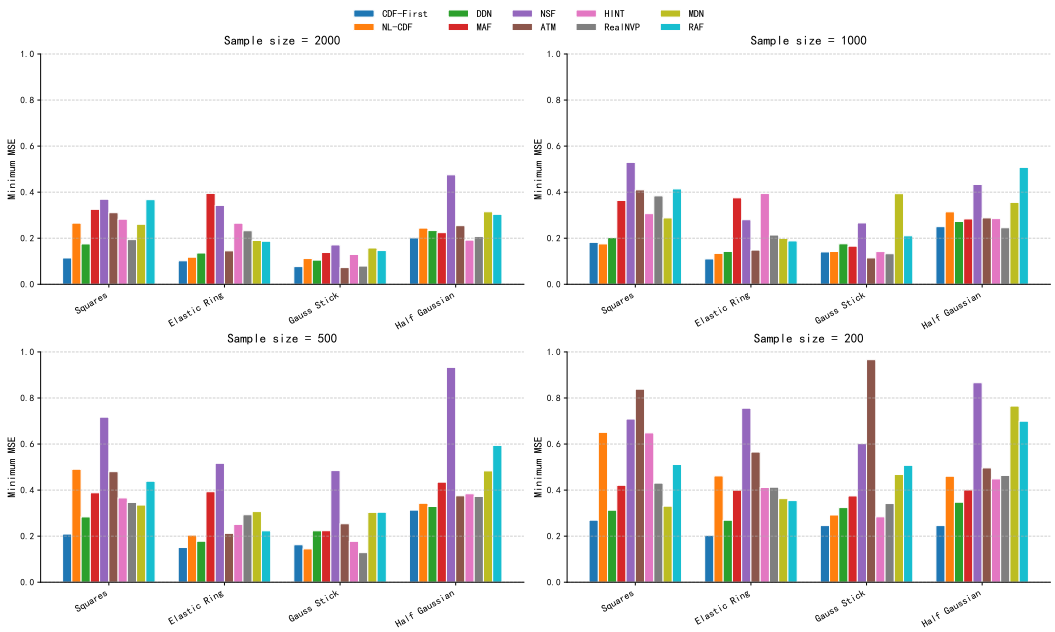


Figure 8: Sample efficiency. Squared error against the known truth for ten estimators on the four synthetic targets — Squares, Elastic Ring, Gauss Stick and Half Gaussian are the disconnected, annular, elongated and half-plane targets of Appendix E.2 — at training-set sizes 2000, 1000, 500 and 200. Bars are grouped by target and colored by method; lower is better, and the vertical scale is shared across panels. CDF-First is lowest on all four targets at $N = 200$ and on two of four at $N = 2000$, while holding the lowest mean over targets at every size. One run per cell at each method's default configuration; no difference is tested.

grows. The aggregate margin at larger N is therefore carried by the size of the wins rather than their number: where CDF-First leads it leads by a wide margin, and where it trails it trails narrowly.

Slope against level. The advantage is one of level, not of slope. As N falls from 2000 to 200, the error of CDF-First grows by a factor of 1.95, which is close to the middle of the range the baselines span: the masked autoregressive flow degrades least, by 1.48, and the adaptive transport map most, by 3.66; the binned model, at 1.93, is almost indistinguishable from ours. Two baselines are worth separating out. The binned model holds a constant ratio to CDF-First, 1.30 at $N = 200$ and 1.31 at $N = 2000$, so it tracks our curve across the whole range. The adaptive transport map does the opposite, moving from 1.58 times our error at $N = 2000$ to 2.97 at $N = 200$, which is the largest change of any method and the reason it is competitive on the UCI benchmarks but not here at small N.

F.2 Qualitative comparison on the synthetic targets

Figure 9 shows the fitted conditional densities of all methods against the known truth, at the four conditioning values used throughout. Squared error summarizes how far each estimate is from the truth; the panels show *how* it is wrong, which the summary cannot.

Disconnected support. The truth is two uniform squares separated by an empty band. Three behaviors appear. CDF-First and the binned model recover both components and keep the band between them empty. The autoregressive and radial flows instead connect the two components with a ridge of mass running diagonally between them, and the coupling flow leaves a thin filament along the same path; the interior of each square is recovered, but the void is not. The mixture places one broad mode on each component and rounds the corners, as a finite Gaussian mixture must. CDF-First's estimate carries a visible high-frequency texture inside each square, a grid pattern at roughly the scale of the stated bandwidth, which none of the baselines shows; it is the cost of representing a piecewise-constant density with a monotone network at $\delta = 10^{-2}$.

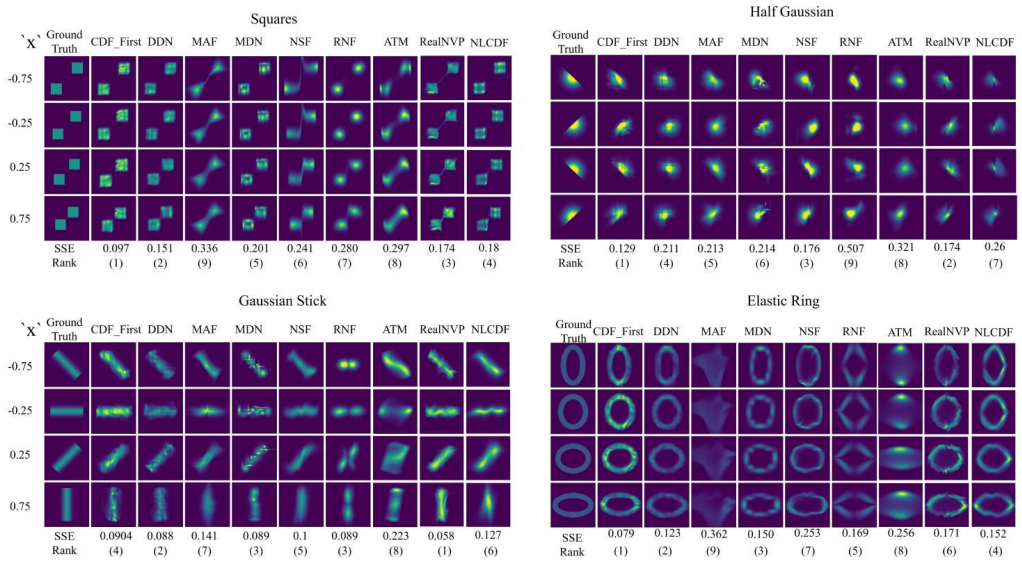


Figure 9: Fitted conditional densities against the known truth on the four synthetic targets, at $x \in \{-0.75, -0.25, 0.25, 0.75\}$ (rows within each block). The leftmost column of each block is the truth; the remaining columns are the estimators, in a fixed order. Colour is density, shared within a row. Squared error and rank are given below each column.

Annular support. The truth is a ring with an empty interior whose eccentricity varies with x. CDF-First, the binned model and the conditional CDF estimator keep the hole open at every conditioning value. The masked autoregressive flow fails outright, spreading mass across the whole region with no ring visible, and the transport map fills the interior with a bright central mode at $x \in \{-0.25, 0.75\}$. The spline and radial flows retain a ring but distort it into a lens or a diamond at the extreme conditioning values, which is the eccentricity being mis-tracked rather than the topology being lost.

Half-plane cut. The truth is a Gaussian sliced through its mode, so the density jumps from its maximum to zero across a straight edge. Every method smooths that edge; the question is by how much. CDF-First retains the sharpest boundary and the correct orientation at all four values of x, at the price of a fan of streaks along the cut. The flows reproduce the bulk but round the edge into a gradient several pixels wide, and the radial flow loses the cut entirely, returning a symmetric blob. The mixture resolves the edge into a cluster of small components rather than a single boundary.

Elongated support. The truth is a Gaussian in one rotated direction and a uniform in the other, so the support is a strip with hard ends. This is the target on which CDF-First does least well, and the figure shows why: the strip is recovered with the right orientation and extent, but with pronounced striping along its length, again at the scale of the stated bandwidth. The coupling flow produces the smoothest strip and the lowest error here. The radial flow splits the strip into two modes at three of the four conditioning values, which is the largest qualitative failure in this panel.

Reading. Two things recur. The methods that normalize through a Jacobian determinant are the ones that bridge the gap in the disconnected target and fill the interior in the annular one. And CDF-First's characteristic failure is texture rather than shape: where it is wrong, it is wrong at the scale of δ, not in the location or extent of the support.

F.3 Calibration Analysis

A proper scoring rule rewards sharpness and calibration together, so a likelihood comparison cannot say whether a model's predictive intervals hold their nominal coverage. We therefore report two calibration diagnostics on two of the UCI benchmarks (Figure 10): a reliability diagram, summarized by expected calibration error, and the histogram of the probability integral transform $\hat{F}(y \mid \mathbf{x})$ evaluated at the held-out responses, which is uniform when the predictive distribution is correct.

We restrict the study to the two univariate datasets, Concrete and Fish, and to the five estimators whose conditional CDF is directly available — ours, the conditional CDF estimator, the hierarchical invertible network, the coupling flow and the adaptive transport map. For the remaining methods the PIT would require numerical inversion or a Monte Carlo estimate, which introduces an error of its own into the diagnostic, and for $d_y > 1$ it would have to be taken coordinate by coordinate under the same ordering the model was trained with. Both are avoidable confounds and we avoid them.

Expected calibration error. CDF-First attains the lowest error on Concrete, 0.0048 against 0.0090 for the next method, and the third lowest of five on Fish, 0.0080 behind 0.0057 and 0.0065. All five methods lie between 0.005 and 0.023 on both datasets, so the reliability curves are close to the diagonal throughout and the differences among them are small in absolute terms.

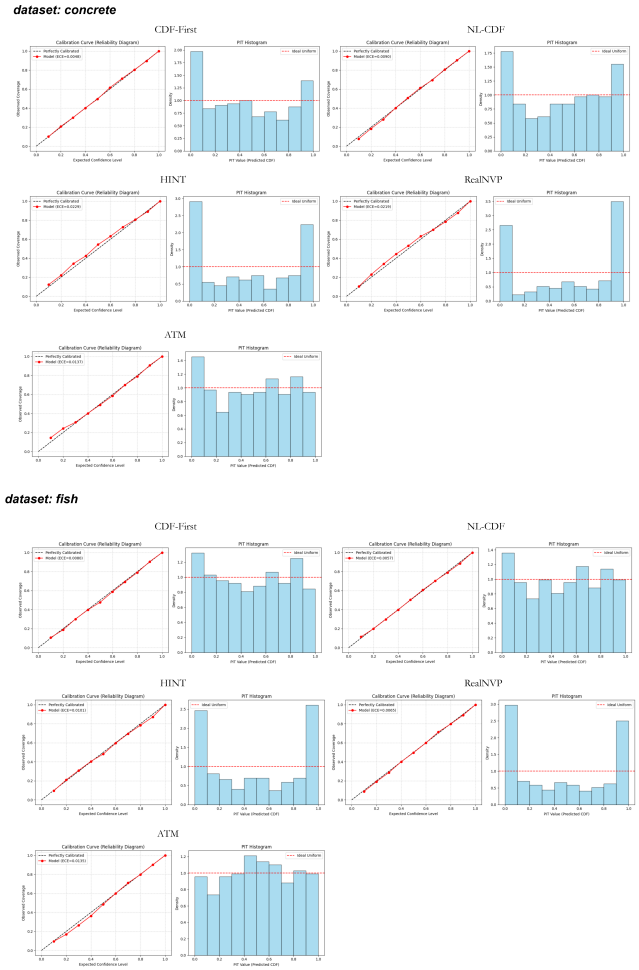


Figure 10: Calibration on Concrete (top) and Fish (bottom). For each estimator: the reliability diagram, with expected calibration error in the legend, and the histogram of the probability integral transform, which is uniform under a correct predictive distribution. A histogram raised at both ends indicates predictive distributions that are too narrow.

Probability integral transform. The histograms separate the methods more sharply than the reliability curves do. On Concrete every method is U-shaped to some degree: the held-out responses fall in the extreme quantiles more often than uniformity predicts, which is the signature of predictive distributions that are too narrow. The effect is mild for CDF-First and the conditional CDF estimator, pronounced for the hierarchical invertible network and the coupling flow, and absent for the transport map, whose histogram is close to flat. On Fish the ordering is the same but the deviations are smaller for the first two methods and unchanged for the other three.

The two diagnostics disagree. This is the substantive finding. On Fish the coupling flow has the second-lowest calibration error of the five and the most U-shaped histogram of any method on either dataset; the transport map has the highest error on both datasets and the flattest histogram. A study reporting expected calibration error alone would rank these two the other way round. The reliability curve aggregates over the predictive distribution and is insensitive to a symmetric excess in both tails, which is exactly the deviation the histogram exposes. We therefore report both, and would caution against reading either alone.

Smoothing does not explain our result. CDF-First estimates a mollified conditional, and mollification widens the predictive distribution. Were the stated bandwidth inflating our intervals, the histogram would be arched rather than U-shaped, with too few observations in the tails. It is mildly U-shaped on both datasets, so at $\delta = 10^{-2}$ the model remains slightly under-dispersed and the smoothing is not the source of whatever calibration advantage the reliability curve shows.

Scope. Two datasets, both univariate, one configuration per method, and no test of any difference. Calibration on the multivariate benchmarks, where the per-coordinate construction of Remark 3.2 would be exercised directly, is not measured here.

G Discussion and Limitations

What the paper establishes. Two properties hold at every parameter value and require nothing of the target beyond absolute continuity: the model is an exactly normalized conditional density, and each returned factor is exactly the model's own factor convolved with a uniform kernel of half-width δ. Both are verified numerically on fitted checkpoints rather than assumed (Section 4.2). A third, weaker statement concerns the objective: it is strictly proper for the mollified conditional and improper for the target. Everything else in the paper is measurement.

The estimand is not the target. This is the cost that qualifies every number we report. CDF-First returns $p * \kappa_{\delta,\alpha}$, not p, and the objective rewards fitting the former. We give no bound on the distance between them. One is available under C^2 smoothness, but all four synthetic targets have jump discontinuities, so such a bound would apply to no experiment in this paper, and we do not state it. What we offer instead is a measurement of how much smoothing the fitted model applies, which is a different and more modest thing.

The domain is declared, and failure outside it is abrupt. The construction assigns zero density outside $\Psi^{-1}(\mathcal{B})$, so a response falling outside the declared domain receives arbitrarily poor likelihood rather than a degraded one. On the synthetic targets the violation is not measurable — two have bounded support and the other two place of order 10^{-23} outside — so this limitation is exercised only by the UCI benchmarks, where we report the fraction of test responses outside the training range. A heavy-tailed synthetic target would exhibit the failure mode directly and we have not run one.

Multivariate scope. For $d_y > 1$ the returned joint is a product of per-coordinate mollifications, not the joint law convolved with a kernel on $\mathbb{R}^{d_y}$, since $\mathbf{y}_{<i}$ is not perturbed. The smoothing therefore accumulates across coordinates, which is one of two candidate explanations for the declining margin with d_y in Section 4.6; the other is that the masked blocks grow as d_y^2, so capacity per conditional falls. Our experiments do not separate them, and a δ sweep on the largest dataset would. We evaluate only to $d_y = 4$ and claim nothing beyond it. The triangular factorization is exact but makes the result order-dependent at finite capacity.

No consistency claim. Consistency would require the true factor to lie in the closure of the model class, control of the empirical process over it, and that the empirical minimizer be found. We assert none of the three. The parameterization is also not identifiable, though the induced law is identified point-wise by $\hat{F}_i$, which is what the objective constrains.

Arithmetic sets a floor on the bandwidth. Equation equation 7 subtracts two values of $\hat{F}_i$ differing by order $2\delta\hat{p}_i$, so a smaller stated bandwidth retains fewer significant digits. In single precision this binds in practice: $\hat{F}_i$ is non-decreasing rather than strictly increasing, and the density floor binds on 22% of grid evaluations at $\delta = 10^{-3}$, falling below 1% at $\delta = 10^{-1}$ and never binding in double precision. Accuracy is flat over four decades above the onset and degrades below it. The construction therefore has a usable range rather than a limit at $\delta \to 0$, and the operating point sits two decades inside it.

The stated bandwidth is measurable only above a transition. Below $\delta \approx 10^{-1}$ no fit is identifiable: smoothing explains too little of the squared error for a width to be read from the output. The upper-bound result therefore holds where the smoothing dominates, and at the operating point used for every accuracy comparison the resolution of CDF-First is no more recoverable than that of any baseline. What distinguishes the stated bandwidth is that it can be moved into the identifiable regime, not that our resolution is known at the operating point.

What the comparisons do and do not control. Parameter counts differ by up to two orders of magnitude and no tuning budget was matched, so the comparisons are between methods at their default configurations. On the synthetic targets each cell is a single run, so we report orderings rather than margins; the UCI results are over ten folds but we do not test them. Replicated seeds with paired rank tests would convert several reported orderings into claims, and a capacity-matched CDF-First at roughly $15{,}000$ parameters would answer the comparison with the conditional CDF estimator directly. Both are cheap and neither is done here.

What the construction gives up. Parameterizing the CDF rather than a transport map removes the Jacobian determinant from the objective and the evaluation path, and costs sampling: drawing from the fitted model requires numerical inversion of $\hat{F}_i$ coordinate by coordinate. Applications that need samples rather than densities are better served by the transport families of Section 2.

The open problem. The classical kernel literature does not merely report a bandwidth, it selects one, by cross-validation, with a theory of the optimal rate. We report a bandwidth and verify that the fitted model respects it. Selecting it — by cross-validating δ against held-out likelihood, or by choosing it to minimize an estimate of the displacement it induces — is the natural next step and is what would make the stated bandwidth a control rather than a declaration.