Super-Level-Set Regression: Conditional Quantiles via Volume Minimization

Sacha Braun[1] **Michael I. Jordan**[1,2] **Francis Bach**[1]
[1]Sierra team, Inria Paris, France [2]Departments of EECS and Statistics, UC Berkeley, USA
{sacha.braun, francis.bach}@inria.fr, jordan@cs.berkeley.edu

Abstract

Constructing minimum-volume prediction regions that satisfy conditional coverage is a fundamental challenge in multivariate regression. Standard approaches rely on explicitly estimating the full conditional density and subsequently thresholding it. This two-step plug-in process is notoriously difficult, sensitive to estimation errors, and computationally expensive. One would like to instead optimize the region directly. Formulating a direct solution is challenging, however, because it requires minimizing a volume objective that is coupled with the conditional quantiles of the model's own estimation error. In this work, we address this challenge. We introduce *super-level-set regression* (SLS), a novel mathematical framework that successfully resolves this implicit coupling, allowing us to directly parameterize and optimize the geometric boundaries of the target conditional level sets. By bypassing full distribution estimation and leveraging flexible volume-preserving frontier functions, our approach natively captures complex, multimodal, and disjoint conditional structures end-to-end. Ultimately, SLS offers a new perspective on multivariate conditional quantile regression, replacing the restrictive assumptions of density-first methods with a direct geometric optimization strategy.

1 Introduction

In many statistical and machine learning applications, accurately characterizing the conditional distribution of a multivariate response variable, $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$, given a feature vector $X \in \mathcal{X}$, is essential. However, in practice, learning the full, often highly complex, conditional density is both statistically challenging and computationally infeasible. For many decision-making tasks, we do not need the entire distribution; rather, it suffices to identify the highest density regions that capture a specific probability mass (Polonik, 1995; Izbicki et al., 2022). These regions are formally known as conditional level sets. Traditionally, extracting these sets relies on a two-step plug-in approach: first estimating the global conditional density, and subsequently thresholding it (Hyndman, 1996). This indirect route forces the model to expend representational capacity on areas of the space that are completely irrelevant to the target quantile. On the other hand, most existing direct estimation strategies are heavily constrained, requiring strong prior assumptions about the shape, location, or topology of the underlying data (Romano et al., 2019; Feldman et al., 2023).

In this work, we introduce *super-level-set regression* (SLS), a novel framework designed to directly parameterize and optimize the boundaries of conditional level sets. Our approach represents a fundamental shift in perspective: rather than attempting to learn the full distribution, our method focuses exclusively and natively on the specific quantile of interest (see Figure 1). By optimizing the boundary directly, we provide an adaptive methodology that dynamically adjusts to the feature vector X, allowing for highly flexible, covariate-dependent conditional level sets.

A key aspect of our method is the utilization of the region's volume as a direct optimization criterion. In our framework, the volume acts as a mechanism to naturally regularize the geometry of the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

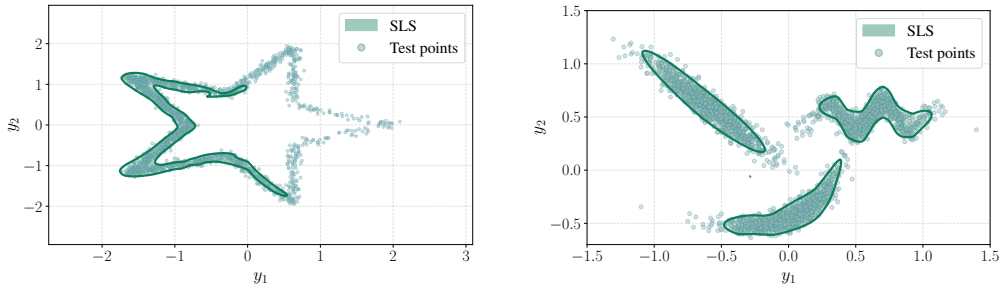

Figure 1: **SLS regression on synthetic 2D distributions**, sampling from $\mathbb{P}_{Y|X}$ for fixed X. **Left:** A single flow-based Mahalanobis frontier capturing an asymmetric, star-shaped density. Target and empirical coverage are both 70%. **Right:** A union of four flow components seamlessly adapting to a disjoint, three-mode distribution. The model successfully allocates mass despite the structural mismatch of components to modes, achieving an exact empirical coverage of 90%.

prediction set, allowing us to bypass the rigid prior assumptions required by classical methods. By penalizing the volume of the set while enforcing a quantile constraint, these sets naturally align with the highest-density region of the data. Through the strategic design of the frontiers of these sets, our approach is capable of estimating level sets with highly intricate geometric structures. This includes handling strongly multimodal distributions, non-convex boundaries, and even complex shapes with internal voids or holes, while conditioning on X.

This approach of directly optimizing the region's volume poses an interesting challenge: it requires minimizing an objective that depends on a conditional quantile implicitly coupled with the model's own predictions. To optimize, one must differentiate through this instance-dependent quantile. While a variety of strategies exist for quantile-constrained optimization, such as differentiable sorting algorithms (Cuturi et al., 2019; Blondel et al., 2020), soft Lagrangian penalties (Cotter et al., 2019), or randomized smoothing (Berthet et al., 2020; Grover et al., 2019), these techniques are designed for marginal or batch-level quantiles. Adapting them to differentiate through *conditional* quantiles that vary dynamically with the covariates X is notoriously difficult, generally forcing a relaxation of the constraint into a weaker marginal guarantee (Braun et al., 2025). Another potential alternative is minimizing the conditional value-at-risk (CVaR) (Rockafellar et al., 2000); however, CVaR focuses solely on minimizing the expected loss of the tail. Consequently, it cannot be used to directly minimize arbitrary functions of the target quantile itself, which is exactly what our framework requires.

As we will show, it is possible to circumvent this implicit coupling through a novel surrogate objective that averages the volume functional over a shrinking probability neighborhood, preserving the differentiability of the objective and capturing the conditional nature of the constraint, modulo finite-sample approximations.

Summary of contributions. Our main contributions are as follows:

- We introduce a general optimization framework capable of minimizing objective functions that are implicitly coupled with the conditional quantiles of the model's own predictions (Proposition 3.1).
- We propose *SLS regression*, a novel approach that focuses exclusively on learning specific conditional level sets. This bypasses the intermediate step of estimating the full conditional distribution and removes the need for restrictive prior assumptions about the data (Section 3).
- We provide an end-to-end framework where the geometric boundary of the level set is directly parameterized and optimized, offering a fundamentally new perspective on multivariate conditional quantile regression (Section 4).
- We demonstrate that by carefully designing flexible parametric frontier functions, our method becomes highly adaptive to the feature vector X, capable of capturing complex conditional structures, including disjoint and multimodal shapes (Section 5).

2 Related work

Highest density regions and predictive sets. The problem of constructing minimum-volume regions achieving a target coverage is solved in theory by the highest density region (HDR) (Hyndman, 1996; Scott and Nowak, 2005). In the context of distribution-free regression, this is often formulated as finding highest predictive density (HPD) sets (Izbicki et al., 2022). Classical methods for extracting HDRs or HPD sets typically follow a plug-in approach: they first train a model to approximate the full conditional density $p(y \mid X)$ and subsequently threshold this density to form regions (Dalmasso et al., 2020; Izbicki et al., 2020; Camehl et al., 2025; Deliu and Liseo, 2026). Normalizing flows (Dinh et al., 2014; Papamakarios et al., 2021) are well suited for modeling complex distributions. In regression, conditional normalizing flows are predominantly trained via maximum likelihood to perform full conditional density estimation (Dinh et al., 2014; Trippe and Turner, 2018; English and Lippert, 2026), after which HDRs can be extracted post-hoc. While statistically sound, learning the entire conditional distribution is notoriously difficult to scale and calibrate (Scott, 2011), often leading to computationally expensive and statistically inefficient prediction sets. Given a complex conditional density estimate, it is difficult to find the optimal conditional threshold yielding the HDR (Izbicki et al., 2022). This often requires Monte Carlo sampling, and an increasing body of work has sought to find ways to compute these sets efficiently, or find distributions for which HDR can be computed in closed form (Wang et al., 2023; Plassier et al., 2025; Dheur et al., 2025; Braun et al., 2025; Dheur and Taieb, 2025) or to rectify its levels post-hoc (Plassier et al., 2025).

Multivariate quantiles and super-level set. Extending quantile regression to multivariate responses $Y \in \mathbb{R}^d$ is fundamentally challenging because there is no natural ordering in multiple dimensions. Optimal transport strategies have also emerged as go-to tools to rank multivariate responses (Thurin et al., 2025; Klein et al., 2025; Ndiaye, 2025), but they fail to go beyond clustering for conditional coverage. Other existing approaches, such as directional quantiles or center-outward quantiles (Hallin and Šiman, 2017; Del Barrio et al., 2024), provide geometrically appealing contours but often fail to guarantee that the enclosed regions correspond to the highest concentration of probability mass, especially for feature-dependent sets.

Bypassing conditional density estimation. To circumvent the challenges of explicit conditional density estimation, a growing body of work attempts to learn prediction sets directly. This shift has been driven by the growing interest in conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Shafer and Vovk, 2008), which allows us to build confidence sets with marginal guarantees. For univariate responses, conformalized quantile regression (CQR) (Romano et al., 2019) and its variants directly estimate lower and upper bounds using the pinball loss. While highly efficient, CQR fundamentally enforces equal-tailed intervals. When the underlying distribution is skewed or multimodal, this restriction displaces the interval away from the true highest density regions, leading to unnecessarily wide and conservative sets. Recent advances, such as conformal thresholded intervals (Luo and Zhou, 2025), attempt to threshold interquantile lengths to dynamically mimic local density. Quantile regression has also been adapted for multivariate settings, such as in the work of Zhou et al. (2024) which produces hyper-rectangles, or that of Feldman et al. (2023), which requires learning multiple directional quantiles. Other methods frame region construction as a direct coverage allocation optimization problem (Sadinle et al., 2019). More recently, Bach (2025) derived a loss function aimed at minimizing set size while maintaining conditional coverage. However, its objective optimizes a global average across all possible coverage levels, meaning it cannot optimize the size for a specific quantile of interest. Closer to our minimum-volume approach, Bars and Humbert (2025) focus on interval length minimization for univariate responses, and Braun et al. (2025) handle more complex forms induced by p-norms in the multivariate setting. However, neither of these works can handle multimodality and, more importantly, they relax the problem to one of marginal coverage. Our work addresses both of these limitations by explicitly targeting conditional coverage.

3 Super-level-set regression

Let $X \in \mathcal{X}$ denote a vector of covariates and let $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$ be a continuous response variable. We assume that $(X, Y) \sim \mathbb{P}_{X,Y}$, where the joint distribution $\mathbb{P}_{X,Y}$ is unknown. Our objective is to construct conditional (super)-level sets for the density $\mathbb{P}_{Y|X}$, given m independent and identically distributed (i.i.d.) samples, $(X_i, Y_i) \sim \mathbb{P}_{X,Y}$.

Assume that the conditional distribution $\mathbb{P}_{Y|X}$ admits a density $p(y \mid X)$ with respect to the Lebesgue measure. For a given probability level $\tau \in (0, 1)$, the conditional super-level set is defined as

$$\mathcal{A}(X) = \{y \in \mathcal{Y} : p(y \mid X) \geq t_\tau(X)\}, \tag{1}$$

where $t_\tau(X)$ is the largest threshold satisfying

$$\mathbb{P}(Y \in \mathcal{A}(X) \mid X) \geq \tau. \tag{2}$$

Our goal is to learn the mapping $\mathcal{A}(\cdot)$. In practice, the true conditional density $p(y \mid X)$ is unknown and notoriously difficult to estimate with high fidelity (Scott, 2011). Traditional strategies rely on negative log-likelihood minimization, but this approach generally requires strong prior assumptions and focuses on learning the entire distribution rather than a specific shape tailored for the target level τ.

The theoretical minimum-volume region achieving at least τ coverage, also known as the highest density region (HDR), can be shown by standard measure-theoretic arguments to be precisely the super-level set of the true conditional distribution, up to null sets (Scott and Nowak, 2005). For every X, the region $\mathcal{A}(X)$ is the solution to the following constrained optimization problem, where $\mathrm{Vol}(\cdot)$ denotes Lebesgue measure:

$$\inf_{\mathcal{B}(X)\subset\mathcal{Y}} \mathrm{Vol}(\mathcal{B}(X)) \quad \text{s.t.} \quad \mathbb{P}(Y \in \mathcal{B}(X)|X) \geq \tau\,. \tag{3}$$

We propose bypassing explicit density estimation to instead search over a hypothesis space of real-valued functions $\mathcal{G} \subset \{G : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$, which we use to parameterize the level sets. We define a candidate confidence region as the sub-level set of one such function:

$$\mathcal{C}_{G,q}(X) = \{y \in \mathcal{Y} : G(X, y) \leq q(X)\}\,. \tag{4}$$

In the following, we refer to a function $G \in \mathcal{G}$ as a *frontier function*. Given a frontier function G, we denote by $\mathrm{Vol}_G(t, X)$ the volume of the set $\{y \in \mathcal{Y} : G(X, y) \leq t\}$ for a threshold $t > 0$.

To strictly satisfy the conditional validity constraint, the boundary threshold $q(X)$ must align with the conditional τ-quantile of the frontier distribution, i.e., $q(X) \geq \text{Quantile}_\tau(\mathbb{P}_{G(X,Y)|X})$.

The problem of finding the optimal confidence region can thus be cast as minimizing the expected volume over the covariates:

$$\inf_{G\in\mathcal{G},q:\mathcal{X}\to\mathbb{R}} \mathbb{E}_X\Big[\mathrm{Vol}_G(q(X), X)\Big] \quad \text{s.t.} \quad q(X) \geq \text{Quantile}_\tau(\mathbb{P}_{G(X,Y)|X}) \quad \mathbb{P}_X\text{-a.s.} \tag{5}$$

If the function class $\mathcal{G}$ is sufficiently expressive, solving Equation (5) exactly recovers the true super-level set $\mathcal{A}(X)$. Crucially, compared to conditional density estimation approaches that typically require restrictive structural assumptions about the data, the only implicit assumption we make is that the true super-level sets can be expressed in the form of Equation (4) for some function $G \in \mathcal{G}$. For instance, with the simple choice $\mathcal{G} = \big\{(x, y) \mapsto \|y - f(x)\|_2 \mid f : \mathcal{X} \to \mathbb{R}^d\big\}$, we can only recover spherical level sets. More expressive frontier functions and a detailed analysis of this property are provided in Section 4.

3.1 Learning strategy

Solving Equation (5) directly is intractable: the volume functional is generally non-differentiable with respect to G, and the quantile constraint $q(X) \geq \mathrm{Quantile}_\tau(\mathbb{P}_{G(X,Y)|X})$ implicitly couples q and G. While the first challenge can be addressed by carefully designing the frontier set $\mathcal{G}$, the implicit link between q and G is harder to resolve.

This problem is a bi-level optimization problem, where the learned parameters are $G(\cdot, \cdot)$ and $q(\cdot)$. The main issue is that for a fixed $q(\cdot)$, Equation (5) tends to reduce the volume uniformly across all samples. One reason for this behavior is that the samples Y are only used to enforce the quantile constraint but do not regularize the volume minimization itself, which remains unconstrained. This leads to degenerate solutions, with the volume collapsing to zero.

To overcome the challenges of optimizing Equation (5) directly, we construct a surrogate objective that bypasses the strict quantile constraint by averaging the volume functional over a shrinking probability neighborhood around the τ-quantile. The intuition is that when $G(X, Y) \approx q(X)$, we can replace $q(X)$ with $G(X, Y)$ in Equation (5).

To formalize the condition $G(X,Y) \approx q(X)$, we introduce a shrinking window around the target quantile. Let $\phi(n)$ and $\psi(n)$ be two positive sequences such that $\phi(n) \to 0$ and $\psi(n) \to 0$ as $n \to +\infty$, where n denotes the number of training steps. We define the quantile neighborhood indicator $K_n(X,g)$ as $K_n(X,g) = \mathbb{1}\left\{q_{\tau-\phi(n)}(X) \le g \le q_{\tau+\psi(n)}(X)\right\}$, where $q_\beta(X) = \mathrm{Quantile}_\beta(\mathbb{P}_{G(X,Y)|X})$ is the conditional β-quantile of the given frontier function G. By the law of total expectation and applying a change of variables to the probability level $\beta = F_{G|X}(g)$, we can rewrite the expectation of the unnormalized surrogate objective as:

$$\mathbb{E}_{X,Y}\left[K_n\big(X,G(X,Y)\big)\mathrm{Vol}_G\big(G(X,Y),X\big)\right] = \mathbb{E}_X\left[\int_{\tau-\phi(n)}^{\tau+\psi(n)} \mathrm{Vol}_G\big(q_\beta(X),X\big)\,d\beta\right]. \quad (6)$$

This explicitly demonstrates that the surrogate objective performs a uniform volume minimization across all quantiles in the interval $[\tau-\phi(n), \tau+\psi(n)]$. Consequently, as the boundaries of this window shrink toward τ, the surrogate increasingly minimizes the volume at the exact target quantile. The following proposition (proof in Appendix A) makes this convergence formal.

Proposition 3.1 (Convergence of the surrogate objective)**.** *Consider the sequence of unconstrained optimization problems:*

$$\inf_{G\in\mathcal{G}} J_n(G) \coloneqq \mathbb{E}_{X,Y}\left[\frac{K_n\big(X,G(X,Y)\big)}{\psi(n)+\phi(n)}\mathrm{Vol}_G\big(G(X,Y),X\big)\right]. \quad (7)$$

Then, under some regularity assumptions (stated in Assumption A.1), as $n \to +\infty$, the sequence of functionals $J_n(G)$ converges uniformly to the true objective $J(G) = \mathbb{E}_X\big[\mathrm{Vol}_G(q_\tau(X),X)\big]$ over $\mathcal{G}$. Furthermore, any limit point of a sequence of optimal solutions $G_n^* \in \arg\min_{G\in\mathcal{G}} J_n(G)$ is an optimal solution to the constrained problem in Equation (5)*.*

This reformulation establishes that minimizing Equation (6), subject to a normalization constraint, is equivalent to the initial objective in Equation (5). Crucially, this equivalence enables our learning procedure by naturally inducing an alternating optimization scheme. To learn the minimal-volume confidence set, we introduce two decay schedules, $\phi(\cdot)$ and $\psi(\cdot)$, and minimize the empirical analogue of Equation (6). We achieve this by alternately updating the frontier function G while holding the quantile estimates q fixed, and subsequently updating q while holding G fixed.

Remark 3.2. *Proposition 3.1 is not strictly tied to the volume functional $\mathrm{Vol}_G(t,X)$. The result naturally generalizes to any continuous objective function $h_G(t,X)$ that satisfies the regularity conditions outlined in Assumption A.1. This point is further discussed in Appendix G.1. For instance, this framework can be used to learn a predictor $f : \mathcal{X} \to \mathbb{R}$ that minimizes the expected conditional median absolute deviation:*

$$\inf_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_X\Big[\mathrm{median}_{Y\sim\mathbb{P}_{Y|X}}\left(|Y-f(X)|\right)\Big].$$

Quantile estimation via pinball loss. Since the true conditional quantile $q_\beta(X) = \mathrm{Quantile}_\beta(\mathbb{P}_{G(X,Y)|X})$ is not available in closed form, we learn it with a separate model $q_\beta : \mathcal{X} \to \mathbb{R}$ trained to minimize the pinball loss (Steinwart and Christmann, 2011) while treating the frontier function G as fixed, as in classical bi-level optimization (Bracken and McGill, 1973). This update only affects the parameters for q, and treats G as a non-differentiable object. This ensures that q tracks the quantile of the current frontier distribution without interfering with the volume objective. Crucially, while this empirical approximation targets the conditional quantile, distribution-free finite-sample guarantees for exact conditional coverage are fundamentally impossible (Foygel Barber et al., 2021); consequently, our proposed method shares this theoretical limitation.

Initialization phase. To initialize the alternating optimization scheme and prevent early instability, we perform an initialization phase by modifying the surrogate objective in Equation (7). Specifically, we replace the localized weighting term $K_n(X,G(X,Y))$ with 1, yielding the simplified, unweighted objective $\mathbb{E}_{X,Y}[\mathrm{Vol}_G(G(X,Y),X)]$. The theoretical justification for this substitution is that computing the expectation over the full conditional distribution $Y \mid X$ is equivalent to integrating the α-quantile volumes over all $\alpha \in [0,1]$. In other words, this unweighted expectation corresponds to $\int_0^1 \mathrm{Vol}_G(q_\alpha(X),X)d\alpha$, as in Bach (2025). By removing the localized window K_n, we essentially

optimize G to minimize the average volume across all possible coverage levels simultaneously, rather than strictly focusing on the τ-quantile. This global minimization provides a robust, non-degenerate initialization for the frontier function G, establishing a strong functional representation before narrowing the focus to the specific τ-quantile neighborhood via the shrinking window.

The bi-level optimization procedure yields our final approach, summarized in Algorithm 1, which we refer to as *super-level-set regression*. We assume that G and the quantile estimators q are parameterized by θ and ω, respectively.

Algorithm 1 SLS regression

Require: Data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^m$, parameterized frontier function G_θ, parameterized quantile networks q_ω, target level τ, decay schedules $\phi(\cdot), \psi(\cdot)$, warm-up steps n_0, total iterations T, learning rates η_θ, η_ω, objective function $h_G(t, X)$ (Volume or log volume for SLS regression).

for $j = 1, \ldots, T$ **do**
  Sample a mini-batch $\mathcal{B} \subseteq \mathcal{D}$
  if $j \leq n_0$ **then**
    $\hat{\mathcal{L}}_\theta \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(X_i,Y_i)\in\mathcal{B}} h_{G_\theta}\big(G_\theta(X_i, Y_i), X_i\big)$
  else
    Calculate window boundaries: $\beta_{\text{low}} = \tau - \phi(j)$ and $\beta_{\text{high}} = \tau + \psi(j)$
    $\hat{\mathcal{L}}_\theta \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(X_i,Y_i)\in\mathcal{B}} \mathbb{1}_{\{q_\omega^{(\beta_{\text{low}})}(X_i) \leq G_\theta(X_i,Y_i) \leq q_\omega^{(\beta_{\text{high}})}(X_i)\}} h_{G_\theta}\big(G_\theta(X_i, Y_i), X_i\big)$
  end if
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \hat{\mathcal{L}}_\theta$
  $\hat{\mathcal{L}}_\omega \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(X_i,Y_i)\in\mathcal{B}} \sum_{\beta\in\{\tau-\phi(j),\tau,\tau+\psi(j)\}} \mathcal{L}^{(i)}_{\text{pinball}}(\omega; \beta)$
  $\omega \leftarrow \omega - \eta_\omega \nabla_\omega \hat{\mathcal{L}}_\omega$
end for

4 Examples of Frontier Functions

We start by restricting our attention to frontier functions $G \in \mathcal{G}$ for which the volume of the sub-level set $\{y \in \mathcal{Y} : G(X, y) \leq t\}$ admits a closed-form expression as a function of the threshold t. We first consider the flow-based Mahalanobis frontier:

$$G_\theta(X, Y) = \big\| L_\theta(X)\big(T_\phi(Y; X) - \mu_\theta(X)\big)\big\|_2^2, \tag{8}$$

where θ and ϕ are learnable parameters. Here, $L_\theta(X)$ is a $d \times d$ lower triangular matrix with strictly positive diagonal entries and $\mu_\theta(X) \in \mathbb{R}^d$. The mapping $T_\phi(\cdot; X) : \mathcal{Y} \to \mathbb{R}^d$ is a conditional volume-preserving normalizing flow (see, e.g., Dinh et al. 2014). By design, this flow is a diffeomorphism constructed to have a Jacobian determinant of exactly 1. This critical property guarantees that the Lebesgue measure is strictly invariant under the transformation T_ϕ, yielding the volume surrogate

$$\mathrm{Vol}(q(X), X) \propto \frac{q(X)^{d/2}}{\det(L_\theta(X))}.$$

Further discussion on losses induced by the frontier in Equation (8) is provided in Appendix B.1. If we were to set $T_\phi(Y; X) = Y$, the confidence regions induced by Equation (4) would be simple rigid ellipsoids. The normalizing flow, however, allows us to apply highly nonlinear transformations to these ellipsoids while exactly preserving their volume in the latent space. This clarifies our previous theoretical assertion: if the function class $\mathcal{G}$ is sufficiently expressive, we can exactly recover the true optimal level sets by minimizing Equation (5). By incorporating the transformation T_ϕ, a natural question arises: what family of geometric sets can be exactly characterized by the sub-level sets of Equation (8)? The following proposition formalizes the expressive power of this formulation.

Proposition 4.1. *Let $\Omega \subset \mathbb{R}^d$ be a compact domain with a smooth boundary, and assume Ω is diffeomorphic to a closed ball. Then, there exists a volume-preserving diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$, a $d \times d$ lower triangular matrix L with strictly positive diagonal entries, a vector $\mu \in \mathbb{R}^d$ and a constant $q > 0$ such that the target set Ω is exactly recovered by:*

$$\Omega = \Big\{ y \in \mathbb{R}^d : \big\| L(T(y) - \mu)\big\|_2^2 \leq q \Big\}.$$

Remark 4.2. *The initialization phase procedure (Section 3.1) yields an important theoretical consequence for the flow-based Mahalanobis frontier: the localized optimization phase using the shrinking window K_n can sometimes be omitted entirely. If a volume-preserving flow T_ϕ maps $Y \mid X$ to an elliptically contoured distribution (e.g., a multivariate Gaussian) with center $\mu_\theta(X)$ and covariance matrix $(L_\theta(X)^\top L_\theta(X))^{-1}$, the sub-level sets of $G(X, Y)$ perfectly match the true minimum-volume sets for* all *coverage levels. In this scenario, the unweighted initialization phase objective is perfectly specified, as it minimizes expected volume globally across all levels. Thus, the initialization phase, followed by an exact calibration of $q(X)$ to the target τ-quantile, is mathematically sufficient to recover the optimal confidence region. An experiment illustrating the usefulness of the shrinking window is provided in Appendix B.*

4.1 Union of normalizing flows

While a single volume-preserving flow is highly expressive, it is fundamentally a diffeomorphism and thus preserves topological connectedness. Consequently, it cannot easily model disjoint, heavily multimodal regions without introducing extreme, physically or biologically implausible distortions in the mapping. A flexible alternative captures complex, disconnected spaces via the union of localized flow-based regions, conceptually mirroring the level sets of a transformed Gaussian mixture, but optimizing boundaries directly by taking the union of flow-based regions.

We define an ensemble of K independent conditional normalizing flows $T_{\phi_k}(Y; X)$, along with their respective localized networks $\mu_{\theta_k}(X)$ and shape matrices $L_{\theta_k}(X)$. We define the individual frontier for the k-th component by using its local latent Mahalanobis distance:

$$G_k(X, Y) = \left\| L_{\theta_k}(X)\big(T_{\phi_k}(Y; X) - \mu_{\theta_k}(X)\big)\right\|_2^2. \tag{9}$$

To ensure all flow components actively participate in covering the target distribution, and to avoid the "dead centers" pathology associated with a hard minimum, we aggregate the scaled base frontiers using a softmax-weighted average (which acts as a differentiable min operator):

$$G_\beta(X, Y) = \sum_{k=1}^{K} w_k G_k(X, Y), \quad \text{where} \quad w_k = \frac{\exp\big(-\beta \cdot G_k(X, Y)\big)}{\sum_{j=1}^{K} \exp\big(-\beta \cdot G_j(X, Y)\big)}, \tag{10}$$

where $\beta > 0$ acts as an inverse temperature controlling the sharpness of the approximation. In practice, β is increased via an annealing scheme so that the frontier function converges to the hard minimum $G_\infty(X, Y) = \min_k G_k(X, Y)$, which is then used to construct the final level set.

A practical limitation of the smooth formulation in Equation (10) is that the exact volume $\mathrm{Vol}_{G_\beta}(t, X)$ is not generally available in closed form. Instead, we optimize a volume proxy Ω_{Vol} defined as the sum of the individual component volumes. While this approximation is not exact when components overlap, in practice, the optimization dynamics naturally drive the components toward configurations with negligible overlap. Further discussion of the training procedure induced by the frontier Equation (10) is presented in Appendix B.2.

The theoretical capacity of these minimum-based frontier to capture disjoint topologies is established in the following proposition, whose proof is a direct consequence of Proposition 4.1.

Proposition 4.3. *Let $\Omega = \bigcup_{k=1}^{K} \Omega_k \subset \mathbb{R}^d$ be a finite union of K disjoint sets, where each Ω_k is a compact domain with a smooth boundary and diffeomorphic to a closed ball. Then, there exists $G_\infty \in \mathcal{G}$ from Equation* (10) *such that (with the X-dependency omitted for clarity):*

$$\Omega = \left\{ y \in \mathbb{R}^d : G_\infty(y) \le q \right\}.$$

Remark 4.4 (Minimum-volume sets beyond highest density regions)**.** *The frontier function G does not need to be related to the conditional density: for any class $\mathcal{G}$, Equation* (5) *returns the set of smallest expected volume among the sets $\mathcal{C}_{G,q}$ with conditional coverage τ. When $\mathcal{G}$ encodes a geometric prior, this is the minimum-volume set of the prescribed shape, which is in general not a level set of the density. For instance, the class $\mathcal{G}$ of the functions $(x, y) \mapsto \|D(x)(y - f(x))\|_\infty$, with $f : \mathcal{X} \to \mathbb{R}^d$ and $D(x)$ a diagonal matrix with positive entries, yields the axis-aligned hyperrectangles $\mathcal{C}_{G,q}(X) = \prod_{i=1}^{d}[f_i(X) \pm q(X)/D_{ii}(X)]$, of volume $\mathrm{Vol}_G(q(X), X) = (2q(X))^d / \det D(X)$. SLS regression then learns the smallest conditional hyperrectangles with coverage τ,* i.e.*, one interval per output coordinate, the d intervals jointly containing Y with probability τ given X, without estimating the density. Such control over the geometry of the sets is valuable when they must follow a prescribed shape, for instance for interpretability.*

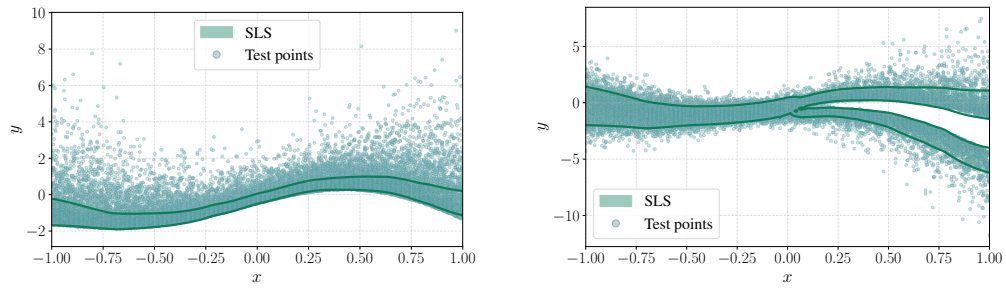

Figure 2: **SLS regression on synthetic 1D conditional distributions. Left:** A single-component asymmetric exponential distribution evaluated at a target coverage of 30%. The shrinking window successfully corrects the sub-optimal centering caused by the initialization phase. **Right:** A mixture of exponentials transitioning from unimodal to bimodal, evaluated at a target coverage of 80%. The union of flows dynamically adapts to the changing topology.

5 Experiments

All of our experiments can be found and reproduced in the associated GitHub repository[1]. Additional information, including the model architecture, learning procedure, and the choice of the shrinking window schedules, is provided in Appendices D & E.

5.1 Synthetic experiments

We start by assessing the robustness of our approach across different synthetic scenarios. Illustrative examples without dependence on the covariates X are provided in Figure 1, and additional figures can be found in Appendix G. These experiments demonstrate that our framework allows us to directly recover complex geometric shapes.

We now focus our analysis on scenarios featuring a strong dependence on the feature vector X. We begin with a univariate example where $X \sim \mathcal{U}([-1, 1])$, and the conditional distribution $Y|X$ follows a mixture of exponential distributions with either one or two components, as shown in Figure 2. For the single-component case, we use the frontier $\mathcal{G} = \{(x, y) \mapsto |y - f(x)| \mid f : \mathcal{X} \to \mathbb{R}\}$, which restricts the prediction sets to intervals. In this example, the unweighted initialization phase, which accounts for all possible volume sizes simultaneously, tends to push $f(x)$ towards the conditional median of $Y|X = x$. Because the exponential noise is highly asymmetric, the median is generally not the optimal center for the highest density region. This highlights the critical importance of the shrinking window, which enables the model to dynamically adjust and learn the precise level set for a specific target quantile.

For the bimodal example, we employ the union of flows frontier (Equation (10)) with $K = 3$ components. The model successfully identifies the unimodal structure for low values of X, and correctly switches to disjoint, multimodal sets as X increases. Crucially, the flow components collaboratively align to form the correct union of intervals, effectively adapting even when the number of parameterized flows ($K = 3$) exceeds the true number of modalities in the data.

We also evaluate our method in a multivariate setting featuring either heteroscedastic Gaussian noise contaminated with 10% outliers, or heteroscedastic exponential noise. In both scenarios, we utilize the flow-based Mahalanobis frontier (Equation (8)). For the Gaussian scenario, we restrict the conditional flow to the identity mapping, thereby limiting the model to recovering rigid ellipsoids. While standard maximum likelihood estimation (minimizing the Gaussian negative log-likelihood) would be heavily skewed by the presence of the outliers, our shrinking window approach learns to completely ignore data points that fall outside the target level set. Similarly, for the exponential noise scenario, our approach consistently and accurately captures the true level sets without being forced to model the entire asymmetric tail.

[1] `https://github.com/ElSacho/super_level_sets_regression`

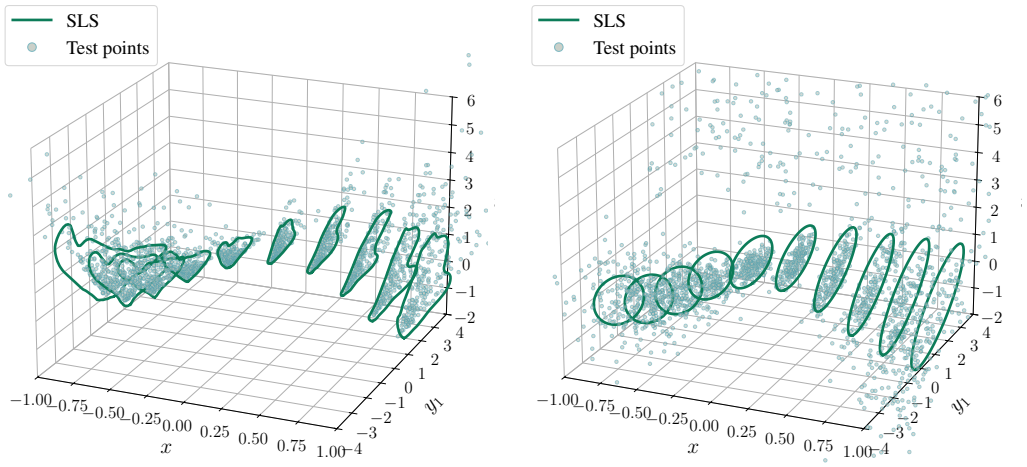

Figure 3: **SLS on synthetic 2D conditional distributions** with a flow-based frontier. **Left:** Heteroscedastic exponential noise at a target coverage of 80%. **Right:** Heteroscedastic Gaussian noise contaminated with 10% outliers, modeled with a rigid ellipsoid frontier (identity flow) at a target coverage of 60%. The model successfully ignores the outliers to capture the highest density region.

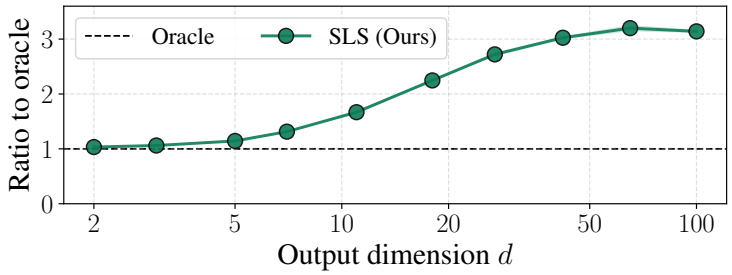

Figure 4: **Robustness to the output dimension** d**.** Average over the test points of $\big(\mathrm{Vol}(\mathcal{C}(X))/\mathrm{Vol}(\mathcal{C}^\star(X))\big)^{1/d}$, where $\mathcal{C}^\star$ is the oracle HDR, for a target coverage of 95% (mean $\pm$ standard error over 10 seeds).

Robustness to the output dimension. We generate $X \sim \mathcal{U}([-1,1])$ and $Y = f(X) + A(X)W$, with $W \sim \mathcal{N}(0, I_d)$, a non-linear mean f and a full matrix $A(X)$ that varies with X, so that the components of Y are heteroscedastic and all correlated. The oracle HDR $\mathcal{C}^\star(X)$ is then an ellipsoid. For $\tau = 0.95$, we learn a single-flow SLS model on 5,000 points, conformalize it, and report the average over 5,000 test points of $\big(\mathrm{Vol}(\mathcal{C}(X))/\mathrm{Vol}(\mathcal{C}^\star(X))\big)^{1/d}$ (details in Appendix G.2). SLS regression remains within 15% of the oracle up to $d = 5$ (Figure 4); the ratio then increases and stabilizes around 3 from $d \approx 40$ to $d = 100$, while the empirical coverage stays between 94% and 96%. We believe that this gap is partly due to the training budget, identical for all d, and to the low-rank precision matrix used for $d \geq 5$ (Appendix D), which cannot represent the full correlation structure of Y.

5.2 Real datasets

For real-world experiments, we compare our approach against a conditional density estimation (CDE) baseline. The goal of this experiment is not to benchmark our strategy, but rather to establish empirically that the results are consistent with those of CDE. The CDE model minimizes the negative log-likelihood of a multivariate quantile function forecaster (Kan et al., 2022), utilizing input convex neural networks (Amos et al., 2017) as implemented by Dheur et al. (2025). To guarantee marginal validity, we conformalize the CDE predictions with the C-HDR non-conformity scores (Izbicki et al., 2022), and our approach with the score detailed in Appendix C. We also compare our approach with the multivariate conformal prediction strategies targeting conditional coverage in Dheur et al. (2025,

Table 1). These are L-CP and C-PCP (Dheur et al., 2025) and CP2-PCP (Plassier et al., 2025). We also compare with PCP (Wang et al., 2023) detailed in Appendix F. Finally, we compare with rectified conformal prediction (RCP) (Plassier et al., 2025), which rectifies a conformity score by learning its conditional quantile, applied to the density-based score of DR-CP (Dheur et al., 2025) and to the PCP score (RCP-DR-CP and RCP-DR-PCP). These two methods learn this quantile on a larger calibration set, the test set having the same size for all methods (Appendix F).

Evaluating these sets is challenging because the true data distribution is unknown. To measure performance, we plot the average volume (scaled by its $(1/d)$-th power, where d is the output dimension) against the estimated conditional coverage deviation, $\mathbb{E}[|\mathbb{P}(Y \in C(X) \mid X) - \tau|]$. We approximate this deviation using the ERT metric (Braun et al., 2026).

The results, averaged across six datasets and ten random seeds, are presented in Figure 5. Overall, SLS regression performs competitively with CDE, operating directly at the Pareto frontier of the inherent trade-off between region size and conditional coverage. When interpreting these results, it is crucial to analyze the trade-off simultaneously: while certain strategies may output smaller sets, they often do so at the cost of significantly degraded conditional coverage.

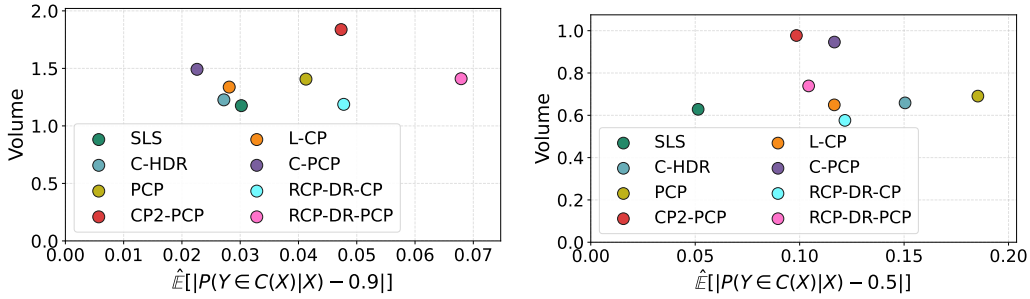

Figure 5: **Performance on real-world datasets.** Average scaled volume versus the estimated conditional coverage deviation. Closer to $(0, 0)$ is better. **Left:** Target coverage 90%. **Right:** Target coverage 50%.

Hyperparameters and computational cost. On the real datasets, the volume of SLS changes by at most 1% when the increment of the inverse temperature β spans three orders of magnitude, and by less than 4% across $K \in \{1, \ldots, 5\}$ flows on three of the six datasets, while the training time grows linearly with K (Appendix G.5). SLS answers queries faster than the sampling-based baselines on four datasets, but is currently the most expensive to train, a comparison that depends heavily on the training budget of each method (Appendix G.6).

6 Conclusion

In this work, we introduced SLS regression, a novel framework that shifts the paradigm of highest density region estimation from an indirect density plug-in approach to direct geometric optimization. By developing a surrogate objective based on a shrinking probability window, we successfully resolved the notoriously difficult implicit coupling between volume minimization and conditional quantile constraints. Its main limitations are its reliance on an accurate estimate of the conditional quantile q, hyperparameters such as the number of flows K to choose, a training cost higher than that of the baselines, and a gap to the oracle that grows with the output dimension. A promising direction for future work is the design of other classes of frontier functions tailored to specific domains, such as incorporating physical priors, handling discrete or mixed-type responses, scaling the geometric optimization to highly structured, high-dimensional spaces.

References

Amos, B., L. Xu, and J. Z. Kolter (2017). Input convex neural networks. *International Conference on Machine Learning*.

Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bach, F. (2025). A convex loss function for set prediction with optimal trade-offs between size and conditional coverage. *arXiv preprint arXiv:2512.19142*.

Bars, B. L. and P. Humbert (2025). On volume minimization in conformal regression. *International Conference on Machine Learning*.

Berthet, Q., M. Blondel, O. Teboul, M. Cuturi, J.-P. Vert, and F. Bach (2020). Learning with differentiable perturbed optimizers. *Advances in Neural Information Processing Systems*.

Blondel, M., O. Teboul, Q. Berthet, and J. Djolonga (2020). Fast differentiable sorting and ranking. In *International Conference on Machine Learning*.

Bracken, J. and J. T. McGill (1973). Mathematical programs with optimization problems in the constraints. *Operations Research 21*(1), 37–44.

Braun, S., L. Aolaritei, M. I. Jordan, and F. Bach (2025). Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*.

Braun, S., E. Berta, M. I. Jordan, and F. Bach (2025). Multivariate standardized residuals for conformal prediction. *arXiv preprint arXiv:2507.20941*.

Braun, S., D. Holzmüller, M. I. Jordan, and F. Bach (2026). Conditional coverage diagnostics for conformal prediction. *International Conference on Machine Learning*.

Camehl, A., D. Fok, and K. Gruber (2025). On superlevel sets of conditional densities and multivariate quantile regression. *Journal of Econometrics 249*, 105807.

Cho, D., C. Yoo, J. Im, and D.-H. Cha (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science 7*(4).

Cotter, A., H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research 20*(172), 1–59.

Cuturi, M., O. Teboul, and J.-P. Vert (2019). Differentiable ranking and sorting using optimal transport. *Advances in Neural Information Processing Systems*.

Dacorogna, B. and J. Moser (1990). On a partial differential equation involving the Jacobian determinant. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, Volume 7, pp. 1–26. Elsevier.

Dalmasso, N., R. Izbicki, and A. Lee (2020). Confidence sets and hypothesis testing in a likelihood-free inference setting. *International Conference on Machine Learning*.

Del Barrio, E., A. G. Sanz, and M. Hallin (2024). Nonparametric multiple-output center-outward quantile regression. *Journal of the American Statistical Association*, 1–15.

Deliu, N. and B. Liseo (2026). Alternative approaches for estimating highest-density regions. *International Statistical Review 94*(1), 97–120.

Dheur, V., M. Fontana, Y. Estievenart, N. Desobry, and S. B. Taieb (2025). Multi-output conformal regression: A unified comparative study with new conformity scores. *International Conference on Machine Learning*.

Dheur, V. and S. B. Taieb (2025). Multivariate latent recalibration for conditional normalizing flows. *Advances in Neural Information Processing Systems*.

Dinh, L., D. Krueger, and Y. Bengio (2014). NICE: Non-linear independent components estimation. *Workshop at the International Conference on Learning Representations*.

English, E. and C. Lippert (2026). JAPAN: Joint adaptive prediction areas with normalising-flows. *International Conference on Learning Representations*.

Feldman, S., S. Bates, and Y. Romano (2023). Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research 24*(24), 1–48.

Foygel Barber, R., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA 10*(2), 455–482.

Grover, A., E. Wang, A. Zweig, and S. Ermon (2019). Stochastic optimization of sorting networks via continuous relaxations. *International Conference on Learning Representations*.

Hallin, M. and M. Šiman (2017). Multiple-output quantile regression. *Handbook of Quantile Regression*, 185–207.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician 50*(2), 120–126.

Izbicki, R., G. Shimizu, and R. B. Stern (2022). CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research 23*(87), 1–32.

Izbicki, R., G. T. Shimizu, and R. B. Stern (2020). Flexible distribution-free conditional predictive bands using density estimators. *International Conference on Artificial Intelligence and Statistics*.

Kan, K., F.-X. Aubet, T. Januschowski, Y. Park, K. Benidis, L. Ruthotto, and J. Gasthaus (2022). Multivariate quantile function forecaster. In *International Conference on Artificial Intelligence and Statistics*.

Klein, M., L. Bethune, E. Ndiaye, and M. Cuturi (2025). Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*.

Luo, R. and Z. Zhou (2025). Conformal thresholded intervals for efficient regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ndiaye, E. (2025). Beyond uncertainty sets: Leveraging optimal transport to extend conformal predictive distribution to multivariate settings. *arXiv preprint arXiv:2511.15146*.

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

Papadopoulos, H., K. Proedrou, V. Vovk, and A. Gammerman (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*.

Papamakarios, G., E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research 22*(57), 1–64.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Plassier, V., A. Fishkov, V. Dheur, M. Guizani, S. B. Taieb, M. Panov, and E. Moulines (2025). Rectifying conformity scores for better conditional coverage. In *International Conference on Machine Learning*.

Plassier, V., A. Fishkov, M. Guizani, M. Panov, and E. Moulines (2025). Probabilistic conformal prediction with approximate conditional validity. In *International Conference on Learning Representations*.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, 855–881.

Qu, J., D. Holzmüller, G. Varoquaux, and M. L. Morvan (2026). TabICLv2: A better, faster, scalable, and open tabular foundation model. *International Conference on Machine Learning*.

Rana, P. (2013). Physicochemical properties of protein tertiary structure. *UCI Machine Learning Repository*.

Rockafellar, R. T., S. Uryasev, et al. (2000). Optimization of conditional value-at-risk. *Journal of Risk 2*, 21–42.

Romano, Y., E. Patterson, and E. Candès (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*.

Sadinle, M., J. Lei, and L. Wasserman (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association 114*(525), 223–234.

Scott, C. and R. Nowak (2005). Learning minimum volume sets. *Advances in Neural Information Processing Systems*.

Scott, D. W. (2011). Multivariate density estimation and visualization. In *Handbook of Computational Statistics: Concepts and Methods*, pp. 549–569. Springer.

Shafer, G. and V. Vovk (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research 9*(3), 371–421.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*(1), 1929–1958.

Steinwart, I. and A. Christmann (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli 17*(1).

Teixeira, P. (2017). Dacorogna-Moser theorem on the Jacobian determinant equation with control of support. *Discrete and Continuous Dynamical Systems 37*(7), 4071–4089.

Thurin, G., K. Nadjahi, and C. Boyer (2025). Optimal transport-based conformal prediction. In *International Conference on Machine Learning*.

Trippe, B. L. and R. E. Turner (2018). Conditional density estimation with Bayesian normalising flows. *Workshop at Advances in Neural Information Processing Systems*.

Tsoumakas, G., E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas (2011). Mulan: A Java library for multi-label learning. *The Journal of Machine Learning Research 12*, 2411–2414.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer.

Wang, J., J. Tong, K. Tan, Y. Vorobeychik, and Y. Kantaros (2023). Conformal temporal logic planning using large language models. *ACM Transactions on Cyber-Physical Systems*.

Wang, Z., R. Gao, M. Yin, M. Zhou, and D. Blei (2023). Probabilistic conformal prediction using conditional random samples. In *International Conference on Artificial Intelligence and Statistics*.

Zhou, Y., L. Lindemann, and M. Sesia (2024). Conformalized adaptive forecasting of heterogeneous trajectories. In *International Conference on Machine Learning*.

Appendices

Appendices

A Proofs **15**

B Alternative Volume Objectives and Theoretical Equivalences **17**

B.1 Single-flow Mahalanobis frontier 17

B.2 Union of normalizing flows 18

C Conformalizing the level sets **19**

D Architecture **19**

D.1 Frontier function architecture 19

D.2 Quantile network 20

E Training strategy **20**

E.1 Shrinking window scheduling 20

E.2 Early stopping 21

E.3 Learning the τ-th quantile 21

F Baselines **21**

G Additional experiments **22**

G.1 Minimizing alternative functions of the conditional quantile 22

G.2 Robustness to the output dimension 23

G.3 Real data 23

G.4 Results per dataset 23

G.5 Sensitivity to the hyperparameters 24

G.6 Computational cost 25

H Declaration of LLM usage **27**

A Proofs

Assumption A.1 (Regularity of the hypothesis class)**.** *The function class $\mathcal{G}$ is a compact metric space. For $\mathbb{P}_X$-almost every $X \in \mathcal{X}$, the following hold uniformly for all $G \in \mathcal{G}$:*

1. ***Uniform density bound:*** *The conditional distribution of $G(X,Y)$ given X admits a density $f_{G|X}(g)$, and there exists a strictly positive function $c(X) > 0$ such that $f_{G|X}(g) \geq c(X)$ in a neighborhood of the τ-quantile $q_\tau(X) = \mathrm{Quantile}_\tau(\mathbb{P}_{G|X})$.*
2. ***Uniform Lipschitz volume:*** *There exists a function $L(X)$ such that the mapping $g \mapsto \mathrm{Vol}_G(g, X)$ is $L(X)$-Lipschitz continuous in the aforementioned neighborhood of $q_\tau(X)$. Furthermore, the objectives $J_n(\cdot)$ and $J(\cdot)$ are continuous on the function class $\mathcal{G}$.*
3. ***Integrability:*** *The true objective $J(G) = \mathbb{E}_X\big[\mathrm{Vol}_G(q_\tau(X), X)\big]$ is finite, and $\mathbb{E}_X\left[\frac{L(X)}{c(X)}\right] < \infty$.*

Proposition A.2 (Convergence of the surrogate objective)**.** *Suppose Assumption A.1 holds. Consider the sequence of unconstrained optimization problems:*

$$\inf_{G\in\mathcal{G}} J_n(G) \coloneqq \mathbb{E}_{X,Y}\left[\frac{K_n\big(X, G(X,Y)\big)}{\psi(n)+\phi(n)}\mathrm{Vol}_G\big(G(X,Y), X\big)\right].$$

Then, as $n \to +\infty$, the sequence of functionals $J_n(G)$ converges uniformly to the true objective $J(G) = \mathbb{E}_X\big[\mathrm{Vol}_G(q_\tau(X), X)\big]$ over $\mathcal{G}$. Furthermore, any limit point of a sequence of optimal solutions $G_n^* \in \arg\min_{G\in\mathcal{G}} J_n(G)$ is an optimal solution to the constrained problem in Equation (5)*.*

Proof. The proof proceeds in two main steps. First, we establish the uniform convergence of the surrogate objective $J_n(G)$ to the true objective $J(G)$. Second, we rely on this uniform convergence and the compactness of the hypothesis class $\mathcal{G}$ to show the convergence of the minimizers.

For a fixed function $G \in \mathcal{G}$, we rewrite the surrogate objective $J_n(G)$ using the law of total expectation:

$$J_n(G) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\frac{K_n\big(X, G(X,Y)\big)}{\psi(n)+\phi(n)}\mathrm{Vol}_G\big(G(X,Y), X\big)\right]\right].$$

Let $F_{G|X}$ denote the conditional cumulative distribution function of $g = G(X,Y)$ given X. By Assumption A.1, $F_{G|X}$ is absolutely continuous in the neighborhood of $q_\tau(X)$. Expressing the inner conditional expectation as an integral with respect to $F_{G|X}$ yields:

$$\frac{1}{\psi(n)+\phi(n)}\int_{q_{\tau-\phi(n)}(X)}^{q_{\tau+\psi(n)}(X)} \mathrm{Vol}_G(g, X)\, dF_{G|X}(g).$$

Using the Lipschitz condition in Assumption A.1, we get

$$\begin{aligned}
\sup_{G\in\mathcal{G}} |J_n(G) - J(G)| &\leq \sup_{G\in\mathcal{G}} \mathbb{E}_X\left[\frac{1}{\psi(n)+\phi(n)}\int_{q_{\tau-\phi(n)}}^{q_{\tau+\psi(n)}} |\mathrm{Vol}_G(g,X) - \mathrm{Vol}_G(q_\tau(X),X)|\, dF_{G|X}(g)\right] \\
&\leq \sup_{G\in\mathcal{G}} \mathbb{E}_X\left[\frac{1}{\psi(n)+\phi(n)}\int_{q_{\tau-\phi(n)}}^{q_{\tau+\psi(n)}} L(X)|g - q_\tau(X)|\, dF_{G|X}(g)\right] \\
&\leq \sup_{G\in\mathcal{G}} \mathbb{E}_X\left[L(X)\cdot\max\Big(\big|q_{\tau+\psi(n)}(X) - q_\tau(X)\big|, \big|q_{\tau-\phi(n)}(X) - q_\tau(X)\big|\Big)\right],
\end{aligned}$$

where we used $\displaystyle\int_{q_{\tau-\phi(n)}}^{q_{\tau+\psi(n)}} dF_{G|X}(g) = \psi(n) + \phi(n)$.

Under the uniform density bound (Assumption A.1.1), the inverse function theorem implies that the conditional quantile function is locally Lipschitz with respect to the probability level, with constant $1/c(X)$. Consequently, for any sufficiently small perturbation ϵ, we have $|q_{\tau+\epsilon}(X) - q_\tau(X)| \leq \epsilon/c(X)$. Applying this to the interval boundaries yields:

$$\max\Big(\big|q_{\tau+\psi(n)}(X) - q_\tau(X)\big|, \big|q_{\tau-\phi(n)}(X) - q_\tau(X)\big|\Big) \leq \frac{\max(\phi(n), \psi(n))}{c(X)}.$$

Substituting this bound into our supremum expectation, we obtain:

$$\sup_{G\in\mathcal{G}} |J_n(G) - J(G)| \leq \sup_{G\in\mathcal{G}} \mathbb{E}_X \left[L(X) \frac{\max(\phi(n), \psi(n))}{c(X)} \right].$$

The term inside the expectation is now independent of G. Factoring out the deterministic sequences gives:

$$\sup_{G\in\mathcal{G}} |J_n(G) - J(G)| \leq \max(\phi(n), \psi(n)) \cdot \mathbb{E}_X \left[\frac{L(X)}{c(X)} \right].$$

By Assumption A.1.3, $\mathbb{E}_X[L(X)/c(X)]$ is finite. As $n \to +\infty$, we have $\phi(n) \to 0$ and $\psi(n) \to 0$, which forces the right-hand side to vanish. This establishes the uniform convergence:

$$\lim_{n\to+\infty} \sup_{G\in\mathcal{G}} |J_n(G) - J(G)| = 0.$$

For the second step, we start by showing the existence of $G_n^* \in \arg\min_{G\in\mathcal{G}} J_n(G)$ for all n.

Because J_n is a continuous real-valued functional operating on the compact topological space $\mathcal{G}$, the extreme value theorem guarantees that it attains its global infimum. Consequently, the set $\arg\min_{G\in\mathcal{G}} J_n(G)$ is non-empty, meaning there exists at least one optimal solution $G_n^* \in \mathcal{G}$ such that $J_n(G_n^*) = \inf_{G\in\mathcal{G}} J_n(G)$.

Let $(G_n^*)_{n\in\mathbb{N}}$ be a sequence of optimal solutions, such that $G_n^* \in \arg\min_{G\in\mathcal{G}} J_n(G)$. Because $\mathcal{G}$ is a compact topological space, this sequence has at least one limit point. Let G^* be such a limit point, meaning there exists a subsequence $(G_{n_k}^*)_{k\in\mathbb{N}}$ that converges to G^*. We aim to show that $G^* \in \arg\min_{G\in\mathcal{G}} J(G)$.

By the definition of the minimizer, for any arbitrary function $G \in \mathcal{G}$, we have:

$$J_{n_k}(G_{n_k}^*) \leq J_{n_k}(G).$$

We take the limit of both sides as $k \to +\infty$. For the right-hand side, the uniform convergence of J_n implies pointwise convergence, so $\lim_{k\to\infty} J_{n_k}(G) = J(G)$.

For the left-hand side, we apply the triangle inequality:

$$|J_{n_k}(G_{n_k}^*) - J(G^*)| \leq |J_{n_k}(G_{n_k}^*) - J(G_{n_k}^*)| + |J(G_{n_k}^*) - J(G^*)|.$$

The first term on the right vanishes as $k \to +\infty$ due to the uniform convergence of J_n to J. The second term vanishes because J is continuous and $G_{n_k}^* \to G^*$.

Thus, $\lim_{k\to\infty} J_{n_k}(G_{n_k}^*) = J(G^*)$. Substituting these limits back into our inequality yields:

$$J(G^*) \leq J(G) \quad \forall\, G \in \mathcal{G}.$$

This concludes the proof that any limit point G^* of the sequence of minimizers is an optimal solution to the true constrained problem. □

Proposition A.3. *Let $\Omega \subset \mathbb{R}^d$ be a compact domain with a smooth boundary, and assume Ω is diffeomorphic to a closed ball. Then, there exists a volume-preserving diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$, a $d \times d$ lower triangular matrix L with strictly positive diagonal entries, a vector $\mu \in \mathbb{R}^d$ and a constant $q > 0$ such that the target set Ω is exactly recovered by:*

$$\Omega = \left\{ y \in \mathbb{R}^d : \left\| L(T(y) - \mu) \right\|_2^2 \leq q \right\}.$$

Proof. Without loss of generality, we can fix $L = I_d$ and $\mu = \mathbf{0}_d$. Let $r > 0$ be the unique radius such that $\mathrm{vol}(\bar{B}(0, r)) = \mathrm{vol}(\Omega)$, i.e., $r = \left(|\Omega|/V_d\right)^{1/d}$, where V_d denotes the volume of the unit d-ball, and set $q := r^2$.

The existence of the mapping T follows from the Dacorogna-Moser theorem (Dacorogna and Moser, 1990, Theorem 1) extended to the whole space $\mathbb{R}^d$ with control of support. Specifically, the proof is analogous to the construction in (Teixeira, 2017, Theorem 1 & Example 1), where the initial domain is Ω and the target domain is the standard Euclidean ball $\bar{B}(0, r)$ of equal volume. Because Ω is diffeomorphic to this closed ball and has the same volume, there exists a global volume-preserving diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T(\Omega) = \bar{B}(0, r)$. □

B Alternative Volume Objectives and Theoretical Equivalences

B.1 Single-flow Mahalanobis frontier

Imposing a determinant constraint. An alternative approach to formulating the objective in Section 4 is to strictly constrain the determinant of the shape matrix. While not the exact method used across all of our experiments, enforcing $\det(L_\theta(X)) = 1$ almost surely for all X acts as a scalar normalization that is seamlessly absorbed by the boundary threshold $q(X)$. Consequently, assuming the response space is d-dimensional, the volume of the candidate region becomes strictly proportional to $q(X)^{d/2}$. Since the mapping $z \mapsto z^{d/2}$ is strictly monotonically increasing for $z \geq 0$, and the optimization decomposes independently for every X, $\mathbb{P}_X$-almost surely; minimizing this expected volume is mathematically equivalent to minimizing the expected threshold $\mathbb{E}_X[q(X)]$ subject to the quantile constraint, a task natively handled by Algorithm 1.

Learning the $\log$-volume objective for stability. We empirically suggest minimizing the expected $\log$ volume rather than the true volume when the output dimension is too large. Due to the quantile constraint, those two objectives are equivalent, because the optimization decomposes independently for every X, $\mathbb{P}_X$-almost surely. However, the $\log$ volume is more stable with the dimension.

Link with conditional density estimation. For elliptically contoured distributions, the highest density regions are natively ellipsoids. As noted in Remark 4.2, our unweighted initialization phase naturally pushes the mapped variable $T_\phi(Y; X)$ toward an elliptical distribution. Therefore, we can expect the loss function during this initialization phase to closely parallel maximum likelihood estimation. The following proposition formalizes this equivalence for a Gaussian distribution.

Proposition B.1. *Let $L(X) \in \mathbb{R}^{d\times d}$ be a parameterized lower triangular matrix, and let the conditional quantile of the scaled squared residuals be $q(X) = \mathrm{Quantile}_\tau(\|L(X)(T_\phi(Y; X) - f(X))\|_2^2)$. Minimizing the objective:*

$$\mathbb{E}_X\left[-\log\det(L(X)) + \frac{d}{2}\log q(X)\right]$$

is equivalent to minimizing the objective:

$$\mathbb{E}_X\left[-\log\det(L(X)) + q(X)\right]$$

up to an additive constant independent of L.

Proof. Consider the integrand of the second objective, $J(L(X)) = -\log\det(L(X)) + q(X)$. Because a scalar multiple of a lower triangular matrix is still a lower triangular matrix, optimizing over $L(X)$ implicitly optimizes over any arbitrary scaling of the matrix. Let us scale $L(X)$ by a strictly positive scalar function $c(X) > 0$, defining $\tilde{L}(X) = c(X)L(X)$.

We evaluate the loss integrand for this scaled matrix. By the properties of the determinant for a $d \times d$ matrix, we have $\det(\tilde{L}(X)) = c(X)^d \det(L(X))$, which yields

$$-\log\det(\tilde{L}(X)) = -d\log c(X) - \log\det(L(X))\,.$$

Furthermore, the squared Euclidean norm scales quadratically as $\|\tilde{L}(X)(T_\phi(Y; X) - f(X))\|_2^2 = c(X)^2\|L(X)(T_\phi(Y; X) - f(X))\|_2^2$. Since $c(X)^2$ is a positive constant conditionally on X, it factors out of the quantile operator, giving the scaled quantile $\tilde{q}(X) = c(X)^2 q(X)$.

Substituting these into the loss integrand yields

$$J(\tilde{L}(X)) = -d\log c(X) - \log\det(L(X)) + c(X)^2 q(X)\,.$$

To find the optimal scaling for any fixed $L(X)$, we take the derivative of $J(\tilde{L}(X))$ with respect to $c(X)$ and set it to zero. This gives $-\frac{d}{c(X)} + 2c(X)q(X) = 0$. Solving for $c(X)$ subject to the constraint $c(X) > 0$, we obtain the optimal scaling factor $c^*(X) = \sqrt{\frac{d}{2q(X)}}$.

We now substitute this optimal scale back into the expression for $J(\tilde{L}(X))$:

$$\min_{c(X)>0} J(c(X)L(X)) = -d\log\left(\sqrt{\frac{d}{2q(X)}}\right) - \log\det(L(X)) + \left(\sqrt{\frac{d}{2q(X)}}\right)^2 q(X)\,.$$

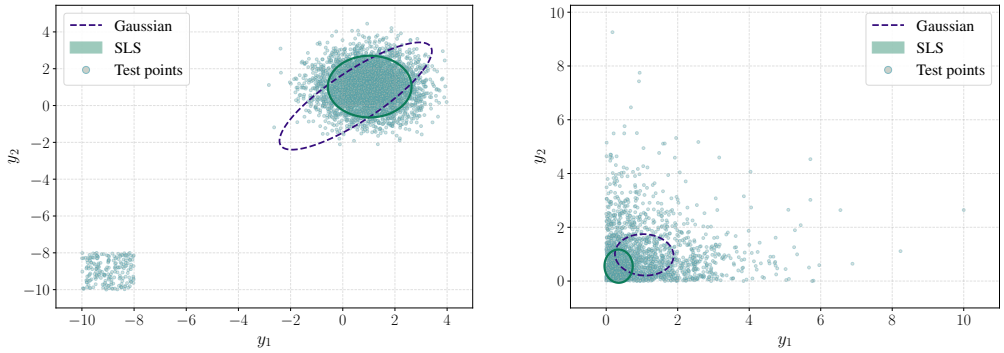

Figure B.1: **SLS regression on synthetic 2D distributions for a fixed covariate X.** All samples are drawn from the true conditional distribution $\mathbb{P}_{Y|X}$. We compare our approach which uses a single-flow frontier fixed to the identity mapping to enforce rigid ellipsoidal sets, against a standard baseline that minimizes the negative log-likelihood (NLL) of a Gaussian distribution. **Left:** A Gaussian distribution contaminated with 5% uniform outliers. For a target coverage of 70%, our shrinking window method successfully ignores the outliers, whereas the NLL baseline is severely distorted by them. **Right:** An asymmetric exponential distribution, illustrating a misspecified setting where the true highest density regions are not elliptical. Despite this structural restriction, for a target coverage of 30%, our approach extracts a significantly smaller and tighter ellipsoid achieving the desired probability mass compared to the Gaussian baseline.

Simplifying the first term gives $-\frac{d}{2}\log\left(\frac{d}{2q(X)}\right) = -\frac{d}{2}\log\left(\frac{d}{2}\right) + \frac{d}{2}\log q(X)$. The final term simplifies to exactly $\frac{d}{2}$. Combining these components, we obtain

$$\min_{c(X)>0} J(c(X)L(X)) = \left[-\log\det(L(X)) + \frac{d}{2}\log q(X)\right] + \frac{d}{2}\left(1 - \log\left(\frac{d}{2}\right)\right).$$

The bracketed expression is precisely the integrand of the first objective. The remaining terms form a constant that depends only on the dimension d and not on the data, the model $f(X)$, or the parameters of L. Taking the expectation over X, the two optimization problems differ only by this additive constant, completing the proof. □

To further contextualize this result, note that the first objective corresponds to minimizing the expectation of the log-volume of the estimated level sets. Because it is theoretically equivalent to the second objective, we can optimize it directly using the alternating scheme in Algorithm 1. During the initialization phase, our procedure minimizes the unweighted proxy:

$$\mathbb{E}_{X,Y}\left[-\log\det(L(X)) + \|L(X)(Y - f(X))\|_2^2\right],$$

which is functionally equivalent to minimizing the negative log-likelihood of a multivariate Gaussian distribution with precision matrix $\Sigma(X)^{-1} = 2L(X)^\top L(X)$.

Crucially, however, our overall framework fundamentally diverges from standard density estimation. Because we explicitly optimize for volume subject to a specific quantile constraint, our method isolates the target level set and ignores distant outliers that would otherwise skew standard maximum likelihood methods. This robustness is illustrated in Figure B.1, where we fix the flow to the identity mapping to recover rigid ellipsoids and further expose this phenomenon. While minimizing the negative log-likelihood fails due to the outliers or a non-elliptical distribution, our shrinking window approach successfully ignores the irrelevant probability mass to find the optimal minimum-volume ellipsoid.

B.2 Union of normalizing flows

To prevent trivial volume collapses, we constrain each shape matrix such that $\det(L_{\theta_k}(X)) = 1$. Under this constraint alone, the resulting predictive region would be restricted to a union of equally sized volumetric shapes. To represent the geometric union of these K flow-transformed ellipsoids

while allowing for varying local volumes, we introduce conditional mixture weights $p_k(X)$ satisfying $p_k(X) > 0$ and $\sum_{k=1}^{K} p_k(X) = 1$.

We define the individual frontier for the k-th component by scaling its latent squared Mahalanobis distance by $p_k(X)^{-2/d}$, where d is the dimensionality of $\mathcal{Y}$:

$$G_k(X, Y) = \left\| p_k(X)^{-\frac{1}{d}} L_{\theta_k}(X) \big(T_{\phi_k}(Y; X) - \mu_{\theta_k}(X) \big) \right\|_2^2 .$$

This score is equivalent to Equation (10), up to the update $L_{\theta_k}(X) \leftarrow p_k(X)^{-\frac{1}{d}} L_{\theta_k}(X)$. This parameterization effectively decouples the global volume allocation from the local geometric shapes dictated by $L_{\theta_k}(X)$. In practice, to prevent premature volume collapse of any single component during early training, we initially freeze the mixture weights uniformly at $p_k(X) = 1/K$ for a set number of epochs. Once the individual flows have stabilized, we unfreeze the weights, allowing the model to dynamically distribute volume across the learned shapes.

C Conformalizing the level sets

In practice, establishing distribution-free finite-sample guarantees for exact conditional coverage is fundamentally impossible without imposing strong assumptions on the underlying data distribution or resorting to trivial prediction sets (Foygel Barber et al., 2021). While our methodology targets conditional adaptivity by learning a single conditional quantile, a statistically more tractable task than estimating the full conditional distribution, this optimization alone cannot strictly guarantee conditional coverage. To provide rigorous statistical validity, we apply a post-hoc corrective procedure to guarantee marginal coverage (Papadopoulos et al., 2002). Specifically, we leverage a split conformal framework (Shafer and Vovk, 2008) using the normalized scores $G(X, Y)/q_\tau(X)$. This leads to prediction sets of the form $C_\tau(X) = \{y \mid G(X, y) \leq r \cdot q_\tau(X)\}$, where r ensures marginal coverage guarantees.

In practice, our method can also be deployed without conformal calibration when the estimator q_τ is sufficiently well learned and exhibits good empirical calibration. In such cases, the learned quantile alone may provide strong conditional adaptivity. The conformal correction serves as a reliable safeguard, acting as a post-hoc adjustment that guarantees marginal coverage even when q_τ is imperfectly estimated.

D Architecture

In this section, we provide the architectural details of the neural network components used in SLS regression, including the volume-preserving flow and the shape matrix parameterization, the quantile networks.

D.1 Frontier function architecture

The frontier function requires parameterizing both a volume-preserving mapping and a localized shape matrix (precision matrix) with a strictly unit determinant.

Conditional volume-preserving flow. The flow $T_\phi(Y; X)$ is constructed by composing multiple conditional additive coupling layers. At each layer, the input vector $y \in \mathbb{R}^d$ is partitioned into two components, y_1 and y_2. To ensure volume preservation (a Jacobian determinant of 1), we apply an affine transformation where only the translation is parameterized by a neural network (see, e.g., Dinh et al. 2014). Specifically, the forward pass for a single layer is defined as:

$$\begin{aligned} z_1 &= y_1 \\ z_2 &= y_2 + \mathcal{F}_\phi(y_1, X) - \mathcal{F}_\phi(\mathbf{0}, X), \end{aligned}$$

where $\mathbf{0}$ is a zero vector of the same dimension as y_1. Subtracting the component evaluated at the origin acts as a dynamic bias correction. The shift function $\mathcal{F}_\phi$ is parameterized by a multi-layer perceptron (MLP). To stabilize training and prevent overfitting, this MLP utilizes layer normalization (Ba et al., 2016) and dropout (Srivastava et al., 2014) between its hidden layers, followed by ReLU

activations. The final linear projection is initialized with weights and biases set strictly to zero, ensuring that the flow acts as an identity mapping at the beginning of training. Successive layers alternate the masking split so that all dimensions are eventually transformed.

Precision matrix parameterization. The shape matrix $L_\theta(X)$ dictates the geometry of the Mahalanobis frontier. We explore two parameterization modes to scale effectively with the dimensionality d of the response space:

- **Full rank:** For moderate dimensions ($d < 5$), we predict all elements of the lower triangular matrix $L_\theta(\cdot)$. The diagonal elements of L are constrained to be strictly positive using a softplus activation.
- **Low-rank plus diagonal:** For higher dimensions, learning all parameters becomes computationally prohibitive and prone to overfitting. Instead, we use a low-rank approximation, defining the precision matrix as $L^\top L = D + VV^\top$, following the implementation of Braun et al. (2025) for a Gaussian model. Here, D is a positive diagonal matrix and $V \in \mathbb{R}^{d\times r}$ is a factor matrix with rank $r = \lceil\sqrt{d}\rceil$. The Mahalanobis distance in Equation (8) is evaluated directly as the quadratic form $z^\top(D + VV^\top)z$ and the volume surrogate is computed using $\sqrt{\det(D + VV^\top)}$.

D.2 Quantile network

The conditional quantile estimator $q_\omega(X)$ must simultaneously output the lower bound, the target, and the upper bound of the shrinking window. To encourage representation sharing and reduce computational overhead, we use a single shared MLP backbone. The backbone maps the covariates X to a hidden representation, which is then passed to three independent linear heads: one for $\tau - \phi(n)$, one for the target τ, and one for $\tau + \psi(n)$. Because the frontier function yields strictly positive distance metrics, the outputs of these linear heads are exponentiated (and shifted by a small $\epsilon = 10^{-6}$) to ensure the predicted quantiles are always strictly positive. We then rank the quantile predictions for each feature X to avoid quantile crossing and ensure consistency.

E Training strategy

E.1 Shrinking window scheduling

The optimization of the volume surrogate relies heavily on the design of the shrinking margin sequences $\phi(n)$ and $\psi(n)$. In our implementation, we parameterize these boundaries using an annealed logistic schedule. Following an initial warm-start phase of n_0 steps where the full volume is penalized, the window begins to shrink from an initial wide margin down to a tight minimum margin.

Let $t = n - n_0$ be the number of steps post-warm-up. The lower margin function is parameterized as:

$$\phi(n) = \text{error}_{\text{init}} + (\text{error}_{\text{min}} - \text{error}_{\text{init}}) \cdot \left(\frac{\sigma(k \cdot (t - t_0)) - \sigma(-k \cdot t_0)}{1 - \sigma(-k \cdot t_0)} \right),$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the sigmoid function, k controls the steepness of the decay, and t_0 is the target step center. This normalized sigmoidal decay ensures a smooth, differentiable transition from the exploratory warm-start phase to the strict exploitation of the target quantile. An identical functional form (with independent hyperparameters) is used for the upper bound schedule $\psi(n)$.

We do not force the decay schedules to shrink to exactly zero (i.e., $\phi(n) \to 0$ and $\psi(n) \to 0$). This relaxation of the initial objective is motivated by improved robustness of the learning procedure. If the window were perfectly thin, minor estimation errors in the quantile networks would cause the model to optimize an incorrect threshold. Furthermore, in a practical finite-sample setting, a zero-width window would result in exploding gradients for the data points that would fall within the targeted neighborhood. Instead, we constrain the decay schedules to converge to strictly positive minimum margins, $\phi(n) \to \varepsilon_{\text{low}}$ and $\psi(n) \to \varepsilon_{\text{high}}$. This results in the minimization of the expected volume integrated over a narrow probability band:

$$\mathbb{E}_X \left[\int_{\tau - \varepsilon_{\text{low}}}^{\tau + \varepsilon_{\text{high}}} \text{Vol}_G(g, X)\, dF_{G|X}(g) \right],$$

which effectively performs a uniform volume minimization across the localized quantile range $[\tau - \varepsilon_{\text{low}}, \tau + \varepsilon_{\text{high}}]$.

E.2 Early stopping

Because the surrogate objective in Equation (7) explicitly depends on the training step via the shrinking window, it cannot be reliably used as a criterion for early stopping. Instead, we select the model checkpoint that yields the minimum average volume on a hold-out validation set while satisfying the target marginal coverage. To ensure a fair volumetric comparison across epochs, we first apply a scalar calibration $q(X) \mapsto r \cdot q_\tau(X)$, where r is dynamically computed on the validation set to strictly guarantee marginal coverage, and subsequently measure the resulting adjusted volume.

E.3 Learning the τ-th quantile

While our alternating optimization scheme effectively aligns the geometric boundary with the target conditional quantile, the quantile network $q_\tau(X)$ is key to improving conditional coverage. To yield the most accurate final conditional highest density regions, once G has been learned, we perform a learning step using TabICLv2 (Qu et al., 2026) on the training set. By leveraging TabICLv2 on a the training set, we refine the final boundary threshold, ensuring a highly accurate estimate of the conditional quantile before extracting the final level set. To obtain a more accurate estimate of the true conditional quantile, one could instead learn this quantity on a held-out dataset. However, to avoid introducing additional methodological complexity and to keep the framework streamlined, we do not adopt this approach. As a consequence, the frontier function G may exhibit some degree of overfitting to the training data.

F Baselines

Among our baseline choices, we choose all strategies targeting conditional coverage from (Dheur et al., 2025, Table 1) in multivariate conformal prediction. The base model for conditional density estimation minimizes the negative log-likelihood of a multivariate quantile function forecaster (Kan et al., 2022), utilizing input convex neural networks (Amos et al., 2017) as implemented by Dheur et al. (2025).

- **C-HDR** (Izbicki et al., 2022): Conformalizes the highest predictive density (HPD) by using the nonconformity score $S_{\text{HDR}}(X, Y) = \mathbb{P}_{y \sim \hat{p}(\cdot|X)}\left(\hat{p}(y|X) \geq \hat{p}(Y|X) \right)$. It then produces regions $C_{\text{C-HDR}}(X) = \{y : \hat{p}(y|X) \geq \hat{t}_q\}$, where $\hat{t}_q$ defines the highest density region (HDR) at level $\hat{q}$.
- **CP2-PCP** (Plassier et al., 2025): Builds predictive sets by using samples from an implicit conditional generative model. For each calibration point it uses two independent draws from the conditional generator to define a conformity score and an inflation parameter τ that accounts for the conditional mass around likely outputs. At prediction time it forms a union of balls around new generated samples, with their size chosen to guarantee marginal validity while improving approximate conditional adaptivity.
- **L-CP** (Dheur et al., 2025): Defines conformity in a latent space using an invertible conditional generative model $\hat{Q} : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$. A latent variable $Z \sim \mathcal{N}(0, I_d)$ is mapped to the output space via $\hat{Q}$, and the conformity score is measured in latent space as

  $$S_{\text{L-CP}}(X, Y) = \|\hat{Q}^{-1}(Y; X)\|.$$

  The prediction region is obtained by taking a ball of radius $\hat{q}$ around the origin in latent space and mapping it back to the output space. This method avoids grid-based directional quantile regression, improving scalability and computational efficiency, and generalizes distributional conformal prediction to multivariate outputs.
- **C-PCP** (Dheur et al., 2025): Estimates the conditional CDF of the conformity score $S(X, Y)$,

  $$S_{\text{CDF}}(x, y) = \mathbb{P}(S_W(X, Y) \leq S_W(x, y) \mid X = x),$$

using a Monte Carlo approximation with K samples

$$S_{\text{ECDF}}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}[S_W(x, \hat{Y}^{(k)}) \leq S_W(x, y)], \quad \hat{Y}^{(k)} \sim \hat{F}_{Y|x}.$$

When $S(x, y) = S_{\text{PCP}}(x, y)$, this yields

$$S_{\text{C-PCP}}(x, y) = \frac{1}{K} \sum_{k \in [K]} 1\left\{ \min_{l \in [L]} \|\hat{Y}^{(k)} - \tilde{Y}^{(l)}\|_2 \leq \min_{l \in [L]} \|y - \tilde{Y}^{(l)}\|_2 \right\}.$$

- **PCP** (Wang et al., 2023): Draws L samples $\tilde{Y}^{(l)} \sim \hat{p}_{Y|x}$ with $\hat{p}_{Y|x}$ the estimated conditional distribution, and defines conformity as the distance to the nearest sample, $S_{\text{PCP}}(X, Y) = \min_{l \in [L]} \|Y - \tilde{Y}^{(l)}\|_2$; the corresponding region is a union of L balls centered at the sampled points.
- **RCP-DR-CP** and **RCP-DR-PCP** (Plassier et al., 2025): Rectified conformal prediction (RCP) improves the conditional coverage of a given conformity score S by estimating its conditional quantile $\hat{t}(X)$ at the target level, and by applying split conformal prediction to the rectified score $S(X, Y) - \hat{t}(X)$. We apply it to the density-based score $S_{\text{DR-CP}}(X, Y) = -\log \hat{p}(Y \mid X)$ of DR-CP (Dheur et al., 2025) (RCP-DR-CP) and to the score S_{PCP} (RCP-DR-PCP), using the code released by the authors and the same conditional density estimator as the other baselines. The calibration set is split into two halves: the first one is used to estimate $\hat{t}$, and the second one for the conformal calibration. Because RCP needs calibration data to learn $\hat{t}$, these two methods use 40%, 10%, 20% and 30% of the data for training, validation, calibration and test, instead of 60%, 5%, 5% and 30% for the other methods, so that all methods are evaluated on test sets of the same size.

G Additional experiments

Hardware. We ran the models on GPUs (NVIDIA V100). The experiment on the output dimension (Appendix G.2) and the sensitivity analysis (Appendix G.5) were run on CPUs.

G.1 Minimizing alternative functions of the conditional quantile

To demonstrate the generality of our framework, we evaluate its ability to minimize functions of the conditional quantile other than the induced volume. Specifically, following the experimental setup from Section 5, we adapt our approach to minimize the expected conditional median absolute deviation:

$$\inf_{f: \mathcal{X} \to \mathbb{R}} \mathbb{E}_X \left[\text{median}_{Y \sim \mathbb{P}_{Y|X}} \left(|Y - f(X)| \right) \right]. \tag{11}$$

We compare our method against two alternatives: (1) a naive batch approach that optimizes the objective solely over the data points responsible for the empirical median, and (2) a Conditional Value-at-Risk (CVaR) approach, which uses a bi-level optimization similar to ours but replaces the shrinking window with a strict tail indicator, $K(X, G(X, Y)) = \mathbb{1}_{\{G(X,Y) \geq q(X)\}}$. Because the true data-generating process is known in this synthetic setup, we also compute and display the theoretically optimal oracle strategy. Results are available in Figure G.1. These alternative are not proper baselines for this task, since they were not designed for it, but are just current alternatives (in particular, the goal of CVaR is to minimize the error on the outliers, which explains why it is not well adapted for this task).

Overall, our approach successfully and accurately estimates the true conditional median absolute deviation. In contrast, the CVaR baseline struggles in this asymmetric noise setting, as its tail-focused objective disproportionately penalizes distances to outliers. While the marginal relaxation approach initially appears effective when the data distribution is relatively well-behaved, the marginal nature of its predictions ultimately leads to vanishing gradients in the exact regions where the model makes the largest errors, a limitation clearly visible in the right-hand scenario.

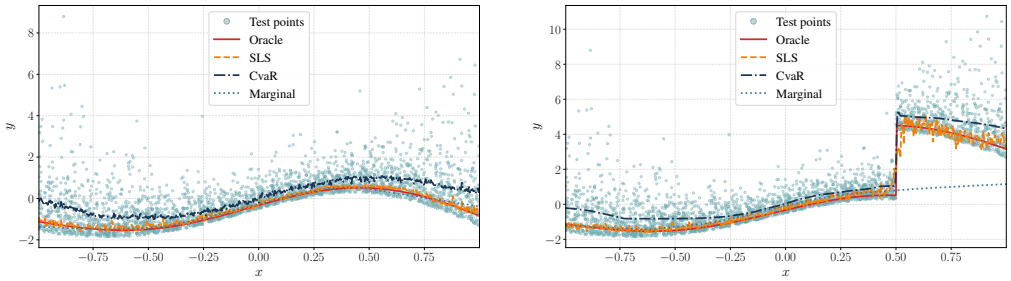

Figure G.1: **Minimization of the conditional median absolute deviation (Equation (11)).** The data are generated as $Y = f(X)+r(X)W$, where $W \sim \mathcal{E}(1)-1$, $X \sim \mathcal{U}([-1,1])$, and heteroscedasticity is introduced via $r(X) = X^2 + 0.5$. **Left:** Continuous base trend with $f(X) = \sin(3X)$. **Right:** Discontinuous base trend with $f(X) = \sin(3X) + 4\mathbb{1}_{\{X>0.5\}}$.

G.2 Robustness to the output dimension

We detail the experiment of Figure 4. The covariate is $X \sim \mathcal{U}([-1,1])$ and $Y = f(X) + A(X)W$, with $W \sim \mathcal{N}(0, I_d)$. The coordinates of the mean $f(X)$ alternate between cosine and affine functions of X, with non-linear couplings between consecutive coordinates, and $A(X) = L(X)L(X)^\top + 10^{-4}I_d$, where $L(X)$ is a lower-triangular matrix whose diagonal and sub-diagonal entries are smooth functions of X (softplus, sine and hyperbolic tangent); the exact expressions are given in the code. Since $Y \mid X \sim \mathcal{N}\big(f(X), A(X)A(X)^\top\big)$, the oracle HDR at level τ is the ellipsoid $\mathcal{C}^\star(X) = \{f(X) + A(X)w : \|w\|_2^2 \le \chi_d^2(\tau)\}$, where $\chi_d^2(\tau)$ is the τ-quantile of the chi-squared distribution with d degrees of freedom. Its volume is $\mathrm{Vol}(\mathcal{C}^\star(X)) = V_d\,\chi_d^2(\tau)^{d/2}\,|\det A(X)|$, where V_d is the volume of the unit ball of $\mathbb{R}^d$, and we report the average over the test points of $\big(\mathrm{Vol}(\mathcal{C}(X))/\mathrm{Vol}(\mathcal{C}^\star(X))\big)^{1/d}$, the ratio between the volume of our set and that of the oracle, normalized by the dimension.

We consider a target coverage $\tau = 0.95$ and ten output dimensions, logarithmically spaced between 2 and 100: $d \in \{2, 3, 5, 7, 11, 18, 27, 42, 65, 100\}$. For each dimension and each of the 10 seeds, we draw 5,000 training points, 1,000 validation points used for early stopping (Appendix E), 5,000 calibration points and 5,000 test points. We use the flow-based Mahalanobis frontier of Equation (8) with a single flow made of one coupling layer, a full-rank precision matrix for $d < 5$ and the low-rank-plus-diagonal parameterization otherwise (Appendix D), and the log-volume objective. The model is trained for 200 epochs of initialization phase, followed by 200 epochs of shrinking-window phase in which the margins ϕ and ψ decrease from 0.5 and 0.2 to 0.03, with a batch size of 256 and a learning rate of 5×10^{-4}. The quantile network is then refitted on the training and validation points with the frontier function fixed, and the sets are conformalized on the calibration points (Appendix C). For large d, the volumes exceed the range of single-precision floating-point numbers, so we compute them in log-space or in double precision.

G.3 Real data

Dataset information. Data are pre-processed using a quantile transformation on both the features (X) and responses (Y) using `scikit-learn` (Pedregosa et al., 2011), ensuring normalization across datasets. See Table G.1 for specific information on the datasets.

G.4 Results per dataset

For these experiments, we utilized a single-flow-based score with the model introduced in Appendix D with 3 layers. With this training protocol, using several flows did not improve the results, which suggests that a single flow is flexible enough for these datasets. The sensitivity analysis of Appendix G.5 shows that additional flows can reduce the volume on rf1 and rf2, but it relies on a lighter protocol, under which the single-flow model is less accurate: even with $K = 5$ flows, the volumes obtained on rf1 and rf2 at $\tau = 0.9$ (0.52 and 0.75) remain well above those of the single-flow model of this benchmark (0.36 and 0.47, Table G.2). The normalizing flows are parameterized using three layers.

Table G.1: Description of the multivariate datasets.

Dataset	Number of samples	Number of features	Dimension of targets
Bias (Cho et al., 2020)	7752	22	2
CASP (Rana, 2013)	45730	8	2
House (Pace and Barry, 1997)	21613	17	2
rf1 (Tsoumakas et al., 2011)	9125	64	8
rf2 (Tsoumakas et al., 2011)	9125	576	8
Taxi (Wang et al., 2023)	61286	6	2

Regarding the covariance parameterization, we adopted a full-rank structure for datasets with fewer than five outputs ($d < 5$), and a low-rank approximation with rank $r = \lceil \sqrt{d} \rceil$ otherwise, where d is the output dimension. For output dimensions $d \geq 5$, we minimize the $\log$ volume instead of the volume. All reported results are averaged across 10 independent trials. To prevent statistical anomalies from skewing the aggregated results, we discard extreme outlier values. Notably, while this instability never occurs with SLS regression, it can manifest in baselines that rely on sampling procedures, such as C-PCP or CP2-PCP. The RCP-based methods use a larger calibration set than the other methods, as detailed in Appendix F.

The performance outcomes are summarized for target coverage rates $\tau = 0.9$ in Table G.2 and $\tau = 0.5$ in Table G.3. These tables report two primary metrics: the volume of the prediction sets and the conditional coverage, evaluated via the ERT metric (Braun et al., 2026) using their default LightGBM classifier. Comparing the efficacy of the different strategies presents an inherent challenge due to the strict trade-off between efficiency and validity: a method achieving a smaller volume often does so at the cost of degraded conditional coverage. Consequently, both the volume and the ERT values must be considered jointly to accurately assess a model's performance.

Table G.2: Performance metrics (mean and standard error) per dataset for $\tau = 0.9$.

Dataset	Metric	SLS	C-HDR	PCP	CP2-PCP	L-CP	C-PCP	RCP-DR-CP	RCP-DR-PCP
biais	Volume	$\mathbf{1.11}_{0.02}$	$1.34_{0.03}$	$1.35_{0.01}$	$1.42_{0.02}$	$1.35_{0.03}$	$1.47_{0.02}$	$1.42_{0.01}$	$1.44_{0.02}$
	ERT	$0.04_{0.00}$	$0.02_{0.01}$	$\mathbf{0.01}_{0.01}$	$0.04_{0.00}$	$0.03_{0.01}$	$0.02_{0.00}$	$0.03_{0.00}$	$0.03_{0.00}$
rf1	Volume	$\mathbf{0.36}_{0.00}$	$0.46_{0.01}$	$0.49_{0.01}$	$0.49_{0.00}$	$0.48_{0.01}$	$0.49_{0.00}$	$0.48_{0.01}$	$0.53_{0.00}$
	ERT	$0.04_{0.00}$	$0.05_{0.00}$	$0.09_{0.00}$	$0.07_{0.00}$	$0.05_{0.01}$	$\mathbf{0.04}_{0.00}$	$0.08_{0.00}$	$0.08_{0.00}$
rf2	Volume	$0.47_{0.01}$	$0.44_{0.00}$	$0.44_{0.01}$	$0.98_{0.53}$	$\mathbf{0.43}_{0.01}$	$0.44_{0.01}$	$0.49_{0.01}$	$0.51_{0.01}$
	ERT	$0.05_{0.00}$	$0.05_{0.00}$	$0.07_{0.01}$	$0.07_{0.00}$	$0.05_{0.00}$	$\mathbf{0.04}_{0.00}$	$0.09_{0.00}$	$0.16_{0.08}$
casp	Volume	$\mathbf{1.15}_{0.01}$	$1.32_{0.01}$	$1.48_{0.01}$	$1.45_{0.01}$	$1.42_{0.01}$	$1.51_{0.01}$	$1.29_{0.01}$	$1.46_{0.00}$
	ERT	$0.02_{0.00}$	$0.02_{0.00}$	$0.02_{0.00}$	$0.04_{0.01}$	$0.02_{0.00}$	$\mathbf{0.01}_{0.00}$	$0.03_{0.00}$	$0.03_{0.00}$
house	Volume	$1.20_{0.01}$	$\mathbf{1.15}_{0.01}$	$1.22_{0.01}$	$1.42_{0.15}$	$1.27_{0.01}$	$1.80_{0.51}$	$1.16_{0.01}$	$1.32_{0.03}$
	ERT	$0.03_{0.00}$	$0.03_{0.00}$	$0.04_{0.00}$	$0.03_{0.00}$	$0.02_{0.00}$	$\mathbf{0.02}_{0.00}$	$0.04_{0.00}$	$0.04_{0.00}$
taxi	Volume	$2.77_{0.02}$	$2.65_{0.11}$	$3.46_{0.28}$	$5.26_{0.79}$	$3.08_{0.02}$	$3.25_{0.01}$	$\mathbf{2.28}_{0.18}$	$3.21_{0.02}$
	ERT	$0.01_{0.00}$	$0.01_{0.00}$	$0.01_{0.00}$	$0.03_{0.02}$	$\mathbf{0.01}_{0.00}$	$0.01_{0.00}$	$0.01_{0.00}$	$0.06_{0.05}$

G.5 Sensitivity to the hyperparameters

We study the sensitivity of SLS regression to the two hyperparameters introduced by the union of normalizing flows (Section 4.1): the number of flows K, and the schedule of the inverse temperature β of Equation (10), which is increased by a fixed increment at each training iteration. We use the six real datasets, a target coverage $\tau = 0.9$, and ten random splits of each dataset into 30%, 10%, 50% and 10% of the data for training, validation, calibration and test. Each flow is made of three coupling layers, the precision matrix uses the low-rank-plus-diagonal parameterization, and the model is trained with the $\log$-volume objective, first with the initialization phase and then with the shrinking window. After training, the quantile network is refitted with the frontier function fixed, and the sets are conformalized (Appendix C). Since the scale of the volume differs across datasets, we report the volume relative to a reference configuration evaluated on the same split, averaged over the splits.

Table G.3: Performance metrics (mean and standard error) per dataset for $\tau = 0.5$. The RCP-based methods use a larger calibration set (Appendix F).

Dataset	Metric	SLS	C-HDR	PCP	CP2-PCP	L-CP	C-PCP	RCP-DR-CP	RCP-DR-PCP
bias	Volume	$\mathbf{0.59}_{0.01}$	$0.69_{0.01}$	$0.77_{0.01}$	$0.79_{0.01}$	$0.70_{0.01}$	$0.80_{0.01}$	$0.70_{0.01}$	$0.82_{0.01}$
	ERT	$0.06_{0.00}$	$0.07_{0.00}$	$0.08_{0.00}$	$\mathbf{0.04}_{0.01}$	$0.07_{0.01}$	$0.05_{0.01}$	$0.07_{0.00}$	$0.05_{0.00}$
rf1	Volume	$\mathbf{0.26}_{0.00}$	$0.29_{0.00}$	$0.32_{0.01}$	$0.36_{0.00}$	$0.31_{0.00}$	$0.36_{0.00}$	$0.32_{0.00}$	$0.39_{0.00}$
	ERT	$\mathbf{0.07}_{0.00}$	$0.22_{0.01}$	$0.29_{0.00}$	$0.12_{0.00}$	$0.22_{0.00}$	$0.13_{0.01}$	$0.20_{0.00}$	$0.15_{0.00}$
rf2	Volume	$0.35_{0.01}$	$\mathbf{0.28}_{0.00}$	$0.30_{0.01}$	$0.33_{0.01}$	$0.28_{0.01}$	$0.33_{0.01}$	$0.32_{0.01}$	$0.37_{0.01}$
	ERT	$\mathbf{0.08}_{0.00}$	$0.20_{0.01}$	$0.26_{0.01}$	$0.12_{0.01}$	$0.20_{0.00}$	$0.12_{0.01}$	$0.21_{0.00}$	$0.17_{0.00}$
house	Volume	$0.66_{0.01}$	$\mathbf{0.52}_{0.00}$	$0.63_{0.01}$	$0.68_{0.01}$	$0.63_{0.01}$	$0.68_{0.01}$	$0.53_{0.01}$	$0.69_{0.01}$
	ERT	$0.05_{0.00}$	$0.09_{0.00}$	$0.13_{0.00}$	$0.04_{0.00}$	$0.08_{0.00}$	$\mathbf{0.04}_{0.00}$	$0.11_{0.00}$	$0.07_{0.00}$
casp	Volume	$0.60_{0.00}$	$\mathbf{0.47}_{0.06}$	$0.63_{0.06}$	$0.72_{0.01}$	$0.56_{0.06}$	$0.80_{0.14}$	$0.50_{0.05}$	$0.64_{0.07}$
	ERT	$\mathbf{0.03}_{0.00}$	$0.11_{0.00}$	$0.15_{0.00}$	$0.05_{0.00}$	$0.10_{0.00}$	$0.14_{0.06}$	$0.10_{0.00}$	$0.11_{0.04}$
taxi	Volume	$1.32_{0.01}$	$1.27_{0.01}$	$1.50_{0.10}$	$1.58_{0.13}$	$1.41_{0.15}$	$1.56_{0.14}$	$\mathbf{1.08}_{0.11}$	$1.52_{0.07}$
	ERT	$0.02_{0.00}$	$0.21_{0.12}$	$0.20_{0.08}$	$\mathbf{0.01}_{0.00}$	$0.03_{0.00}$	$0.33_{0.11}$	$0.04_{0.00}$	$0.07_{0.04}$

Number of flows. For each $K \in \{1, \ldots, 5\}$, the model is trained from scratch for 200 epochs of each phase, with an increment of β of 10^{-3}. Figure G.2 (left) shows that the volume varies by less than 4% across K on taxi, casp and house, so that additional flows neither help nor hurt on these datasets. On rf1 and rf2, whose output dimension is $d = 8$, additional flows reduce the volume, by 17% and 22% for $K = 5$; the resulting volumes nevertheless remain above those of the single-flow model of the benchmark, trained with a heavier protocol (Table G.2). On bias, they increase it by up to 19%. The training time grows linearly with K (Figure G.2, right): for $K = 5$, training is between 3.0 and 3.9 times longer than for $K = 1$. In the absence of prior knowledge on the number of modes, K can thus be selected like any other hyperparameter, by comparing the volumes obtained on a validation set, keeping in mind that the training cost grows with K.

Increment of β. A model with $K = 3$ flows is first trained for 250 epochs of initialization phase. For each increment of β in $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, a copy of this model is then trained for 250 epochs of shrinking-window phase, on the same mini-batches. Figure G.3 shows that the volume changes by at most 1% when the increment ranges over three orders of magnitude: the method is insensitive to this hyperparameter on these datasets.

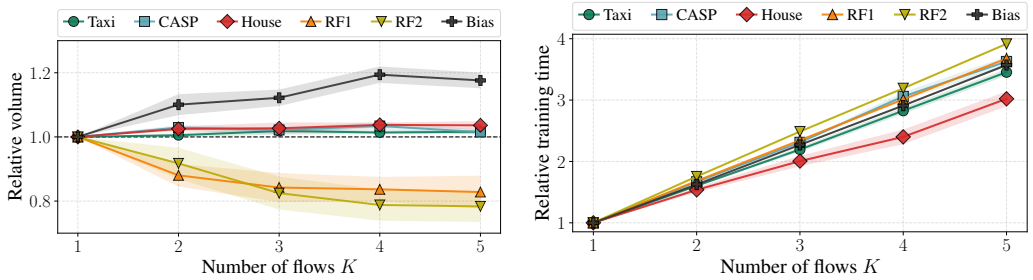

Figure G.2: **Sensitivity to the number of flows** K ($\tau = 0.9$). **Left:** Volume relative to the volume obtained with $K = 1$ on the same split. **Right:** Total training time relative to $K = 1$. Mean $\pm$ standard error over 10 splits.

G.6 Computational cost

Table G.4 reports the query time of each method, *i.e.*, the time needed to decide whether the test points belong to their prediction sets, and Table G.5 its total time, which adds the training time. For the baselines, the training time is that of the conditional density estimator; for SLS, it includes both training phases and the TabICLv2 estimate of the conditional quantile (Appendix E). All the methods of a run are timed on the same machine.

At query time, SLS is faster than the sampling-based methods C-HDR, PCP, C-PCP and CP2-PCP on bias, rf1, casp and house, by a factor between 5 and 46. On rf2 and taxi, it is slower than PCP, and

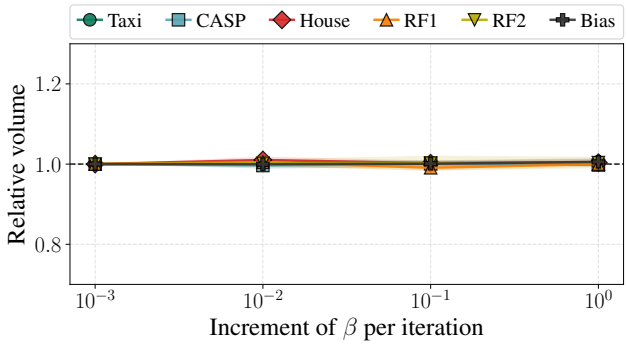

Figure G.3: **Sensitivity to the increment of the inverse temperature** β ($K = 3$, $\tau = 0.9$). Volume relative to the volume obtained with an increment of 10^{-3} on the same split. Mean $\pm$ standard error over 10 splits.

than C-HDR on taxi: its query time includes the prediction of the TabICLv2 quantile estimate, an in-context learner that uses the training set as context, whose cost grows with the number of training points and features. L-CP and RCP-DR-CP, whose scores do not require sampling from the density estimator, are the fastest at query time.

Training is currently the main cost of SLS: its total time is between 1.1 and 10.8 times that of the other methods. This comparison should however be interpreted with care, since it mostly reflects training budgets that were not tuned for speed. In our experiments, SLS is trained for a fixed number of epochs in each of its two phases (250, or 500 on rf1 and rf2), the best checkpoint being only selected a posteriori (Appendix E). The training time of SLS can be reduced directly, for instance by decreasing the number of epochs, by stopping the training once the validation volume no longer improves, or by increasing the batch size together with the learning rate, possibly at the cost of slightly larger sets. Conversely, a single density estimator can be shared by several baselines. A fair comparison of training costs would thus require tuning the budget of every method for a given accuracy, which is beyond the scope of this work.

Table G.4: Query time (in seconds), *i.e.*, time needed to decide whether each test point belongs to its prediction set, averaged over the runs and $\tau \in \{0.5, 0.9\}$.

Dataset	SLS	C-HDR	PCP	CP2-PCP	L-CP	C-PCP	RCP-DR-CP
biais	1.40	14.63	13.58	37.52	0.05	40.65	1.05
rf1	5.94	43.44	38.00	94.06	0.04	107.23	1.88
rf2	42.00	43.34	36.62	92.53	0.06	120.37	0.60
casp	18.64	94.78	91.46	458.40	0.03	280.98	2.68
house	3.99	46.03	38.19	182.74	0.15	110.27	0.95
taxi	250.03	125.26	116.50	320.22	0.05	324.01	3.60

Table G.5: Total time (in seconds), *i.e.*, training time plus query time, averaged over the runs and $\tau \in \{0.5, 0.9\}$.

Dataset	SLS	C-HDR	PCP	CP2-PCP	L-CP	C-PCP	RCP-DR-CP
biais	543.17	114.28	113.23	137.17	99.70	140.29	131.50
rf1	986.52	475.64	470.19	526.26	432.23	539.43	434.07
rf2	770.42	642.53	635.81	691.72	147.89	719.56	148.43
casp	2842.67	411.20	407.88	774.82	316.44	597.40	319.10
house	1558.61	201.69	193.85	338.39	155.80	265.93	144.43
taxi	6860.97	995.87	987.11	1190.82	870.66	1194.62	874.21

H Declaration of LLM usage

In this work, we used generative AI tools to provide feedback on theoretical results and experiments. Specifically, we asked Gemini 3 Flash to provide feedback on the first version of the paper. The same model has been used to help for coding experiments, including support qualitative and thematic data analysis and help interpret results, after a first human analysis performed by the first author. For the mathematical contributions, AI has been used to verify our proofs and claims, help gather references for existing proof arguments (such as in Theorem A.3). The core contributions of the paper only come from human ideas. Additionally, we used generative AI tools to generate code for matplotlib figures. We used LLM feedback to identify initial paper weaknesses and provided human responses to those points.

We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, clean and reformat datasets, or prove our mathematical claims.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the first author.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Experiments and theoretical guarantees support our claims.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is discussed in the conclusion and in the main text.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs are in appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the code alongside our paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data are available on dedicated website, and on our repository.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: This is done in Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All standard errors are reported in Appendix.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Those information are available in Appendix.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This project conform the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact for this work as it tackles uncertainty quantification.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This is done in Appendix.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code comes with explanation.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Declaration is in Appendix.