Relaxed Quantile Regression: Prediction Intervals for Asymmetric Noise

Thomas Pouplin [*1] **Alan Jeffares** [*1] **Nabeel Seedat** [1] **Mihaela van der Schaar** [1]

Abstract

Constructing valid prediction intervals rather than point estimates is a well-established approach for uncertainty quantification in the regression setting. Models equipped with this capacity output an interval of values in which the ground truth target will fall with some prespecified probability. This is an essential requirement in many real-world applications where simple point predictions' inability to convey the magnitude and frequency of errors renders them insufficient for high-stakes decisions. Quantile regression is a leading approach for obtaining such intervals via the empirical estimation of quantiles in the (non-parametric) distribution of outputs. This method is simple, computationally inexpensive, interpretable, assumption-free, and effective. However, it does require that the specific quantiles being learned are chosen a priori. This results in (a) intervals that are arbitrarily symmetric around the median which is sub-optimal for realistic skewed distributions, or (b) learning an excessive number of intervals. In this work, we propose Relaxed Quantile Regression (RQR), a direct alternative to quantile regression based interval construction that removes this arbitrary constraint whilst maintaining its strengths. We demonstrate that this added flexibility results in intervals with an improvement in desirable qualities (e.g. mean width) whilst retaining the essential coverage guarantees of quantile regression.

1. Introduction

Reliable uncertainty estimation is an essential requirement for safely and robustly deploying neural networks in real-world applications (Amodei et al., 2016; Dietterich, 2017; Kompa et al., 2021). However, research has consistently shown this to be a challenging problem in practice (Guo et al., 2017; Yao et al., 2019; Ayhan & Berens, 2022). Therefore, significant efforts have been made to address this task in order to contribute towards more reliable and trustworthy models (see e.g. Gawlikowski et al. (2023)). A significant aspect of this effort is developing regression methods that output predictive intervals rather than point predictions. This has proven to be a crucial requirement in high-stakes applications including medical decision-making (Begoli et al., 2019), autonomous driving (Su et al., 2023), and energy forecasting (Wang et al., 2022).

Especially in the case of neural networks, *quantile regression* (Koenker & Bassett Jr, 1978) has emerged as a powerful method for obtaining such intervals. This approach requires the model to output estimates of two quantiles rather than a single point prediction, which is easily optimized in practice by a simple change in loss function. These quantiles may then be used to construct an interval (μ_1, μ_2) within which the true label will lie with probability α (a formal description of quantile regression is provided in Section 2). Obtaining predictive intervals via quantile regression has earned substantial popularity in both research and practice (Koenker & Hallock, 2001; Koenker, 2017; Yu et al., 2003; Fitzenberger et al., 2001). This uptake can be attributed to several factors, including (a) methodological simplicity requiring minimal changes to the modeling procedure, (b) negligible increased computational cost (in contrast to e.g. ensemble methods (Lakshminarayanan et al., 2017)), (c) a simple, easily interpreted characterization of uncertainty (Savelli & Joslyn, 2013; Goodwin et al., 2010), (d) lack of parametric assumptions on the data-generating process, and (e) enduring empirical effectiveness (Chung et al., 2021; Tagasovska & Lopez-Paz, 2019). Furthermore, quantile regression methods can be wrapped in the conformal prediction procedure of Vovk et al. (2005) to additionally provide finite sample coverage guarantees, as demonstrated in Romano et al. (2019).

This work primarily focuses on the standard task of outputting predictive intervals that obtain a *prespecified* level of coverage (but the method could also be applied to the related non-prespecified task e.g. Chung et al. (2021)). This problem does not have a unique solution (consider e.g. trivial solutions in which some percentage of predic-

[*]Equal contribution [1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK. Correspondence to: Thomas Pouplin <tp531@cam.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

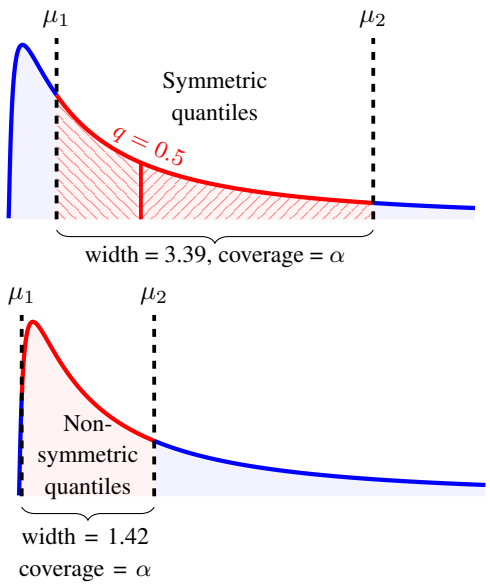

Figure 1. **Symmetric quantiles.** We compare two pairs of intervals on an identical (non-symmetric) log-normal probability distribution where in both cases a fixed level of coverage α is obtained. In the upper figure, the intervals are selected to be symmetric in terms of probability mass around the median $q = 0.5$ (i.e. the two lined regions and contain equal probability mass) as in the case of quantile regression. In the lower figure, we remove this constraint and obtain a much narrower interval with identical coverage.

tions are given infinite width intervals). Therefore, several variations of quantile regression have been introduced in recent years (see Section 2) which are typically also evaluated based on *additional* desirable properties of their resulting prediction intervals. These include minimizing interval width and achieving improved conditional coverage (see e.g. Pearce et al. (2018); Feldman et al. (2021) respectively, and further discussion in Section 2).

In this work our **contributions** are threefold: **(1)** In Section 3.1 we identify a substantial inefficiency in the standard procedure of first estimating quantiles from which predictive intervals are then derived. We show that this typically results in intervals with their midpoint fixed at the median which is undesirable for non-symmetric distributions (Figure 1) *or* requires learning more quantiles than necessary resulting in a more complex learning problem and, therefore, sub-optimal performance (see e.g. SQR in Section 2); **(2)** In Section 3.2 we propose a novel objective which directly learns intervals *without* a priori specifying particular quantiles. We then equip this function with a regularization term that aids in selecting among possible interval choices by rewarding desirable properties such as narrower intervals or improved conditional coverage. This results in a method we term Relaxed Quantile Regression (RQR); **(3)** We theoretically show that the solution of our proposed objective achieves valid coverage in expectation with bounded variance (Section 3.2). Empirically, we find that it results in superior performance to existing methods when evaluated on standard benchmarks (Section 4).

2. Background

In this section, we provide a summary of relevant existing works that convert neural networks from outputting point estimates to outputting predictive intervals (i.e. *single model approaches*). In Appendix A we provide a broad summary of predictive interval generation more generally and highlight some of the unique advantages of the single model approach that we consider in this text.

Deriving intervals from quantiles. Throughout this work we consider the standard regression task consisting of input/target pairs $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ with $d \in \mathbb{N}$ where bold denotes vectors and non-bold denotes scalars. We express realizations of these random variables (i.e. data) using lower-case $(\mathbf{x}, y)$. Denoting the cumulative distribution function of a probability distribution with $\mathbb{F}$, we recall that a *quantile function* is given by $\mathbb{F}^{-1}(p) = \inf\{q \in \mathbb{R} : p \leq \mathbb{F}(q)\}$ where $p \in (0, 1)$ is the desired *quantile probability*. Specifically, this provides some *quantile value* q such that $\mathbb{P}(Y \leq q) = p$. In the machine learning setting, we are generally interested in the probability distribution of Y *conditional* on a given input $\mathbf{X} = \mathbf{x}$. Throughout this work, we refer to the task of estimating a quantile value q corresponding to a particular quantile probability p from data as *estimating quantiles*. Once we have some function $\mu : \mathbb{R}^d \to \mathbb{R}$ for estimating quantiles (e.g. a neural network), we might wish to construct a conditional *interval* such that $\mathbb{P}(\mu_1(\mathbf{x}) \leq Y \leq \mu_2(\mathbf{x})) = \alpha$ with $\mu_1(\mathbf{x}) < \mu_2(\mathbf{x})$[1] and $\alpha \in (0, 1)$. In other words, a pair of bounds between which the target will lie with some desired probability α. We will generally drop the dependence of μ_1 and μ_2 on $\mathbf{x}$ for ease of notation. Clearly, this interval can be easily derived from the quantile function by simply noting that $\mathbb{P}(\mu_1 \leq Y \leq \mu_2) = \mathbb{P}(Y \leq \mu_2) - \mathbb{P}(Y \leq \mu_1)$. Therefore the problem of constructing valid intervals may be solved by approximating the quantile function to estimate the appropriate quantiles *and then* constructing an interval from these quantile values. However, this assumes that we know which quantile probabilities we should use in advance as any pair of quantile probabilities $p_l < p_u$ such that $p_u - p_l = \alpha$ will result in an interval with α coverage. Therefore, existing approaches typically select symmetric quantile probabilities or estimate an infinite number

[1] As a minor technical note, when a method is *permutation invariant* (as with our proposed method described in Section 3.2), this expression becomes $\mathbb{P}(\min(\mu_1(\mathbf{x}), \mu_2(\mathbf{x})) \leq Y \leq \max(\mu_1(\mathbf{x}), \mu_2(\mathbf{x}))) = \alpha$ and thus no longer assumes one bound to be greater than the other.

Table 1. **Quantile regression methods.** Several desirable properties of these methods are considered. Note that ½ denotes partially satisfying a property as the SQR & IR objectives can trade-off between coverage and interval width but do not directly consider conditional coverage.

Property	QR	SQR	IR	Ours
Suitable for non-centered distributions	✗	✓	✓	✓
Avoids explicitly learning all quantiles	✓	✗	✓	✓
Dynamically controls the trade-off between any desirable objectives	✗	½	½	✓
Asymptotic coverage guarantees	✓	✓	✗	✓
No Gaussian approx. or assumption of iid miscoverage of instances	✓	✓	✗	✓

of quantile probabilities from which intervals can be constructed. We discuss the consequences of this fact in Section 3.1.

Quantile Regression (QR). Here we refer to such methods that aim to provide intervals by accurately estimating quantiles. In the case of neural networks, the key distinction from standard point estimation methods which predict the expected value $\mathbb{E}(Y|\mathbf{X})$ lies in the choice of loss function. We require a loss function $\mathcal{L} : \cdot \to \mathbb{R}^+$ mapping from a quantile estimate to a scalar loss upon which we can apply gradient descent. Perhaps the most widely known approach is that of the *pinball loss* function (also known as *quantile loss*) of Koenker & Bassett Jr (1978); Steinwart & Christmann (2011). For a quantile estimator $\mu : \mathbb{R}^d \to \mathbb{R}$, the pinball loss expression[2]

$$\rho_q(\mu, \mathbf{x}, y) = \begin{cases} q(y - \mu(\mathbf{x})) & \text{if } y - \mu(\mathbf{x}) \geq 0 \\ (q-1)(y - \mu(\mathbf{x})) & \text{if } y - \mu(\mathbf{x}) < 0\,. \end{cases}$$

Then a strategy to construct an interval of targeted coverage level of α involves estimating two specific conditional quantiles, denoted as q_l and q_u, where q_l corresponds to the $\frac{1-\alpha}{2}$ quantile, and q_u corresponds to the $1-\frac{1-\alpha}{2}$ quantile. Thus, the loss function optimized by the neural network is given by

$$\mathcal{L}_\alpha^{\text{QR}}((\mu_1, \mu_2), \mathbf{x}, y) = \rho_{\frac{1-\alpha}{2}}(\mu_1, \mathbf{x}, y) + \rho_{1-\frac{1-\alpha}{2}}(\mu_2, \mathbf{x}, y).$$

This methodology ensures that the probability of the ground truth target y falling within the interval $[\mu_1, \mu_2]$ is α, thereby establishing the desired mean coverage. In practice, two particular quantiles are typically predefined and learned using a single neural network with two outputs.

[2]We note that the Winkler Score (or interval score) objective (Dunsmore, 1968; Winkler, 1972), which also regularly appears in the literature, is proportional to the QR objective as shown in Appendix B and learns all quantiles like SQR.

We refer to this antecedent approach as *quantile regression (QR)* throughout this work.

Simultaneous Quantile Regression (SQR). Rather than predefining two particular quantiles, Tagasovska & Lopez-Paz (2019) propose to learn *all* possible quantiles with a single output model by augmenting the neural network with an additional input for the desired quantile. We express this simultaneous quantile regressor as $\mu_q(\mathbf{x})$. Throughout training the quantile q is stochastically selected from a uniform distribution where any quantile loss function may be applied (e.g. pinball).

A model trained using SQR is underspecified in the sense that there are infinite quantile pairs that can be used to construct a valid interval of a given coverage level at test time. Therefore we consider two strategies for selecting a particular interval: (a) *SQR-C* selects the centered interval (as in QR) assuming that jointly learning all quantiles will produce a more accurate estimator; (b) *SQR-N* selects the pair of (potentially non-centered) quantiles that produce the narrowest interval.

Interval Regression (IR). Learning the intervals directly without the intermediate step of first learning quantiles is an alternative approach that has emerged somewhat independently of the quantile regression literature. A method proposed in Pearce et al. (2018) achieves best-in-class empirical performance by introducing a loss function that attempts to balance coverage with interval width. By making the strong assumptions that (a) the cases of miscoverage are *iid* and (b) batch sizes are sufficiently large for the binomial distribution to be well approximated by a Gaussian, the authors derive the following objective

$$\mathcal{L}_\alpha^{\text{IR}}((\mu_1, \mu_2), \mathbf{x}, y) = \frac{n}{c}(\mu_2 - \mu_1) \cdot \mathbb{1}_{\mu_1 \leq y \leq \mu_2} + \lambda \frac{n}{\alpha(1-\alpha)} \max(0, \alpha - \frac{c}{n})$$

Here $\lambda \in \mathbb{R}$ denotes a hyperparameter weighting term and n denotes the batch size, and $c := \sum_{i=1}^{n} \mathbb{1}_{\mu_1 \leq y \leq \mu_2}$. Unfortunately, unlike quantile regression methods (and our proposed method later in this work), it is unclear if this method achieves theoretical coverage guarantees.

Evaluation. Competing approaches to obtaining prediction intervals are typically compared across a range of well-established desirable properties. As we only have access to a finite dataset in practice, asymptotic coverage is not guaranteed resulting in the need to evaluate *calibration*. Ideally, we would produce intervals that accurately model the *conditional* probability $P(\mu_1(\mathbf{x}) < y < \mu_2(\mathbf{x})|\mathbf{x})$. Unfortunately, in the standard setting, we cannot directly estimate this quantity (Zhao et al., 2020) and instead often consider the *marginal* probability $P(\mu_1(\mathbf{x}) < y < \mu_2(\mathbf{x}))$.

We can estimate the marginal calibration on the test set using the Prediction Interval Coverage Probability (PICP) which simply measures the ratio of observations falling inside their intervals (Kuleshov et al., 2018; Tagasovska & Lopez-Paz, 2019).

Despite the aforementioned impossibility of exactly estimating the former conditional quantity, several proxy metrics have been proposed that test for independence between examples $\mathbf{x}$ and instances of miscoverage. These include using Pearson's correlation between interval width and miscoverage cases (Feldman et al., 2021) and the independence rewarding Hilbert-Schmidt independence criterion (HSIC) (Greenfeld & Shalit, 2020).

However, probably the most common criteria of evaluation considers aggregate *interval width*. For a fixed level of coverage, narrower intervals are typically considered preferable. This can also prevent trivial solutions with some potentially infinite width intervals, which may still satisfy empirical tests of coverage. Sometimes referred to as *sharpness* (Gneiting et al., 2007) or *adaptive coverage* (Seedat et al., 2023), we primarily consider Mean Prediction Interval Width (MPIW) which measures the mean interval width across the test data (Tagasovska & Lopez-Paz, 2019). We provide (a) some extended discussion on the motivation for these objectives in Appendix F and (b) a formal description of all evaluation metrics in Appendix D to ensure that this work is self-contained.

3. Relaxed Quantile Regression

3.1. Highlighting the Limitation of Existing Quantile Regression Methods

Using quantile estimation as a means for obtaining predictive intervals may be viewed as a victim of Vapnik's famous heuristic that "when solving a problem of interest, do not solve a more general problem as an intermediate step" (Vapnik, 2006). The fundamental limitation of estimating predictive quantiles as an intermediate step toward estimating predictive intervals becomes apparent by closely considering the quantile regression approach.

The standard quantile regression approach consists of selecting the two quantiles a priori such that the region between them results in a predictive interval with a desired level of coverage. In practice, for a desired coverage level α, the standard approach is to select the $\frac{1-\alpha}{2}$ and $1-\frac{1-\alpha}{2}$ quantiles. However, selecting a specific pair of *non-symmetric* quantiles would require knowledge of the underlying noise distribution which is unknown. As illustrated in Figure 1, when the underlying noise distribution around Y is, in fact, non-symmetric this results in wider than necessary intervals due to being arbitrarily centered (in terms of probability mass) around the median[3]. On real-world tasks, we should expect non-symmetric noise distributions to be ubiquitous. This has been highlighted in previous work (Tagasovska & Lopez-Paz, 2019) and we further illustrate this by including histograms of the target distributions of popular, real-world datasets in Table 8 and their summary statistics in Table 6. Whilst the true noise distribution cannot be known on real data, the shapes of these empirical distributions suggest that perfect symmetry is a very strong assumption to hold over natural phenomena.

The existing resolution to this issue, as introduced by Tagasovska & Lopez-Paz (2019), is to learn *all* possible quantile probabilities in $(0, 1)$ (see e.g. SQR in Section 2). This is explicitly aimed at rectifying the aforementioned limitation as the authors note that it enables them to "model non-Gaussian, skewed, asymmetric, multimodal, and heteroskedastic aleatoric noise in data". The idea being that, once all quantiles are learned, any pair may be selected such that they satisfy α coverage in addition to other qualities (e.g. narrower intervals). However, this introduces a significantly more challenging learning problem of estimating an infinite number of quantiles rather than exactly two for each example which can negatively impact the performance of the underlying quantile estimator. Furthermore, it is not obvious how to select a specific interval at test time as simply selecting the narrowest valid interval is likely to induce a bias that can negatively impact empirical coverage. In our experiments in Section 4, we show that these drawbacks result in this approach generally achieving inferior performance when compared against the former approach despite its added flexibility.

In Section 3.2 we introduce an alternative approach that solves the interval estimation problem whilst removing the median-centering constraint. This is achieved by relaxing the requirement to select symmetric quantiles a priori from which intervals can be constructed and, instead, estimating a pair of potentially asymmetric intervals *directly*. Although the previous work of Pearce et al. (2018) also considers a direct interval regression approach, ours is the first to build upon the quantile regression literature, thereby maintaining a strong theoretical foundation and converging to a solution that achieves coverage guarantees in expectation. This is reflected in superior coverage when empirically evaluated in Section 4. Table 1 summarizes the key differences between our proposed method and those of previous works.

[3]Of course, if centered intervals are *required* then quantile regression is appropriate. However, this is unlikely to be a wise requirement in the prevalent case of a skewed target distribution.

3.2. Proposed Resolution: Relaxed Quantile Regression

Relaxed Quantile Regression (RQR). We begin by introducing a novel objective which *directly* learns intervals without the intermediate step of prespecifying quantiles (i.e. *relaxing* the quantile learning requirement). For a targeted coverage level α and a neural network outputting two interval bounds (μ_1, μ_2) we minimize

$$\mathcal{L}_\alpha^{\text{RQR}}((\mu_1, \mu_2), \mathbf{x}, y) = \begin{cases} \alpha\kappa & \text{if } \kappa \geq 0 \\ (\alpha - 1)\kappa & \text{if } \kappa < 0 \end{cases} \quad (1)$$
$$\text{with } \kappa = (y - \mu_1)(y - \mu_2).$$

The key intuition is that $\text{sign}(\kappa)$ informs us whether the target falls between the two bounds, thus allowing us to optimize for our desired coverage level α. This expression makes no assumptions about the shape of the target noise distribution. It does not require the interval bounds to be placed at specific quantile values, nor does it require the neural network to explicitly model all quantiles. As neither μ_1 nor μ_2 are explicitly tied to being the upper or lower bound, we select these bounds to be $\max(\mu_1, \mu_2)$ and $\min(\mu_1, \mu_2)$ respectively. This makes the expression *permutation invariant* and thus avoids the crossing interval problem faced by quantile regression methods (Park et al., 2022) (see Appendix E for further discussion on this point). Theorem 3.1 provides formal guarantees that the minimization of this loss function in expectation yields a valid interval i.e. the coverage rate α is achieved. Note that proofs for all theorems and propositions are provided in Appendix B.

Theorem 3.1 (RQR In-sample Coverage)**.** *For any random variable Y associated with an input x, $\forall\, \alpha \in [0, 1]$,*

$$(\mu_1^*(x), \mu_2^*(x)) = \underset{\mu_1, \mu_2}{\arg\min} \{\mathbb{E}_Y(\mathcal{L}_\alpha^{\text{RQR}}((\mu_1, \mu_2), x, Y))\}$$
$$\implies \mathbb{P}(\mu_1^*(x) < Y < \mu_2^*(x)) = \alpha.$$

Furthermore, in Theorem 3.3 we derive additional desirable properties of the RQR objective. Specifically, that it obtains correct coverage in expectation with a variance bound that decreases in proportion to the inverse data size (i.e. increasing data size implies lower variance). This proof relies on first showing that this objective achieves the correct finite in-sample coverage – which we provide in Theorem 3.2 of Appendix B.

Theorem 3.3 (RQR is Unbiased and Consistent)**.** *For any random variable Y associated with an input x, we consider N realizations of this random variable : $\{y_i\}_{i=1}^N$. $(\mu_{1,N}, \mu_{2,N})$ are the bounds of our estimator trained on these N samples. $\forall\, \alpha \in [0, 1]$, we name Q_N the absolute true miscoverage of our estimator*

$$Q_N = |\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \alpha|$$

Then $E[Q_N] = 0$ and the variance of Q_N is bounded by $\frac{1}{4N}$.

These theoretical properties ensure that the RQR expression is well-motivated and provides a suitable objective for the goal of generating well-calibrated prediction intervals in practice. We further illustrate this by analyzing the objective's gradients in Section 3.3.

Rewarding preferable solutions. Given the added flexibility of shifting intervals rather than bounding them around the median, there now exists a potentially infinite number of competing solutions that achieve the desired level of coverage. We note that the RQR expression is largely agnostic to solutions that obtain additional desirable properties such as narrower interval widths or improved conditional coverage which have been discussed extensively in previous works (see e.g. Feldman et al., 2021; Tagasovska & Lopez-Paz, 2019). Therefore, we would like to induce some preference among solutions. A natural strategy is to upweight preferable solutions (e.g. narrower intervals or improved conditional coverage) via an additive regularization term $\mathcal{R}$ and a scalar weighting term $\lambda \in \mathbb{R}^+$. Thereby we provide practitioners with the flexibility to choose whichever interval properties provide the most utility. Therefore, the complete RQR-$\mathcal{R}$ (regularized) objective takes the form

$$\mathcal{L}_\alpha^{\text{RQR-}\mathcal{R}}((\mu_1, \mu_2), \mathbf{x}, y) = \mathcal{L}_\alpha^{\text{RQR}}((\mu_1, \mu_2), \mathbf{x}, y) + \lambda \cdot \mathcal{R}. \quad (2)$$

Given this structure, we now introduce two specific choices for $\mathcal{R}$.

(1) RQR-W: Width Minimizing $\mathcal{R}$. As discussed in Section 2 & Appendix F, interval width is a principal criterion for evaluating methods for obtaining predictive intervals (Tagasovska & Lopez-Paz, 2019; Feldman et al., 2021; Romano et al., 2019). A direct approach for minimizing interval width is to penalize the sample-wise squared interval width such that

$$\mathcal{L}_\alpha^{\text{RQR-W}}((\mu_1, \mu_2), \mathbf{x}, y) = \mathcal{L}_\alpha^{\text{RQR}}((\mu_1, \mu_2), \mathbf{x}, y) + \lambda \frac{(\mu_2 - \mu_1)^2}{2}.$$

By integrating this penalty term, we aim to effectively navigate the landscape of potential solutions, encouraging the model to prioritize intervals of reduced length whilst still maintaining the targeted level of coverage. Analogous to the approach taken in Theorem 3.1, we can extend our analysis to consider this complete loss function. We find that the introduced penalty term induces a bias to the coverage of the interval estimator – the result is now $\mathbb{P}(\mu_1^* < Y < \mu_2^*) = \alpha - 2\lambda$. However, we can easily remove the

bias by modifying the targeted coverage rate. By choosing $\hat{\alpha} = \alpha + 2\lambda$, we obtain $\mathbb{P}(\mu_1^* < Y < \mu_2^*) = \hat{\alpha} - 2\lambda = \alpha$ as desired.

Theorem 3.4 (RQR-W In-sample Coverage)**.** *For any random variable Y associated with an input x, $\forall\, \alpha \in [0, 1]$,*

$$(\mu_1^*(x), \mu_2^*(x)) = \underset{\mu_1,\mu_2}{\arg\min} \{\mathbb{E}_Y(\mathcal{L}^{\textit{RQR-W}}_{\alpha+2\lambda}(\mu_1, \mu_2), x, Y))\}$$
$$\implies \mathbb{P}(\mu_1^*(x) < Y < \mu_2^*(x)) = \alpha$$

In the supplemental material we also show that, analogous to RQR, RQR-W obtains correct finite in-sample coverage (Theorem 3.5) and is unbiased and consistent (Theorem 3.6). An important additional benefit of the penalty term, presented in Proposition 3.7, is that it makes the loss function convex for a wide range of target distributions and λ. Hence, leading to a welcome additional result: the existence and uniqueness of its minimum.

Proposition 3.7 (Existence and Uniqueness of Solution)**.** *μ_1^{min} and μ_2^{max} denote the bounds of our optimization problem. For a target distribution Y with a cumulative distribution function that is k-Lipschitz continuous with $k < 1 + \frac{\alpha}{\mu_2^{max} - \mu_1^{min}}$, when $\lambda > max(0, \int_{\mu_1^{min}}^{\mu_2^{max}} d\mathbb{P}_Y(y) - \alpha)$, the minimum of $\mathcal{L}^{\text{RQR-W}}_{\alpha+2\lambda}$ exists and is unique.*

We later empirically verify (see Section 4) that this objective does perform well on real-world data.

(2) RQR-O: Width-Coverage Independence $\mathcal{R}$. As discussed in Section 2, whilst interval construction methods are most commonly evaluated based on their resulting interval width for a realized coverage level (sometimes referred to as the high-quality principle (Pearce et al., 2018)), alternative objectives may also be desirable. Feldman et al. (2021) introduced *orthogonal quantile regression* (OQR) which instead optimized for a notion of *conditional coverage* rather than minimizing interval width. Specifically, the authors introduce a regularization term that promotes independence between the size of the intervals and occurrences of a (mis)coverage event. They combine their proposed regularization term with QR and report significant gains in measures of conditional coverage. We can easily combine their term with our RQR instead by setting it as the regularization term $\mathcal{R}$ in Equation (2) where

$$\mathcal{R}(\cdot) = \left| \frac{\text{Cov}(\mathbf{w}, \mathbf{m})}{\text{Var}(\mathbf{w})\text{Var}(\mathbf{m})} \right|.$$

With $\mathbf{w}$ denoting the vector of interval widths where $w_i = |\mu_2(\mathbf{x}_i) - \mu_1(\mathbf{x}_i)|$ and $\mathbf{m}$ denoting the indicator vector of coverage events where $m_i = \mathbb{I}_{y_i \in [\mu_1(\mathbf{x}_i), \mu_2(\mathbf{x}_i)]}$ – both calculated on the training data. Thus, this regularization term can simply be interpreted as the Pearson correlation between the interval widths and instances of coverage or miscoverage[4]. We refer to this complete objective as RQR-O (orthogonal) and, as with the other objectives, we provide proof for its expected coverage in Appendix B.

Trading-off width and orthogonality. We note that there is typically a trade-off between minimizing interval width and maximizing conditional coverage. As we later observe in Figure 3, obtaining near optimal conditional coverage generally requires wider intervals than is strictly necessary for obtaining valid marginal coverage. We emphasize that in this work we are agnostic as to which qualities are preferable and, instead, we enable the practitioner to make this decision based on their specific application. In Section 4, we will empirically verify that our proposed objective behaves as expected and successfully utilizes its added flexibility to outperform benchmark methods at achieving their respective goals whilst maintaining empirical coverage.

3.3. Gradient Analysis of the RQR objective

We now further analyze how our proposed objective in Equation (1) behaves analytically by investigating a concrete setting. In particular, let us consider a specific interval for a given observation x and suppose that the bounds $(\mu_1, \mu_2) = (0, 1)$ are output by a model. We are now interested in how these two bounds will be updated by gradient descent if we observe various potential values of the ground truth label y. This is captured by the gradients $\frac{\partial \mathcal{L}}{\partial \mu_1}$ and $\frac{\partial \mathcal{L}}{\partial \mu_2}$ which describe the *direction* and *relative magnitude* in which each bound will move.

In Figure 2 (upper) we provide a plot capturing the analytic value of these respective gradients for RQR at a range of potential values of the ground truth y. Positive gradient values (i.e. the upper half of the subplot) indicate that a given bound shifts to the left on the number line while, conversely, negative gradient values (i.e. the lower half of the subplot) indicate that a given bound shifts to the right. Then the absolute values capture the magnitude of how much that bound will shift.

In the left-hand side region ($y \leq 0$) where the y values fall outside on the left of the interval, we note that both bounds will shift to the left, with the nearer (lower) bound moving by the greatest magnitude. Similarly, in the right-hand side region ($y \geq 1$) we observe the opposite where both bounds move to the right. Finally, when the target does fall inside the interval ($y \in (0, 1)$) the bounds are adjusted by a smaller magnitude and are narrowed. When applied collectively across the entire training set, this results in the intervals being widened for miscoverage events and narrowed for coverage events as we might desire.

[4]Note that taking the absolute value results in this term penalizing correlation between interval width and instances of *either* coverage or miscoverage (as we would desire).

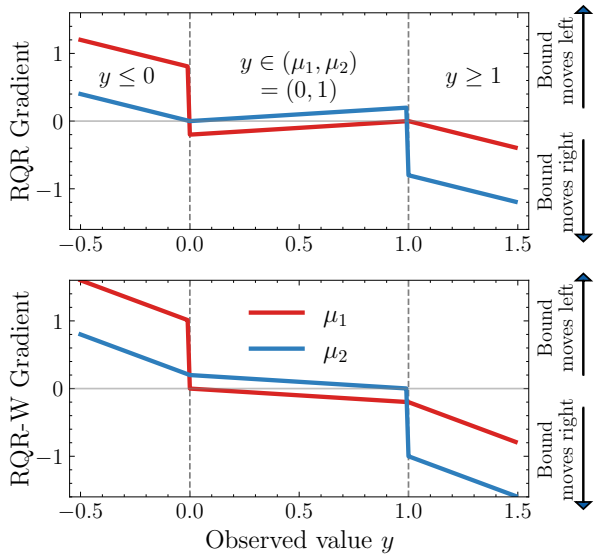

Figure 2. **Gradient analysis.** The gradients with respect to the RQR objective (*upper*) and RQR-W objective (*lower*) for a single x over a range of potential values of y. The current predicted bounds $(\mu_1, \mu_2) = (0, 1)$ are updated using these gradients.

In contrast, in Figure 2 (lower), when we include the width-reducing regularization term (i.e. the RQR-W objective with $\lambda = 0.1$) this plot changes. The key difference in this case is in the center region where the target falls inside the interval ($y \in (0, 1)$). As with the upper figure, the interval still narrows – but now the further bound from the target y has the larger gradient allowing the interval to narrow by a greater amount before a miscoverage event will occur. In other words, while previously the nearest bound to the target y narrowed the most, now the further bound narrows by a greater magnitude.

Given that this analysis shows a narrowing for coverage events and a widening for miscoverage events *for a single example*, we might ask how the solution of the respective objectives differ *in aggregate* across the entire training data in which we observe both cases. In Appendix G we provide a special case where we can derive a closed-form solution in which we find that the RQR-W objective does indeed obtain a narrower solution than RQR with equal coverage.

4. Experiments

Empirical verification.[5] We begin by empirically verifying the predicted gains of the RQR objective over quantile regression on non-symmetric noise distributions. As the noise distribution cannot be known on real-world data, here we generate synthetic data according to a known process. The data is generated according to a data-generating process in which the label is determined according to $Y = X + \epsilon$, where $X \in \mathbb{R}$ represents a deterministic component which is set as constant and $\epsilon \in \mathbb{R}$ is the noise component. Then the noise distribution is selected as either a (symmetric) Gaussian or a (non-symmetric) truncated Gaussian. We fit a simple linear neural network consisting of just a single layer. As illustrated in Table 2, the empirical results match our theoretical expectations. All methods perform equally well on the symmetric Gaussian noise where intervals centered at the median are optimal. However, QR fails to achieve optimal width on the truncated Gaussian due to being arbitrarily centered at the median. Whilst RQR (w/o reg) is unbiased, it is not sufficiently incentivized to produce narrower intervals and thus performs similarly to QR in that regard. In line with our analytic observations that motivated this objective, only RQR-W (with reg) achieves the optimally narrow interval solution.

[5]Code at https://github.com/TPouplin/RQR.

Table 2. **Empirical verification**. Coverage and width (± standard errors) are assessed for producing predictive intervals on symmetric and non-symmetric noise distributions with $\alpha = 0.8$. As expected, for realistic, skewed noise distributions, width-minimizing RQR (RQR-W) produces narrower intervals.

	Coverage	Width	Dist. Histogram
QR	0.81 (±0.007)	0.26 (±0.003)	
RQR (w/o reg)	0.81 (±0.003)	0.26 (±0.002)	
RQR-W (with reg)	0.82 (±0.004)	0.26 (±0.002)	
QR	0.80 (±0.002)	1.61 (±0.007)	
RQR (w/o reg)	0.80 (±0.007)	1.72 (±0.010)	
RQR-W (with reg)	0.81 (±0.002)	1.50 (±0.005)	

Applying RQR-W & RQR-O in practice. We now turn our attention to applying our proposed interval prediction approach to practical tasks in high-stakes domains. To illustrate, we consider a task in the robotics domain in which safety and reliability are essential, especially when applied in close proximity to humans (Baek & Kröger, 2023). We consider the task of estimating a robot arm-effector's distance from a target given noisy measurements of inputs such as joint positions or twist angles. The *kin8nm* dataset provides an example of such a task consisting of 8192 measurements (Ghahramani, 1996). For this problem, we desire accurate intervals such that the arm may be used safely and effectively and, depending on the specifics of the application, either interval width or conditional coverage may be important. In Figure 3 we provide the results of applying RQR to this problem using the same experimental setup as described later in the benchmarking experiments.

In the *upper* subplot, we optimize for interval width using RQR-W and compare the resulting intervals to those using QR. For presentation we sort the test set points according to increasing target magnitude and apply a Savitzky–Golay filter to smooth the intervals (see Figure 6 for a version without smoothing applied). We observe that the

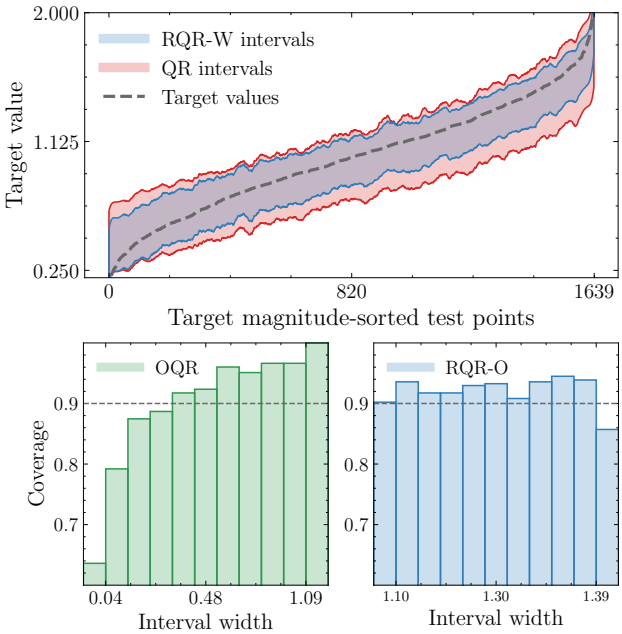

Figure 3. **Applying RQR-W & RQR-O in practice.** Resulting intervals on robotics distance estimation task. **Upper:** The RQR-W objective achieves generally narrower intervals across the 1639 test examples. **Lower:** The RQR-O objective achieves more consistent coverage across different interval widths.

RQR-W intervals are noticeably narrower despite providing the same marginal coverage. Then, in the *lower* subplot, we demonstrate the effects of instead optimizing for conditional coverage using RQR-O and evaluate the empirical coverage when the test set is divided into subgroups based on interval width. In the left of the two histograms, we observe that the baseline method, OQR, under-covers for narrower intervals whilst over-covering for wider intervals. This is in contrast to RQR-O on the right which achieves generally balanced coverage across all interval widths. Given that conditional coverage is the exclusive goal in this case, the RQR-O solution discovers that wider intervals are necessary to achieve this. As a result, it is able to ensure that the probability of error is not dependent on interval width resulting in a more consistent coverage across the data.

Therefore, unlike previous works, RQR may be viewed as a general-purpose approach to constructing intervals in which the practitioner may choose to prioritize among additional interval properties depending on a particular application. In this particular robotics example, minimizing interval width may result in a more accurate estimate of the object's distance while improving conditional coverage may help prevent specific failure modes.

Benchmarking RQR-W. We now proceed to investigate the performance of RQR-W more broadly on the standard quantile regression benchmark tasks used in Tagasovska & Lopez-Paz (2019); Chung et al. (2021); Pearce et al. (2018) consisting of nine datasets from Asuncion & Newman (2007). We extend this benchmark, as suggested by the conference reviewers, with three additional datasets from Grinsztajn et al. (2022). We follow the preprocessing and experimental protocol described in Appendix C in line with previous works. To summarize, we train two-layer neural networks using a grid search to find optimal hyperparameters. All experiments are repeated over 10 seeds for each dataset. For building SQR-C intervals, we followed the method prescribed by the authors which consists of selecting the symmetric (0.05, 0.95) intervals. We provide a summary of our results in Table 3. We report the proportion of datasets in which each method obtains coverage (within a 2.5% margin of error). We then report the mean absolute miscoverage distance from the desired level (90%) across the 12 datasets. Finally, we also provide the frequency with which each method achieves the (possibly joint) narrowest mean interval width. These results illustrate that, on aggregate, the added flexibility of RQR-W is better able to balance our desire for narrow intervals *whilst also* obtaining our desired level of coverage than existing baselines. Gains in realized coverage are a convenient side effect of two factors: (1) quantile regression introduces two sources of error (i.e. estimating each of the two quantiles) in which an error on *either* estimate can result in miscoverage. This is not the case for methods that estimate the interval directly[6]; (2) minimizing interval width selects intervals that are likely to lie in denser regions and are likely to exhibit lower variance in their estimates (see Appendix F for an extended discussion on this point).

Table 3. **Benchmarking RQR-W**. We evaluate each baseline method against RQR-W across 12 datasets with the aggregated results presented. RQR-W obtains the desired coverage level more frequently (coverage obtained) with a lower magnitude of miscoverage from the desired level (mean miscoverage) and achieves the lowest or joint lowest ($\pm$ 1 standard error) mean interval width more frequently (narrowest intervals) across the benchmark tasks.

	RQR-W	QR	SQR-C	SQR-N	IR
Coverage obtained	**12/12**	10/12	9/12	3/12	11/12
Mean miscoverage (%)	**1.63**	2.45	2.13	4.02	2.73
Narrowest intervals	**8/12**	6/12	1/12	2/12	0/12

The complete results for each dataset are included in Appendix H.1 where we provide the resulting coverage and MPIW with means and standard errors of means reported. There we also include histograms of the target distributions for each dataset to highlight the point that non-symmetric distributions should be the *standard expectation* on real-world regression tasks. Note that the target coverage level

[6]Note that we are not the first to notice this fact, see e.g. Sec. 2 of Takeuchi et al. (2006) for a much earlier work that has commented on this limitation.

is set to 90% throughout our experiments and we ablate for other coverage levels in Appendix H.2 where we observe similar results. Overall, the results in this section demonstrate that the added flexibility of circumventing the standard step of learning predefined quantiles can be utilized to obtain narrower intervals in practice.

Benchmarking RQR-O. In a similar vein, we evaluate RQR-O for its effectiveness in improving measures of conditional coverage. In this case, we compare to the original OQR work of Feldman et al. (2021). We evaluate both methods on the same benchmark datasets as previously and follow the same experimental protocol as in Feldman et al. (2021). The key distinction in this experiment is that the coefficient of the regularization parameter for both methods is incrementally reduced until empirical coverage is achieved. This is because a comparison of conditional coverage using these metrics is only meaningful if both methods achieve a similar level of empirical marginal coverage. Again, the experimental setup is described in detail in Appendix C. We investigate this regularization term's effectiveness at achieving its stated objective of enforcing orthogonality between interval width and instances of miscoverage. Since both methods use an identical regularization expression, the gains obtained by RQR-O in this section are due to pairing this term with our RQR objective from Equation (1) rather than existing quantile regression objectives that suffer from the limitations discussed in Section 3.1. The results are provided in Table 4 where we compare performance based on a test set evaluation of coverage, % improvement in Pearson correlation over OQR, and % improvement in HSIC over OQR. We generally find that RQR-O achieves a significant improvement in these measures of conditional coverage, again indicating that the added flexibility of this approach enables a more favorable solution to be found. An important takeaway from these results is that, since Pearson's correlation is the regularization objective used by both methods, the gains in performance when considering it as an evaluation metric provide direct evidence that the added flexibility provided by the RQR loss function (due to not being centered around the median) enables it to find a better solution in the auxiliary task (in this case, minimizing Pearson's correlation between instances of miscoverage and interval width).

Table 4. **Benchmarking RQR-O.** We compare our proposed loss function to those used in OQR for achieving improved conditional coverage as evaluated using standard metrics (see Section 2). We report the empirical coverage achieved and the % improvement obtained over OQR in Pearson correlation and HSIC ($\pm$ a standard error).

Dataset	RQR-O (ours) coverage	OQR coverage	Pearson correlation	HSIC
concrete	88.75 (0.14)	87.54 (0.12)	+80.19 (0.72)	+98.01 (0.12)
power	90.00 (0.05)	91.86 (0.05)	+86.49 (0.63)	+77.79 (1.19)
wine	89.15 (0.27)	88.59 (0.10)	+71.42 (1.14)	+96.77 (0.29)
yacht	88.71 (0.99)	89.07 (0.20)	-45.80 (11.28)	-18.49 (10.7)
naval	90.58 (0.07)	90.12 (0.07)	+87.83 (0.48)	+16.87 (2.95)
energy	89.77 (0.18)	90.42 (0.09)	+25.04 (2.70)	+76.57 (2.12)
boston	90.49 (0.18)	91.97 (0.15)	+75.97 (0.80)	+85.67 (1.54)
kin8nm	90.40 (0.07)	89.94 (0.05)	+89.16 (0.30)	+99.91 (0.01)
protein	89.87 (0.02)	89.45 (0.03)	+80.29 (0.58)	+99.91 (0.01)

5. Conclusion

In this work we have introduced *Relaxed Quantile Regression*, a direct alternative to quantile regression that circumvents the requirement to prespecify the exact quantiles being learned whilst maintaining its attractive coverage properties. We then demonstrated that this new loss can be easily combined with user-specified regularization terms to obtain a solution suitable for a given application (i.e. narrower intervals or improved conditional coverage). Finally, we evaluated the method against state-of-the-art single-model methods across standard benchmark tasks demonstrating that this added flexibility is converted into improved performance in terms of either narrower intervals or better conditional coverage, depending on the practitioner's preference.

For future work, devising new approaches to evaluating predictive intervals and developing complimentary regularization terms such that they can be paired with RQR provides an opportunity to extend upon this work. Additionally, applying a post hoc conformalization procedure to an underlying quantile regression predictor has already been demonstrated to be highly effective (e.g. Romano et al. (2019); Sesia & Romano (2021)). Although the choice of base estimator is acknowledged to impact the performance of the overall procedure (Sesia & Romano, 2021), to the best of our knowledge, no work has systematically investigated the interaction between the two[7] – thus providing a valuable direction for future research.

[7]One work that is relevant to this direction is Sesia & Candès (2020) which does consider different conformalization procedures with a fixed quantile regression method but not vice versa.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

We would like to thank Fergus Imrie, Julianna Piskorz, Johannes Vallikivi, Alicia Curth, and the anonymous reviewers for very insightful comments and discussions on earlier drafts of this paper. TP would like to thank AstraZeneca for their sponsorship and support. AJ and NS gratefully acknowledge funding from the Cystic Fibrosis Trust. This work was supported by Azure sponsorship credits granted by Microsoft's AI for Good Research Lab.

References

Abdullah, A. A., Hassan, M. M., and Mustafa, Y. T. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 10:36538–36562, 2022.

Abe, T., Buchanan, E. K., Pleiss, G., Zemel, R., and Cunningham, J. P. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660, 2022.

Abe, T., Buchanan, E. K., Pleiss, G., and Cunningham, J. P. Pathologies of predictive diversity in deep ensembles. *arXiv preprint arXiv:2302.00704*, 2023.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Ayhan, M. S. and Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2022.

Baek, W.-J. and Kröger, T. Safety evaluation of robot systems via uncertainty quantification. *arXiv preprint arXiv:2302.10644*, 2023.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1): 20–23, 2019.

Brando, A., Center, B. S., Rodriguez-Serrano, J., Vitrià, J., et al. Deep non-crossing quantiles through the partial derivative. In *International Conference on Artificial Intelligence and Statistics*, pp. 7902–7914. PMLR, 2022.

Caflisch, R. E. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49, 1998.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Dewolf, N., Baets, B. D., and Waegeman, W. Valid prediction intervals for regression problems. *Artificial Intelligence Review*, 56(1):577–613, 2023.

Dietterich, T. G. Steps toward robust artificial intelligence. *Ai Magazine*, 38(3):3–24, 2017.

Dunsmore, I. A bayesian approach to calibration. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):396–405, 1968.

Feldman, S., Bates, S., and Romano, Y. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34: 2060–2071, 2021.

Fitzenberger, B., Koenker, R., and Machado, J. A. *Economic applications of quantile regression*. Springer Science & Business Media, 2001.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1901–1910. PMLR, 2019.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pp. 1–77, 2023.

Ghahramani, Z. The kin datasets, 1996.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Goodwin, P., Önkal, D., and Thomson, M. Do forecasts expressed as prediction intervals improve production planning decisions? *European Journal of Operational Research*, 205(1):195–201, 2010.

Graves, A. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Greenfeld, D. and Shalit, U. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pp. 3759–3768. PMLR, 2020.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hekler, A., Brinker, T. J., and Buettner, F. Test time augmentation meets post-hoc calibration: uncertainty quantification under real-world conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14856–14864, 2023.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Hunter, D. R. and Lange, K. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

Hwang, J. G. and Ding, A. A. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438):748–757, 1997.

Jeffares, A., Liu, T., Crabbé, J., and van der Schaar, M. Joint training of deep ensembles fails due to learner collusion. *arXiv preprint arXiv:2301.11323*, 2023.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22(3):337–346, 2010.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356, 2011.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koenker, R. Quantile regression: 40 years on. *Annual review of economics*, 9:155–176, 2017.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Koenker, R. and Hallock, K. F. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32, 2019.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Park, Y., Maddix, D., Aubet, F.-X., Kan, K., Gasthaus, J., and Wang, Y. Learning quantile functions without

quantile crossing for distribution-free time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pp. 8127–8150. PMLR, 2022.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pp. 4075–4084. PMLR, 2018.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., and Hinton, G. Regularizing neural networks by penalizing confident output distributions, 2017. URL https://openreview.net/forum?id=HkCjNI5ex.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Rahaman, R. et al. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.

Rasmussen, C. E., Williams, C. K., et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Sagi, O. and Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Savelli, S. and Joslyn, S. The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, 27(4):527–541, 2013.

Seber, G. A. and Lee, A. J. *Linear regression analysis*, volume 330. John Wiley & Sons, 2003.

Seedat, N., Jeffares, A., Imrie, F., and van der Schaar, M. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 10160–10177. PMLR, 2023.

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

Skafte, N., Jørgensen, M., and Hauberg, S. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Steinwart, I. and Christmann, A. Estimating conditional quantiles with the help of the pinball loss. 2011.

Su, S., Li, Y., He, S., Han, S., Feng, C., Ding, C., and Miao, F. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5588–5594. IEEE, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tagasovska, N. and Lopez-Paz, D. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Takeuchi, I., Le, Q., Sears, T., Smola, A., et al. Nonparametric quantile estimation. 2006.

Tierney, L. and Kadane, J. B. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., and Buettner, F. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132, 2021.

Vapnik, V. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, Y., Xu, H., Zou, R., Zhang, L., and Zhang, F. A deep asymmetric laplace neural network for deterministic and probabilistic wind power forecasting. *Renewable Energy*, 196:497–517, 2022.

Wen, Y., Tran, D., and Ba, J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2019.

Wilson, A. G. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33: 4697–4708, 2020.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial intelligence and statistics*, pp. 370–378. PMLR, 2016a.

Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. Stochastic variational deep kernel learning. *Advances in neural information processing systems*, 29, 2016b.

Winkler, R. L. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Xie, L., Wang, J., Wei, Z., Wang, M., and Tian, Q. Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4753–4762, 2016.

Yao, J., Pan, W., Ghosh, S., and Doshi-Velez, F. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

Yu, K. and Jones, M. Local linear quantile regression. *Journal of the American statistical Association*, 93(441): 228–237, 1998.

Yu, K., Lu, Z., and Stander, J. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D: The Statistician*, 52(3):331–350, 2003.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 609–616, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Zhang, Y., Wen, H., and Wu, Q. A contextual bandit approach for value-oriented prediction interval forecasting. *IEEE Transactions on Smart Grid*, 2023.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pp. 11387–11397. PMLR, 2020.

A. Extended Related Work

In this work, we have focused on *single model approaches* (i.e. quantile/interval regression) for obtaining predictive intervals from neural networks with the intention of developing more effective methods within this category. Due to the vital importance of uncertainty quantification, a vast literature of disparate alternative approaches has emerged, each representing interesting research directions with their own respective strengths and weaknesses. In this section, we provide a broad overview of the leading methods for obtaining predictive intervals from neural networks. Given the extensive nature of this topic, this summary is not exhaustive and we refer the reader to the more comprehensive references cited within each topic for a more complete overview. For a recent survey on predictive intervals in regression problems more generally (i.e. beyond neural networks) we refer the reader to Dewolf et al. (2023) or for general neural network uncertainty quantification to Gawlikowski et al. (2023).

Single model approaches. The subject of this work is individual models that output intervals rather than point estimates which we refer to as single model approaches. An in-depth description of the state-of-the-art methods within this category is provided in Section 2. Whilst these methods typically estimate aleatoric uncertainty, it is quite straightforward to combine them with Bayesian or ensemble methods to simultaneously account for epistemic uncertainty (Tagasovska & Lopez-Paz, 2019). As previously discussed, quantile regression (Koenker & Bassett Jr, 1978) has excelled in this category – including prior to and outside of the deep learning regime (Koenker & Hallock, 2001; Meinshausen & Ridgeway, 2006; Yu & Jones, 1998). Similarly, the evaluated direct interval estimation method of Pearce et al. (2018) was also an evolution of the foundational work of Khosravi et al. (2010). Chung et al. (2021) primarily focuses on the task of training models that output the full predictive distribution (rather than intervals for a preselected coverage level) which suffers from the same limitations as SQR that were discussed in the main text. Elsewhere in the time series setting, Gasthaus et al. (2019) estimate quantiles using monotonic regression splines. Whilst strictly speaking a dual model approach, variance networks fit a second neural network to estimate the variance of a prediction (Skafte et al., 2019). A very simple approach that can be easily combined with most other approaches is to regularize the neural network such that measures of calibration are optimized using e.g. confident output penalization (Pereyra et al., 2017), label smoothing (Szegedy et al., 2016), or induced label noise (Xie et al., 2016).

Bayesian methods. Bayesian neural networks attempt to model the target probability distribution for a given test example x given some observed data $\mathcal{D}$ by marginalizing over a distribution of network parameters θ such that $P(y|x, \mathcal{D}) = \int P(y|x, \theta)P(\theta|\mathcal{D})$ from which predictive intervals can be derived. This expression requires intractable calculations which may be approximated in practice using various techniques. A simple approach consists of taking a second-order Taylor expansion around the maximum a posteriori estimate of θ to produce a Gaussian approximation of $P(\theta|\mathcal{D})$ known as the *Laplace approximation* (Tierney & Kadane, 1986). In practice, further scaling efforts have been required to apply this method in the modern deep learning context (Daxberger et al., 2021). Alternatively, *variational inference* substitutes $P(\theta|\mathcal{D}) \approx q(\theta)$ where $q(\theta)$ denotes some tractable parametric approximation such as a Gaussian distribution (Hinton & Van Camp, 1993). Significant research has investigated methods for extending this approach to account for modern datasets and architectures (Graves, 2011; Zhang et al., 2018). *Monte Carlo integration* instead approximates the integral over parameters with a finite sum such that $\int P(y|x, \theta)P(\theta|\mathcal{D}) \approx \frac{1}{M}\sum_{j=1}^{M} P(y|x, \theta_j)$ (Caflisch, 1998). Then different choices of selecting a subset of M weight parameterizations result in alternative instantiations of this approximation (see e.g. Ch. 17 of Goodfellow et al. (2016)). *Monte Carlo dropout*, which drops neurons at test time according to a Bernoulli distribution to estimate uncertainty, has become popular due to its conceptual and implementation simplicity (Gal & Ghahramani, 2016). A somewhat distinct Bayesian approach is the *Gaussian process* which is a collection of random variables of which any finite sample is Gaussian distributed specified by a specific mean function and covariance kernel (Rasmussen et al., 2006). Although this method suffers from important limitations (e.g. scaling), its adaption to the deep learning setting has achieved notable performance on benchmark tasks (Wilson et al., 2016a;b).

Deep ensembles. Aggregating outputs over a set of neural networks has emerged as a simple but effective method that accounts for epistemic uncertainty (Lakshminarayanan et al., 2017). Whilst this approach can be considered studied through a Bayesian perspective (Wilson & Izmailov, 2020; Wilson, 2020), it has primarily developed from the classical ensembling literature (Sagi & Rokach, 2018). It is hypothesised that the empirical success of deep ensembles is due to their better exploration of the loss landscape (Fort et al., 2019) – with diversity typically achieved through random initialization and batching due to known challenges in optimizing for exploration (Jeffares et al., 2023; Abe et al., 2023). However, recent works have suggested that their improved calibration may be overstated (Rahaman et al., 2021) and that performance increases may be better understood as being due to an increased model capacity (Abe et al., 2022). These gains also come at the cost of a significant computational overhead with the relative computational cost growing linearly with ensemble

Table 5. **Categories of interval construction approaches.** A broad comparison of different categories of interval construction highlighting the key distinctions of single model approaches. Detailed descriptions and discussion is provided in the text.

	(1)	**(2)**	**(3)**	**(4)**
Single model approaches	✓	✓	✓	✓
Bayesian methods	✗	✗	✓	1/2
Deep ensembles	✗	✗	✗	✓
Parametric approaches	1/2	✗	✓	1/2
Post-hoc methods	✓	✓	✗	✓

size. More efficient approaches to deep ensembling have been proposed in recent years to reduce this cost (e.g. Wen et al., 2019).

Parametric approaches. The classical statistics literature has a long history of developing principled estimates of predictive intervals derived from parametric assumptions in the data generating process (see e.g. Ch. 5.3 Seber & Lee, 2003). One such approach in the neural network setting is the delta method which makes a linearity assumption in the region around a prediction paired with a Gaussian assumption on the noise distribution (Hwang & Ding, 1997; Khosravi et al., 2011). Another example is Nix & Weigend (1994) who also assume the noise distribution to be Gaussian and derive a cost function to estimate its value with an auxiliary output to the network.

Post-hoc methods. Several methods exist in which calibrated intervals are constructed or updated as a post-processing step for an existing point predictor or interval estimator respectively. Perhaps the most notable of these is conformal prediction (Vovk et al., 2005) and, in particular, inductive conformal prediction (Papadopoulos, 2008), which provides prediction intervals with finite sample marginal coverage guarantees. Romano et al. (2019) further developed an approach that also performs well conditionally (i.e. intervals where width is adaptive to a given example). As noted by the authors, this method "can wrap around any algorithm for quantile regression", thus making it a complimentary post-hoc approach. Similarly, Sesia & Romano (2021) use (multiple) quantile regression as a base estimator for conformalization which they report to "work particularly well". Other methods typically focus on improving the calibration of an underlying point predictor. Approaches include Platt scaling (Platt et al., 1999), temperature scaling (Tomani et al., 2021), histogram binning (Zadrozny & Elkan, 2001), test time augmentation (Hekler et al., 2023), and isotonic regression (Zadrozny & Elkan, 2002).

Comparing categories of interval construction. We now provide some high level distinctions between these categories of approaches for constructing intervals. This is *not* intended as a complete evaluation of competing uncertainty quantification approaches, rather we wish to highlight the advantages of single model approaches to emphasize the significance of developments *within* this category. Furthermore, due to the broadness of these categories and the lack of clear boundaries between them, the following distinctions act as generalizations for which some exceptions exist. A discussion of these distinctions is provided in the next paragraph with a summary provided in Table 5.

(1) *Minimal computational overhead* - Deep ensembles require training a neural network from scratch M times while Bayesian methods typically require approximations to produce tractable algorithms for large-scale models (Abdullah et al., 2022; Osawa et al., 2019). Parametric methods require some overhead with the specific amount method dependent. In contrast, the other categories, including single model approaches, generally only require at most a change of loss function at training time. **(2)** *Data-assumption-free valid intervals* - Parametric approaches and Bayesian methods generally require assumptions on the data-generating process to provide validity guarantees on their intervals. Deep ensembles don't provide such guarantees on derived intervals. However, many single model and post-hoc methods provide asymptotic or even finite sample guarantees of valid intervals. **(3)** *Distinguishes between aleatoric & epistemic uncertainty* - Explicitly differentiating between aleatoric (irreducible) and epistemic (reducible) uncertainty can be valuable (Hüllermeier & Waegeman, 2021). As discussed in (Tagasovska & Lopez-Paz, 2019; Pearce et al., 2018), the loss function of single model approaches generally estimates aleatoric uncertainty while epistemic uncertainty can also be accounted for by applying

e.g. orthonormal certificates or interval ensembling. Explicitly modeling uncertainties in this way also tends to be at the heart of Bayesian and parametric methods (see e.g. Kendall & Gal (2017)). Deep ensembles and post-hoc methods do not typically make this distinction. **(4)** *Directly applicable to any loss-based algorithm* - Single-model approaches are typically characterized by simply replacing a point estimate loss function with a quantile regression loss (e.g. mean squared error $\rightarrow$ pinball loss). Similarly, deep ensembles only require running an algorithm multiple times and post-hoc methods typically wrap around or recallibrate arbitrary models. Some Bayesian and parametric methods are generally applicable (e.g. Monte Carlo dropout) however others require more substantial changes resulting in different algorithms when applied to neural networks (e.g. a Gaussian process).

B. Theory and Proofs

Equivalence of the Winkler Score and the Pinball Loss.

Proof. For a targeted coverage level $1-\alpha$ and the bounds (μ_1,μ_2), we demonstrate that the Winkler score is proportional to the Pinball Loss by a factor $\frac{2}{\alpha}$.

$$\begin{aligned}
\frac{2}{\alpha}\mathcal{L}_\alpha^{\text{QR}}((\mu_1,\mu_2,\mathbf{x},y) &= \frac{2\rho_{\frac{\alpha}{2}}(\mu_1(x),\mathbf{x},y)+2\rho_{1-\frac{\alpha}{2}}(\mu_2(x),\mathbf{x},y)}{\alpha} \\
&= \frac{2(\mathbb{I}_{y\leq\mu_1(x)}-\frac{\alpha}{2})(\mu_1(x)-y)+2(\mathbb{I}_{y\leq\mu_2(x)}-1+\frac{\alpha}{2})(\mu_2(x)-y)}{\alpha} \\
&= \frac{2(\mathbb{I}_{y\leq\mu_1(x)}-\frac{\alpha}{2})(\mu_1(x)-y)+2(1-\mathbb{I}_{y\leq\mu_2(x)}-\frac{\alpha}{2})(y-\mu_2(x))}{\alpha} \\
&= \frac{2(\mathbb{I}_{y\leq\mu_1(x)}-\frac{\alpha}{2})(\mu_1(x)-y)+2(\mathbb{I}_{y\geq\mu_2(x)}-\frac{\alpha}{2})(y-\mu_2(x))}{\alpha} \\
&= (\mu_2(x)-\mu_1(x))+\frac{2}{\alpha}\mathbb{I}_{y\leq\mu_1(x)}(q_{\frac{\alpha}{2}}-y)+\frac{2}{\alpha}\mathbb{I}_{y\geq\mu_2(x)}(y-\mu_2(x))
\end{aligned}$$

We obtain the expression of the Winkler Score (Winkler, 1972). □

Theorem 3.1 (RQR In-sample Coverage)**.** *For any random variable Y associated with an input x, $\forall\,\alpha\in[0,1]$,*

$$(\mu_1^*(x),\mu_2^*(x)) = \underset{\mu_1,\mu_2}{\arg\min}\,\{\mathbb{E}_Y(\mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2),x,Y))\} \implies \mathbb{P}(\mu_1^*(x)<Y<\mu_2^*(x))=\alpha.$$

Proof. In these proofs, we omit explicitly including the input x for clarity. However, the reader should recall that Y is the random variable associated with the input x.

First, we can rewrite our new loss with an indicator function:

$$\mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2),y)) = (y-\mu_1)(y-\mu_2)(\alpha-\mathbb{I}_{y\in[\mu_1,\mu_2]})$$

Then, we consider the expectation of the loss:

$$\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2),Y))) = \alpha\int_{-\infty}^{\infty}(y-\mu_1)(y-\mu_2)d\mathbb{P}_Y(y) - \int_{\mu_1}^{\mu_2}(y-\mu_1)(y-\mu_2)d\mathbb{P}_Y(y)$$

We find the expression of its partial derivatives with respect to the bounds:

$$\begin{aligned}
\frac{\partial\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR}})}{\partial\mu_1} &= -\alpha\int_{-\infty}^{\infty}(y-\mu_2)d\mathbb{P}_Y(y)+\int_{\mu_1}^{\mu_2}(y-\mu_2)d\mathbb{P}_Y(y) \\
\frac{\partial\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR}})}{\partial\mu_2} &= -\alpha\int_{-\infty}^{\infty}(y-\mu_1)d\mathbb{P}_Y(y)+\int_{\mu_1}^{\mu_2}(y-\mu_1)d\mathbb{P}_Y(y)
\end{aligned}$$

At the minimum, the gradient of the expected loss is null. Thus, if there is a minimum at the point (μ_1^*, μ_2^*) with $\mu_2^* > \mu_1^*$,

$$\begin{aligned}
&\frac{\partial \mathbb{E}(\mathcal{L}_\alpha^{\text{RQR}})}{\partial \mu_1}\Big|_{\mu_1^*,\mu_2^*} - \frac{\partial \mathbb{E}(\mathcal{L}_\alpha^{\text{RQR}})}{\partial \mu_2}\Big|_{\mu_1^*,\mu_2^*} = 0 \\
\implies &-\alpha \int_{-\infty}^{\infty} (\mu_1^* - \mu_2^*) d\mathbb{P}_Y(y) + \int_{\mu_1^*}^{\mu_2^*} (\mu_1^* - \mu_2^*) d\mathbb{P}_Y(y) = 0 \\
\implies &-\alpha(\mu_1^* - \mu_2^*) \int_{-\infty}^{\infty} d\mathbb{P}_Y(y) + (\mu_1^* - \mu_2^*) \int_{\mu_1^*}^{\mu_2^*} d\mathbb{P}_Y(y) = 0 \\
\implies &-\alpha \int_{-\infty}^{\infty} d\mathbb{P}_Y(y) + \int_{\mu_1^*}^{\mu_2^*} d\mathbb{P}_Y(y) = 0 \\
\implies &\mathcal{P}(\mu_1^* < Y < \mu_2^*) = \alpha
\end{aligned}$$

□

Theorem 3.2 (RQR with Finite Samples)**.** *For any random variable Y associated with an input x, we consider N realizations of this random variable : $\{y_i\}_{i=1,N}$. $\forall\, \alpha \in [0,1]$ such that $\alpha \cdot N \in \mathbb{N}$,*

$$(\mu_1^*(x), \mu_2^*(x)) = \underset{\mu_1,\mu_2}{\arg\min} \{\sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i)\} \implies \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_1 *, \mu_2 *]} = \alpha.$$

Proof. We rewrite our sum as an integral with discrete density $\mathbb{I}_{y\in\{y_i\}} dy$:

$$\sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i) = \alpha \int_{-\infty}^{\infty} (y-\mu_1)(y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy - \int_{\mu_1}^{\mu_2} (y-\mu_1)(y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy$$

We find the expression of its partial derivatives with respect to the bounds:

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_1} &= -\alpha \int_{-\infty}^{\infty} (y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy + \int_{\mu_1}^{\mu_2} (y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy \\
\frac{\partial \sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_2} &= -\alpha \int_{-\infty}^{\infty} (y-\mu_1)\mathbb{I}_{y\in\{y_i\}} dy + \int_{\mu_1}^{\mu_2} (y-\mu_1)\mathbb{I}_{y\in\{y_i\}} dy
\end{aligned}$$

At the minimum, the gradient of the expected loss is null. Thus, if there is a minimum at the point (μ_1^*, μ_2^*) with $\mu_2^* > \mu_1^*$,

$$\begin{aligned}
&\frac{\partial \sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_1}\Big|_{\mu_1^*,\mu_2^*} - \frac{\partial \sum_{i=1}^{N} \mathcal{L}_\alpha^{\text{RQR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_2}\Big|_{\mu_1^*,\mu_2^*} = 0 \\
\implies &-\alpha \int_{-\infty}^{\infty} (\mu_1^* - \mu_2^*)\mathbb{I}_{y\in\{y_i\}} dy + \int_{\mu_1^*}^{\mu_2^*} (\mu_1^* - \mu_2^*)\mathbb{I}_{y\in\{y_i\}} dy = 0 \\
\implies &-\alpha(\mu_1^* - \mu_2^*) \sum_{i=1}^{N} 1 + (\mu_1^* - \mu_2^*) \sum_{y_i \in [\mu_1^*, \mu_2^*]} 1 = 0 \\
\implies &\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_1 *, \mu_2 *]} = \alpha
\end{aligned}$$

□

Theorem 3.3 (RQR is Unbiased and Consistent)**.** *For any random variable Y associated with an input x, we consider N realizations of this random variable : $\{y_i\}_{i=1}^N$. $(\mu_{1,N}, \mu_{2,N})$ are the bounds of our estimator trained on these N samples. $\forall\, \alpha \in [0,1]$, we name Q_N the absolute true miscoverage of our estimator*

$$Q_N = |\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \alpha|$$

Then $E[Q_N] = 0$ and the variance of Q_N is bounded by $\frac{1}{4N}$.

Proof. We can decompose Q_N in two terms :

$$Q_N \leq |\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}| + |\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]} - \alpha|$$

The theorem 3.2 proves that $\forall N \in \mathbb{N}$ the second term is null.

Then, the first term is simply the difference between an integral and its Monte Carlo estimator. Hence, we can derive well-known results about its mean $\mathbb{E}[Q_N]$ and variance $\mathbb{V}[Q_N]$.

$$\begin{aligned}
\mathbb{E}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}] &= \int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}] \\
&= \int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \int_{\mu_{1,N}}^{\mu_{2,N}} dPy \\
&= 0
\end{aligned}$$

Hence, $\mathbb{E}[Q_N] \leq \mathbb{E}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}] = 0$.

$$\mathbb{V}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}] = \frac{1}{N}\mathbb{V}[\mathbb{I}_{Y \in [\mu_{1,N}, \mu_{2,N}]}]$$

Thus, $\mathbb{V}[Q_N] \leq \frac{1}{N}\mathbb{V}[\mathbb{I}_{Y \in [\mu_{1,N}, \mu_{2,N}]}]$
Moreover, $\mathbb{V}[\mathbb{I}_{Y \in [\mu_{1,N}, \mu_{2,N}]}] = P(Y \in [\mu_{1,N}, \mu_{2,N}])(1 - P(Y \in [\mu_{1,N}, \mu_{2,N}]))$ and the study of the function $f : x \mapsto x(1-x)$ for $x \in [0,1]$ shows that this function is bounded by $\frac{1}{4}$.
Thus, $\mathbb{V}[Q_N] \leq \mathbb{V}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_{1,N}, \mu_{2,N}]}] \leq \frac{1}{4N}$ □

Theorem 3.4 (RQR-W In-sample Coverage)**.** *For any random variable Y associated with an input x, $\forall\, \alpha \in [0,1]$,*

$$(\mu_1^*(x), \mu_2^*(x)) \quad = \quad \underset{\mu_1, \mu_2}{\arg\min}\, \{\mathbb{E}_Y((\mu_1, \mu_2), x, Y))\} \qquad \Longrightarrow \qquad \mathbb{P}(\mu_1^*(x) \quad < \quad Y \quad < \quad \mu_2^*(x)) \quad = \quad \alpha$$

Proof. Similarly, we are starting by rewriting the RQR-W loss with indicator functions.

$$\begin{aligned}
\text{RQR-W}_\alpha(\mu_1, \mu_2, y) &= \text{RQR}_\alpha(\mu_1, \mu_2, y) + \frac{\lambda(\mu_2 - \mu_1)^2}{2} \\
&= (y - \mu_1)(y - \mu_2)(\alpha - \mathbb{I}_{y \in [\mu_1, \mu_2]}) + \frac{\lambda(\mu_2 - \mu_1)^2}{2}
\end{aligned}$$

Then, we consider the expectation of the loss :

$$\begin{aligned}
&\mathbb{E}(\text{RQR-W}_\alpha(\mu_1, \mu_2, Y)) = \\
&\alpha \int_{-\infty}^{\infty} (y - \mu_1)(y - \mu_2) d\mathbb{P}_Y(y) - \int_{\mu_1}^{\mu_2} (y - \mu_1)(y - \mu_2) d\mathbb{P}_Y(y) + \frac{\lambda(\mu_2 - \mu_1)^2}{2}
\end{aligned}$$

We find the expression of the partial derivatives with respect to the bounds:

$$\frac{\partial \mathbb{E}(\text{RQR-W}_\alpha)}{\partial \mu_1} = -\alpha \int_{-\infty}^{\infty} (y-\mu_2) d\mathbb{P}_Y(y) + \int_{\mu_1}^{\mu_2} (y-\mu_2) d\mathbb{P}_Y(y) - \lambda(\mu_2-\mu_1)$$
$$\frac{\partial \mathbb{E}(\text{RQR-W}_\alpha)}{\partial \mu_2} = -\alpha \int_{-\infty}^{\infty} (y-\mu_1) d\mathbb{P}_Y(y) + \int_{\mu_1}^{\mu_2} (y-\mu_1) d\mathbb{P}_Y(y) + \lambda(\mu_2-\mu_1)$$

At the minimum, the gradient of the expected loss is null. Thus, if there is a minimum at the point (μ_1^*, μ_2^*) with $\mu_2^* > \mu_1^*$,

$$\begin{aligned}
&\frac{\partial \mathbb{E}(\text{RQR-W}_\alpha)}{\partial \mu_1}\Big|_{\mu_1^*,\mu_2^*} - \frac{\partial \mathbb{E}(\text{RQR-W}_\alpha)}{\partial \mu_2}\Big|_{\mu_1^*,\mu_2^*} = 0 \\
\implies & -\alpha \int_{-\infty}^{\infty} (\mu_1^*-\mu_2^*) d\mathbb{P}_Y(y) + \int_{\mu_1^*}^{\mu_2^*} (\mu_1^*-\mu_2^*) d\mathbb{P}_Y(y) + 2\lambda(\mu_1^*-\mu_2^*) = 0 \\
\implies & -\alpha(\mu_1^*-\mu_2^*) \int_{-\infty}^{\infty} d\mathbb{P}_Y(y) + (\mu_1^*-\mu_2^*) \int_{\mu_1^*}^{\mu_2^*} d\mathbb{P}_Y(y) + 2\lambda(\mu_1^*-\mu_2^*) = 0 \\
\implies & -\alpha \int_{-\infty}^{\infty} d\mathbb{P}_Y(y) + \int_{\mu_1^*}^{\mu_2^*} d\mathbb{P}_Y(y) + 2\lambda = 0 \\
\implies & \mathcal{P}(\mu_1^* < Y < \mu_2^*) = \alpha - 2\lambda
\end{aligned}$$

Then, as α is a constant, we can replace it with a corrected term. When choosing $\hat{\alpha} = \alpha + 2\lambda$, we obtain

$$(\mu_1^*, \mu_2^*) = \underset{\mu_1,\mu_2}{\arg\min} \{\mathbb{E}(\text{RQR-W}_{\hat{\alpha}}(\mu_1, \mu_2, Y))\} \implies \mathcal{P}(\mu_1^* < Y < \mu_2^*) = \hat{\alpha} - 2\lambda = \alpha$$

□

Theorem 3.5 (RQR-W with Finite Samples)**.** *For any random variable Y associated with an input x, we consider N realizations of this random variable: $\{y_i\}_{i=1,N}$. $\forall\, \alpha \in [0,1]$ such that $\alpha \cdot N \in \mathbb{N}$,*

$$(\mu_1^*(x), \mu_2^*(x)) = \underset{\mu_1,\mu_2}{\arg\min} \{\sum_{i=1}^{N} \mathcal{L}_{\alpha+2\lambda}^{\text{RQR-W}}((\mu_1,\mu_2), x, y_i)\} \implies \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{y_i \in [\mu_1*, \mu_2*]} = \alpha$$

Proof. We rewrite our sum as an integral with discrete density $\mathbb{I}_{y\in\{y_i\}} dy$:

$$\sum_{i=1}^{N} \mathcal{L}_{\alpha+2\lambda}^{\text{RQR-W}}((\mu_1,\mu_2), x, y_i) = \alpha \int_{-\infty}^{\infty} (y-\mu_1)(y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy - \int_{\mu_1}^{\mu_2} (y-\mu_1)(y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy + N\frac{\lambda(\mu_2-\mu_1)^2}{2}$$

We find the expression of its partial derivatives with respect to the bounds:

$$\frac{\partial \sum_{i=1}^{N} \mathcal{L}_{\alpha+2\lambda}^{\text{QFR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_1} = -(\alpha+2\lambda) \int_{-\infty}^{\infty} (y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy + \int_{\mu_1}^{\mu_2} (y-\mu_2)\mathbb{I}_{y\in\{y_i\}} dy - N\lambda(\mu_2-\mu_1)$$
$$\frac{\partial \sum_{i=1}^{N} \mathcal{L}_{\alpha}^{\text{QFR}}((\mu_1,\mu_2), x, y_i)}{\partial \mu_2} = -(\alpha+2\lambda) \int_{-\infty}^{\infty} (y-\mu_1)\mathbb{I}_{y\in\{y_i\}} dy + \int_{\mu_1}^{\mu_2} (y-\mu_1)\mathbb{I}_{y\in\{y_i\}} dy + N\lambda(\mu_2-\mu_1)$$

At the minimum, the gradient of the expected loss is null. Thus, if there is a minimum at the point (μ_1^*, μ_2^*) with $\mu_2^* > \mu_1^*$,

$$\frac{\partial \sum_{i=1}^N \mathcal{L}^{\text{RQR-W}}_{(\alpha+2\lambda)}((\mu_1,\mu_2),x,y_i)}{\partial \mu_1}\Big|_{\mu_1^*,\mu_2^*} - \frac{\partial \sum_{i=1}^N \mathcal{L}^{\text{RQR-W}}_{\alpha+2\lambda}((\mu_1,\mu_2),x,y_i)}{\partial \mu_2}\Big|_{\mu_1^*,\mu_2^*} = 0$$

$$\implies -(\alpha+2\lambda)\int_{-\infty}^{\infty}(\mu_1^*-\mu_2^*)\mathbb{I}_{y\in\{y_i\}}dy + \int_{\mu_1^*}^{\mu_2^*}(\mu_1^*-\mu_2^*)\mathbb{I}_{y\in\{y_i\}}dy + 2N\lambda(\mu_2^*-\mu_1^*) = 0$$

$$\implies -(\alpha+2\lambda)(\mu_1^*-\mu_2^*)\sum_{i=1}^N 1 + (\mu_1^*-\mu_2^*)\sum_{y_i\in[\mu_1^*,\mu_2^*]} 1 + 2N\lambda(\mu_2^*-\mu_1^*) = 0$$

$$\implies \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_1*,\mu_2*]} = \alpha$$

□

Theorem 3.6 (RQR-W is Unbiased and Consistant)**.** *For any random variable Y associated with an input x, we consider N realizations of this random variable : $\{y_i\}_{i=1,N}$. $(\mu_{1,N}, \mu_{2,N})$ are the bounds of our estimator trained on these N samples. $\forall\, \alpha \in [0,1]$, we name Q_N the absolute true miscoverage of our estimator*

$$Q_N = |\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \alpha|$$

Then $E[Q_N] = 0$ and the variance of Q_N is bounded by $\frac{1}{4N}$

Proof. We can decompose Q_N in two terms :

$$Q_N \leq |\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}| + |\frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]} - \alpha|$$

The theorem 3.5 proves that $\forall N \in \mathbb{N}$ the second term is null.

Then, the first term is simply the difference between an integral and its Monte Carlo estimator. Hence, we can derive well-known results about its mean $\mathbb{E}[Q_N]$ and variance $\mathbb{V}[Q_N]$.

$$\begin{aligned}\mathbb{E}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}] &= \int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}] \\ &= \int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \int_{\mu_{1,N}}^{\mu_{2,N}} dPy \\ &= 0\end{aligned}$$

Hence, $\mathbb{E}[Q_N] \leq \mathbb{E}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}] = 0$.

$$\mathbb{V}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}] = \frac{1}{N}\mathbb{V}[\mathbb{I}_{Y\in[\mu_{1,N},\mu_{2,N}]}]$$

Thus, $\mathbb{V}[Q_N] \leq \frac{1}{N}\mathbb{V}[\mathbb{I}_{Y\in[\mu_{1,N},\mu_{2,N}]}]$
Moreover, $\mathbb{V}[\mathbb{I}_{Y\in[\mu_{1,N},\mu_{2,N}]}] = P(Y\in[\mu_{1,N},\mu_{2,N}])(1-P(Y\in[\mu_{1,N},\mu_{2,N}]))$ and the study of the function $f : x \mapsto x(1-x)$ for $x \in [0,1]$ shows that this function is bounded by $\frac{1}{4}$.
Thus, $\mathbb{V}[Q_N] \leq \mathbb{V}[\int_{\mu_{1,N}}^{\mu_{2,N}} dPy - \frac{1}{N}\sum_{i=1}^N \mathbb{I}_{y_i\in[\mu_{1,N},\mu_{2,N}]}] \leq \frac{1}{4N}$

□

Proposition 3.7 (Existence and Uniqueness of Solution)**.** *μ_1^{min} and μ_2^{max} denote the bounds of our optimization problem. For a target distribution Y with a cumulative distribution function that is k-Lipschitz continuous with $k < 1 + \frac{\alpha}{\mu_2^{max} - \mu_1^{min}}$, when $\lambda > max(0, \int_{\mu_1^{min}}^{\mu_2^{max}} d\mathbb{P}_Y(y) - \alpha)$, the minimum of $\mathcal{L}_{\alpha+2\lambda}^{\text{RQR-W}}$ exists and is unique.*

Proof. We study the function $\mu_1, \mu_2 \mapsto \mathbb{E}(\mathcal{L}_\alpha^{\text{RQR-W}}((\mu_1, \mu_2), Y)$ in a closed subset of $\mathbb{R}^2$ where $\mu_1 < \mu_2$. We named this subset $\mathbb{D}$. On this closed subset of the space to say, it exists μ_2^{max} and μ_1^{min} such that $\forall(\mu_1, \mu_2) \in \mathbb{D}$ $\mu_2 < \mu_2^{max}$ and $\mu_1 > \mu_1^{min}$.

Moreover, we assume that Y can be associated with a probability density function $d\mathbb{P}_Y$ and that its cumulative distribution function is k-Lipschitz continuous with $k < 1 + \frac{\alpha}{\mu_2^{max} - \mu_1^{min}}$

$$\forall(\mu_2, \mu_1) \in [\mu_1^{min}, \mu_2^{max}]^2 \mid \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y)| < (1 + \frac{\alpha}{\mu_2^{max} - \mu_1^{min}})|(\mu_2 - \mu_1)|$$

The non-negativity of the studied function gives us the existence of the minimum.

To demonstrate the uniqueness of the minimum, we show that the eigenvalues of the Hessian matrix are positive. We start by computing the gradient of the expected loss :

$$\begin{aligned}\nabla\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR-W}}) &= \begin{bmatrix} \frac{\partial\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR-W}})}{\partial\mu_1} \\ \frac{\partial\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR-W}})}{\partial\mu_2} \end{bmatrix} \\ &= \begin{bmatrix} -\alpha\int_{-\infty}^{\infty}(y-\mu_2)d\mathbb{P}_Y(y) + \int_{\mu_1}^{\mu_2}(y-\mu_2)d\mathbb{P}_Y(y) - \lambda(\mu_2-\mu_1) \\ -\alpha\int_{-\infty}^{\infty}(y-\mu_1)d\mathbb{P}_Y(y) + \int_{\mu_1}^{\mu_2}(y-\mu_1)d\mathbb{P}_Y(y) + \lambda(\mu_2-\mu_1) \end{bmatrix}\end{aligned}$$

Then, we compute the Hessian matrix:

$$\nabla^2\mathbb{E}(\mathcal{L}_\alpha^{\text{RQR-W}}) = \begin{bmatrix} \mu_2 - \mu_1 + \lambda & \alpha - \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y) - \lambda \\ \alpha - \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y) - \lambda & \mu_2 - \mu_1 + \lambda \end{bmatrix}$$

The eigenvalues of the Hessian matrix are given by $\lambda_\pm = \mu_2 - \mu_1 + \lambda \pm |\alpha - \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y) - \lambda|$

From that, we get that $\lambda > \int_{\mu_1^{min}}^{\mu_2^{max}} d\mathbb{P}_Y(y) - \alpha \implies \forall(\mu_1, \mu_2) \in \mathbb{D}$ $\lambda > \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y) - \alpha$ (1) because the probability density function $d\mathbb{P}_Y$ is positive.

Additionally, we use the assumption of the k-Lipschitz continuity of the CDF and we obtain the following inequality :

$$\forall(\mu_1, \mu_2) \in \mathbb{D}\ \mu_2 - \mu_1 \geq \int_{\mu_1}^{\mu_2}(1 + \frac{\alpha}{\mu_2 - \mu_1}) - \alpha \geq \int_{\mu_1}^{\mu_2}(1 + \frac{\alpha}{\mu_2^{max} - \mu_1^{min}}) - \alpha \geq \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y - \alpha$$

Therefore, under the condition $\lambda > \max(0, \int_{\mu_1^{min}}^{\mu_2^{max}} d\mathbb{P}_Y(y) - \alpha)$, we obtain the positiveness of λ_- :

$$\begin{aligned}\lambda_- &= \mu_2 - \mu_1 + \lambda - |\alpha - \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y) - \lambda| = \mu_2 - \mu_1 + \lambda - (\lambda - \alpha + \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y))\ (1) \\ \implies \lambda_- &= \mu_2 - \mu_1 - (-\alpha + \int_{\mu_1}^{\mu_2} d\mathbb{P}_Y(y)) > 0\end{aligned}$$

The first eigenvalue λ_+ is obviously non-negative, thus both eigenvalues are non-negative which means that the Hessian matrix is semi-definite positive.

In conclusion, when the condition $\lambda > \max(0, \int_{\mu_1^{min}}^{\mu_2^{max}} d\mathbb{P}_Y(y) - \alpha)$ is respected, our optimal interval prediction loss is convex. Hence, it has a unique minimum.

Both the existence and the uniqueness of the minimum have been proven. □

Theorem 3.8 (Validity of RQR-O - A variation of the validity of orthogonal quantile regression theorem from Feldman et al. (2021))**.** *Suppose $Y|X = x$ follows a continuous distribution for each $x \in \mathcal{X}$, and suppose that $\mu_1(X), \mu_2(X) \in \mathcal{F}$. Consider the infinite-data version of the RQR-O optimization :*

$$\arg\min_{\mu_1,\mu_2 \in \mathbb{F}} \{\mathbb{E}(\mathcal{L}_\alpha^{RQR}((\mu_1, \mu_2), X, Y) + \gamma \mathbb{R}(w, m))\}$$

Then, true conditional intervals with α coverage are solutions to the above optimization problem.

Proof. We note $m = \mathbb{I}_{y \in [\mu_1, \mu_2]}$ the coverage function and $w = |\mu_2 - \mu_1|$ the interval length function. We consider the true conditional intervals that satisfied $\mathbb{P}(\mu_1(X) < Y < \mu_2(X)|X = x) = \alpha$. Feldman et al. (2021) has shown that for such intervals, the coverage and interval length functions are independent.

Therefore, similarly to vanilla QR or the Winkler score, these true conditional intervals are a solution to the RQR problem. Moreover, by definition, the independence of w and m fixes the Pearson correlation (or the HSIC score) of these functions to 0. Thus, the orthogonal penalty term that is based on these metrics is also minimized. Hence, the true conditional intervals with α coverage are a solution to the RQR-0 optimization problem. □

C. Experimental Details

We follow standard preprocessing on all datasets with features standardized such that they have zero mean and unit variance and targets divided by their mean. All experiments are repeated over 10 random seeds with means and standard errors of means reported throughout. We train two-layer neural networks with 64 hidden units and ReLU activations throughout (consistent with Feldman et al. (2021)). We also use the Adam optimizer (Kingma & Ba, 2014).

In the benchmarking of RQR-W, we applied the following experimental design. A training-validation-testing split with a ratio [0.6,0.2,0.2] is applied. Then we perform a hyperparameter grid search for all methods. Each combination is first fit to the training data with the model selection based on evaluations on the validation set. For a given hyperparameter combination, the best model is selected based on the epoch that achieves the best interval length (such that target coverage is achieved). The reported values are the evaluations on the test set. All methods evaluated follow this protocol. The grid search considers the following hyperparameters: dropout probability $\in \{0.1, 0.2, 0.3\}$, learning rate $\in [0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001]$, and regularization coefficient $\in [0.01, 0.1, 1, 5, 10, 20, 30, 40, 50]$. Other hyperparameters are fixed: number of epochs 400, batch size 10000.

In the benchmarking of RQR-O, we followed the exact procedure and used the implementation of the OQR baseline in Feldman et al. (2021). In what follows, we will describe this procedure. A training-validation-testing split with a ratio of 0.4 for testing and a further split of 0.9-0.1 for training-validation is applied. The default hyperparameters are used for both methods with only the regularization coefficient tuned. Specifically, it is set to 1 and then decreased in increments following $[1, 0.5, 0.1, 0.05, \ldots]$ until the desired coverage is achieved. Both the RQR-O and OQR hyperparameters are - learning rate: 1e-3, maximum number of epochs: 10000, dropout probability: 0, and batch size: 1024. Early stopping patience is set to 200 epochs.

D. Formal Description of Evaluation Metrics

To ensure this work is self-contained, in this section we include a formal description of the evaluation metrics introduced in Section 2 and used in Section 4. We implement these metrics as described in the referenced works. In all cases, we evaluate on some evaluation dataset $\mathcal{D}$ consisting of N input/target pairs $(\mathbf{x}, y)$.

Definition D.1 (Prediction Interval Coverage Probability (PICP) (Tagasovska & Lopez-Paz, 2019))**.** Defined as the number of true observations falling inside the estimated prediction interval, this is calculated as

$$PICP := \frac{1}{N} \sum_{(\mathbf{x},y)\sim\mathcal{D}} \mathbb{I}_{\mu_1 \leq y \leq \mu_2}.$$

Table 6. Summary statistics of the standard benchmark datasets for quantile regression.

Dataset	Mean	Variance	Skewness	Kurtosis
concrete	35.82	279.08	0.42	-0.32
wine	5.64	0.65	0.22	0.29
yacht	10.50	229.84	1.75	2.00
energy	22.31	101.81	0.36	-1.25
kin8nm	0.71	0.07	0.09	-0.53
naval	0.99	0.00	-0.00	-1.20
power	454.37	291.28	0.31	-1.05
boston	22.53	84.59	1.10	1.47
protein	7.75	37.43	0.57	-1.14

Definition D.2 (Mean Prediction Interval Width (MPIW) (Tagasovska & Lopez-Paz, 2019))**.** Defined as the average interval width across the evaluation dataset, this is calculated as

$$MPIW := \frac{1}{N} \sum_{(\mathbf{x},y)\sim\mathcal{D}} |\mu_2 - \mu_1|.$$

Definition D.3 (Width-coverage Pearson's Correlation (WCPC) (Feldman et al., 2021))**.** Denoting $\mathbf{w}$ as the vector of interval widths where $w_i = |\mu_2(\mathbf{x}_i) - \mu_1(\mathbf{x}_i)|$ and $\mathbf{m}$ as the indicator vector of coverage events where $m_i = \mathbb{I}_{y_i \in [\mu_1(\mathbf{x}_i), \mu_2(\mathbf{x}_i)]}$ for $i \in \{1, 2, \ldots, N\}$, then we define

$$WCPC := \left| \frac{\text{Cov}(\mathbf{w}, \mathbf{m})}{\text{Var}(\mathbf{w})\text{Var}(\mathbf{m})} \right|.$$

Definition D.4 (Hilbert-Schmidt Independence Criterion (HSIC) (Greenfeld & Shalit, 2020; Feldman et al., 2021))**.** In this definition, upper case bold letters denote matrices. Using the same definitions of $\mathbf{w}$ and $\mathbf{m}$ from Definition D.3, the coverage kernel matrix is given by $\mathbf{R}_{i,j} = k(\mathbf{m}_i, \mathbf{m}_j)$ and the width kernel is given by $\mathbf{K}_{i,j} = k(\mathbf{w}_i, \mathbf{w}_j)$ where k denotes the Gaussian kernel. We also introduce a centering matrix $\mathbf{H}_{i,j} = \delta_{i,j} - \frac{1}{N}$ where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Then we can calculate HSIC as

$$HSIC := \sqrt{\frac{tr(\mathbf{KHRH})}{(N-1)^2}}$$

E. Crossing Bounds

A well-known limitation of the quantile regression approach to constructing prediction intervals is any estimation error of the population level intervals can result in *crossing bounds* where the upper bound of the interval falls *below* the lower bound of the interval (see e.g. Brando et al. (2022); Park et al. (2022)). Apart from being a conceptual limitation, this can also affect the users' trust in the system. A key advantage of directly estimating the interval as proposed in this work is that the raw model outputs are not strictly associated with being a particular bound. In other words, our objective is invariant to permutations of the upper and lower bounds, and simply sets the lower bound to be the minimum and the upper bound to be the maximum of the two model outputs. Therefore, *which* output neuron acts as either bound can even change on a per-example basis. Formally, this can be observed in the $\kappa = (y - \mu_1)(y - \mu_2)$ term in our loss function (Equation (1)) which is clearly invariant to permutations between μ_1 and μ_2. In contrast, crossing bounds in the case of quantile regression will result in miscoverage if the bounds are permuted to avoid a negative interval. This may be considered a distinct advantage of the direct interval prediction approach which sidesteps the conceptual issue of crossing quantiles which occurs due to the independent estimation of the upper and lower bounds in quantile-based approaches.

Recent work in Brando et al. (2022) has proposed methods for preventing crossing bounds in the case of quantile regression by using Chebyshev polynomials to add additional constraints into the the objective. In what follows we compare our proposed RQR-W method to this non-crossing quantile method using the same experimental setup as in Table 3. As this method is a quantile estimation method and does not prescribe how to construct intervals we include two approaches (a) "narrowest" where we bin search the narrowest interval at each new inference and (b) "symmetric" where we use the symmetric quantiles (0.05, 0.95). The results are included in Table 7.

Table 7. **Comparison between RQR-W vs Deep Non-crossing Quantiles**. We present two different versions of the deep non-crossing quantiles method: "Narrowest" and "Symmetric" (see text for details). For the 9 datasets, we display interval width and the coverage level achieved in parentheses.

Dataset	RQR-W	Narrowest	Symmetric
concrete	0.43 (89.95)	0.20 (57.33)	0.30 (73.30)
boston	0.51 (89.8)	0.26 (64.90)	0.36 (78.53)
naval	0.01 (90.95)	0.01 (97.43)	0.02 (99.04)
energy	0.13 (89.87)	0.07 (50.39)	0.14 (72.86)
wine	0.35 (89.44)	0.19 (62.56)	0.28 (76.75)
power	0.03 (90.73)	0.02 (73.70)	0.03 (89.83)
protein	1.59 (90.07)	1.13 (78.04)	1.50 (89.26)
kin8nm	0.34 (90.59)	0.25 (70.96)	0.38 (85.87)
yacht	0.27 (93.23)	0.29 (51.29)	0.35 (61.61)

In these results we note that, as with the previously evaluated methods for estimating all quantiles simultaneously, the non-crossing quantile approach consistently struggles to maintain coverage at test time. We seek solutions that minimize interval width *for a fixed level of coverage* which is not achieved by this baseline which we attribute to two limitations. Firstly, when we select the narrowest possible interval from a set of possible intervals that each obtain marginal coverage, it is not a random choice and it is *more* likely that we are selecting a particular interval in which coverage is not maintained. Therefore, analogous to the effect of *overfitting* in the standard point prediction setting, we would expect that picking the narrowest intervals is likely to result in more miscoverage events than selecting a fixed interval (as is reflected in these results). Secondly, learning all quantiles simultaneously is a more challenging learning problem than simply learning two quantiles for a fixed capacity model. This results in poorer predictive performance at the task in hand.

F. Motivation for Optimizing Interval Width & Conditional Coverage

Although several methods might achieve a desired level of marginal coverage with their intervals, the task of finding a set of intervals that obtain such coverage is generally an underspecified problem with a potentially infinite number of admissible solutions. Given this, any interval-producing method is required to introduce some additional regularization (either implicit or explicit) in order to select among the possible solutions. In the case of quantile regression, the pair of symmetric quantiles are selected. However, given that other attributes are typically desired in these intervals (as illustrated by the large body of work that attempts to minimize interval width (Chung et al., 2021; Tagasovska & Lopez-Paz, 2019; Pearce et al., 2018) or maximize conditional coverage (Feldman et al., 2021; Hunter & Lange, 2000)), for many applications there are likely more preferable solutions than the symmetric quantiles found via quantile regression. This has been observed in real-world applications such as renewable energy forecasting where it has been noted that "the probability distribution of the renewable energy source's power output is generally skewed, thereby the width of central prediction intervals is often unnecessarily wide" (Zhang et al., 2023).

One additional motivation for minimizing the width of intervals beyond the advantage of narrower predictive intervals is that the minimal width intervals generally lie in denser regions of the underlying probability distribution. Because we estimate these statistics using a sample from the population distribution we typically incur some variance in this estimate which is typically lower in the more dense regions of the probability space. To illustrate this, consider the distribution we presented in Figure 1 of the main text where the ground truth intervals are known. Now suppose we take a sample from this distribution (i.e. a training set) and estimate both the empirical minimum width intervals and the empirical quantiles. In both cases there is likely to be some estimation error. If we were to repeat this process multiple times we note that the variance around the estimates in lower-density regions (where the symmetric quantiles lie) is likely to be greater than in the high-density regions (where the minimum width intervals lie). As a consequence of this, quantile estimation methods may suffer from larger errors in estimating true quantiles (particularly in small sample regimes) resulting in more difficulties in obtaining exact coverage. We verify this claim empirically in Figure 4.

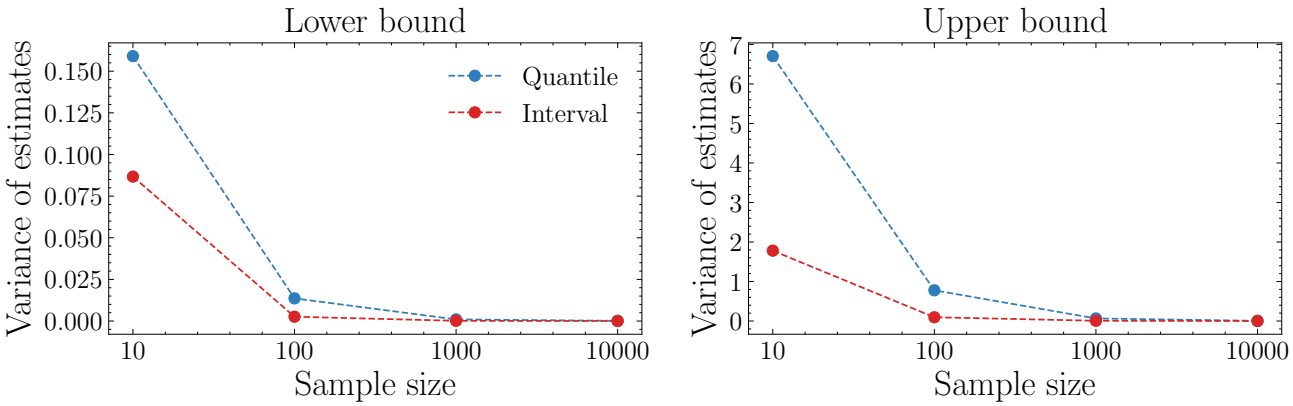

Figure 4. **Variance of the estimated bounds.** Estimating the minimal width interval rather than symmetric quantiles is also likely to result in lower variance estimates of the two bounds due to being estimated in more dense regions of the distribution. This is illustrated on the log-normal distribution example from Figure 1. We take samples of various sizes from this distribution and estimate both the quantiles and the minimum width bounds. We find that the variance of these estimates is significantly larger when estimating the former.

G. A Special Case: The Gumbel Distribution

Given the analysis provided in Section 3.3, where we illustrate that both the RQR and RQR-W objectives can narrow for coverage events and widen for miscoverage events for a single example, we might wish to better understand how they differ across a full data distribution (i.e. considering a mixture of coverage and miscoverage events). In Section 4 we demonstrate empirically that RQR-W does indeed find narrower solutions in aggregate, but investigating this analytically is typically not tractable. In this section, we provide the special case of a target that follows the Gumbel distribution in which we can derive a solution in closed form. Specifically, we compare the optimal solutions found by RQR and RQR-W and show that while both solutions achieve target coverage, the latter obtains narrower aggregate intervals.

We consider the Gumbel distribution $\mathcal{G}$ (with probability density function (PDF): $x \to e^{-(x+e^{-x})}$ and cumulative distribution function (CDF): $x \to e^{-e^{(-x)}}$) as this distribution has a closed form inverse CDF: $Q : p \to -ln(-ln(p))$ which makes the calculus tractable.

Given the random variable $Y \sim \mathcal{G}$, for a coverage α, the subset of valid intervals (μ_l, μ_u) is characterised by the following equality $P(\mu_l \leq Y <= \mu_u) = \alpha$.

Thus, in this subset, we can express μ_u with respect to μ_l :

$$
\begin{aligned}
CDF(\mu_u) - CDF(\mu_l) &= \int_{\mu_l}^{\mu_u} yPdy = \alpha \\
\implies \mu_u &= Q(\alpha + CDF(\mu_l)) \\
\implies \mu_u &= -ln(-ln(\alpha + e^{-e^{-\mu_l}}))
\end{aligned}
$$

Hence, on the subset of valid intervals, our 2-dimensional optimization problem (finding μ_u and μ_l) becomes a 1-dimensional problem (finding μ_l). Moreover, we can express the interval width $W = \mu_u - \mu_l$ as a function of μ_l such that $W : \mu_l \to -ln(-ln(\alpha + e^{-e^{-\mu_l}})) - \mu_l$ and both losses as a function of μ_l in a similar manner ($\mathcal{L}^{\text{RQR}}(\mu_l, \mu_u) = \mathcal{L}^{\text{RQR}}(\mu_l)$ and $\mathcal{L}^{\text{RQR-W}}(\mu_l, \mu_u) = \mathcal{L}^{\text{RQR-W}}(\mu_l)$).

Finally, we want to show that the width of the interval found by optimizing $\mathcal{L}^{\text{RQR-W}}$ is less than the one obtained by optimizing $\mathcal{L}^{\text{RQR}}$. Thus, in Figure 5, for different values of μ_l, we plot the expected value of $\mathcal{L}^{\text{RQR}}$, $\mathcal{L}^{\text{RQR-W}}$, and the respective widths of the resulting intervals (i.e. having μ_l as a lower bound).

We name μ_l^{RQR} (respectively μ_l^{RQR-W}), the lower bound that characterized the interval obtained by optimizing for the RQR objective (respectively the RQR-W objective), i.e μ_l^{RQR} such that $\mu_l^{RQR} = \arg\min_{\mu_l} \mathcal{L}^{\text{RQR}}(\mu_l)$.

In Figure 5, we observe that $\mu_l^{RQR} > \mu_l^{RQR-W}$ as $\mu_l^{RQR} \approx -0.9$ and $\mu_l^{RQR} \approx -1.2$. Given the increasing width of

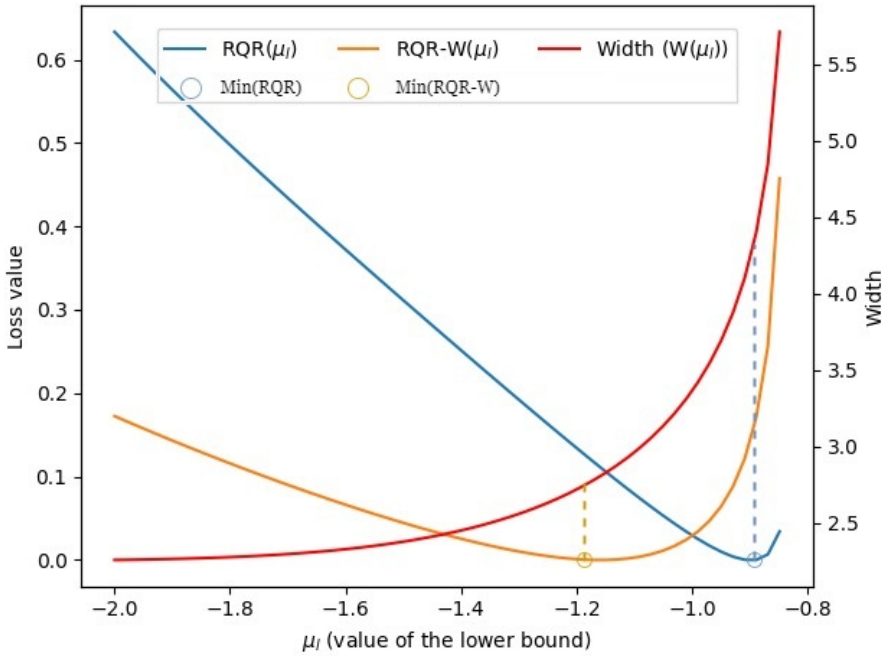

Figure 5. **A closed form solution for the Gumbel distribution.** Comparison of the minimum reached by the one-dimensional losses $\mathcal{L}^{\text{RQR}}(\mu_l)$ and $\mathcal{L}^{\text{RQR-W}}(\mu_l)$. Note that the dashed lines visually map each loss functions minimum loss to their corresponding interval width.

the interval with respect to μ_l (red curve), we conclude that $W(\mu_l^{RQR}) > W(\mu_l^{RQR-W})$. **Hence, the interval found by optimizing RQR is wider than the interval found by optimizing RQR-W**. This demonstrates that including the additional regularization term in RQR-W does indeed find a solution with narrower intervals in aggregate.

H. Additional Results

H.1. Complete RQR-W Results

In Table 8 we present the complete results across the 12 datasets evaluating RQR-W against standard baselines (as summarized in Table 3). For each dataset, we report both the mean coverage obtained (where the target was 90%) and the mean interval width over 10 runs. One standard error is included in parentheses. Of course, shorter intervals are only desirable when the target level of coverage is maintained. Therefore, we exclude results that fail to achieve coverage which we indicate with a ~~strikethrough~~. In line with previous work of Tagasovska & Lopez-Paz (2019), we consider coverage to be met if the empirical coverage lies within 2.5% of the target level α (after accounting for uncertainty). While obtaining empirical coverage that is greater than the desired coverage level may often be less harmful than obtaining less than the desired coverage level, at a minimum this still reflects an inefficiency. Additionally, there are many applications in which we are primarily interested in the miscoverage cases (e.g. extreme events) and, therefore, *overcoverage* may be problematic in addition to being inefficient. Thus, this protocol symmetrically discards intervals that undercover and overcover by a certain margin.

Table 8. **Benchmarking RQR-W**. We evaluate each method across 12 datasets where the first row reports the marginal coverage obtained and the second row reports interval width as measured using MPIW. All results report the test set mean over 10 runs ($\pm$ a standard error). We ~~strikethrough~~ results where desired coverage is not achieved (see text for details).

Dataset	Ours	QR	SQR-C	SQR-N	IR	Dist. Hist.
concrete	87.33 (1.08) **1.36** (0.02)	~~85.83~~ (1.04) 1.44 (0.02)	86.31 (1.44) 1.40 (0.04)	86.26 (1.61) **1.35** (0.03)	90.00 (0.79) 1.65 (0.12)	
power	88.83 (0.25) **0.11** (0.00)	87.48 (0.30) **0.11** (0.00)	89.39 (0.97) 0.16 (0.03)	~~86.44~~ (1.15) 0.16 (0.03)	~~62.18~~ (13.58) 1.09 (0.22)	
wine	89.78 (0.42) 0.42 (0.01)	92.97 (0.82) **0.37** (0.01)	90.03 (0.88) 0.46 (0.03)	~~86.88~~ (0.76) 0.45 (0.03)	90.44 (0.51) 1.07 (0.16)	
yacht	88.55 (0.88) **3.44** (0.33)	88.23 (1.63) **3.44** (0.44)	~~86.29~~ (1.05) 3.61 (0.39)	~~85.65~~ (1.57) 2.91 (0.25)	88.71 (1.42) 4.99 (1.21)	
naval	89.01 (0.37) 0.03 (0.00)	88.06 (0.37) **0.02** (0.00)	89.26 (1.48) 0.32 (0.02)	88.47 (1.13) 0.32 (0.02)	89.77 (0.38) 2.56 (0.21)	
energy	88.57 (0.76) **1.24** (0.02)	88.05 (1.57) **1.25** (0.02)	88.57 (1.04) 1.30 (0.01)	~~86.30~~ (1.18) 1.24 (0.01)	88.83 (0.67) 1.58 (0.15)	
boston	87.16 (0.64) **1.08** (0.01)	~~85.39~~ (1.07) 1.21 (0.02)	~~84.12~~ (0.78) 1.12 (0.04)	~~82.94~~ (0.79) 1.03 (0.03)	88.82 (0.61) 1.37 (0.06)	
kin8nm	87.63 (0.19) **1.15** (0.00)	87.45 (0.32) 1.16 (0.00)	87.59 (0.43) **1.15** (0.01)	~~85.85~~ (0.57) 1.12 (0.01)	90.23 (0.28) 1.36 (0.05)	
protein	88.15 (0.10) **2.12** (0.01)	89.07 (0.25) 2.18 (0.00)	88.57 (0.16) 2.17 (0.02)	~~85.61~~ (0.27) 2.04 (0.01)	89.88 (0.32) 2.26 (0.08)	
sulfur	87.69 (0.28) 1.11 (0.02)	87.94 (0.31) 1.15 (0.01)	87.65 (0.46) 1.09 (0.02)	87.39 (0.50) **1.02** (0.01)	89.98 (0.34) 1.18 (0.04)	
cpu_act	89.51 (0.34) 0.72 (0.02)	88.20 (0.34) **0.67** (0.00)	89.52 (0.74) 0.77 (0.01)	~~83.86~~ (1.08) 0.75 (0.02)	90.13 (0.35) 1.82 (0.23)	
miami	88.22 (0.21) **1.43** (0.02)	87.93 (0.30) 1.60 (0.03)	~~87.12~~ (0.46) 1.91 (0.05)	~~86.05~~ (0.44) 1.55 (0.03)	89.89 (0.28) 1.85 (0.05)	

H.2. Evaluating Different Coverage Levels

Lower coverage intervals (i.e. narrower intervals) can lead to greater instability in the estimates. Thus, it is valuable to evaluate our method in the low coverage probability regime to ensure this increased instability doesn't have a disproportionally large effect on our method. We empirically analyze the relative performance of our method for 70%, 50%, and 30% coverage levels in Table 9, Table 10 and Table 11,.

Table 9. **Comparison on 70% targeted coverage.** For the 9 datasets, we display interval width and the coverage level achieved in parentheses ± a standard error for both. All results report the test set mean over 10 runs.

Dataset	Ours	QR	SQR-C	SQR-N	IR
boston	66.76 (1.20)	67.35 (1.72)	68.63 (2.04)	65.20 (2.51)	69.90 (0.53)
	0.65 (0.02)	0.66 (0.01)	**0.70 (0.02)**	0.62 (0.02)	0.89 (0.10)
protein	68.95 (0.24)	68.83 (0.48)	70.60 (0.44)	61.98 (1.63)	69.09 (0.30)
	1.58 (0.00)	**1.53 (0.02)**	1.75 (0.03)	1.36 (0.04)	1.50 (0.06)
concrete	67.18 (1.24)	66.50 (1.51)	68.50 (1.42)	63.16 (1.94)	68.98 (0.34)
	0.94 (0.02)	0.95 (0.02)	0.97 (0.02)	0.89 (0.02)	1.05 (0.06)
energy	68.18 (0.86)	66.69 (1.08)	70.58 (2.02)	57.27 (1.94)	69.87 (0.49)
	0.88 (0.02)	0.90 (0.01)	1.02 (0.02)	0.80 (0.03)	1.09 (0.09)
kin8nm	68.16 (0.13)	68.04 (0.43)	69.27 (0.41)	62.20 (1.06)	69.91 (0.33)
	0.78 (0.00)	0.79 (0.00)	0.81 (0.01)	0.72 (0.02)	0.86 (0.03)
naval	69.07 (0.18)	68.58 (0.61)	70.56 (1.15)	60.65 (3.02)	70.34 (0.28)
	0.02 (0.00)	**0.02 (0.00)**	0.12 (0.03)	0.11 (0.03)	1.16 (0.17)
power	69.14 (0.23)	68.31 (0.83)	69.99 (1.17)	56.77 (1.45)	49.11 (10.72)
	0.09 (0.00)	**0.08 (0.00)**	0.11 (0.02)	0.09 (0.02)	3e8 (1.5e8)
wine	69.78 (0.24)	72.44 (1.40)	76.59 (1.58)	72.09 (3.41)	70.78 (0.41)
	0.26 (0.02)	0.23 (0.01)	0.27 (0.02)	**0.22 (0.02)**	0.94 (0.14)
yacht	69.68 (0.63)	68.55 (3.32)	72.10 (2.37)	62.26 (4.83)	69.84 (1.75)
	2.04 (0.19)	1.52 (0.20)	1.63 (0.17)	1.07 (0.05)	**1.35 (0.24)**

As we might expect, both methods have increased variance in their estimates at this level. The mean standard errors for the 90% coverage rate are 0.5 and 0.8 for RQR-W and QR respectively while they are 0.7 (RQR-W) and 1.1 (QR) for the 50% coverage rate. Using the same definition of achieving coverage as in Table 3, we find that over the 3 different lower coverage rates, RQR-W outperforms other methods in both coverage rate and interval width. RQR-W reaches the coverage targeted coverage rate in 23/27 cases and outputs the smallest intervals in 14/27 cases. By comparison, the second best performing loss is QR which only reaches the desired coverage in 17/27 cases and outputs the smallest interval in 10/27 cases. Overall this is consistent with previous results indicating that RQR-W also obtains more narrow intervals when we consider alternative coverage levels.

Table 10. **Comparison on 50% targeted coverage.** For the 9 datasets, we display interval width and the coverage level achieved in parentheses ± a standard error for both. All results report the test set mean over 10 runs.

Dataset	Ours	QR	SQR-C	SQR-N	IR
boston	45.10 (1.48)	46.27 (1.29)	49.31 (1.70)	35.59 (2.33)	49.51 (0.62)
	0.34 (0.01)	0.33 (0.01)	**0.38 (0.02)**	0.34 (0.02)	0.65 (0.13)
protein	49.82 (0.23)	49.20 (0.32)	50.22 (0.29)	44.56 (1.87)	49.41 (0.25)
	1.28 (0.05)	1.04 (0.05)	1.37 (0.01)	0.54 (0.05)	**0.65 (0.04)**
concrete	48.30 (0.95)	48.30 (1.83)	50.78 (1.89)	36.94 (3.37)	50.10 (0.35)
	0.60 (0.01)	**0.59 (0.01)**	0.64 (0.02)	0.52 (0.05)	0.75 (0.06)
energy	49.74 (1.09)	46.36 (1.73)	53.57 (1.29)	45.26 (2.80)	48.51 (1.05)
	0.57 (0.04)	0.57 (0.03)	0.78 (0.01)	0.49 (0.05)	0.76 (0.08)
kin8nm	48.61 (0.40)	48.23 (0.47)	50.04 (0.65)	41.64 (3.07)	49.82 (0.16)
	0.51 (0.00)	**0.51 (0.00)**	0.53 (0.01)	0.47 (0.03)	0.92 (0.29)
naval	49.83 (0.26)	49.46 (0.35)	49.30 (1.21)	33.89 (1.15)	49.33 (0.27)
	0.01 (0.00)	**0.01 (0.00)**	0.10 (0.02)	0.09 (0.02)	2.48 (1.13)
power	50.11 (0.32)	48.96 (0.38)	50.84 (1.18)	42.07 (2.02)	34.93 (7.63)
	0.06 (0.00)	**0.06 (0.00)**	0.06 (0.00)	0.05 (0.00)	3e8 (2e8)
wine	49.97 (0.26)	47.63 (1.54)	55.41 (2.68)	40.97 (1.32)	49.97 (0.75)
	0.18 (0.01)	**0.18 (0.00)**	0.20 (0.01)	0.14 (0.02)	0.81 (0.13)
yacht	50.00 (1.18)	49.19 (2.68)	53.71 (4.19)	53.39 (4.45)	49.03 (1.63)
	1.76 (0.11)	0.78 (0.07)	1.25 (0.06)	**0.50 (0.07)**	**0.50 (0.05)**

Table 11. **Comparison on 30% targeted coverage.** For the 9 datasets, we display interval width and the coverage level achieved in parentheses ± a standard error for both. All results report the test set mean over 10 runs.

Dataset	Ours	QR	SQR-C	SQR-N	IR
boston	28.53 (0.69)	25.39 (2.31)	32.55 (2.14)	15.98 (2.69)	30.20 (0.43)
	0.18 (0.01)	0.16 (0.00)	**0.20 (0.02)**	0.15 (0.02)	4.37 (3.87)
protein	29.97 (0.19)	29.85 (0.35)	29.65 (0.28)	15.60 (2.24)	29.76 (0.09)
	1.01 (0.06)	0.58 (0.03)	0.87 (0.01)	0.11 (0.02)	**0.32 (0.04)**
concrete	28.20 (0.88)	27.52 (1.13)	29.81 (1.70)	16.36 (4.30)	30.00 (0.38)
	0.31 (0.01)	0.32 (0.01)	**0.34 (0.01)**	0.20 (0.05)	0.61 (0.09)
energy	29.35 (0.63)	27.92 (1.08)	31.82 (1.31)	17.60 (2.71)	29.61 (0.90)
	0.29 (0.04)	0.27 (0.01)	0.54 (0.01)	0.14 (0.02)	0.64 (0.18)
kin8nm	29.30 (0.39)	28.17 (0.70)	30.45 (0.61)	22.15 (2.79)	30.25 (0.43)
	0.30 (0.01)	0.29 (0.00)	0.31 (0.00)	0.25 (0.03)	3.61 (1.80)
naval	29.88 (0.13)	29.88 (0.24)	29.36 (1.18)	12.39 (1.65)	30.00 (0.13)
	0.01 (0.00)	**0.01 (0.00)**	0.05 (0.02)	0.04 (0.01)	0.69 (0.15)
power	30.01 (0.19)	29.44 (0.35)	31.05 (0.84)	19.04 (1.42)	20.85 (4.55)
	0.03 (0.00)	**0.03 (0.00)**	0.07 (0.02)	0.05 (0.01)	3e8 (2e8)
wine	29.59 (0.18)	29.88 (1.53)	31.53 (1.32)	24.63 (3.24)	29.88 (0.33)
	0.13 (0.01)	**0.13 (0.00)**	0.16 (0.00)	0.11 (0.02)	0.95 (0.26)
yacht	24.84 (1.63)	25.32 (2.65)	30.81 (3.10)	20.81 (2.92)	29.68 (0.77)
	0.71 (0.19)	0.30 (0.05)	0.63 (0.05)	0.11 (0.02)	**0.20 (0.03)**

H.3. Non-Smoothed kin8nm Intervals

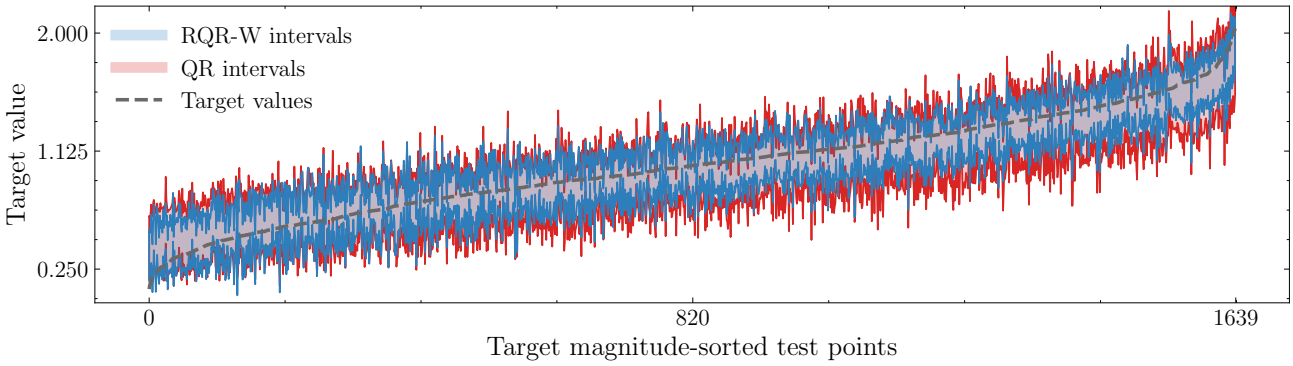

Figure 6. **Non-smoothed kin8nm intervals.** Figure 3 (left) without Savitzky–Golay filter applied. In this version it is apparent that micoverage events do occur as we would expect while the difference in interval width is less legible.