CaliPSo: Calibrated Predictive Models with Sharpness as Loss Function

Anonymous Authors[1]

Abstract

Conformal prediction methods have become increasingly common for accurately capturing uncertainty with machine learning models. However, conformal prediction typically recalibrates an existing model, making it heavily reliant on the quality of the uncalibrated model. Moreover, they either enforce marginal calibration strictly, yielding potentially coarse predictive intervals, or attempt to strike a balance between interval coarseness and calibration. Motivated by these shortcomings, we present CaliPSo a neural network model that is marginally calibrated out-of-the-box and stays so throughout training. This property is achieved by adding a model-dependent constant to the model prediction that shifts it in a way that ensures calibration. During training, we then leverage this to focus exclusively on sharpness - the property of returning tight predictive intervals - rendering the model more useful at test time. We show thorough experimental results, where our method exhibits superior performance compared to several state-of-the-art approaches.

1. Introduction

Though conformal prediction has seen increasing interest in recent years (Zhan et al., 2022; Ren et al., 2023; Sun et al., 2024), the underlying methods present several drawbacks. Conformal prediction methods recalibrate an existing model by relabeling the quantiles according to their actual coverage on a holdout dataset (Kuleshov et al., 2018). Although this yields a calibrated model, it comes at the expense of poorer sharpness, meaning that the model is potentially less informative in certain regions of the input space. To address these shortcomings, an increasing amount of algorithms have attempted to trade off calibration and sharpness during training. These include minimizing a weighted sum of sharpness and calibration terms (Chung et al., 2021), considering alternative losses with weaker calibration guarantees but better sharpness (Song et al., 2019; Kuleshov & Deshpande, 2022), and backpropagating through a differentiable recalibration procedure during training (Dheur & Ben taieb, 2024). However, while some of these methods do not enforce calibration, none optimize sharpness directly.

Motivated by the shortcomings mentioned above, we present **cali**brated **p**redictions using **s**harpness as a l**o**ss function (CaliPSo). Our method guarantees marginal calibration on the training data at all times, allowing us to optimize exclusively for sharpness during training. Our approach relies on two novel paradigms for training quantiles. Our approach is easy to implement and can extend any conventional regression approach to obtain sharply calibrated quantiles. Firstly, we employ additive terms to enforce calibration during training. This contrasts with conventional recalibration techniques, which relabel quantiles based on their coverage. Secondly, we use a different subset of the data to train each quantile, where each subset is computed sequentially using quantiles corresponding to diminishing coverage intervals.

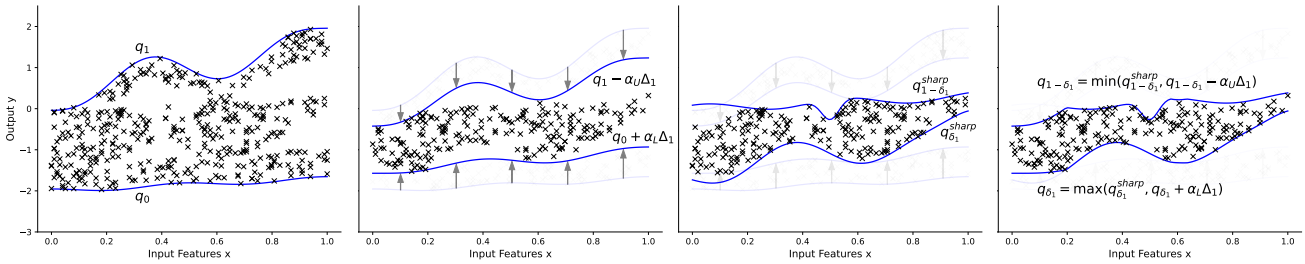
2. Preliminaries and Problem Statement

This section introduces the regression problem considered in this paper along with a formal definition of calibration for regression models.

We consider a regression setting, where the inputs X and targets Y are random variables taking values $x \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$. We use $F_X : \mathcal{X} \to [0,1]$, $F_{Y|x} : \mathcal{Y} \to [0,1]$ and $F_Y : \mathcal{Y} \to [0,1]$ to denote the cumulative density function

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) (b) (c) (d)

Figure 1. Data partition and quantile generation procedure. The 0 and 1 quantiles are trained by minimizing sharpness while containing the full data set (a). Given $\Delta_1 = q_0 - q_1$, we then compute the differences $\alpha_L\Delta_1$ and $\alpha_U\Delta_1$ by which we must move inward from q_0 and q_1 until $1 - 2\delta_1$ of the data is contained (b). The remaining data is used to train tight bounding functions by minimizing the L_1 loss while guaranteeing coverage (c). At test time, we apply minimum and maximum operations to the models optimized for sharpness, guaranteeing calibration (d). This operation is repeated for an arbitrary number of quantile pairs.

(CDF) of X, the CDF of Y conditioned on $X = x$, and the marginal CDF of Y, respectively.

We aim to obtain a model of the conditional CDF $F_{Y|x}$, which we denote as $\hat{F}_{Y|x}$. For any given $\delta \in [0, 1]$, we denote the corresponding modeled quantile function by $q_\delta(x) \coloneqq \hat{F}^{-1}_{Y|x}(\delta)$. To derive a model, we assume to have an iid data set $\mathcal{D} = \{x_i, y_i\}_{i=1,\dots,N}$, where $x_i \sim F_X$ and $y_i \sim F_{Y|x_i}$.

2.1. Model Calibration

Marginal calibration measures the accuracy of probabilistic models. Intuitively, we say that a model is marginally calibrated if a model of a conditional δ-quantile covers δ of the observed data.

Definition 2.1 (Marginal calibration)**.** The model $\hat{F}$ is said to be marginally calibrated if the corresponding quantile function $\hat{q}$ satisfies

$$\mathbb{P}_{x\sim F_X,\, y\sim F_{Y|x}}\left(q_\delta(x) \leq y\right) = \delta, \qquad \forall\, \delta \in [0, 1]. \tag{1}$$

Marginal calibration, commonly known simply as calibration, holds if the model outputs correct marginal quantiles. This notion of calibration implies that the conditional quantiles are correct on average, and can be approximately obtained by recalibrating any existing model (Kuleshov et al., 2018). Note that it is possible to have a marginally calibrated yet uninformative model, e.g., if a model always outputs the marginal quantiles, independently of the input x. For this reason, it is desirable to obtain marginally calibrated models that produce concentrated (low-entropy) conditional distributions. This property is typically known as sharpness (Gneiting et al., 2007), and corresponds to having the shortest possible distance between quantiles. Sharpness is measured in terms of the average covariance of the conditional distribution, but heuristics are also frequently used, e.g., the average length of centered 95% confidence intervals (Kuleshov et al., 2018).

In this work, we design a model that guarantees marginal calibration on the training data throughout training. This allows us to optimize it for sharpness.

3. Sharp Marginally Calibrated Model

We now introduce our model and describe how it is trained. Our method sequentially computes shrinking confidence intervals, starting with the zero and one quantiles. We begin by describing how they are obtained, then introduce the computation rule for quantiles for smaller confidence intervals.

3.1. Sharply Calibrated Zero and One Quantiles

We first model sharp and well-calibrated zero and one quantiles of the ground truth CDF F. For an arbitrary pair of regression models $\hat{q}_0^{\text{uncal}}, \hat{q}_1^{\text{uncal}} : \mathcal{X} \to \mathcal{Y}$, we obtain marginally calibrated quantile models $\hat{q}_0$, $\hat{q}_1$ by shifting the output such

that they cover the full dataset, i.e.,

$$\hat{q}_0(x) = \min_{x_i, y_i \in \mathcal{D}} (y_i - \hat{q}_0^{\text{uncal}}(x_i)) + \hat{q}_0^{\text{uncal}}(x), \qquad \hat{q}_1(x) = \max_{x_i, y_i \in \mathcal{D}} (y_i - \hat{q}_1^{\text{uncal}}(x_i)) + \hat{q}_1^{\text{uncal}}(x).$$

However, though $\hat{q}_0$ and $\hat{q}_1$ produce valid zero and one quantiles, they will likely be very coarse, i.e., the corresponding conditional intervals will be unnecessarily large. To avoid this, we can train the models to maximize sharpness - the average interval length produced by the quantile model pair. This is achieved by minimizing the average errors

$$\mathbb{E}_{x_i, y_i \sim \mathcal{D}} \left(\|\hat{q}_0(x_i) - y_i\|_1 \right) \qquad \mathbb{E}_{x_i, y_i \sim \mathcal{D}} \left(\|\hat{q}_1(x_i) - y_i\|_1 \right).$$

This modeling procedure has the advantage that the resulting quantiles are perfectly calibrated on the training data throughout training, allowing us to focus exclusively on sharpness.

3.2. Sharply Calibrated δ and $1 - \delta$ Quantiles

We now consider quantiles $\hat{q}_\delta$ and $\hat{q}_{1-\delta}$, where $0 < \delta < 0.5$. We obtain sharp calibrated models by computing the maximum/minimum over two models for each quantile:

$$\hat{q}_\delta(x) = \max \left(\hat{q}_\delta^{\text{cal}}(x), \hat{q}_\delta^{\text{sharp}}(x) \right), \qquad \hat{q}_{1-\delta} = \min \left(\hat{q}_{1-\delta}^{\text{cal}}(x), \hat{q}_{1-\delta}^{\text{sharp}}(x) \right), \tag{2}$$

where the superscripts cal and sharp denote models aimed at enforcing calibration and sharpness, respectively. We describe how each term is obtained in the following.

We compute the calibrated but potentially coarse components $\hat{q}_\delta^{\text{cal}}(x)$ and $\hat{q}_{1-\delta}^{\text{cal}}(x)$ by appropriately interpolating between $\hat{q}_0(x)$ and $\hat{q}_1(x)$:

$$\hat{q}_\delta^{\text{cal}}(x) = \hat{q}_0(x) + \alpha_L(\hat{q}_1(x) - \hat{q}_0(x)), \qquad \hat{q}_{1-\delta}^{\text{cal}}(x) = \hat{q}_0(x) + \alpha_U(\hat{q}_1(x) - \hat{q}_0(x)), \tag{3}$$

where α_L and α_U satisfy

$$\alpha_L = \text{quantile} \left(\delta, \frac{y - \hat{q}_0(x))}{\hat{q}_1(x)) - \hat{q}_0(x))} \right), \qquad \alpha_U = \text{quantile} \left(1 - \delta, \frac{y - \hat{q}_0(x))}{\hat{q}_1(x)) - \hat{q}_0(x))} \right).$$

This results in a model that is calibrated but not necessarily sharp. To obtain the sharp components $\hat{q}_\delta^{\text{sharp}}$ and $\hat{q}_{1-\delta}^{\text{sharp}}$, we employ a procedure similar to that used for the ones and zero quantiles. First, we use the coarse quantiles $\hat{q}_\delta^{\text{cal}}$ and $\hat{q}_{1-\delta}^{\text{cal}}$ to obtain the subset of the data $\mathcal{D}_{2\delta}$ within the corresponding confidence intervals:

$$\mathcal{D}_{2\delta} = \left\{ (x_i, y_i) \in \mathcal{D} \;\middle|\; \hat{q}_\delta^{\text{cal}}(x_i) \leq y_i \leq \hat{q}_{1-\delta}^{\text{cal}}(x_i) \right\}. \tag{4}$$

We then compute sharp quantiles, calibrated on $\mathcal{D}_{2\delta}$, using the same procedure as for the zero and one quantiles:

$$\hat{q}_\delta^{\text{sharp}}(x) = \min_{x_i, y_i \in \mathcal{D}_{2\delta}} (y_i - \hat{q}_\delta^{\text{uncal}}(x_i)) + \hat{q}_\delta^{\text{uncal}}(x), \qquad \hat{q}_{1-\delta}^{\text{sharp}}(x) = \max_{x_i, y_i \in \mathcal{D}_{2\delta}} (y_i - \hat{q}_{1-\delta}^{\text{uncal}}(x_i)) + \hat{q}_{1-\delta}^{\text{uncal}}(x),$$

where $\hat{q}_\delta^{\text{uncal}}$ and $\hat{q}_{1-\delta}^{\text{uncal}}$ are neural networks. We train these models by minimizing the average errors on $\mathcal{D}_{2\delta}$:

$$\mathbb{E}_{x_i, y_i \sim \mathcal{D}_{2\delta}} \left(\|\hat{q}_\delta^{\text{sharp}}(x_i) - y_i\|_1 \right) \qquad \mathbb{E}_{x_i, y_i \sim \mathcal{D}_{2\delta}} \left(\|\hat{q}_{1-\delta}^{\text{sharp}}(x_i) - y_i\|_1 \right).$$

This procedure is illustrated in Figure 1.

When retraining the zero and one quantiles, $\mathcal{D}_{2\delta}$ may change. However, this is not reflected in the gradients of $\hat{q}_\delta^{\text{sharp}}$ and $\hat{q}_{1-\delta}^{\text{sharp}}$. In practice, we did not observe this to be a concern, since convergence of the zero and one quantiles implies convergence of $\mathcal{D}_{2\delta}$, allowing for stable joint training. This procedure can be repeated arbitrarily frequently for increasing values of δ, where we use $\hat{q}_\delta$ and $\hat{q}_{1-\delta}$ to obtain calibrated components for subsequent models, allowing for more refined models.

Table 1. Expected calibration error and sharpness of different methods over 5 repetitions per experiment. We report the expected calibration error (ECE) and 95% confidence interval width (95% CI) obtained with our approach, CaliPSo, along with Model Agnostic Quantile Regression (Chung et al., 2021) (MAQR), QRTC (Dheur & Ben taieb, 2024). Lower is better for all metrics. The results are presented as the mean over 5 trials ± 1 standard error. The best mean result across the methods is highlighted in bold for each metric and dataset. If a method achieves the best performance on both metrics for a dataset, it is additionally highlighted in green, and if it fails to achieve the best in either metric for a dataset it is highlighted red. Our method is always the best in at least one metric.

DATA SET	METRIC	CALIPSO	MAQR	QRTC
BOSTON	ECE	0.0975 ± 0.0035	0.0926 ± 0.0147	$\mathbf{0.0723 \pm 0.0074}$
	SHARPNESS	$\mathbf{0.1222 \pm 0.0062}$	0.1603 ± 0.0143	0.2882 ± 0.0052
YACHT	ECE	0.1248 ± 0.0165	0.1233 ± 0.0086	$\mathbf{0.0879 \pm 0.0145}$
	SHARPNESS	$\mathbf{0.0213 \pm 0.0025}$	0.0216 ± 0.0027	0.3267 ± 0.0347
WINE	ECE	0.0473 ± 0.0088	0.0419 ± 0.0040	$\mathbf{0.0329 \pm 0.0082}$
	SHARPNESS	$\mathbf{0.3152 \pm 0.0345}$	0.3976 ± 0.0099	0.3774 ± 0.0039
CONCRETE	ECE	$\mathbf{0.0534 \pm 0.0096}$	0.0963 ± 0.0062	0.0544 ± 0.0027
	SHARPNESS	$\mathbf{0.1469 \pm 0.0106}$	0.1507 ± 0.0100	0.3274 ± 0.0045
KIN8NM	ECE	$\mathbf{0.0290 \pm 0.0033}$	0.0484 ± 0.0038	0.0340 ± 0.0031
	SHARPNESS	0.1499 ± 0.0051	$\mathbf{0.1483 \pm 0.0013}$	0.2068 ± 0.0060
ENERGY	ECE	$\mathbf{0.0736 \pm 0.0075}$	0.1279 ± 0.0092	0.0748 ± 0.0055
	SHARPNESS	0.0469 ± 0.0068	$\mathbf{0.0336 \pm 0.0022}$	0.1737 ± 0.0014
POWER	ECE	0.0144 ± 0.0017	0.0226 ± 0.0019	$\mathbf{0.0100 \pm 0.0010}$
	SHARPNESS	$\mathbf{0.1508 \pm 0.0021}$	0.1797 ± 0.0029	0.2016 ± 0.0011
NAVAL	ECE	$\mathbf{0.0096 \pm 0.0021}$	0.0194 ± 0.0049	0.0139 ± 0.0030
	SHARPNESS	0.0185 ± 0.0010	$\mathbf{0.0134 \pm 0.0007}$	0.2025 ± 0.0020

3.3. Early stopping

Although the proposed method enforces calibration on the training data, it risks overfitting if we do not define an adequate early-stopping criterion. To this end, we keep a heldout validation data set $\mathcal{D}^{\text{val}}$, such that $\mathcal{D}^{\text{val}} \cap \mathcal{D} = \emptyset$, which we use to estimate calibration at test time. Specifically, we only consider models where the expected calibration error $\mathcal{D}^{\text{val}}$ does not exceed a pre-specified hyperparameter $\epsilon > 0$. The expected calibration error (ECE), which measures calibration, is computed as (Kuleshov et al., 2018)

$$l_{\text{ece}} = \frac{1}{n_p} \sum_{j=1}^{n_p} |p_j - \hat{p}_j|, \qquad \hat{p}_j = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i=1}^{N_{\text{val}}} \mathbb{I}\Big(\hat{q}_{p_j}(x_i) \leq y_i\Big), \tag{5}$$

where $(x_i, y_i) \in \mathcal{D}_{\text{val}}$.

4. Experiments

Here we report experimental results on benchmark data sets from the UCI repository. We compare against two state-of-the-art methods: Model Agnostic Quantile Regression (MAQR) (Chung et al., 2021), QRTC (Dheur & Ben taieb, 2024). To assess the performance of each method using the expected calibration error (Kuleshov et al., 2018) and average length of the centered 95% intervals, a proxy metric for sharpness (Gneiting et al., 2007). The results are shown in Table 1.

We use 10 models in our method to predict the following quantiles: 0, 0.025, 0.05, 0.1, 0.2, 0.8, 0.9, 0.95, 0.975 and 1. We run 5 seeds for each experiment and use a train/validation/test split of 72%/18%/10%.

5. Conclusion

We have presented CaliPSo, a regression modeling approach that outputs marginally calibrated models, allowing us to train them exclusively by maximizing sharpness. Our approach is competitive with different state-of-the-art approaches on various UCI datasets, achieving the best performance on the concrete dataset.

References

Capone, A., Hirche, S., and Pleiss, G. Sharp calibrated gaussian processes. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011.

Dheur, V. and Ben taieb, S. A large-scale study of probabilistic calibration in neural network regression. In *The 40th International Conference on Machine Learning*, 2023.

Dheur, V. and Ben taieb, S. Probabilistic calibration by design for neural network regression. In *The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf.

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.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kuleshov, V. and Deshpande, S. Calibrated and sharp uncertainties in deep learning via density estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11683–11693. PMLR, 2022. URL https://proceedings.mlr.press/v162/kuleshov22a.html.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

MacKay, D. J. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

Marx, C., Zhao, S., Neiswanger, W., and Ermon, S. Modular conformal calibration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15180–15195. PMLR, 2022. URL https://proceedings.mlr.press/v162/marx22a.html.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

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.

Rasmussen, C. E. and Williams, C. K. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 2006.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5897–5906. PMLR, 2019. URL https://proceedings.mlr.press/v97/song19a.html.

Sun, J., Jiang, Y., Qiu, J., Nobel, P., Kochenderfer, M. J., and Schwager, M. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36, 2024.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005. doi: 10.1007/b106715.

Zhan, X., Wang, F., and Gevaert, O. Reliably filter drug-induced liver injury literature with natural language processing and conformal prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(10):5033–5041, 2022.

A. Related Work

A.1. Uncalibrated probabilistic models

The need to accurately quantify uncertainty has motivated the use and development of various probabilistic models for regression. Gaussian processes have been extensively used to quantify uncertainty due to their ability to differentiate between epistemic and aleatoric uncertainty (Rasmussen & Williams, 2006; Deisenroth & Rasmussen, 2011). Similarly, Bayesian neural networks (MacKay, 1995) aim to capture epistemic and aleatoric uncertainty by placing an appropriate prior over network parameters. More recently, other techniques have been developed for obtaining probabilistic deep neural networks, such as ensembles Lakshminarayanan et al. (2017) and dropout Gal et al. (2017).

A.2. Calibrated Classification

Most predictive models are not calibrated after training and need to undergo recalibration. Calibration was initially developed in the context of classification. To achieve calibration, tools from conformal prediction were commonly used, notably Platt scaling (Platt et al., 1999) and isotonic regression (Niculescu-Mizil & Caruana, 2005). There has been extensive work on obtaining calibrated models in the domain of classification. While there are many methods that do not employ post-processing, we only focus here on methods that employ some form of post-processing. Most forms of post-processing-based calibration for classification fall into the category of conformal methods (Vovk et al., 2005), which, given an input, aim to produce sets of labels that contain the true label with a pre-specified probability. Arguably the two most common forms of calibration are isotonic regression (Niculescu-Mizil & Caruana, 2005) and Platt scaling (Platt et al., 1999). In Niculescu-Mizil & Caruana (2005), Platt scaling and isotonic regression are analyzed extensively for different types of predictive models. In Guo et al. (2017), a modified form of Platt scaling for modern classification neural networks is proposed.

A.3. Calibrated Regression

More recently, conformal calibration techniques have been proposed to recalibrate machine learning models for regression (Kuleshov et al., 2018). A general overview of basic recalibration methods plus theoretical guarantees is provided in Marx et al. (2022). While these methods have been shown to yield well-calibrated models, the resulting predictive quantiles are potentially much too crude, resulting in predictions that perform poorly in terms of sharpness, i.e., the corresponding confidence intervals will overestimate the model error by a very large margin. As an alternative to purely achieving calibration, there have been works that explicitly attempt to balance calibration and sharpness. The works of Song et al. (2019); Kuleshov & Deshpande (2022) attempt to achieve distribution calibration, corresponding to minimizing the average pinball loss. However, while Song et al. (2019) relies on complex approximations and provides no theoretical guarantees, Kuleshov & Deshpande (2022) does not allow for optimizing for sharpness directly, and calibration is only guaranteed asymptotically as the number of data points grows. In Chung et al. (2021), the authors propose different losses to approximate the conditional distribution of the data, which is generally more challenging to achieve than marginal calibration. In Dheur & Ben taieb (2024), the authors propose a differentiable recalibration approach, which allows them to enforce calibration during training. Though this approach is similar to CaliPSo in some respects, the way calibration is enforced does not allow optimizing for sharpness. Hence, the authors minimize the negative log-likelihood. In a similar vein, (Capone et al., 2024) exploit properties of kernel models to formulate a flexible calibration approach that can be optimized for sharpness.

However, the model employed therein is very rigid and highly dependent on a potentially poorly trained base model. A large-scale survey of calibrated regression methods is given in Dheur & Ben taieb (2023).

B. Discussion

B.1. Limitations

Since the marginally calibrating elements are computed by applying the quantile function to a vector-valued difference function, their derivative is not well-defined whenever a switch between entries occurs, potentially exhibiting strong changes over small changes in the input space. Though we did not observe this issue during our experiments, it could pose problems, particularly if the learning rate is too high.

The proposed method trains a separate model to predict each quantile, which may pose a challenge should the models be resource-intensive to train, or if many quantiles need to be modeled. To reduce the number of models, it may be possible to adapt the method to work with a single model which predicts multiple quantile levels as output.

B.2. Selecting the Number of Quantiles m

A potentially important quantity when employing CaliPSo is the number of quantiles m used to design the model. Arguably the most important factor to consider here is the computational cost, as it potentially becomes large as m increases. In our experiments, we observed that the results only improved marginally beyond $m = 5$ quantiles. However, this may well be due to characteristics of the datasets considered here.

B.3. Theoretical Guarantees

Much like any other regression method, the proposed approach allows for recalibration, yielding typical conformal prediction guarantees, which state that the calibration error is inversely proportional to the amount of training data (Marx et al., 2022). However, this theoretical guarantee requires the model to be recalibrated using data that allows an interchangeability assumption. Since we use the observation data both to train the model and discard poorly calibrated parameters, it does not satisfy this assumption. Hence, strictly speaking, we would have to employ a randomly sampled subset of the data, which is only used for calibration at the end of training. However, we observe that recalibrating the model on the training and validation data does not significantly deteriorate calibration performance compared to calibrating on a holdout data set, which is consistent with remarks from other works (Kuleshov et al., 2018).