Temperature Scaling for Quantile Calibration

Saiteja Utpala
saitejautpala@gmail.com

Piyush Rai
Department of Computer Science
Indian Institute of Technology Kanpur, India
piyush@cse.iitk.ac.in

Abstract

Deep learning models are often poorly calibrated, i.e., they may produce overconfident predictions that are wrong, implying that their uncertainty estimates are unreliable. While a number of approaches have been proposed recently to calibrate classification models, relatively little work exists on calibrating regression models. Temperature Scaling is one of the most popular methods for *classification calibration*, often performing better than or comparably to more sophisticated methods. We investigate the use of Temperature Scaling for *regression calibration* under notion of quantile calibration.

1 Introduction

One of main reasons probabilistic machine learning models are important is that they provide uncertainty estimates. Uncertainty quantification enables informed decision making. The caveat however is that these decisions are reliable only if uncertainty is "reliable". Calibration offers a precise mathematical definition of what reliability means. Calibration is important for critical applications like healthcare, self-driving cars, etc. Calibration in context of classification models has been studied extensively. [16, 19, 20, 21, 15, 11, 10, 18, 12, 6]. Recently, [9] proposed a new notion of calibration for regression called *Quantile Calibration*. We investigate the use of temperature scaling, which is one of the most popular *classification calibration* methods, for *regression calibration* under notion of quantile calibration.

2 Background and Definitions

2.1 Classification Calibration

Definition 1 (Binary Classification Calibration)**.** Given $\mathsf{M} : \mathcal{X} \rightarrow [0,1]$, we say that M is calibrated if the following holds

$$\mathbb{P}\Big[Y = 1 \,\Big|\, \mathsf{M}[X] = p\Big] = p \;\; \forall p \in [0,1] \tag{1}$$

If we pick a value between $[0,1]$, say 0.8, then among all the examples whose predicted probability of belonging to class 1 is 0.8, the proportion of examples that actually belong to class 1 should be 0.8. The objective of post-hoc calibration is to learn a mapping R s.t the new model $\mathsf{R} \circ \mathsf{M}$ is calibrated [17].

Usually, the calibration mapping is learned on the training set or the validation set. However, ideally, it should be done on a separate calibration dataset. Given such a calibration dataset $\{\mathbf{x}_k, y_k\}$, a mapping R is learned on a re-calibration dataset $\left\{ \;\mathsf{M}[\mathbf{x}_k]\;\; , \;\; \frac{\sum_{i=1}^{m} \mathbb{I}\big[\big(\mathsf{M}[\mathbf{x}_k] = \mathsf{M}[\mathbf{x}_i]\big) \wedge \big(y_i = 1\big)\big]}{\sum_{i=1}^{m} \mathbb{I}\big[\mathsf{M}[\mathbf{x}_k] = \mathsf{M}[\mathbf{x}_i]\big]} \;\; \right\}$,

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

which is essentially empirical approximation to Eq.1. Based on the mappings considered, we can get different calibration methods. With Logistic mapping we get Platt Scaling [16]; with Isotonic mapping, we get Isotonic Calibration [20]. etc. The notion of calibration in classification models has been extended to multi-class settings as well [10, 12, 17])

2.2 Temperature Scaling

Temperature Scaling is one of the state-of-art methods for classification calibration. Temperature Scaling was originally conceived in context of knowledge distillation [8]. Despite its simplicity, it performs better than or comparaby to more sophisticated methods, like Bayesian Binning into Quantiles, Isotonic Regression, Dirichlet Calibration, etc. [10, 6]. Essentially, temperature scaling learns a mapping of form $\mathsf{R}(p) = p^T$

2.3 Quantile Calibration

Unlike classification calibration, notion of calibration for regression is relatively new. In one of the earliest attempts in this direction, [5] proposed various notions of calibration for regression but didn't propose algorithms to recalibrate a miscalibrated model. Recently, [9] proposed a new notion of calibration called *Quantile Calibration* based on *Probabilistic Calibration* in [5] and applied it to calibrate regression models. A probabilistic regression model can be seen as conditional PDF/conditional CDF. In the rest of the paper, we express it as conditional CDF $\mathsf{M} : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$. So, $\mathsf{M}(x)$ denotes model's predicted CDF for $x \in \mathcal{X}$ denoted as F_x

Definition 2 (**Quantile Calibration**). Given a regression model $\mathsf{M} : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$ and X, Y jointly distributed as $\mathbf{P}$, the model M is said to be Quantile Calibrated *iff*

$$\mathbb{P}\Big[\,[\,\mathsf{M}(X)\,](Y) \leq p\Big] = p \;\; \forall p \in [0, 1] \tag{2}$$

An appealing aspect of quantile calibration is that we get calibrated confidence intervals. Just like post-hoc classification calibration, the objective of post-hoc regression calibration is to learn a mapping R s.t. $\mathsf{R} \circ \mathsf{M}$ is quantile calibrated. The mapping R to be learned is given by following observation.

Theorem 1. For any Model $\mathsf{M} : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$, and given canonical calibration mapping $\mathsf{R}(p) = \mathbb{P}\big[[\mathsf{M}(X)](Y) \leq p\big]$, $\mathsf{R} \circ \mathsf{M}$ is quantile calibrated

In addition to proposing above definition of calibration, [9] suggested use of Isotonic Calibration, well known technique for classification calibration. Given calibration dataset $\{\mathbf{x}_i, y_i\}_{i=1}^m$ Isotonic Calibration for quantile calibration is obtained by using isotonic regression on a re-calibration dataset $\mathcal{D} = \Big\{ \Big(\; M(\mathbf{x}_i)[y_i] \;\; , \;\; \frac{1}{m}\sum_{j=1}^m \mathbb{I}\big[\; M(\mathbf{x}_j)[y_j] \leq M(\mathbf{x}_i)[y_i] \;\big] \; \Big) \Big\}_{i=1}^m$. Note that the only difference between isotonic calibration in classification calibration and regression calibration is how recalibration dataset is constructed. Quantitatively, calibration is measured by ℓ_2 quantile calibration error. Given a test set $\{x_n, y_n\}_{n=1}^N$, with predictions are $F_n = \mathsf{M}(x_n)$ and m equidistant points $\{p_m\}_{m=1}^M$ in $(0, 1]$

$$\mathcal{CE}(F) = \frac{1}{M}\sum_{i=1}^{M}\Big[\sum_{j=1}^{N}\frac{1}{N}\mathrm{I}[F_j(y_j) \leq p_i] - p_i\Big]^2 \tag{3}$$

3 Temperature Scaling for Quantile Calibration

We propose to learn canonical calibration mapping of quantile calibration by temperature scaling $\mathsf{R}(x) = x^p$. Our proposal is justified by following simple lemma.

Claim 1. let F be CDF of any r.v then for any $\alpha > 0$ we have that F^α is valid CDF again.

Proof.

1. Non-decreasing and right continuous : $G'(x) = \alpha F(x)^{\alpha-1} > 0$ as $\alpha > 0$ and right continuous because it is polynomial

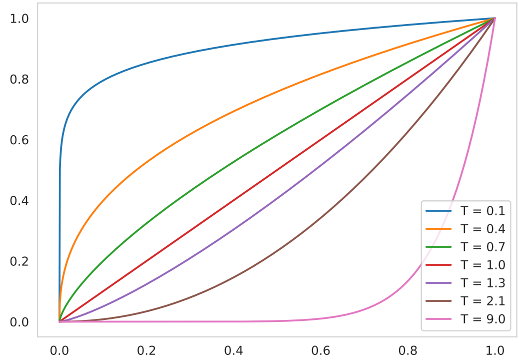


Figure 1: plot of x^T for different values of T

2. $\lim_{x \to \infty} G(x) = \lim_{x \to \infty} F(x)^{\alpha} = \left[\lim_{x \to \infty} F(x)\right]^{\alpha} = 1^{\alpha} = 1$
3. $\lim_{x \to -\infty} G(x) = \lim_{x \to -\infty} F(x)^{\alpha} = \left[\lim_{x \to -\infty} F(x)\right]^{\alpha} = 0^{\alpha} = 0$

□

Such family of distributions are called *exponentiated distributions* which have been well-studied in the Statistics literature [7, 1, 2, 14]

An ideal desirable for the family of mappings is that it should be flexible enough to correct wide ranges of mis-calibrations. Importantly, it should contain $y = x$, because it shouldn't harm already well-calibrated model. Fig. 1 shows that the temperature scaling family is an ideal candidate. Now for fitting the parameter T, we use Eq. 3

$$T^* = \arg\min_{T} \frac{1}{M} \sum_{i=1}^{M} \left[\left(\sum_{j=1}^{N} \frac{1}{N} \mathrm{I}[F_j(y_j) \leq p_i] \right)^{T} - p_i \right]^2$$

A couple of key advantages of temperature scaling over isotonic regression are as follows

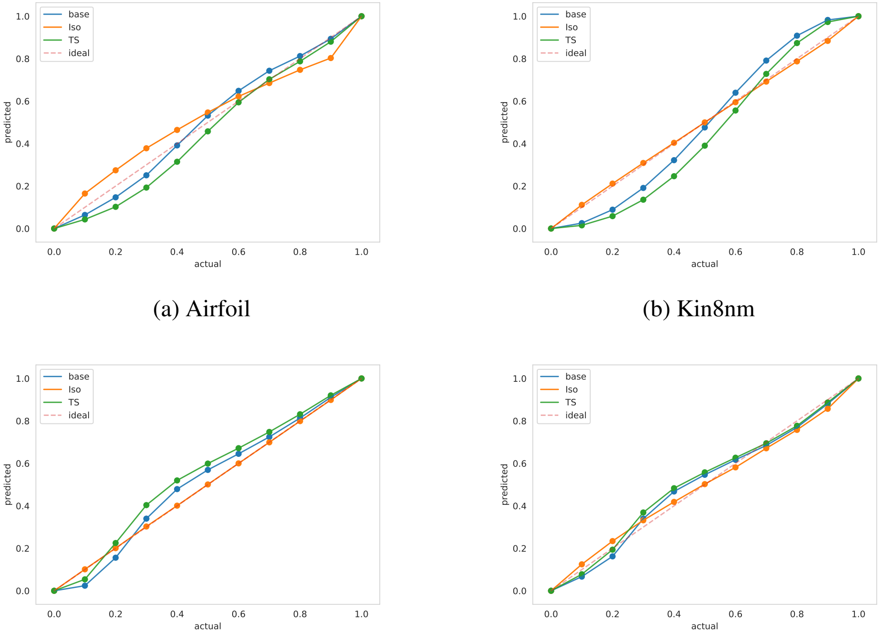
1. One of important drawbacks of isotonic regression is that, after isotonic calibration, the CDF losses its smoothness and the PDF becomes discontinuous, which is undesirable. With temperature scaling, the smoothness is preserved
2. It is much simpler and easier to use than isotonic regression as we are just fitting a single parameter T. In particular, the updated PDF and CDF values can be obtained in $\mathcal{O}(1)$ time, unlike isotonic regression

4 Experiments

We consider two different architectures - Dropout VI [3, 4] and Deep Ensembles [13]. The dataset sizes ranges from 308 to 515345 and input feature dimensions ranges from 6 to 91. Every dataset, except Year Prediction MSD, is divided into 5 splits whereas for Year Prediction MSD there is a single split where we train on 463715 points and test on 51630 points. This experiment is repeated 5 times and averages are reported except for year prediction MSD. We use 2 hidden layer network with 128 units with ReLU activation, and trained with Adam Optimizer with a learning rate of 10^{-2} for 64 epochs. For Temperature Scaling we use the LBFGS optimizer and run it for 50 epochs, and for calibration dataset we use the training dataset. The results are presented in Tab. 1 and Tab. 2. The calibration plots are shown in Fig. 2 and Fig. 3. The average temperature is shown in Tab. 3 and the plot of calibration loss vs temperature is shown in Fig. 4

Dataset	Heteroscedastic Dropout VI					
	Calibration Error(%)			NLL		
	base	Iso	TS	base	Iso	TS
Air Foil	**12.72** ± **1.99**	17.44 ± 2.91	15.23 ± 2.72	2.71 ± 0.02	2.30 ± 0.05	**-0.18** ± **0.10**
Boston Housing	**23.30** ± **3.83**	30.17 ± 4.99	31.57 ± 10.05	3.23 ± 0.03	2.68 ± 0.09	**0.20** ± **0.17**
Concrete Strength	**29.75** ± **2.32**	34.71 ± 3.97	40.46 ± 6.99	3.65 ± 0.02	3.34 ± 0.06	**0.13** ± **0.21**
Fish Toxicity	3.05 ± 0.36	**1.40** ± **0.16**	5.18 ± 0.76	1.25 ± 0.01	0.64 ± 0.02	**-0.01** ± **0.01**
Kin8nm	7.26 ± 0.22	**0.22** ± **0.02**	14.88 ± 3.04	-0.87 ± 0.01	**-1.60** ± **0.02**	0.31 ± 0.07
Protein Structure	3.04 ± 0.42	**0.05** ± **0.00**	8.23 ± 2.02	2.89 ± 0.00	2.21 ± 0.01	**0.32** ± **0.06**
Red Wine	3.23 ± 0.73	**2.96** ± **0.24**	4.65 ± 2.03	0.97 ± 0.00	0.37 ± 0.03	**0.10** ± **0.00**
White Wine	**4.02** ± **0.41**	4.38 ± 0.17	4.80 ± 0.78	1.10 ± 0.00	0.52 ± 0.03	**0.02** ± **0.01**
Year Prediction MSD	3.83 ± NA	**0.02** ± **NA**	9.96 ± NA	3.47 ± NA	3.59 ± NA	**-0.38** ± **NA**

Table 1: Base denotes the base model without post-hoc calibration. Iso denotes the model after isotonic calibration and TS denotes the model after Temperature scaling



(a) Airfoil

(b) Kin8nm

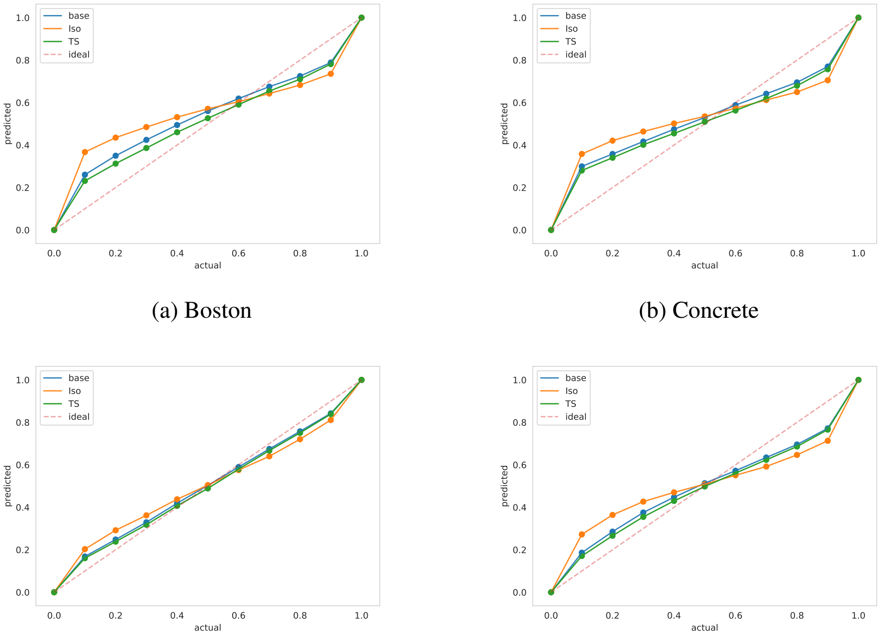
(c) Protein

(d) Red

Figure 2: Dashed line (y=x) indicates perfect calibration. The closer to dashed line, the better

Dataset	Deep Ensembles with Adversarial Training					
	Calibration Error(%)			NLL		
	base	Iso	TS	base	Iso	TS
Air Foil	**23.93** ± **2.92**	38.70 ± 2.64	**23.93** ± **5.22**	2.92 ± 0.10	2.96 ± 0.10	**-0.23** ± **0.11**
Boston Housing	**37.61** ± **7.82**	50.71 ± 5.02	40.01 ± 16.61	4.45 ± 0.44	4.32 ± 0.39	**0.09** ± **0.62**
Concrete Strength	39.97 ± 4.06	51.05 ± 3.86	**37.95** ± **7.27**	4.91 ± 0.20	4.86 ± 0.22	**-0.23** ± **0.27**
Fish Toxicity	**3.50** ± **0.43**	6.34 ± 0.13	5.10 ± 0.96	1.64 ± 0.03	1.18 ± 0.03	**0.10** ± **0.06**
Kin8nm	**0.64** ± **0.36**	5.36 ± 0.08	3.44 ±2.03	-1.34 ± 0.00	**-1.66** ± **0.02**	0.18 ± 0.04
Protein Structure	2,37 ± 0.16	**0.07** ± **0.01**	3.19 ± 0.32	2.60 ± 0.00	1.72 ±0.00	**0.08** ± **0.02**
Red Wine	**7.95** ± **0.36**	16.75 ± 0.82	10.06 ± 1.38	1.98 ± 0.07	1.15 ± 0.18	**-0.01** ± **0.07**
White Wine	**8.71** ± **1.29**	19.51 ± 0.59	9.40 ± 3.15	1.64 ± 0.04	0.73 ± 0.11	**-0.04** ± **0.02**
Year Prediction MSD	1.31 ± NA	**0.07** ± **NA**	3.05 ± NA	3.34 ± NA	3.67 ± NA	**-0.21** ± **NA**

Table 2: Base denotes the base model without post-hoc calibration. Iso denotes the model after isotonic calibration and TS denotes the model after Temperature scaling



(a) Boston

(b) Concrete

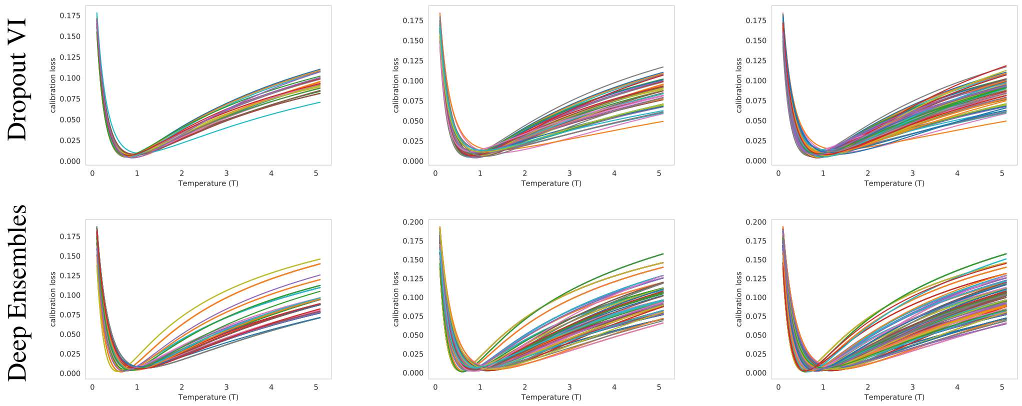
(c) Fish

(d) Red

Figure 3: Dashed line (y=x) indicates perfect calibration. The closer to dashed line, the better

Dataset	dimensions		RMSE		Avg temperature	
	N	D	Dropout	Ensembles	Dropout	Ensembles
Air Foil	1503	5	3.61 ± 0.06	**3.15 ± 0.07**	0.87 ± 0.02	0.88 ± 0.07
Boston Housing	506	13	**4.64 ± 0.19**	4.87 ± 0.16	1.07 ± 0.07	0.96 ± 0.13
Concrete Strength	1030	8	**9.00 ± 0.18**	9.11 ± 0.25	1.04 ± 0.07	0.93 ± 0.07
Fish Toxicity	908	6	**0.93 ± 0.00**	**0.93 ± 0.01**	0.98 ± 0.05	0.98 ± 0.07
Kin8nm	8182	8	0.09 ± 0.00	**0.07 ± 0.00**	0.88 ± 0.02	0.93 ± 0.02
Protein Structure	45730	9	4.63 ± 0.01	**4.11 ± 0.27**	1.16 ± 0.03	1.05 ± 0.01
Red Wine	1599	11	**0.65 ± 0.00**	0.69 ± 0.00	1.07 ± 0.03	0.98 ± 0.06
White Wine	4898	11	**0.73 ± 0.00**	0.76 ± 0.01	1.01 ± 0.01	0.94 ± 0.02
Year Prediction MSD	515345	90	9.12 ± NA	**8.70 ± NA**	0.84 ± NA	0.91 ± NA

Table 3: Base denotes the base model without post-hoc calibration. Iso denotes the model after isotonic calibration and TS denotes the model after Temperature scaling



(a) Aifoil

(b) Boston

(c) Concrete

Figure 4: plots showing varying T and calibration loss on calibration dataset

5 Discussion

Tab. 1 and Tab. 2 show that Temperature Scaling doesn't perform as well as expected in case of quantile calibration. One would think one reason for this is because we are using training data as calibration dataset. But this is not the case because in case of Isotonic regression there is *two orders* of magnitude improvement for large datasets like Protein,Year prediction MSD. We conjecture two hypotheses as to why Temperature Scaling is not performing as expected:

1. Temperature Scaling is *flexible* but may not be *flexible enough* because we are just using single parameter (T), while isotonic regression is non-parametric method.
2. Using Calibration error as objective for fitting T may be another reason. If this is the case, using better suited and properly regularized objective could alleviate the problem.

6 Conclusion

We investigated the performance of Temperature Scaling for regression calibration in context of quantile calibration and found that it doesn't perform as well as it does for classification calibration. We have identified some potential reasons for this and it would be interesting to investigate them further to see whether this simple method can be useful for regression calibration.

References

[1] Essam K Al-Hussaini and Mohammad Ahsanullah. Exponentiated distributions. *Atlantis Studies in Probability and Statistics. Atlantis Press, Paris, France*, 2015.

[2] M Masoom Ali, Manisha Pal, and Jung-Soo Woo. Some exponentiated distributions. *Communications for Statistical Applications and Methods*, 14(1):93–109, 2007.

[3] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[5] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[7] Rameshwar D Gupta and Debasis Kundu. Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(1):117–130, 2001.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[9] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

[10] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12295–12305, 2019.

[11] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631, 2017.

[12] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3787–3798, 2019.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[14] Saralees Nadarajah and Samuel Kotz. The exponentiated type distributions. *Acta Applicandae Mathematica*, 92(2):97–111, 2006.

[15] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 2901. NIH Public Access, 2015.

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

[17] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. *arXiv preprint arXiv:1902.06977*, 2019.

[18] Yongqiao Wang and Xudong Liu. Multivariate probability calibration with isotonic bernstein polynomials. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2547–2553. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[19] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001.

[20] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[21] Wenliang Zhong and James T Kwok. Accurate probability calibration for multiple classifiers. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer, 2013.