Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration

Supplementary material

Meelis Kull
Department of Computer Science
University of Tartu
meelis.kull@ut.ee

Miquel Perello-Nieto
Department of Computer Science
University of Bristol
miquel.perellonieto@bris.ac.uk

Markus Kängsepp
Department of Computer Science
University of Tartu
markus.kangsepp@ut.ee

Telmo Silva Filho
Department of Statistics
Universidade Federal da Paraíba
telmo@de.ufpb.br

Hao Song
Department of Computer Science
University of Bristol
hao.song@bristol.ac.uk

Peter Flach
Department of Computer Science
University of Bristol and
The Alan Turing Institute
peter.flach@bristol.ac.uk

Contents

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

A Source code

The instructions and code for the experiments can be found on https://dirichletcal.github.io/.

B Proofs

Theorem 1 (Equivalence of generative, linear and canonical parametrisations)**.** *The parametric families $\hat{\boldsymbol{\mu}}_{DirGen}(\mathbf{q};\boldsymbol{\alpha},\boldsymbol{\pi})$, $\hat{\boldsymbol{\mu}}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b})$ and $\hat{\boldsymbol{\mu}}_{Dir}(\mathbf{q};\mathbf{A},\mathbf{c})$ are equal, i.e. they contain exactly the same calibration maps.*

Proof. We will prove that:

1. every function in $\hat{\boldsymbol{\mu}}_{DirGen}(\mathbf{q};\boldsymbol{\alpha},\boldsymbol{\pi})$ belongs to $\hat{\boldsymbol{\mu}}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b})$;
2. every function in $\hat{\boldsymbol{\mu}}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b})$ belongs to $\hat{\boldsymbol{\mu}}_{Dir}(\mathbf{q};\mathbf{A},\mathbf{c})$;
3. every function in $\hat{\boldsymbol{\mu}}_{Dir}(\mathbf{q};\mathbf{A},\mathbf{c})$ belongs to $\hat{\boldsymbol{\mu}}_{DirGen}(\mathbf{q};\boldsymbol{\alpha},\boldsymbol{\pi})$.

1. Consider a function $\hat{\mu}(\mathbf{q}) = \hat{\boldsymbol{\mu}}_{DirGen}(\mathbf{q};\boldsymbol{\alpha},\boldsymbol{\pi})$. Let us start with an observation that any vector $\mathbf{x} = (x_1,\ldots,x_k) \in (0,\infty)^k$ with only positive elements can be renormalised to add up to 1 using the expression $\boldsymbol{\sigma}(\mathbf{ln}(\mathbf{x}))$, since:

$$\boldsymbol{\sigma}(\mathbf{ln}(\mathbf{x})) = \mathbf{exp}(\mathbf{ln}(\mathbf{x}))/(\sum_i \exp(\ln(x_i))) = \mathbf{x}/(\sum_i x_i)$$

where $\mathbf{exp}$ is an operator applying exponentiation element-wise. Therefore,

$$\hat{\mu}(\mathbf{q}) = \boldsymbol{\sigma}(\mathbf{ln}(\pi_1 f_1(\mathbf{q}),\ldots,\pi_k f_k(\mathbf{q})))$$

where $f_i(\mathbf{q})$ is the probability density function of the distribution $Dir(\boldsymbol{\alpha}^{(i)})$ where $\boldsymbol{\alpha}^{(i)}$ is the i-th row of matrix $\boldsymbol{\alpha}$. Hence, $f_i(\mathbf{q}) = \frac{1}{B(\boldsymbol{\alpha}^{(i)})}\prod_{j=1}^k q_j^{\alpha_{ij}-1}$, where $B(\cdot)$ denotes the multivariate beta function. Let us define a matrix $\mathbf{W}$ and vector $\mathbf{b}$ as follows:

$$w_{ij} = \alpha_{ij} - 1, \qquad b_i = \ln(\pi_i) - \ln(B(\boldsymbol{\alpha}^{(i)}))$$

with w_{ij} and α_{ij} denoting elements of matrices $\mathbf{W}$ and $\boldsymbol{\alpha}$, respectively, and b_i, π_i denoting elements of vectors $\mathbf{b}$ and $\boldsymbol{\pi}$. Now we can write

$$\begin{aligned}
\ln(\pi_i f_i(\mathbf{q})) &= \ln(\pi_i) - \ln(B(\boldsymbol{\alpha}^{(i)})) + \ln\prod_{j=1}^k q_j^{\alpha_{ij}-1} \\
&= \ln(\pi_i) - \ln(B(\boldsymbol{\alpha}^{(i)})) + \sum_{j=1}^k (\alpha_{ij}-1)\ln(q_j) \\
&= b_i + \sum_{j=1}^k w_{ij}\ln(q_j)
\end{aligned}$$

and substituting this back into $\hat{\mu}(\mathbf{q})$ we get:

$$\begin{aligned}
\hat{\mu}(\mathbf{q}) &= \boldsymbol{\sigma}(\mathbf{ln}(\pi_1 f_1(\mathbf{q}),\ldots,\pi_k f_k(\mathbf{q}))) \\
&= \boldsymbol{\sigma}(\mathbf{b} + \mathbf{W}\mathbf{ln}(\mathbf{q})) = \hat{\mu}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b})
\end{aligned}$$

2. Consider a function $\hat{\mu}(\mathbf{q}) = \hat{\boldsymbol{\mu}}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b})$. Let us define a matrix $\mathbf{A}$ and vector $\mathbf{c}$ as follows:

$$a_{ij} = w_{ij} - \min_i w_{ij}, \qquad \mathbf{c} = \boldsymbol{\sigma}(\mathbf{W}\,\mathbf{ln}\,\mathbf{u} + \mathbf{b})$$

with a_{ij} and w_{ij} denoting elements of matrices $\mathbf{A}$ and $\mathbf{W}$, respectively, and $\mathbf{u} = (1/k,\ldots,1/k)$ is a column vector of length k. Note that $\mathbf{A}\,\mathbf{x} = \mathbf{W}\,\mathbf{x} + const_1$ and $\mathbf{ln}\,\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{x} + const_2$ for any x where $const_1$ and $const_2$ are constant vectors (all elements are equal), but the constant depends on

$\mathbf{x}$. Taking into account that $\boldsymbol{\sigma}(\mathbf{v} + const) = \boldsymbol{\sigma}(\mathbf{v})$ for any vector $\mathbf{v}$ and constant vector $const$, we obtain:

$$\begin{aligned}\hat{\mu}_{Dir}(\mathbf{q}; \mathbf{A}, \mathbf{c}) &= \boldsymbol{\sigma}(\mathbf{A} \ln \frac{\mathbf{q}}{1/k} + \ln \mathbf{c}) = \boldsymbol{\sigma}(\mathbf{W} \ln \frac{\mathbf{q}}{1/k} + const_1 + \ln \mathbf{c}) \\ &= \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{q} - \mathbf{W} \ln \mathbf{u} + const_1 + \ln \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{u} + \mathbf{b})) \\ &= \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{q} - \mathbf{W} \ln \mathbf{u} + const_1 + \mathbf{W} \ln \mathbf{u} + \mathbf{b} + const_2) \\ &= \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{q} + \mathbf{b} + const_1 + const_2) = \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{q} + \mathbf{b}) = \hat{\mu}_{DirLin}(\mathbf{q}; \mathbf{W}, \mathbf{b}) \\ &= \hat{\mu}(\mathbf{q})\end{aligned}$$

3. Consider a function $\hat{\mu}(\mathbf{q}) = \hat{\boldsymbol{\mu}}_{Dir}(\mathbf{q}; \mathbf{A}, \mathbf{c})$. Let us define a matrix $\boldsymbol{\alpha}$, vector $\mathbf{b}$ and vector π as follows:

$$\alpha_{ij} = a_{ij} + 1, \qquad \mathbf{b} = \ln \mathbf{c} - \mathbf{A} \ln \mathbf{u}, \qquad \pi_i = \exp(b_i) \cdot B(\boldsymbol{\alpha}^{(i)})$$

with α_{ij} and a_{ij} denoting elements of matrices $\boldsymbol{\alpha}$ and $\mathbf{A}$, respectively, and $\mathbf{u} = (1/k, \ldots, 1/k)$ is a column vector of length k. We can now write:

$$\begin{aligned}\hat{\mu}(\mathbf{q}) = \hat{\mu}_{Dir}(\mathbf{q}; \mathbf{A}, \mathbf{c}) &= \boldsymbol{\sigma}(\mathbf{A} \ln \frac{\mathbf{q}}{1/k} + \ln \mathbf{c}) = \boldsymbol{\sigma}(\mathbf{A} \ln \mathbf{q} - \mathbf{A} \ln \mathbf{u} + \ln \mathbf{c}) \\ &= \boldsymbol{\sigma}((\boldsymbol{\alpha} - 1) \ln \mathbf{q} + \mathbf{b})\end{aligned}$$

Element i in the vector within the softmax is equal to:

$$\begin{aligned}\sum_{j=1}^{k} (\boldsymbol{\alpha}_{ij} - 1) \ln(q_j) + b_j &= \sum_{j=1}^{k} (\boldsymbol{\alpha}_{ij} - 1) \ln(q_j) + \ln(\pi_i \cdot \frac{1}{B(\boldsymbol{\alpha}^{(i)})}) \\ &= \ln(\pi_i \cdot \frac{1}{B(\boldsymbol{\alpha}^{(i)})} \prod_{j=1}^{k} q_j^{\boldsymbol{\alpha}_{ij} - 1}) \\ &= \ln(\pi_i \cdot f_i(\boldsymbol{\alpha}^{(i)}))\end{aligned}$$

and therefore:

$$\hat{\mu}(\mathbf{q}) = \boldsymbol{\sigma}((\boldsymbol{\alpha} - 1) \ln(\mathbf{q}) + \mathbf{b}) = \boldsymbol{\sigma}(\ln(\pi_i \cdot f_i(\boldsymbol{\alpha}^{(i)}))) = \hat{\boldsymbol{\mu}}_{DirGen}(\mathbf{q}; \boldsymbol{\alpha}, \boldsymbol{\pi})$$

□

The following proposition proves that temperature scaling can be viewed as a general-purpose calibration method, being a special case within the Dirichlet calibration map family.

Proposition 1. *Let us denote the temperature scaling family by $\hat{\mu}'_{TempS}(\mathbf{z}; t) = \boldsymbol{\sigma}(\mathbf{z}/t)$ where $\mathbf{z}$ are the logits. Then for any t, temperature scaling can be expressed as*

$$\hat{\mu}'_{TempS}(\mathbf{z}; t) = \hat{\mu}_{DirLin}(\boldsymbol{\sigma}(\mathbf{z}); \frac{1}{t}\mathbf{I}, \mathbf{0})$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{0}$ is the vector of zeros.

Proof. Let us first observe that for any $\mathbf{x} \in \mathbb{R}^k$ there exists a constant vector $const$ (all elements are equal) such that $\ln \boldsymbol{\sigma}(\mathbf{x}) = \mathbf{x} + const$. Furthermore, $\boldsymbol{\sigma}(\mathbf{v} + const) = \boldsymbol{\sigma}(\mathbf{v})$ for any vector $\mathbf{v}$ and any constant vector $const$. Therefore,

$$\begin{aligned}\hat{\mu}_{DirLin}(\boldsymbol{\sigma}(\mathbf{z}); \frac{1}{t}\mathbf{I}, \mathbf{0}) &= \boldsymbol{\sigma}(\frac{1}{t} \mathbf{I} \ln \boldsymbol{\sigma}(\mathbf{z}))) \\ &= \boldsymbol{\sigma}(\frac{1}{t} \mathbf{I} (\mathbf{z} + const)) \\ &= \boldsymbol{\sigma}(\frac{1}{t} \mathbf{I} \mathbf{z} + \frac{1}{t} \mathbf{I} const) \\ &= \boldsymbol{\sigma}(\mathbf{z}/t + const') \\ &= \boldsymbol{\sigma}(\mathbf{z}/t) \\ &= \hat{\mu}'_{TempS}(\mathbf{z}; t)\end{aligned}$$

where $const' = \frac{1}{t} \mathbf{I} const$ is a constant vector as a product of a diagonal matrix with a constant vector. □

C Dirichlet calibration

In this section we show some examples of reliability diagrams and other plots that can help to understand the representational power of Dirichlet calibration compared with other calibration methods.

C.1 Reliability diagram examples

We will look at two examples of reliability diagrams on the original classifier and after applying 6 calibration methods. Figure 1 shows the first example for the 3 class classification dataset *balance-scale* and the classifier MLP. This figure shows the confidence-reliability diagram in the first column and the classwise-reliability diagrams in the other columns. Figure 1a shows how posterior probabilities from the MLP have small gaps between the true class proportions and the predicted means. This visualisation may indicate that the original classifier is already well calibrated. However, when we separate the reliability diagram per class, we notice that the predictions for the first class are underconfident, as indicated by high proportions of the true class at low mean predictions. On the other hand, the predictions on classes 2 and 3 are overconfident in the regions of posterior probabilities compressed between $[0.2, 0.5]$ while being underconfident in higher regions.

Table 1: Averaged results for the confidence-ECE and classwise-ECE metrics of 6 calibrators applied on an MLP trained on the *balance-scale* dataset.

	DirL2	Beta	FreqB	Isot	WidB	TempS	Uncal
conf-ECE	$\mathbf{0.04_1}$	0.05_3	0.13_7	0.05_2	0.08_6	0.05_4	0.08_5
cw-ECE	0.12_2	0.13_3	0.29_7	$\mathbf{0.12_1}$	0.17_5	0.15_4	0.20_6

The following subfigures show how the different calibration methods try to reduce ECE, occasionally increasing the error. As can be seen in Table 1, Dirichlet L2 and One-vs.Rest isotonic regression obtain the lowest ECE while One-vs.Rest frequency binning makes the original calibration worse. Looking at Figure 1i it is possible to see how temperature scaling manages to reduce the overall overconfidence in the higher range of probabilities for classes 2 and 3, but makes the situation worse in the interval $[0.2, 0.6]$. However, it manages to reduce the overall ECE.

In the second example we show 3 calibration methods for a 4 class classification problem (car dataset) applied on the scores of an Adaboost SAMME classifier. Figure 2 shows one reliability diagram per class (C_1 *acceptable*, C_2 *good*, C_3 *unacceptable*, and C_4 *very good*).

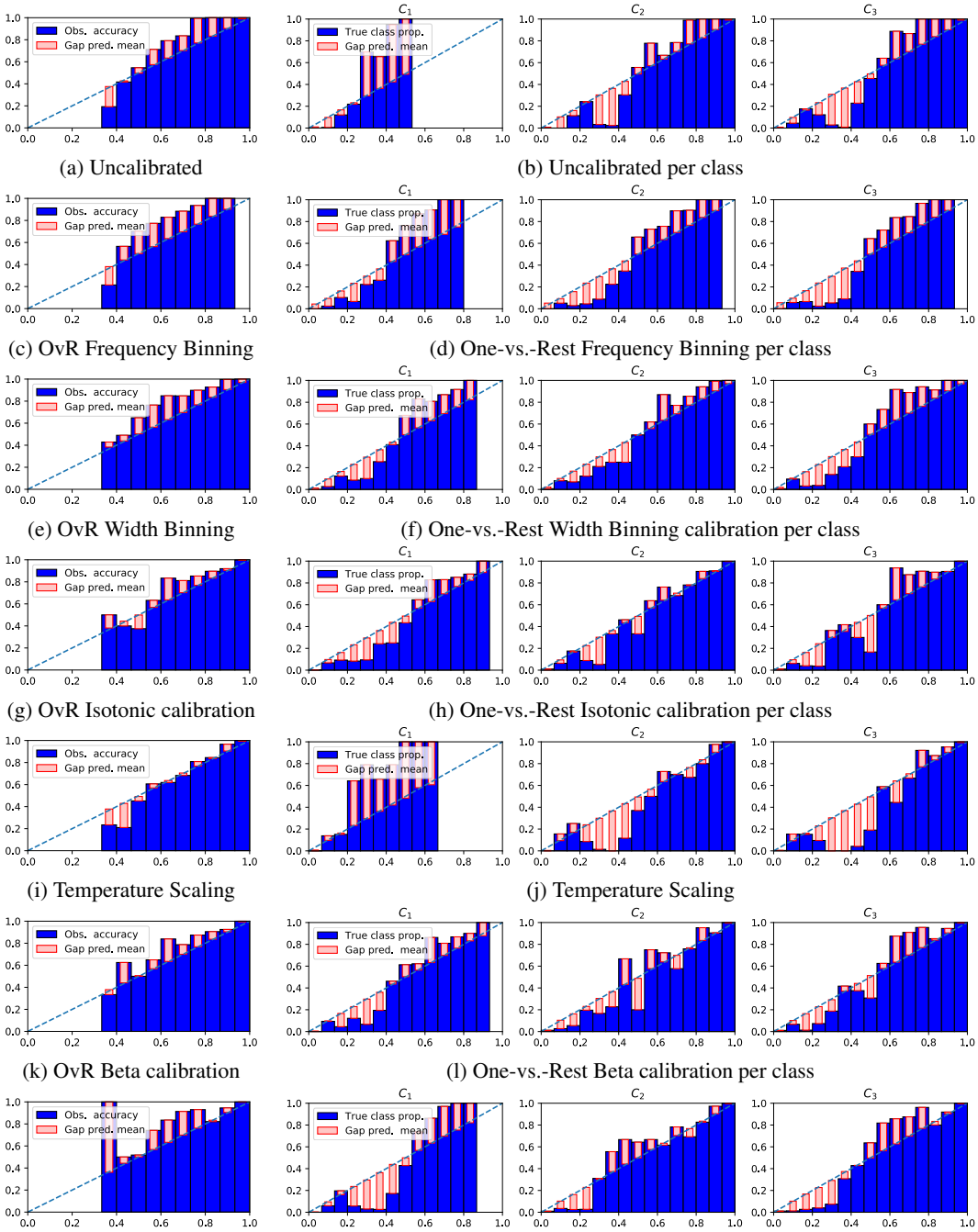
From this Figure we can see that the uncalibrated model is underconfident for classes 1, 2 and 3, showing posterior probabilities never higher than 0.7, while having true class proportions higher than 0.7 in the mentioned interval. We can see that after applying some of the calibration models the posterior probabilities reach higher probability values. As can be seen in Table 2, Dirichlet L2

Table 2: Averaged results for the confidence-ECE and classwise-ECE metrics of 6 calibrators applied on an Adaboost SAMME trained on the *car* dataset.

	DirL2	Beta	FreqB	Isot	WidB	TempS	Uncal
conf-ECE	$\mathbf{0.07_1}$	0.10_4	0.12_5	0.07_2	0.09_3	0.14_7	0.14_6
cw-ECE	0.18_2	0.23_3	0.29_5	$\mathbf{0.18_1}$	0.25_4	0.32_7	0.29_6

and One-vs.Rest Isotonic Regression obtain the lowest ECE while Temperature Scaling makes the original calibration worse. Figure 2d shows how Dirichlet calibration with L2 regularisation achieved the largest spread of probabilities, also reducing the error mean gap with the predictions and the true class proportions. On the other hand, temperature scaling reduced ECE for class 1, but hurt the overall performance for the other classes.

A more detailed depiction of the previous reliability diagrams can be seen in Figure 3. In this case, the posterior probabilities are not introduced in bins, but a boxplot summarises their full distribution. The first observation here is, for the *good* and *very good* classes, the uncalibrated model tends to predict probability vectors with small variance, i.e. the outputs do not change much among different instances. Among the calibration approaches, temperature scaling still maintains this low level of



(a) Uncalibrated

(b) Uncalibrated per class

(c) OvR Frequency Binning

(d) One-vs.-Rest Frequency Binning per class

(e) OvR Width Binning

(f) One-vs.-Rest Width Binning calibration per class

(g) OvR Isotonic calibration

(h) One-vs.-Rest Isotonic calibration per class

(i) Temperature Scaling

(j) Temperature Scaling

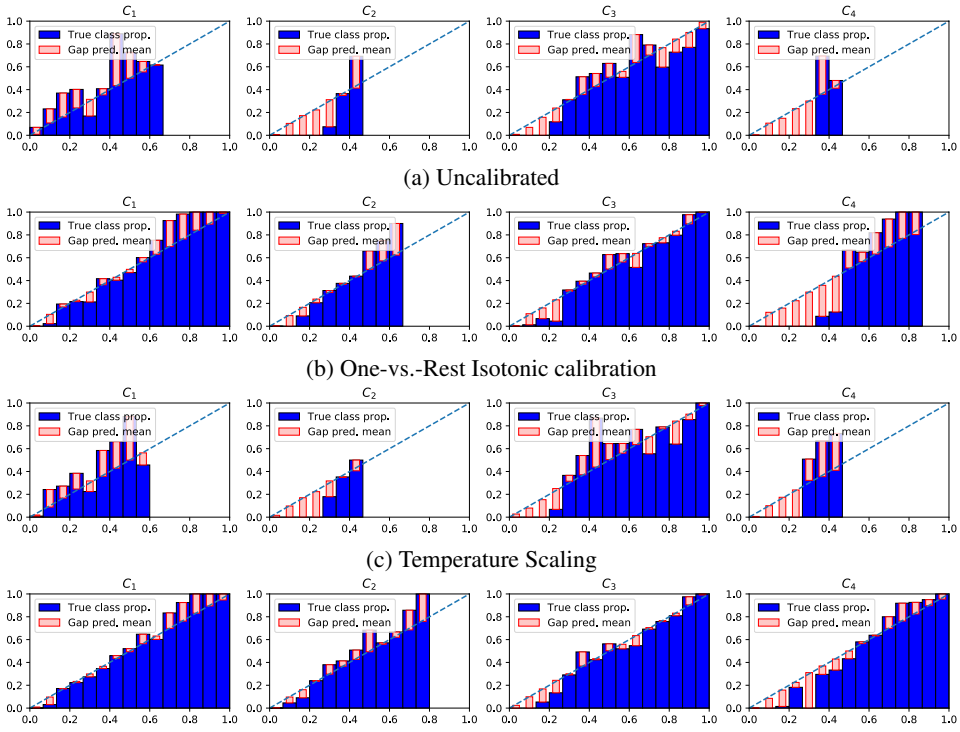
(k) OvR Beta calibration

(l) One-vs.-Rest Beta calibration per class

(m) Dirichlet L2

(n) Dirichlet with L2 regularisation per class

Figure 1: Confidence-reliability diagrams in the first column and classwise-reliability diagrams in the remaining columns, for a real experiment with the multilayer perceptron classifier on the balance-scale dataset and a subset of the calibrators. All the test partitions from the 5 times 5-fold-cross-validation have been aggregated to draw every plot.



(a) Uncalibrated

(b) One-vs.-Rest Isotonic calibration

(c) Temperature Scaling

(d) Dirichlet with L2 regularisation

Figure 2: Reliability diagrams per class for a real experiment with the classifier Ada boost SAMME on the car dataset and 3 calibrators. The test partitions from the 5 times 5-fold-cross-validation have been aggregated to draw every plot.

variance, while both isotonic and Dirichlet L2 manage to show a higher variance on the outputs. While this observation cannot be justified here without quantitative analysis, another observation clearly shows an advantage of using Dirichlet L2. For the *acceptable* class, only Dirichlet L2 is capable of providing the highest mean probability for the correct class, while the other three methods tend to put higher probability mass on the *unacceptable* class on average.

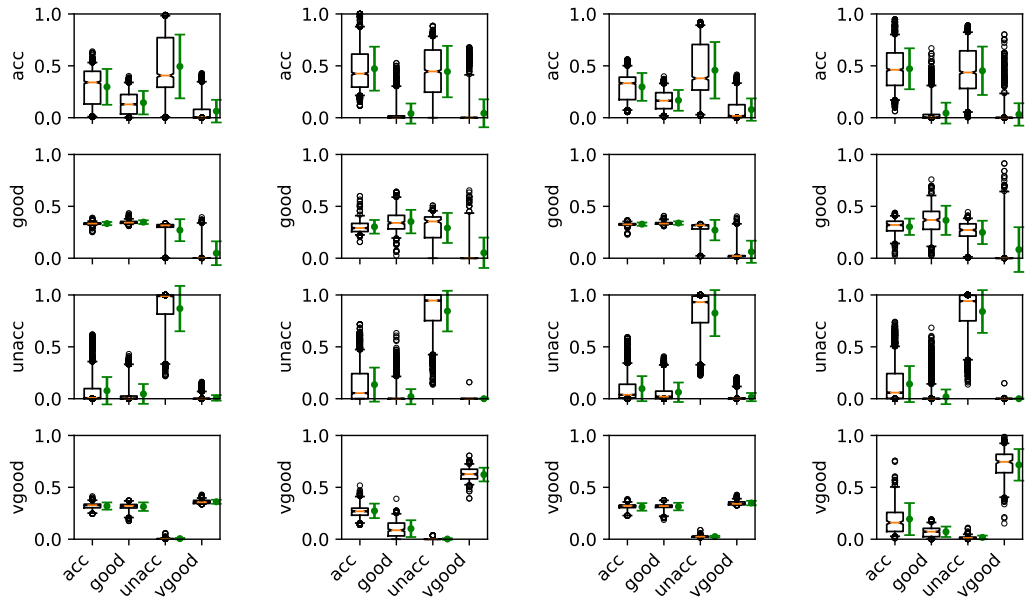
D Experimental setup

In this section we provide the detailed description of the experimental setup on a variety of non-neural classifiers and datasets. While our implementation of Dirichlet calibration is based on standard Newton-Raphson with multinomial logistic loss and L2 regularisation, as mentioned at the end of Section 3, existing implementations of logistic regression (e.g. scikit-learn) with the log transformed predicted probabilities can also be easily applied.

D.1 Datasets and performance estimation

The full list of datasets, and a brief description of each one including the number of samples, features and classes is presented in Table 3.

Figure 4 shows how every dataset was divided in order to get an estimated performance for every combination of dataset, classifier and calibrator. Each dataset was divided using 5 times 5-fold-cross-validation to create 25 test partitions. For each of the 25 partitions the corresponding training set was divided further with a 3-fold-cross-validation for wich the bigger portions were used to train the classifiers (and validate the calibratiors if they had hyperparameters), and the small portion was used to train the calibrators. The 3 calibrators trained in the inner 3-folds were used to predict the corresponding test partition, and their predictions were averaged in order to obtain better estimates of their performance with the 7 different metrics (accuracy, Brier score, log-loss, maximum calibration



(a) Uncalibrated (b) Isotonic (c) Temperature scaling (d) Dirichlet L2

Figure 3: Effect of Dirichlet Calibration on the scores of Ada boost SAMME on the *car* dataset which is composed of 4 classes (*acceptable*, *good*, *unacceptable*, and *very good*). The whiskers of each box indicate the 5th and 95th percentile, the notch around the median indicates the confidence interval. The green error bar to the right of each box indicates one standard deviation on each side of the mean. In each subfigure, the first boxplot corresponds to the posterior probabilities for the samples of class 1, divided in 4 boxes representing the posterior probabilities for each class. A good classifier should have the highest posterior probabilities in the box corresponding to the true class. In Figure 3a it is possible to see that the first class (*acceptable*) is missclassified as belonging to the third class (*unacceptable*) with high probability values, while Dirichlet Calibration is able to alleviate that problem. Also, for the second and fourth true classes (*good*, and *very good*) the original classifier uses a reduced domain of probabilities (indicative of underconfidence), while Dirichlet calibration is able to spread these probabilities with more meaningful values (as indicated by a reduction of the calibration losses; See Figure 2).

error, confidence-ECE, classwise-ECE and the p test statistic of the ECE metrics). Finally, the 25 resulting measures were averaged.

D.2 Full example of statistical analysis

The following is a full example of how the final rankings and statistical tests are computed. For this example, we will focus on the metric log-loss, and we will start with the naive Bayes classifier. Table 4 shows the estimated log-loss by averaging the 5-times 5-fold cross-validation log-losses of the inner 3-fold aggregated predictions. The sub-indices are the ranking of every calibrator for each dataset (ties in the ranking share the averaged rank). The resulting table of sub-indices is used to compute the Friedman test statistic, resulting in a value of 97.9 and a p-value of $1.14e^{-17}$ indicating statistical difference between the different calibration methods. The last row contains the average ranks of the full table, which is shown in the corresponding critical difference diagram in Figure 5a. The critical difference uses the Bonferroni-Dunn one-tailed statistical test to compute the minimum ranking distance that is shown in the Figure, indicating that for this particular classifier and metric the Dirichlet calibrator with L2 regularisation is significantly better than the other methods, with the exception of Dirichlet with ODIR regularisation.

The same process is applied to each of the 11 classifiers for every metric. Table 6 shows the final average results of all classifiers. Notice that the row corresponding to naive Bayes has the rounded average rankings from Figure 5a.

dataset	n_samples	n_features	n_classes
abalone	4177	8	3
balance-scale	625	4	3
car	1728	6	4
cleveland	297	13	5
dermatology	358	34	6
glass	214	9	6
iris	150	4	3
landsat-satellite	6435	36	6
libras-movement	360	90	15
mfeat-karhunen	2000	64	10
mfeat-morphological	2000	6	10
mfeat-zernike	2000	47	10
optdigits	5620	64	10
page-blocks	5473	10	5
pendigits	10992	16	10
segment	2310	19	7
shuttle	101500	9	7
vehicle	846	18	4
vowel	990	10	11
waveform-5000	5000	40	3
yeast	1484	8	10

Table 3: Datasets used for the large-scale empirical comparison.

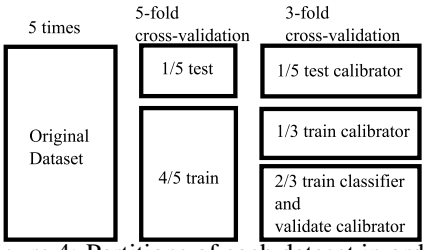
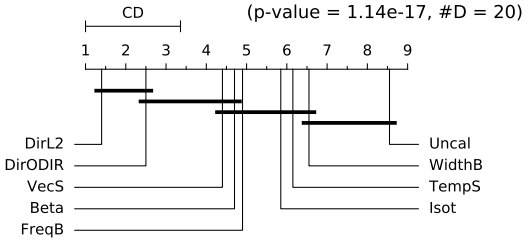


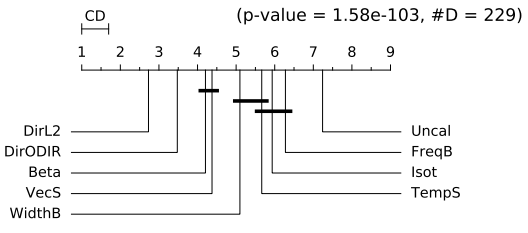
Figure 4: Partitions of each dataset in order to estimate out-of-sample measures.

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
abalone	1.95_{9}	$\mathbf{0.89_{1}}$	0.89_{2}	0.89_{5}	0.89_{6}	0.89_{4}	0.90_{7}	0.89_{3}	0.92_{8}
balance-sc	0.48_{9}	0.22_{2}	$\mathbf{0.22_{1}}$	0.31_{4}	0.41_{8}	0.23_{3}	0.38_{7}	0.36_{6}	0.35_{5}
car	1.57_{9}	0.38_{2}	$\mathbf{0.38_{1}}$	0.59_{6}	0.92_{8}	0.43_{3}	0.56_{5}	0.55_{4}	0.67_{7}
cleveland	2.72_{9}	1.02_{2}	$\mathbf{1.00_{1}}$	1.30_{6}	1.36_{8}	1.08_{3}	1.32_{7}	1.12_{4}	1.14_{5}
dermatolog	2.57_{9}	$\mathbf{0.20_{1}}$	0.41_{7}	0.36_{5}	0.31_{3}	0.44_{8}	0.33_{4}	0.23_{2}	0.40_{6}
glass	2.93_{9}	$\mathbf{1.12_{1}}$	1.57_{6}	1.62_{8}	1.35_{4}	1.48_{5}	1.58_{7}	1.13_{3}	1.12_{2}
iris	0.13_{4}	$\mathbf{0.11_{1}}$	0.12_{3}	0.13_{5}	0.12_{2}	0.13_{6}	0.30_{8}	0.33_{9}	0.21_{7}
landsat-sa	3.87_{9}	$\mathbf{0.35_{1}}$	0.36_{2}	0.55_{4}	0.61_{7}	0.55_{3}	0.58_{5}	0.58_{6}	0.74_{8}
libras-mov	4.99_{9}	$\mathbf{0.94_{1}}$	1.73_{7}	1.32_{3}	1.16_{2}	1.33_{4}	2.04_{8}	1.66_{6}	1.43_{5}
mfeat-karh	0.45_{9}	$\mathbf{0.20_{1}}$	0.20_{2}	0.23_{3}	0.24_{5}	0.23_{4}	0.39_{8}	0.38_{7}	0.29_{6}
mfeat-morp	1.78_{9}	$\mathbf{0.71_{1}}$	0.79_{2}	0.91_{7}	1.01_{8}	0.86_{5}	0.85_{4}	0.82_{3}	0.88_{6}
mfeat-zern	1.73_{9}	0.60_{2}	$\mathbf{0.59_{1}}$	0.71_{3}	0.76_{5}	0.71_{4}	0.85_{8}	0.82_{6}	0.84_{7}
optdigits	3.23_{9}	0.46_{3}	0.46_{2}	0.54_{6}	0.83_{7}	0.52_{5}	$\mathbf{0.43_{1}}$	0.47_{4}	0.84_{8}
page-block	0.76_{9}	$\mathbf{0.17_{1}}$	0.17_{3}	0.19_{4}	0.48_{8}	0.21_{6}	0.17_{2}	0.20_{5}	0.21_{7}
pendigits	1.29_{9}	0.19_{2}	$\mathbf{0.19_{1}}$	0.46_{3}	0.53_{7}	0.46_{4}	0.47_{5}	0.48_{6}	0.58_{8}
segment	1.39_{9}	$\mathbf{0.28_{1}}$	0.31_{2}	0.46_{4}	0.53_{7}	0.46_{5}	0.47_{6}	0.46_{3}	0.56_{8}
vehicle	2.34_{9}	$\mathbf{0.98_{1}}$	0.99_{2}	1.08_{5}	1.16_{7}	1.07_{4}	1.17_{8}	1.04_{3}	1.10_{6}
vowel	0.83_{5}	$\mathbf{0.55_{1}}$	0.62_{2}	0.80_{3}	0.85_{6}	0.80_{4}	1.04_{8}	1.06_{9}	0.89_{7}
waveform-5	0.80_{9}	$\mathbf{0.33_{1}}$	0.33_{2}	0.37_{4}	0.43_{7}	0.35_{3}	0.39_{6}	0.38_{5}	0.46_{8}
yeast	5.12_{9}	1.26_{2}	$\mathbf{1.22_{1}}$	1.33_{6}	2.06_{8}	1.32_{5}	1.29_{3}	1.31_{4}	1.43_{7}
avg rank	8.55	**1.40**	2.50	4.70	6.15	4.40	5.85	4.90	6.55

Table 4: Ranking of calibration methods applied on the classifier nbayes with the measure=loss(Friedman statistic test = 9.79E+01, p-value = 1.14E-17)



(a) Average over all datasets for Naive Bayes classifier



(b) Average over all classifiers

Figure 5: Critical Difference diagrams for the averaged ranking results of the metric Log-loss.

E Results

In this Section we present all the final results, including ranking tables for every metric, critical difference diagrams, the best hyperparameters selected for Dirichlet calibration with L2 regularisation, Frequency binning and Width binning; a comparison of how calibrated the 11 classifiers are, and additional results on deep neural networks.

E.1 Final ranking tables for all metrics

We present here all the final ranking tables for all metrics (Tables 5, 6, 7, 8, 9, 10, 11, and 12). For each ranking, a lower value is indicative of a better metric value (eg. a higher accuracy corresponds to a lower ranking, while a lower log-loss corresponds to a lower ranking as well). Additional details on how to interpret the tables can be found in Section D.2.

Table 5: Rankings for Accuracy

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	7.5	3.1	**2.5**	5.7	7.5	5.5	4.5	4.2	4.4
forest	5.5	5.4	5.7	**3.6**	4.8	3.9	4.5	6.9	4.7
knn	5.8	6.4	4.4	4.3	6.4	4.1	3.7	6.4	**3.5**
lda	6.6	3.5	**2.9**	5.1	6.1	4.7	4.4	6.5	5.2
logistic	6.5	3.1	**2.7**	4.4	6.4	5.7	5.1	6.2	4.9
mlp	4.9	4.9	4.6	**3.7**	5.0	4.6	4.8	7.0	5.5
nbayes	8.2	2.2	**1.6**	5.5	7.8	3.9	4.3	5.0	6.5
qda	7.3	3.5	**2.6**	4.6	7.2	4.6	3.4	5.7	6.0
svc-linear	6.1	**3.1**	3.3	4.6	6.6	4.8	4.2	6.7	5.5
svc-rbf	7.4	4.9	4.5	4.1	6.7	3.7	**3.1**	4.9	5.7
tree	3.8	6.9	5.5	4.6	**3.6**	5.5	4.5	5.3	5.3
avg rank	6.31	4.30	**3.67**	4.56	6.18	4.64	4.24	5.90	5.20

Table 6: Rankings for log-loss

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	8.4	2.1	**2.0**	4.7	7.8	5.1	4.9	4.8	5.1
forest	7.0	5.7	4.4	**3.1**	3.9	3.4	5.8	6.8	4.8
knn	8.2	2.8	4.9	6.4	7.2	5.4	**2.7**	4.4	3.0
lda	7.4	**1.9**	3.0	3.5	5.6	4.0	7.2	7.1	5.2
logistic	7.6	**1.7**	2.5	3.2	5.3	4.0	8.1	7.6	5.0
mlp	4.5	**2.7**	4.5	3.1	3.5	4.0	8.0	8.6	6.2
nbayes	8.6	**1.4**	2.5	4.7	6.2	4.4	5.8	4.9	6.5
qda	7.5	**2.2**	3.0	3.8	5.4	4.5	6.7	7.3	4.6
svc-linear	6.7	**1.9**	2.3	3.7	4.4	3.7	8.1	8.1	6.0
svc-rbf	7.6	4.0	3.4	3.7	5.7	**2.9**	6.0	5.4	6.3
tree	6.2	3.6	5.7	6.2	7.4	6.7	**2.0**	4.0	3.2
avg rank	7.24	**2.73**	3.48	4.21	5.66	4.38	5.93	6.28	5.10

Table 7: Rankings for Brier score

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	8.5	2.8	**2.0**	4.5	8.3	4.8	4.0	4.7	5.5
forest	7.3	6.1	4.9	2.7	4.4	**2.2**	3.5	7.8	6.2
knn	6.4	5.5	4.2	3.8	5.9	**2.8**	3.6	7.3	5.5
lda	6.9	3.0	**2.3**	3.6	6.8	4.1	4.0	8.0	6.2
logistic	8.3	**2.3**	2.4	3.9	6.2	4.3	3.6	8.0	6.0
mlp	5.5	3.8	3.9	**3.5**	5.4	3.8	3.6	8.7	7.0
nbayes	8.5	2.0	**1.4**	5.0	7.2	3.4	4.0	6.3	7.1
qda	7.5	3.0	**2.0**	4.2	6.8	4.1	3.4	7.8	6.2
svc-linear	7.7	2.4	**2.3**	4.3	5.5	4.0	4.0	8.6	6.3
svc-rbf	7.8	4.4	3.8	3.7	6.4	**3.1**	3.5	5.8	6.5
tree	**1.7**	7.1	5.1	3.9	2.8	3.6	5.1	8.4	7.1
avg rank	6.90	3.88	**3.13**	3.90	5.97	3.65	3.85	7.40	6.32

Table 8: Rankings for MCE

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	8.2	3.8	**3.7**	4.8	6.2	4.3	5.2	4.0	4.7
forest	6.3	5.9	4.8	4.0	4.9	5.4	4.7	4.9	**4.0**
knn	3.5	5.4	6.4	4.7	6.7	6.5	3.9	4.6	**3.4**
lda	6.6	5.1	4.2	5.1	4.6	**4.1**	5.6	5.0	4.7
logistic	5.2	4.3	4.5	5.8	4.7	5.6	6.1	4.7	**4.0**
mlp	5.6	5.0	6.4	4.0	5.2	4.5	5.6	5.2	**3.4**
nbayes	7.8	5.5	5.5	4.6	4.5	4.0	5.0	**4.0**	4.1
qda	6.8	4.9	5.2	5.3	4.0	5.1	4.8	5.6	**3.4**
svc-linear	5.3	5.0	5.0	5.0	4.8	4.9	5.9	4.9	**4.2**
svc-rbf	4.4	5.9	5.7	5.4	4.4	6.0	**3.9**	4.1	5.1
tree	5.4	6.3	5.0	5.1	5.8	5.6	4.0	4.0	**3.8**
avg rank	5.91	5.18	5.14	4.90	5.09	5.11	4.97	4.64	**4.07**

Table 9: Rankings for confidence-ECE

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	8.2	3.1	**2.5**	4.0	7.3	2.9	4.2	6.4	6.3
forest	7.3	5.9	3.0	4.0	4.4	**2.5**	2.8	8.3	6.7
knn	4.0	4.7	**3.2**	4.5	4.1	3.4	5.5	8.1	7.5
lda	6.7	4.4	**3.0**	3.4	6.0	4.0	3.3	8.0	6.2
logistic	8.0	3.6	3.4	4.0	5.0	3.5	**3.2**	8.0	6.3
mlp	5.7	4.2	3.2	3.8	5.6	3.8	**2.3**	8.7	7.7
nbayes	8.0	3.8	**2.3**	4.3	5.6	3.5	3.6	7.5	6.3
qda	6.5	4.3	2.9	4.0	6.2	4.0	**2.2**	8.4	6.5
svc-linear	7.0	4.0	3.2	4.2	4.6	3.2	3.2	8.5	6.9
svc-rbf	5.7	4.5	4.4	4.6	4.0	4.0	**3.2**	8.6	6.0
tree	**2.0**	6.2	4.1	3.0	3.7	3.7	5.9	8.8	7.6
avg rank	6.27	4.44	**3.20**	4.00	5.12	3.50	3.59	8.12	6.75

Table 10: Rankings for classwise-ECE

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	6.7	3.3	**2.9**	5.4	6.3	3.4	4.0	6.4	6.5
forest	7.6	5.6	3.4	3.6	5.0	**2.1**	2.7	8.3	6.7
knn	4.1	4.6	3.1	4.2	5.5	**2.5**	5.4	7.9	7.6
lda	6.3	4.1	3.4	3.4	6.8	3.8	**3.0**	7.9	6.3
logistic	8.0	3.1	3.6	3.8	6.1	3.4	3.1	7.8	5.9
mlp	5.5	4.1	3.2	3.2	6.3	3.8	**2.7**	8.6	7.6
nbayes	7.4	2.8	**1.9**	5.0	7.8	3.4	3.8	7.0	6.0
qda	6.5	3.6	2.6	4.0	7.0	4.0	**2.5**	8.3	6.3
svc-linear	7.1	3.7	3.8	4.0	4.8	3.2	3.2	8.5	6.7
svc-rbf	6.0	4.0	4.0	4.2	5.0	3.7	**3.3**	8.6	6.2
tree	**2.2**	6.3	4.0	3.7	4.2	3.4	5.0	8.5	7.7
avg rank	6.12	4.12	**3.28**	4.05	5.89	3.34	3.54	7.99	6.67

E.2 Final critical difference diagrams for every metric

In order to perform a final comparison between calibration methods, we considered every combination of dataset and classifier as a group $n = \#datasets \times \#classifiers$, and ranked the results of the k calibration methods. With this setting, we have performed the Friedman statistical test followed by the one-tailed Bonferroni-Dunn test to obtain critical differences (CDs) for every metric (See Figure

Table 11: Rankings for p-confidence-ECE

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	7.9	2.8	**2.5**	4.3	6.9	3.8	4.5	6.3	6.1
forest	7.0	6.0	**2.9**	3.6	3.6	3.2	3.2	8.4	7.1
knn	6.2	3.0	**2.9**	4.2	4.8	3.9	4.6	7.9	7.5
lda	7.3	3.7	**3.1**	3.9	5.3	4.1	3.5	8.0	6.2
logistic	7.8	3.5	4.0	3.7	4.7	3.6	**3.4**	8.0	6.4
mlp	5.1	4.3	3.7	3.6	4.7	4.1	**3.2**	8.7	7.6
nbayes	8.1	3.2	**2.5**	4.4	5.3	3.6	3.3	7.6	6.9
qda	6.8	3.9	3.4	3.6	5.5	4.3	**2.6**	8.3	6.6
svc-linear	6.8	4.0	3.5	3.8	4.3	**3.3**	4.1	8.5	6.8
svc-rbf	5.6	4.5	4.2	4.6	4.2	4.0	**3.2**	8.4	6.2
tree	**2.9**	5.3	4.0	3.3	4.0	4.5	5.2	8.2	7.5
avg rank	6.50	4.01	**3.33**	3.90	4.84	3.87	3.71	8.03	6.81

Table 12: Rankings for p-classwise-ECE

	Uncal	DirL2	DirODIR	Beta	TempS	VecS	Isot	FreqB	WidthB
adas	7.4	3.1	**3.0**	5.1	6.8	3.7	4.1	6.0	5.7
forest	6.1	4.9	4.7	**3.2**	4.4	4.3	3.6	8.0	5.8
knn	7.7	**2.0**	4.2	5.0	6.1	4.6	3.0	7.0	5.5
lda	7.3	**2.9**	4.1	3.5	6.0	3.8	3.7	7.9	5.8
logistic	6.9	**3.0**	4.3	3.6	5.1	4.3	4.1	8.0	5.6
mlp	5.0	**2.5**	5.2	2.9	4.3	4.5	5.0	8.6	6.9
nbayes	8.2	**1.6**	2.3	4.9	7.0	3.2	4.5	6.8	6.5
qda	7.5	**2.2**	3.6	3.8	6.0	4.6	3.7	8.2	5.5
svc-linear	6.2	**3.0**	4.4	4.0	4.5	4.0	4.8	8.5	5.5
svc-rbf	5.5	3.8	4.1	4.4	5.0	4.8	**3.4**	8.3	5.8
tree	3.7	4.6	**3.6**	5.0	4.2	4.4	4.2	8.0	7.3
avg rank	6.50	**3.07**	3.97	4.12	5.40	4.21	4.01	7.75	5.98

6). The results showed Dirichlet L2 as the best calibration method for the measures accuracy, log-loss and p-cw-ece with statistical significance (See Figures 6a 6c, and 6h), and in the group of the best calibration methods in the rest of the metrics with statistical significance, but no difference within the group. It is worth mentioning that Figure 6c showed statistical difference between Dirichlet L2, OvR Beta, OvR width binning, and the rest of the calibrators in one group; in the mentioned order.

E.3 Best calibrator hyperparameters

Figure 8 shows the best hyperparameters for every inner 3-fold-cross-validation. Dirichlet L2 (Figure 8a) shows a preference for regularisation hyperparameter $\lambda = 1e^{-3}$ and lower values. Our current minimum regularisation value of $1e^{-7}$ is also being selected multiple times, indicating that lower values may be optimal in several occasions. However, this fact did not seem to hurt the overall good results in our experiments. One-vs.-Rest frequency binning tends to prefer 10 bins of equal number of samples, while One-vs.Rest width binning prefers 5 equal sized bins (See Figures 8b and 8c respectively).

E.4 Comparison of classifiers

In this Section we compare all the classifiers without post-hoc calibration on 17 of the datasets; from the total of 21 datasets *shuttle*, *yeast*, *mfeat-karhunen* and *libras-movement* were removed from this analysis as at least one classifier was not able to complete the experiment.

Figure 9 shows the Critical Difference diagram for all the 8 metrics. In particular, the MLP and the SVC with linear kernel are always in the group with the higher rankings and never in the lowest. Similarly, random forest is consistently in the best group, but in the worst group as well in 4 of the measures. SVC with radial basis kernel is in the best group 6 times, but 3 times in the worst. On the other hand, naive Bayes and Adaboost SAMME are consistently in the worst group and never in the best one. The rest of the classifiers did not show a clear ranking position.

Figures 10b and 10a show the proportion of times each classifier passed the p-conf-ECE and p-cw-ECE statistical test for all datasets and cross-validation folds.

E.5 Deep neural networks

In this section, we provide further discussion about results from the deep networks experiments. These are given in the form of critical difference diagrams (Figure 11) and tables (Tables 13-20) both including the following measures: error rate, log-loss, Brier score, maximum calibration error (MCE), confidence-ECE (conf-ECE), classwise-ECE (cw-ECE), as well as significance measures p-conf-ECE and p-cw-ECE.

In addition, Table 21 compares MS-ODIR and vector scaling on log-loss. On the table, we also added MS-ODIR-zero which was obtained from the respective MS-ODIR model by replacing the off-diagonal entries with zeroes. Each experiment is replicated three times with different splits on datasets. This is done to compare the stability of the methods. In each replication, the best scoring model is written in bold.

Finally, Figure 12 shows that temperature scaling systematically under-estimates class 4 probabilities on the model `c10_resnet_wide32` on CIFAR-10.

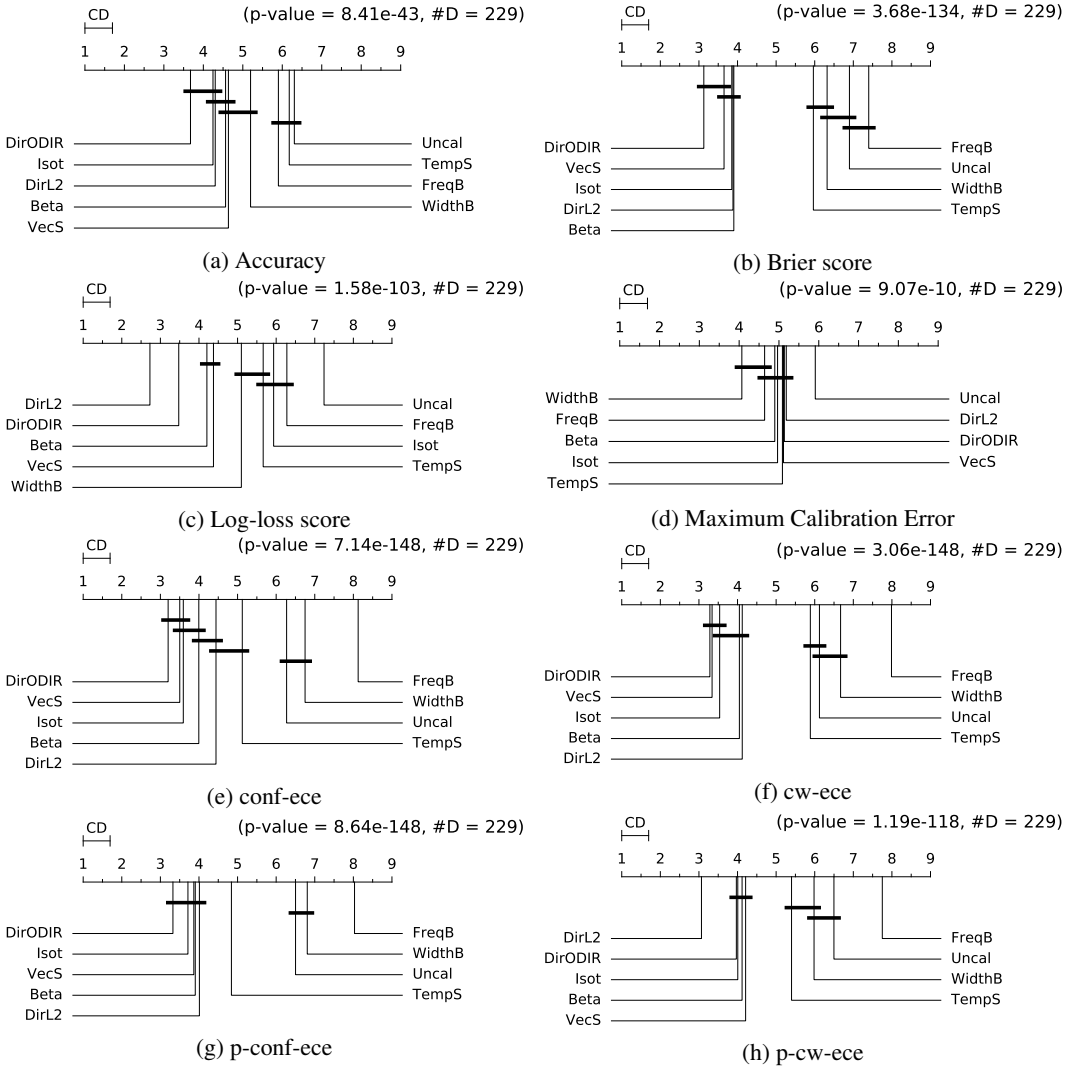


Figure 6: Critical difference of the average of multiclass classifiers.

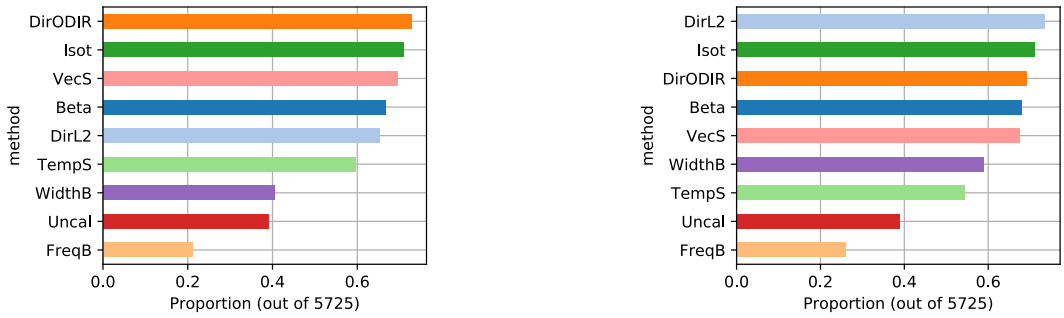
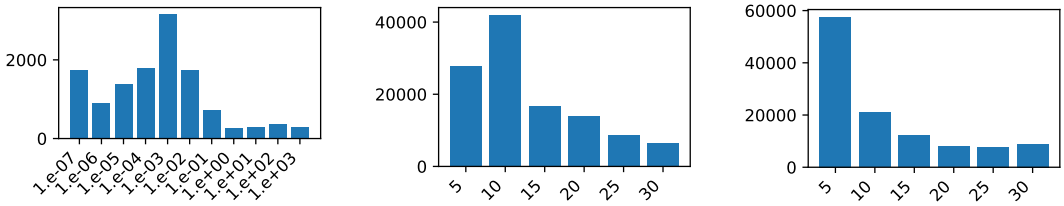
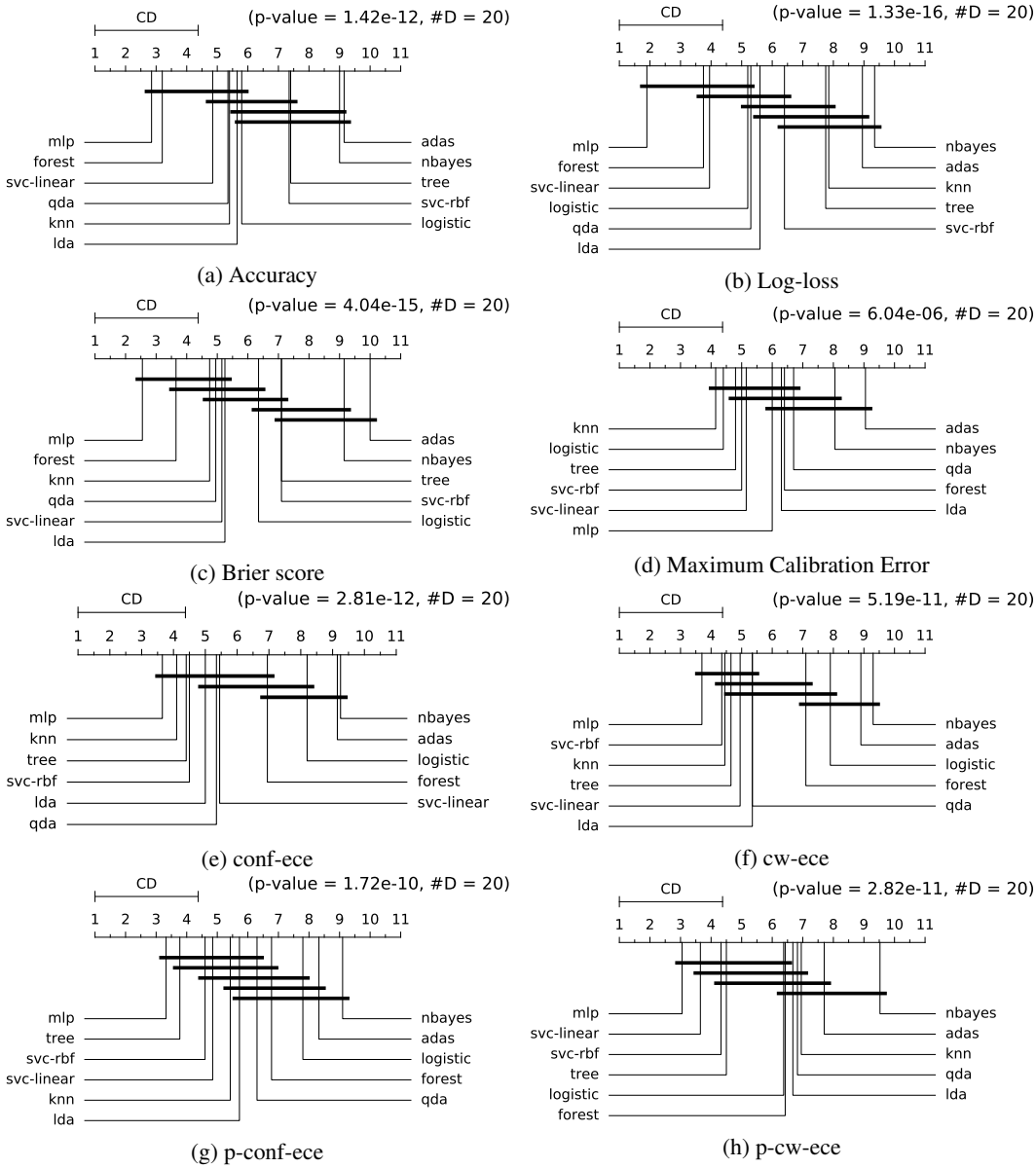


Figure 7: Proportion of times each calibrator passes a calibration p-test with a p-value higher than 0.05.



(a) Dirichlet L2 λ (b) OvR Frequency binning #bins (c) OvR Width binning #bins

Figure 8: Histogram of the selected hyperparameters during the inner 3-fold-cross-validation



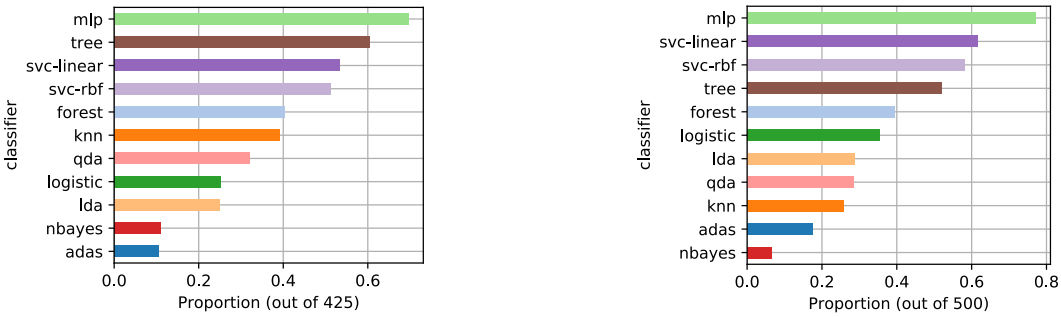
(a) Accuracy (b) Log-loss

(c) Brier score (d) Maximum Calibration Error

(e) conf-ece (f) cw-ece

(g) p-conf-ece (h) p-cw-ece

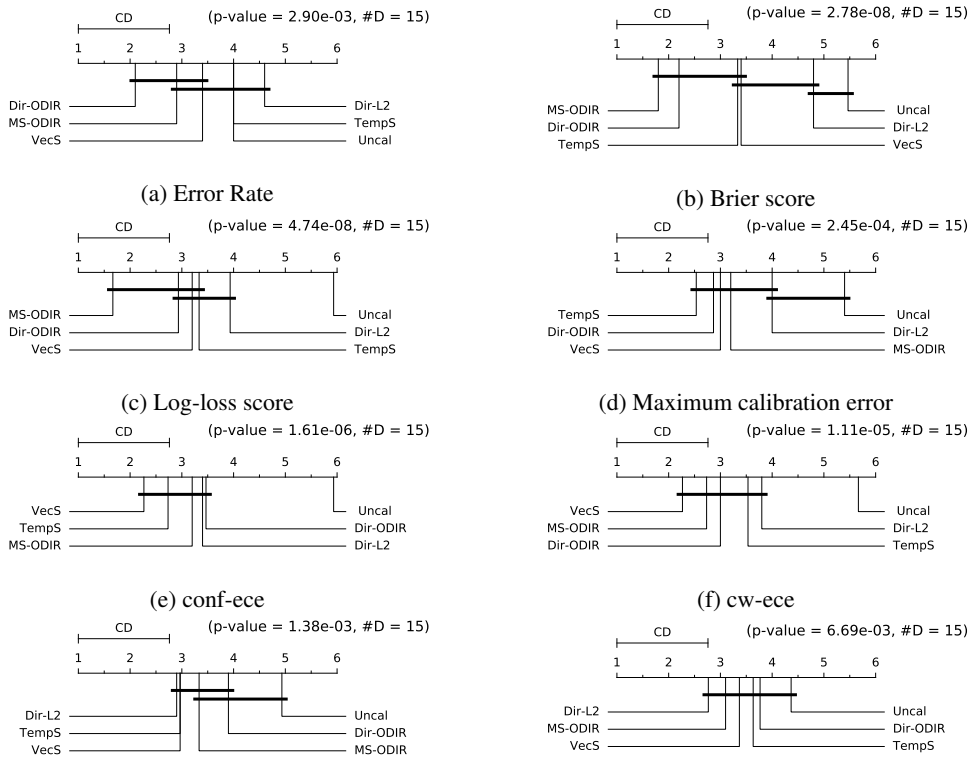
Figure 9: Critical difference of uncalibrated classifiers.



(a) p-conf-ece

(b) p-cw-ece

Figure 10: Proportion of times each classifier is already calibrated with different p-tests.



(a) Error Rate

(b) Brier score

(c) Log-loss score

(d) Maximum calibration error

(e) conf-ece

(f) cw-ece

(g) p-conf-ece

(h) p-cw-ece

Figure 11: Critical difference of the deep neural networks.

Table 13: Scores and ranking of calibration methods for **log-loss**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.39098_{6}	$\mathbf{0.19497_{1}}$	0.19692_{4}	0.19536_{2}	0.19743_{5}	0.19634_{3}
c10_densenet40	0.42821_{6}	0.22509_{5}	$\mathbf{0.22048_{1}}$	0.22371_{4}	0.22270_{3}	0.22240_{2}
c10_lenet5	0.82326_{6}	0.80031_{5}	0.74418_{2}	0.74441_{3}	0.74704_{4}	$\mathbf{0.74262_{1}}$
c10_resnet110	0.35827_{6}	0.20926_{5}	$\mathbf{0.20303_{1}}$	0.20511_{3}	0.20595_{4}	0.20375_{2}
c10_resnet110_SD	0.30325_{6}	0.17760_{5}	0.17694_{4}	0.17608_{3}	0.17549_{2}	$\mathbf{0.17537_{1}}$
c10_resnet_wide32	0.38170_{6}	0.19148_{5}	0.18464_{4}	0.18203_{2}	0.18276_{3}	$\mathbf{0.18165_{1}}$
c100_convnet	1.64120_{6}	$\mathbf{0.94162_{1}}$	1.18945_{5}	0.96121_{2}	0.96369_{4}	0.96141_{3}
c100_densenet40	2.01740_{6}	1.05713_{2}	1.25293_{5}	1.05909_{4}	1.05831_{3}	$\mathbf{1.05084_{1}}$
c100_lenet5	2.78365_{6}	2.64979_{5}	2.59482_{4}	2.48951_{2}	2.51590_{3}	$\mathbf{2.48670_{1}}$
c100_resnet110	1.69371_{6}	1.09169_{3}	1.21239_{5}	1.09607_{4}	1.08916_{2}	$\mathbf{1.07370_{1}}$
c100_resnet110_SD	1.35250_{6}	0.94214_{3}	1.19837_{5}	0.94477_{4}	$\mathbf{0.92341_{1}}$	0.92731_{2}
c100_resnet_wide32	1.80215_{6}	0.94453_{3}	1.08711_{5}	0.95288_{4}	0.93650_{2}	$\mathbf{0.93273_{1}}$
SVHN_convnet	0.20460_{6}	0.15142_{5}	0.14246_{3}	0.13791_{2}	0.14388_{4}	$\mathbf{0.13760_{1}}$
SVHN_resnet152_SD	0.08542_{6}	$\mathbf{0.07861_{1}}$	0.08463_{5}	0.08038_{2}	0.08124_{4}	0.08100_{3}
avg rank	6.0	3.5	3.79	2.93	3.14	1.64

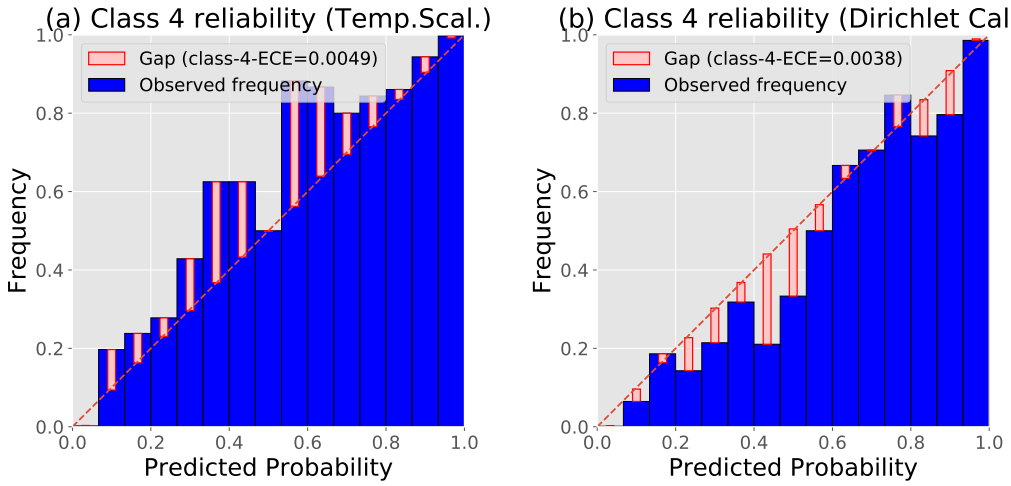


Figure 12: Reliability diagrams of c10_resnet_wide32 on CIFAR-10: (a) classwise-reliability for class 4 after temperature scaling; (b) classwise-reliability for class 4 after Dirichlet calibration.

Table 14: Scores and ranking of calibration methods for **Brier score**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.01090_{6}	$\mathbf{0.00952_{1}}$	0.00969_{5}	0.00955_{3}	0.00958_{4}	0.00953_{2}
c10_densenet40	0.01274_{6}	0.01100_{4}	0.01102_{5}	0.01097_{2}	0.01097_{3}	$\mathbf{0.01097_{1}}$
c10_lenet5	0.03788_{6}	0.03748_{5}	0.03510_{2}	0.03511_{3}	0.03523_{4}	$\mathbf{0.03502_{1}}$
c10_resnet110	0.01102_{6}	0.00979_{4}	0.00979_{5}	0.00977_{2}	0.00978_{3}	$\mathbf{0.00976_{1}}$
c10_resnet110_SD	0.00981_{6}	0.00874_{4}	0.00877_{5}	0.00867_{3}	0.00867_{2}	$\mathbf{0.00866_{1}}$
c10_resnet_wide32	0.01047_{6}	0.00924_{5}	0.00909_{4}	$\mathbf{0.00888_{1}}$	0.00891_{3}	0.00889_{2}
c100_convnet	0.00425_{5}	$\mathbf{0.00358_{1}}$	0.00441_{6}	0.00358_{2}	0.00362_{4}	0.00361_{3}
c100_densenet40	0.00491_{6}	0.00401_{3}	0.00468_{5}	0.00400_{2}	0.00403_{4}	$\mathbf{0.00400_{1}}$
c100_lenet5	0.00813_{6}	0.00792_{5}	0.00786_{4}	0.00760_{2}	0.00767_{3}	$\mathbf{0.00760_{1}}$
c100_resnet110	0.00453_{6}	0.00392_{3}	0.00438_{5}	0.00391_{2}	0.00393_{4}	$\mathbf{0.00391_{1}}$
c100_resnet110_SD	0.00418_{5}	0.00367_{4}	0.00456_{6}	0.00364_{3}	$\mathbf{0.00360_{1}}$	0.00361_{2}
c100_resnet_wide32	0.00432_{6}	0.00355_{4}	0.00401_{5}	0.00354_{3}	0.00352_{2}	$\mathbf{0.00351_{1}}$
SVHN_convnet	0.00776_{6}	0.00598_{5}	0.00555_{3}	$\mathbf{0.00530_{1}}$	0.00561_{4}	0.00532_{2}
SVHN_resnet152_SD	0.00297_{3}	$\mathbf{0.00291_{1}}$	0.00305_{6}	0.00293_{2}	0.00299_{5}	0.00298_{4}
avg rank	5.64	3.5	4.71	2.21	3.29	1.64

Table 15: Scores and ranking of calibration methods for **confidence-ECE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.04760_{6}	0.01065_{5}	0.00769_{2}	0.00960_{4}	$\mathbf{0.00740_{1}}$	0.00782_{3}
c10_densenet40	0.05500_{6}	0.00946_{2}	$\mathbf{0.00568_{1}}$	0.01097_{5}	0.01018_{4}	0.00988_{3}
c10_lenet5	0.05180_{6}	0.01665_{5}	0.01383_{3}	0.01367_{2}	$\mathbf{0.01310_{1}}$	0.01468_{4}
c10_resnet110	0.04750_{6}	0.01132_{5}	$\mathbf{0.00680_{1}}$	0.01086_{3}	0.01130_{4}	0.01059_{2}
c10_resnet110_SD	0.04113_{6}	$\mathbf{0.00555_{1}}$	0.00646_{4}	0.00815_{5}	0.00579_{3}	0.00566_{2}
c10_resnet_wide32	0.04505_{6}	0.00784_{4}	$\mathbf{0.00524_{1}}$	0.00837_{5}	0.00769_{3}	0.00727_{2}
c100_convnet	0.17614_{6}	$\mathbf{0.01367_{1}}$	0.14347_{5}	0.02069_{3}	0.01965_{2}	0.02660_{4}
c100_densenet40	0.21156_{6}	$\mathbf{0.00902_{1}}$	0.12380_{5}	0.01138_{2}	0.01224_{3}	0.02197_{4}
c100_lenet5	0.12125_{6}	0.01499_{4}	0.01369_{2}	0.02003_{5}	$\mathbf{0.01294_{1}}$	0.01407_{3}
c100_resnet110	0.18480_{6}	$\mathbf{0.02380_{1}}$	0.14535_{5}	0.02822_{4}	0.02693_{2}	0.02735_{3}
c100_resnet110_SD	0.15861_{5}	$\mathbf{0.01214_{1}}$	0.15920_{6}	0.02283_{4}	0.01296_{2}	0.02246_{3}
c100_resnet_wide32	0.18784_{6}	$\mathbf{0.01472_{1}}$	0.13509_{5}	0.01891_{3}	0.01718_{2}	0.02581_{4}
SVHN_convnet	0.07755_{6}	0.01179_{4}	0.01910_{5}	0.00997_{2}	$\mathbf{0.00934_{1}}$	0.01037_{3}
SVHN_resnet152_SD	0.00862_{6}	0.00607_{4}	0.00691_{5}	$\mathbf{0.00582_{1}}$	0.00595_{2}	0.00604_{3}
avg rank	5.93	2.79	3.57	3.43	2.21	3.07

Table 16: Scores and ranking of calibration methods for **classwise-ECE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.10375_{6}	0.04423_{4}	0.04262_{2}	0.04507_{5}	$\mathbf{0.04259_{1}}$	0.04352_{3}
c10_densenet40	0.11430_{6}	0.03977_{5}	$\mathbf{0.03412_{1}}$	0.03687_{4}	0.03609_{2}	0.03678_{3}
c10_lenet5	0.19849_{6}	0.17141_{5}	$\mathbf{0.05185_{1}}$	0.05891_{4}	0.05705_{2}	0.05862_{3}
c10_resnet110	0.09846_{6}	0.04344_{5}	$\mathbf{0.03206_{1}}$	0.03950_{4}	0.03653_{3}	0.03615_{2}
c10_resnet110_SD	0.08647_{6}	0.03071_{4}	0.03148_{5}	0.02937_{3}	0.02713_{2}	$\mathbf{0.02681_{1}}$
c10_resnet_wide32	0.09530_{6}	0.04775_{5}	0.03153_{3}	0.02947_{2}	0.03164_{4}	$\mathbf{0.02921_{1}}$
c100_convnet	0.42414_{6}	$\mathbf{0.22683_{1}}$	0.40185_{5}	0.24041_{3}	0.24063_{4}	0.23958_{2}
c100_densenet40	0.47026_{6}	0.18664_{2}	0.32985_{5}	$\mathbf{0.18630_{1}}$	0.18879_{3}	0.19112_{4}
c100_lenet5	0.47264_{6}	0.38481_{5}	0.21865_{4}	0.21348_{2}	$\mathbf{0.20293_{1}}$	0.21379_{3}
c100_resnet110	0.41644_{6}	0.20095_{3}	0.35885_{5}	$\mathbf{0.18639_{1}}$	0.19442_{2}	0.20270_{4}
c100_resnet110_SD	0.37518_{6}	0.20310_{4}	0.37346_{5}	0.18895_{3}	$\mathbf{0.17015_{1}}$	0.18552_{2}
c100_resnet_wide32	0.42027_{6}	0.18573_{4}	0.33258_{5}	0.17951_{2}	$\mathbf{0.17082_{1}}$	0.17966_{3}
SVHN_convnet	0.15935_{6}	0.03830_{4}	0.04276_{5}	0.02638_{2}	$\mathbf{0.02480_{1}}$	0.02750_{3}
SVHN_resnet152_SD	0.01940_{2}	$\mathbf{0.01849_{1}}$	0.02184_{6}	0.01988_{3}	0.02120_{5}	0.02088_{4}
avg rank	5.71	3.71	3.79	2.79	2.29	2.71

Table 17: Scores and ranking of calibration methods for **MCE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.59173_{6}	0.23150_{4}	0.12432_{2}	0.24830_{5}	0.12831_{3}	$\mathbf{0.07621_{1}}$
c10_densenet40	0.33396_{6}	0.09929_{2}	0.11679_{4}	$\mathbf{0.07858_{1}}$	0.12046_{5}	0.11297_{3}
c10_lenet5	0.11281_{6}	0.09158_{3}	$\mathbf{0.05112_{1}}$	0.09009_{2}	0.09996_{4}	0.10061_{5}
c10_resnet110	0.29580_{6}	0.23639_{4}	0.24405_{5}	$\mathbf{0.08331_{1}}$	0.13130_{2}	0.22678_{3}
c10_resnet110_SD	0.32484_{6}	$\mathbf{0.07823_{1}}$	0.23064_{5}	0.13309_{3}	0.14276_{4}	0.08422_{2}
c10_resnet_wide32	0.37215_{4}	$\mathbf{0.07060_{1}}$	0.49283_{6}	0.41567_{5}	0.26539_{3}	0.26372_{2}
c100_convnet	0.36391_{6}	0.13689_{4}	0.23333_{5}	0.07235_{2}	$\mathbf{0.07043_{1}}$	0.08171_{3}
c100_densenet40	0.45400_{6}	$\mathbf{0.02213_{1}}$	0.19748_{5}	0.04074_{2}	0.04293_{3}	0.05004_{4}
c100_lenet5	0.20097_{6}	0.05836_{3}	$\mathbf{0.05678_{1}}$	0.06774_{4}	0.05749_{2}	0.08939_{5}
c100_resnet110	0.39882_{6}	0.07099_{2}	0.20732_{5}	0.08026_{4}	0.07354_{3}	$\mathbf{0.06678_{1}}$
c100_resnet110_SD	0.48291_{6}	0.04099_{2}	0.24578_{5}	0.05979_{3}	$\mathbf{0.04038_{1}}$	0.06612_{4}
c100_resnet_wide32	0.45639_{6}	$\mathbf{0.03606_{1}}$	0.19370_{5}	0.05521_{2}	0.06605_{4}	0.06468_{3}
SVHN_convnet	0.30011_{5}	0.40691_{6}	$\mathbf{0.16154_{1}}$	0.18458_{3}	0.16312_{2}	0.18588_{4}
SVHN_resnet152_SD	0.25032_{5}	$\mathbf{0.18244_{1}}$	0.23895_{4}	0.19649_{2}	0.23092_{3}	0.80082_{6}
avg rank	5.71	2.5	3.86	2.79	2.86	3.29

Table 18: Scores and ranking of calibration methods for **error rate (%)**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	6.18000_{2}	6.18000_{2}	6.38000_{6}	$\mathbf{6.12000_{1}}$	6.36000_{5}	6.32000_{4}
c10_densenet40	7.58000_{5}	7.58000_{5}	$\mathbf{7.49000_{1}}$	7.53000_{4}	7.52000_{3}	7.50000_{2}
c10_lenet5	27.26000_{5}	27.26000_{5}	$\mathbf{25.25000_{1}}$	25.44000_{2}	25.49000_{3}	25.50000_{4}
c10_resnet110	6.44000_{1}	6.44000_{1}	6.54000_{6}	6.49000_{4}	6.47000_{3}	6.49000_{4}
c10_resnet110_SD	5.96000_{5}	5.96000_{5}	5.90000_{4}	$\mathbf{5.77000_{1}}$	5.83000_{3}	5.81000_{2}
c10_resnet_wide32	6.07000_{5}	6.07000_{5}	5.94000_{4}	5.76000_{2}	$\mathbf{5.74000_{1}}$	5.81000_{3}
c100_convnet	26.12000_{1}	26.12000_{1}	30.96000_{6}	26.22000_{3}	26.56000_{4}	26.60000_{5}
c100_densenet40	30.00000_{3}	30.00000_{3}	33.48000_{6}	29.87000_{2}	30.16000_{5}	$\mathbf{29.61000_{1}}$
c100_lenet5	66.41000_{5}	66.41000_{5}	65.97000_{4}	62.53000_{2}	63.59000_{3}	$\mathbf{62.44000_{1}}$
c100_resnet110	28.52000_{4}	28.52000_{4}	30.04000_{6}	$\mathbf{28.36000_{1}}$	28.40000_{2}	28.45000_{3}
c100_resnet110_SD	27.17000_{4}	27.17000_{4}	31.43000_{6}	26.96000_{3}	26.50000_{2}	$\mathbf{26.42000_{1}}$
c100_resnet_wide32	26.18000_{4}	26.18000_{4}	27.69000_{6}	26.07000_{2}	26.08000_{3}	$\mathbf{26.06000_{1}}$
SVHN_convnet	3.82750_{5}	3.82750_{5}	3.42811_{3}	$\mathbf{3.34728_{1}}$	3.51845_{4}	3.37105_{2}
SVHN_resnet152_SD	1.84773_{2}	1.84773_{2}	1.90535_{6}	$\mathbf{1.80547_{1}}$	1.87462_{4}	1.87462_{4}
avg rank	4.14	4.14	4.64	2.11	3.25	2.71

Table 19: Scores and ranking of calibration methods for **p-confidence-ECE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.0_{6}	0.032_{4}	0.363_{2}	0.019_{5}	$\mathbf{0.461_{1}}$	0.052_{3}
c10_densenet40	0.0_{4}	0.002_{2}	$\mathbf{0.525_{1}}$	0.000_{4}	0.000_{4}	0.000_{4}
c10_lenet5	0.0_{6}	0.008_{5}	0.027_{4}	0.084_{3}	$\mathbf{0.155_{1}}$	0.144_{2}
c10_resnet110	0.0_{4}	0.000_{4}	$\mathbf{0.246_{1}}$	0.000_{4}	0.000_{4}	0.000_{4}
c10_resnet110_SD	0.0_{6}	0.105_{4}	$\mathbf{0.179_{1}}$	0.003_{5}	0.114_{3}	0.124_{2}
c10_resnet_wide32	0.0_{6}	0.017_{3}	$\mathbf{0.281_{1}}$	0.005_{4}	0.005_{4}	0.076_{2}
c100_convnet	0.0_{5}	$\mathbf{0.174_{1}}$	0.000_{5}	0.049_{2}	0.021_{3}	0.000_{5}
c100_densenet40	0.0_{5}	$\mathbf{0.817_{1}}$	0.000_{5}	0.617_{2}	0.238_{3}	0.000_{5}
c100_lenet5	0.0_{6}	0.153_{4}	0.217_{3}	0.001_{5}	0.395_{2}	$\mathbf{0.422_{1}}$
c100_resnet110	0.0_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}
c100_resnet110_SD	0.0_{4}	0.009_{2}	0.000_{4}	0.000_{4}	$\mathbf{0.060_{1}}$	0.000_{4}
c100_resnet_wide32	0.0_{4}	$\mathbf{0.022_{1}}$	0.000_{4}	0.000_{4}	0.001_{2}	0.000_{4}
mnist_mlp	0.0_{6}	0.616_{3}	$\mathbf{0.948_{1}}$	0.486_{4}	0.455_{5}	0.677_{2}
SVHN_convnet	0.0_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}
SVHN_resnet152_SD	0.0_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}	0.000_{3}
avg rank	4.93	2.97	2.9	3.9	2.97	3.33

Table 20: Scores and ranking of calibration methods for **p-classwise-ECE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.0_6	0.0104_4	$\mathbf{0.1276_1}$	0.0038_5	0.0340_2	0.0114_3
c10_densenet40	0.0_4	0.0000_4	$\mathbf{0.0093_1}$	0.0000_4	0.0000_4	0.0000_4
c10_lenet5	0.0_5	0.0000_5	$\mathbf{0.6014_1}$	0.0390_4	0.1230_2	0.0501_3
c10_resnet110	0.0_4	0.0000_4	$\mathbf{0.0088_1}$	0.0000_4	0.0000_4	0.0000_4
c10_resnet110_SD	0.0_6	0.0058_5	0.0105_3	0.0077_4	0.1816_2	$\mathbf{0.2196_1}$
c10_resnet_wide32	0.0_5	0.0000_5	0.0096_3	0.0158_2	0.0006_4	$\mathbf{0.0249_1}$
c100_convnet	0.0_4	$\mathbf{0.0770_1}$	0.0000_4	0.0000_4	0.0000_4	0.0000_4
c100_densenet40	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
c100_lenet5	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
c100_resnet110	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
c100_resnet110_SD	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
c100_resnet_wide32	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
mnist_mlp	0.0_6	$\mathbf{0.5669_1}$	0.0842_3	0.0022_5	0.0280_4	0.1178_2
SVHN_convnet	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
SVHN_resnet152_SD	0.0_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3	0.0000_3
avg rank	4.37	3.63	2.77	3.77	3.37	3.1

Table 21: Comparison of MS-ODIR and vector scaling for **log-loss**.

	Replication 1			Replication 2			Replication 3		
	VecS	MS-ODIR	MS-ODIR-zero	VecS	MS-ODIR	MS-ODIR-zero	VecS	MS-ODIR	MS-ODIR-zero
c10_convnet	0.19774	**0.19632**	0.19632	–	–	–	–	–	–
c10_densenet40	**0.22240**	0.22240	0.22240	0.21316	**0.21186**	0.21366	0.21350	**0.21325**	0.21327
c10_lenet5	0.74688	**0.74262**	0.74830	0.69392	**0.69287**	0.69335	**0.67955**	0.67974	0.68127
c10_resnet110	0.20624	**0.20375**	0.20537	0.20064	**0.19803**	0.20040	0.19655	**0.19536**	0.19739
c10_resnet110_SD	0.17545	**0.17537**	0.17539	0.18123	**0.18094**	0.18097	**0.17799**	0.17829	0.17829
c10_resnet_wide32	0.18274	**0.18165**	0.18302	0.18522	**0.18364**	0.18546	0.17431	**0.17274**	0.17448
c100_convnet	0.96311	**0.96141**	0.96149	–	–	–	–	–	–
c100_densenet40	1.05714	**1.05084**	1.06804	1.06366	**1.05456**	1.07107	1.07704	**1.06918**	1.08559
c100_lenet5	2.51695	**2.48670**	2.57932	2.21546	**2.20054**	2.22360	2.28054	**2.27887**	2.29485
c100_resnet110	1.08824	**1.07370**	1.10137	1.09066	**1.08267**	1.11116	1.11977	**1.10672**	1.13900
c100_resnet110_SD	**0.92275**	0.92731	0.92730	0.87758	**0.87698**	0.87701	**0.88523**	0.88731	0.88727
c100_resnet_wide32	0.93724	**0.93273**	0.94060	0.93291	**0.92531**	0.94854	0.93183	**0.92439**	0.94568
SVHN_convnet	0.14392	**0.13760**	0.14507	–	–	–	–	–	–
SVHN_resnet152_SD	0.08131	**0.08100**	0.08100	0.12728	**0.12723**	0.12639	0.12559	**0.12453**	0.12381