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

UNIVERSITY OF TARTU
University of BRISTOL
sphere
The Alan Turing Institute
Universidade Federal da Paraíba ESTATÍSTICA

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, Peter Flach

Class probabilities predicted by most multiclass classifiers are uncalibrated, often tending towards over-confidence. With neural networks, calibration can be improved by temperature scaling, a method to learn a single corrective multiplicative factor for inputs to the last softmax layer. On non-neural models the existing methods apply binary calibration in a pairwise or one-vs-rest fashion. We propose a natively multiclass calibration method applicable to classifiers from any model class, derived from Dirichlet distributions and generalising the beta calibration method from binary classification. It is easily implemented with neural nets since it is equivalent to log-transforming the uncalibrated probabilities, followed by one linear layer and softmax.

Notion of calibration:

- A probabilistic classifier $\hat{\mathbf{p}}$ is multiclass-calibrated if for any prediction vector $\mathbf{q} = (q_1, \ldots, q_k) \in \Delta_k$, the proportions of classes among all possible instances x getting the same prediction $\hat{\mathbf{p}}(\mathbf{x}) = \mathbf{q}$ are:

$$P(Y = i \mid \hat{\mathbf{p}}_i(X) = \mathbf{q}) = q_i \text{ for } i = 1, \ldots, k.$$

- Classwise-calibrated if:

$$P(Y = i \mid \hat{\mathbf{p}}_i(X) = q_i) = q_i.$$

- Confidence-calibrated if:

$$P(Y = \arg\max(\hat{\mathbf{p}}(X)) \mid \max(\hat{\mathbf{p}}(X)) = c) = c.$$

Evaluation of calibration:

- Some empirical measures of calibration include:

$$\text{confidence-ECE} = \sum_{i=1}^{m} \frac{|B_i|}{n} |y_j(B_i) - \hat{p}_j(B_i)|$$

$$\text{classwise-ECE} = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{|B_{i,j}|}{n} |y_j(B_{i,j}) - \hat{p}_j(B_{i,j})|$$

- Other measures include well-known proper scoring rules, such as Brier score and log-loss
- Every proper loss is minimised by the canonical calibration function

$$\mu(\mathbf{q}) = (P(Y = 1 \mid \hat{\mathbf{p}}(X) = \mathbf{q}), \ldots, P(Y = k \mid \hat{\mathbf{p}}(X) = \mathbf{q}))$$

Temperature scaling:

- Temperature scaling learns a temperature parameter t > 0 which decreases (t > 1) or increases (t < 1) the confidence of a model with softmax output

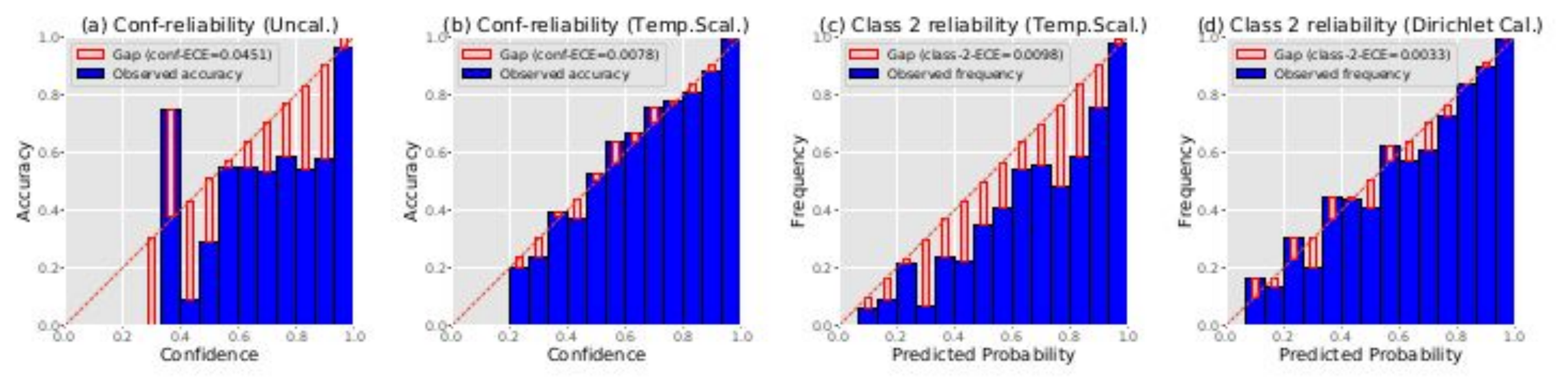


Figure 1: Reliability diagrams of c10_resnet_wide32 on CIFAR-10: (a) confidence-reliability before calibration; (b) confidence-reliability after temperature scaling; (c) classwise-reliability for class 2 after temperature scaling; (d) classwise-reliability for class 2 after Dirichlet calibration.

Dirichlet calibration:

- We consider the distribution of prediction vectors $\hat{\mathbf{p}}(\mathbf{x})$ separately on instances of each class, and assume these are Dirichlet distributions with different parameters:

$$\hat{\mathbf{p}}(X) \mid Y = j \sim \mathrm{Dir}(\alpha^{(j)})$$

generative parametrisation: $\hat{\mu}_{DirGen}(\mathbf{q}; \alpha, \pi) = (\pi_1 f_1(\mathbf{q}), \ldots, \pi_k f_k(\mathbf{q}))/z$

linear parametrisation: $\hat{\mu}_{DirLin}(\mathbf{q}; \mathbf{W}, \mathbf{b}) = \sigma(\mathbf{W} \ln \mathbf{q} + \mathbf{b})$

canonical parametrisation: $\hat{\mu}_{Dir}(\mathbf{q}; \mathbf{A}, \mathbf{c}) = \sigma(\mathbf{A} \ln \frac{\mathbf{q}}{1/k} + \ln \mathbf{c})$

- All parametrisations are equal, i.e. they contain exactly the same calibration maps.

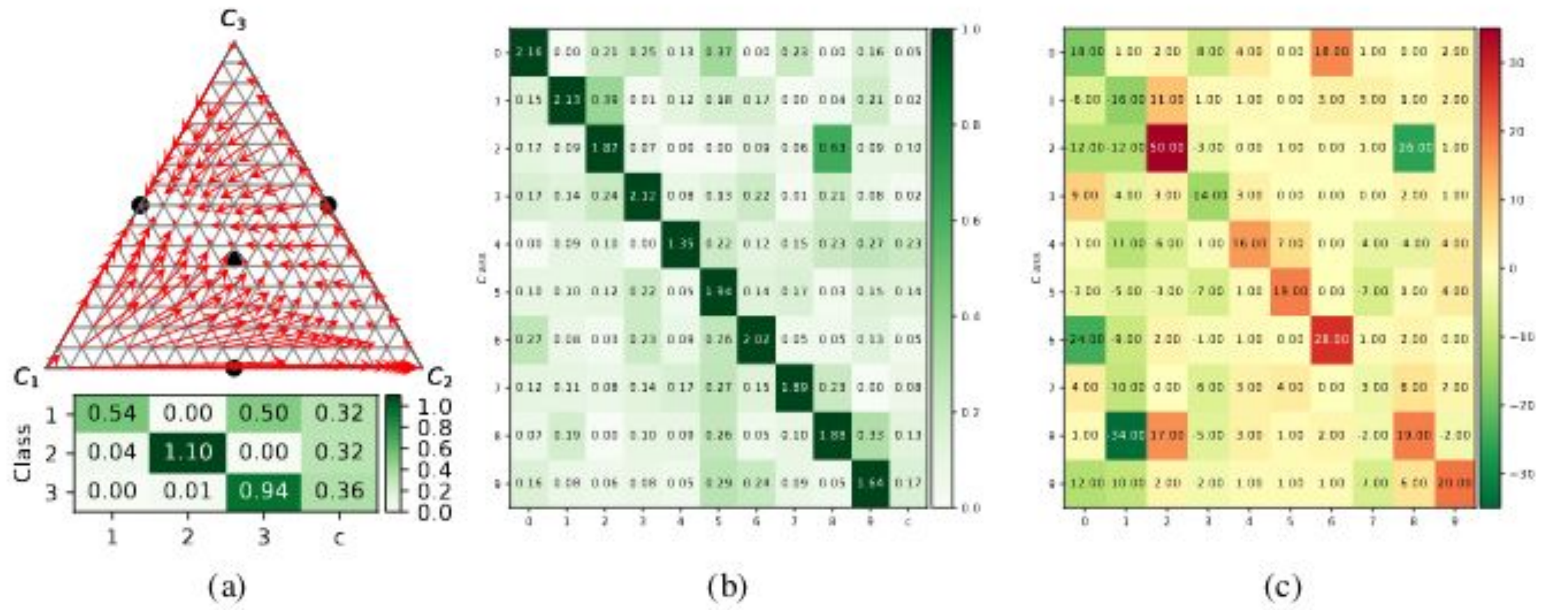
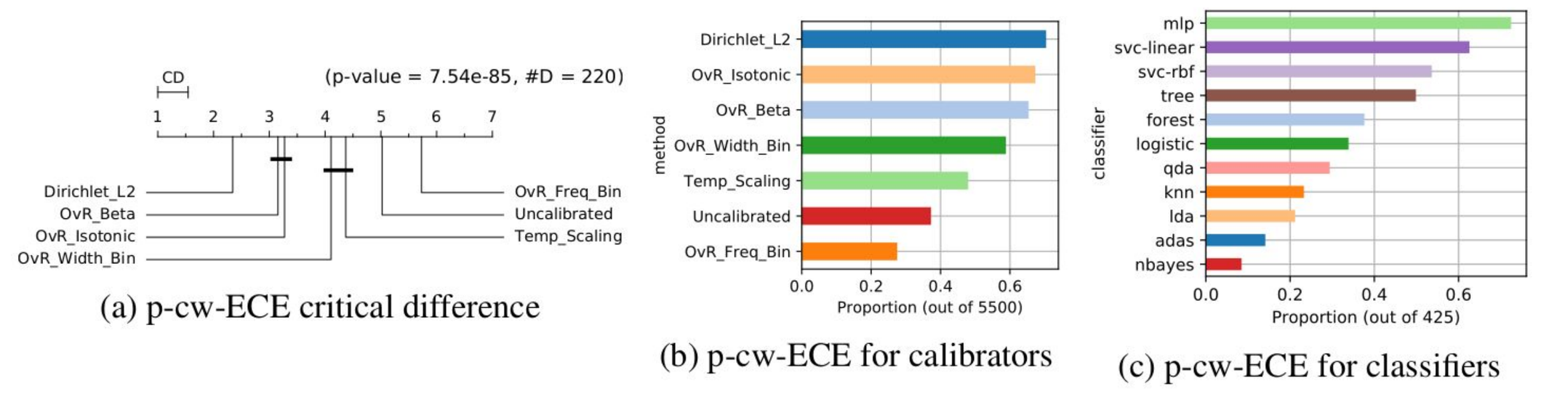


Figure 2: Interpretation of Dirichlet calibration maps: (a) calibration map for MLP on the abalone dataset, 4 interpretation points shown by black dots, and canonical parametrisation as a matrix with $\mathbf{A}, \mathbf{c}$; (b) canonical parametrisation of a map on SVHN_convnet; (c) changes to the confusion matrix after applying this calibration map.

Non-neural experiment:

- 21 datasets and 11 classifiers = 231 settings
- 6 calibration methods:
- Best or tied best performance



Deep Neural experiment:

- 3 datasets CIFAR-10, CIFAR-100, SVHN
- 11 convolutional NNs
- 3 pretrained CNNs
- 5 calibration methods:
- 8 evaluation measures
- 5-fold-crossval.

Table 3: Scores and ranking of calibration methods for **cw-ECE**.

	Uncal	general-purpose calibrators TempS	Dir-L2	Dir-ODIR	calibrators using logits VecS	MS-ODIR
c10_convnet	0.104_6	0.044_4	0.043_2	0.045_5	$\mathbf{0.043_1}$	0.044_3
c10_densenet40	0.114_6	0.040_5	$\mathbf{0.034_1}$	0.037_4	0.036_2	0.037_3
c10_lenet5	0.198_6	0.171_5	$\mathbf{0.052_1}$	0.059_4	0.057_2	0.059_3
c10_resnet110	0.098_6	0.043_5	$\mathbf{0.032_1}$	0.039_4	0.037_3	0.036_2
c10_resnet110_SD	0.086_6	0.031_4	0.031_5	0.029_3	0.027_2	$\mathbf{0.027_1}$
c10_resnet_wide32	0.095_6	0.048_5	0.032_3	0.029_2	0.032_4	$\mathbf{0.029_1}$
c100_convnet	0.424_6	$\mathbf{0.227_1}$	0.402_5	0.240_3	0.241_4	0.240_2
c100_densenet40	0.470_6	0.187_2	0.330_5	$\mathbf{0.186_1}$	0.189_3	0.191_4
c100_lenet5	0.473_6	0.385_5	0.219_4	0.213_2	$\mathbf{0.203_1}$	0.214_3
c100_resnet110	0.416_6	0.201_3	0.359_5	$\mathbf{0.186_1}$	0.194_2	0.203_4
c100_resnet110_SD	0.375_6	0.203_4	0.373_5	0.189_3	$\mathbf{0.170_1}$	0.186_2
c100_resnet_wide32	0.420_6	0.186_4	0.333_5	0.180_2	$\mathbf{0.171_1}$	0.180_3
SVHN_convnet	0.159_6	0.038_4	0.043_5	0.026_2	$\mathbf{0.025_1}$	0.027_3
SVHN_resnet152_SD	0.019_2	$\mathbf{0.018_1}$	0.022_6	0.020_3	0.021_5	0.021_4
Average rank	5.71	3.71	3.79	2.79	2.29	2.71

Table 4: Scores and ranking of calibration methods for **log-loss**.

Uncal	general-purpose calibrators TempS	Dir-L2	Dir-ODIR	calibrators using logits VecS	MS-ODIR
0.391_6	$\mathbf{0.195_1}$	0.197_4	0.195_2	0.197_5	0.196_3
0.428_6	0.225_5	$\mathbf{0.220_1}$	0.224_4	0.223_3	0.222_2
0.823_6	0.800_5	0.744_2	0.744_3	0.747_4	$\mathbf{0.743_1}$
0.358_6	0.209_5	$\mathbf{0.203_1}$	0.205_3	0.206_4	0.204_2
0.303_6	0.178_5	0.177_4	0.176_3	0.175_2	$\mathbf{0.175_1}$
0.382_6	0.191_5	0.185_4	0.182_2	0.183_3	$\mathbf{0.182_1}$
1.641_6	$\mathbf{0.942_1}$	1.189_5	0.961_2	0.964_4	0.961_3
2.017_6	1.057_2	1.253_5	1.059_4	1.058_3	$\mathbf{1.051_1}$
2.784_6	2.650_5	2.595_4	2.490_2	2.516_3	$\mathbf{2.487_1}$
1.694_6	1.092_3	1.212_5	1.096_4	1.089_2	$\mathbf{1.074_1}$
1.353_6	0.942_3	1.198_5	0.945_4	$\mathbf{0.923_1}$	0.927_2
1.802_6	0.945_3	1.087_5	0.953_4	0.937_2	$\mathbf{0.933_1}$
0.205_6	0.151_5	0.142_3	0.138_2	0.144_4	$\mathbf{0.138_1}$
0.085_6	$\mathbf{0.079_1}$	0.085_5	0.080_2	0.081_4	0.081_3
6.0	3.5	3.79	2.93	3.14	1.64

Conclusion:

- Dirichlet calibration: New parametric general-purpose multiclass calibration method
  - Natural extension of two-class Beta calibration
  - Easy to implement with multinomial logistic regression on log-transformed class probabilities
- Advances state-of-the-art on Neural Networks by introducing ODIR regularisation

Future work:

- Which neural architectures and training methods have temperature scaling as a canonical calibration function
- Use other distributions of the exponential family
- Investigate scores coming from mixtures of distributions per class