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

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

NeurIPS 2019

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

Contributions

- New parametric calibration method:

	Logit space	Class probability space
Binary classification	Derived from Gaussian distribution Platt scaling[1]	Derived from Beta distribution Beta calibration[2] (+ constrained variants)
Multi-class classification	Matrix scaling[3] (+ vector scaling, temperature scaling)	Derived from Dirichlet distribution Dirichlet calibration (+ constrained variants)

- New regularization method for matrix scaling (and for Dirichlet calibration):

  ODIR – Off-Diagonal and Intercept Regularisation

- Multi-class classifier evaluation:

  Confidence-calibrated
  Classwise-calibrated
  Multiclass-calibrated

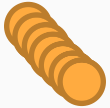
  Confidence-reliability diagram
  Classwise-reliability diagrams

  Confidence-ECE
  Classwise-ECE

Making classifiers more trustworthy

Making classifiers more trustworthy

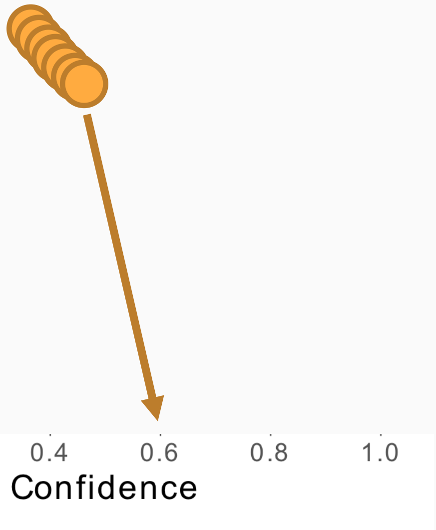
a classifier with 60% accuracy

on a set of instances 

Making classifiers more trustworthy

a classifier with 60% accuracy

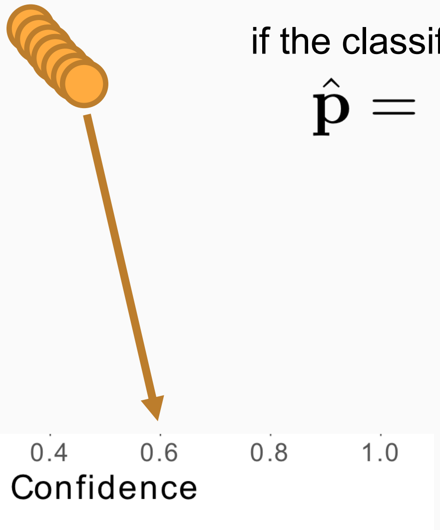
on a set of instances



Making classifiers more trustworthy

a classifier with 60% accuracy

on a set of instances



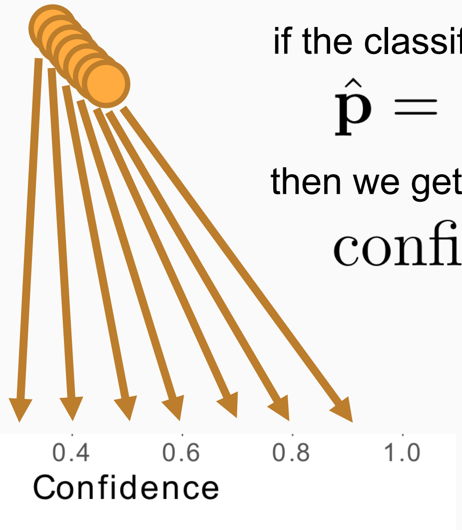
if the classifier reports class probabilities

$$\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$$

Making classifiers more trustworthy

a classifier with 60% accuracy

on a set of instances



if the classifier reports class probabilities

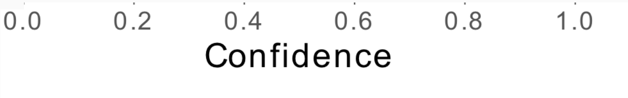
$$\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$$

then we get instance-specific

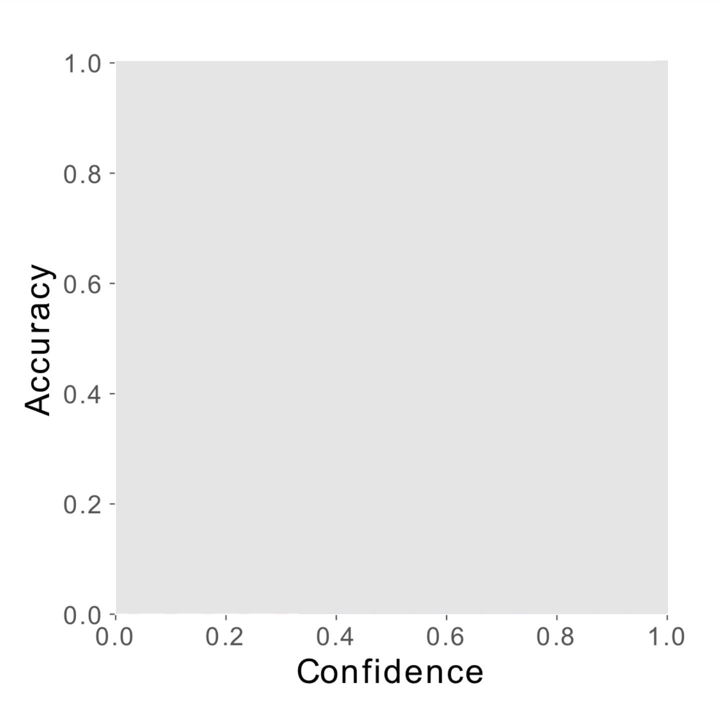
$$\text{confidence} = \max(\hat{\mathbf{p}})$$

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

Trustworthy if confidence-calibrated

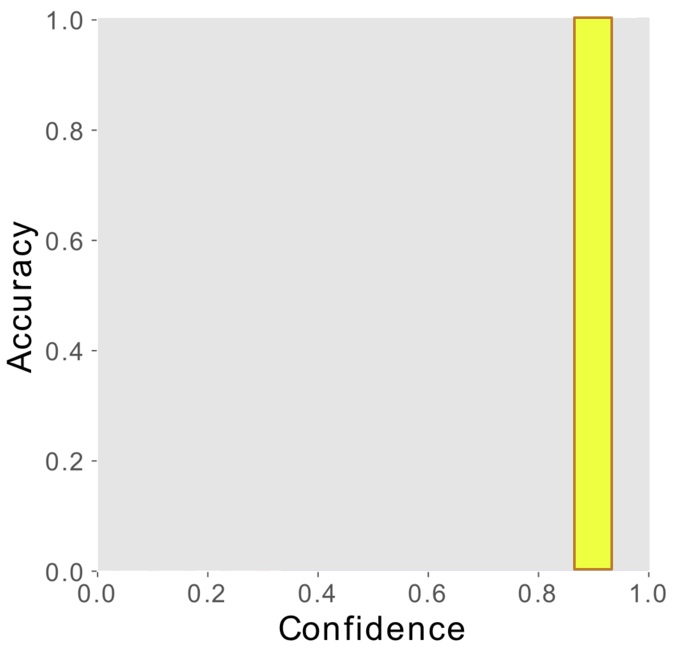


Trustworthy if confidence-calibrated

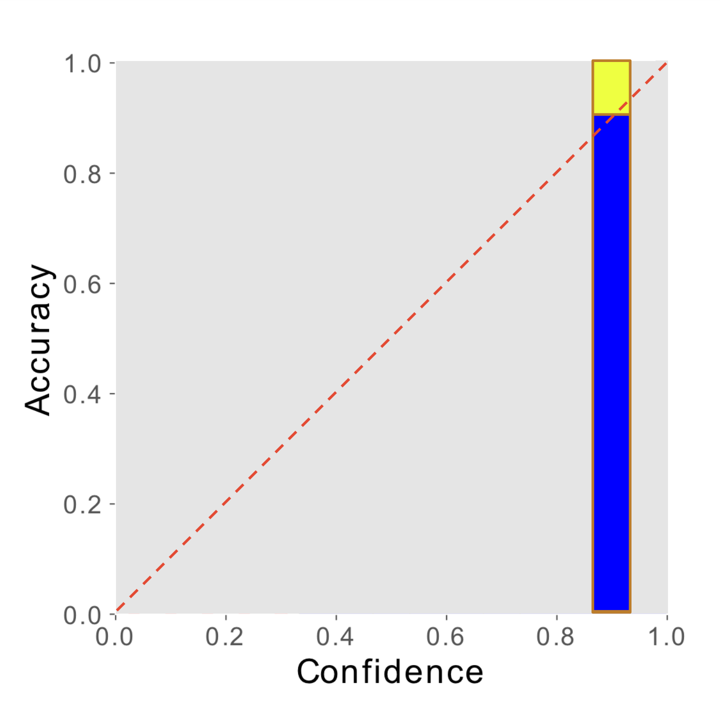


UNIVERSITY OF TARTU
University of BRISTOL
Universidade Federal da Paraíba ESTATÍSTICA
sphere
The Alan Turing Institute

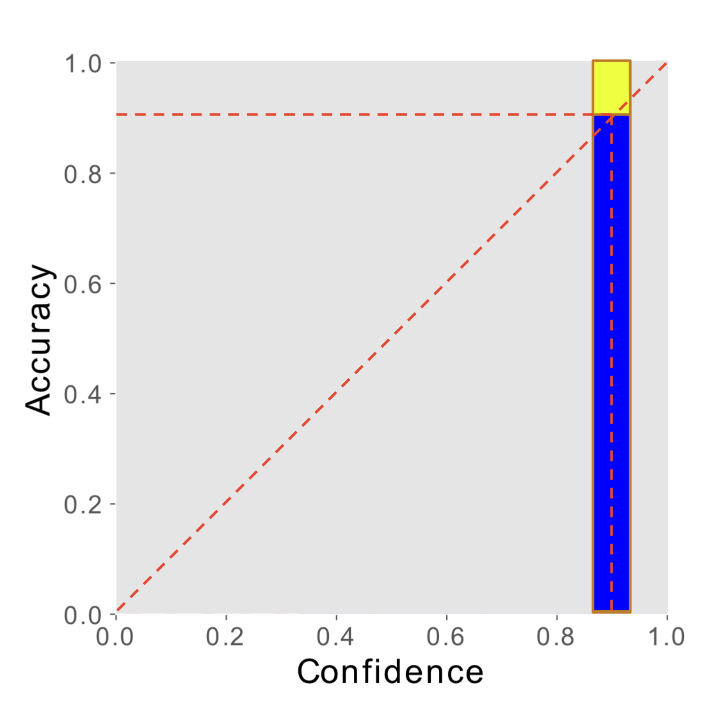
Trustworthy if confidence-calibrated



Trustworthy if confidence-calibrated

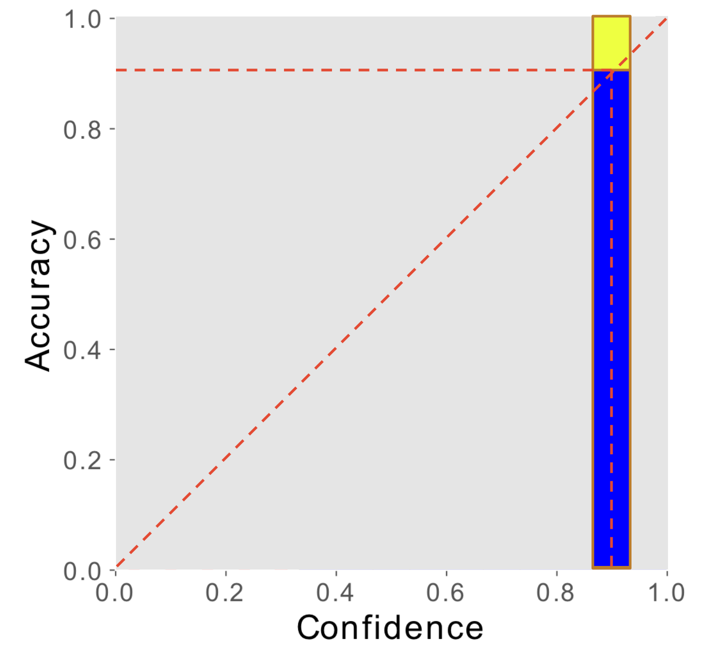


Trustworthy if confidence-calibrated



Trustworthy if confidence-calibrated

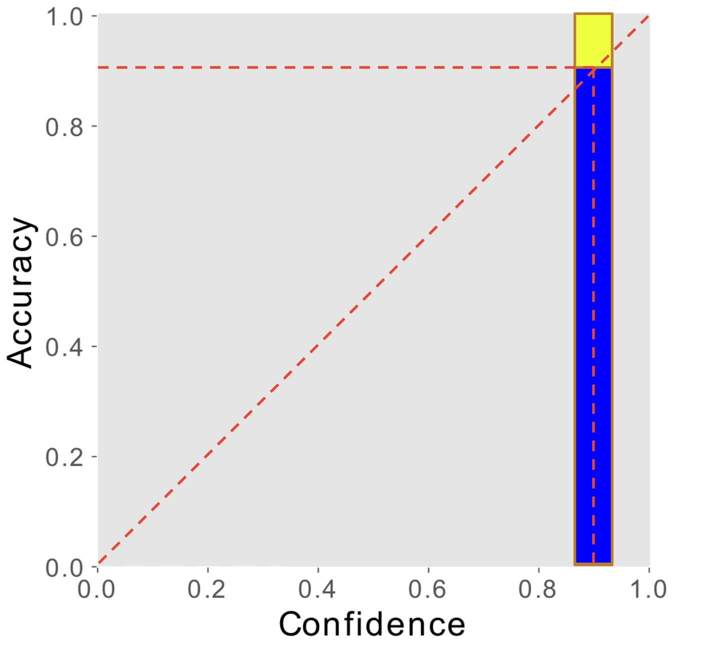
$$\max \hat{\mathbf{p}}(X) = 0.9$$



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

Trustworthy if confidence-calibrated

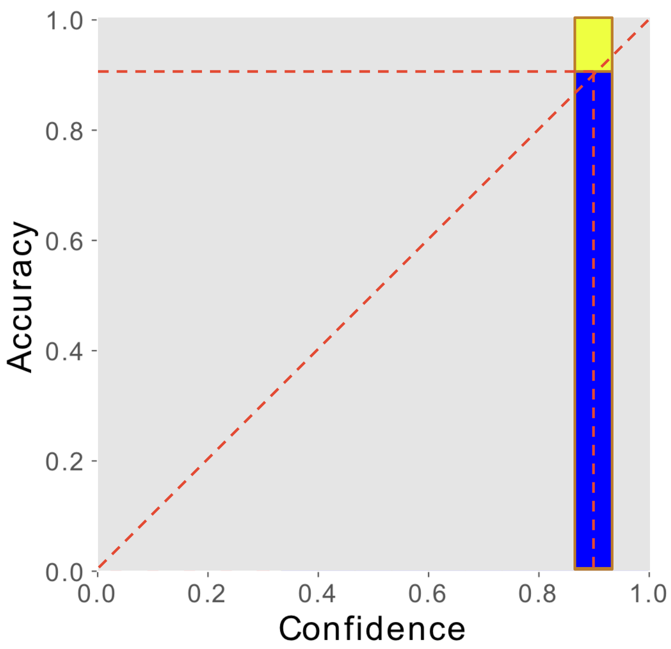
$$Y = \arg\max \hat{\mathbf{p}}(X) \mid \max \hat{\mathbf{p}}(X) = 0.9$$



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

Trustworthy if confidence-calibrated

$$P(\boxed{\boxed{Y = \arg\max \hat{\mathbf{p}}(X)} \mid \max \hat{\mathbf{p}}(X) = 0.9}) = 0.9$$

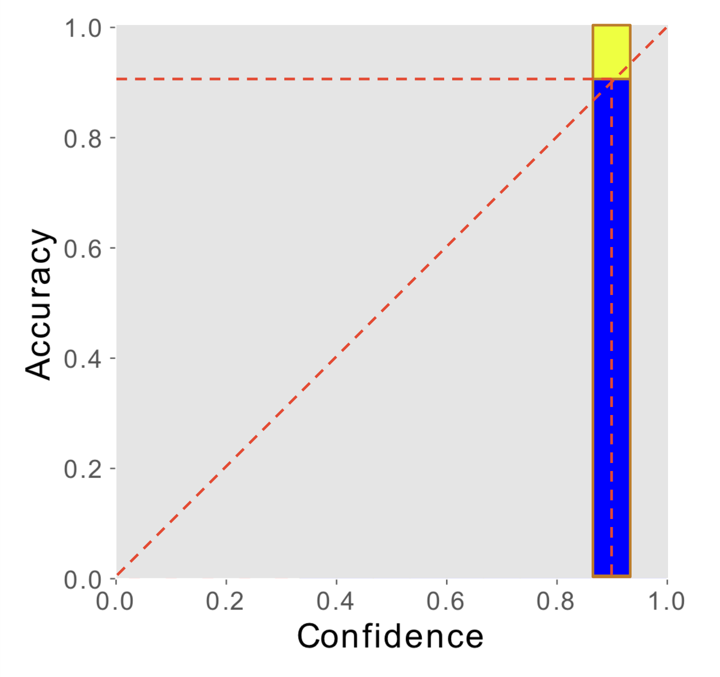


Trustworthy if confidence-calibrated

$$P(Y = \arg\max \hat{\mathbf{p}}(X) \mid \max \hat{\mathbf{p}}(X) = 0.9) = 0.9$$

Confidence-calibrated:

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

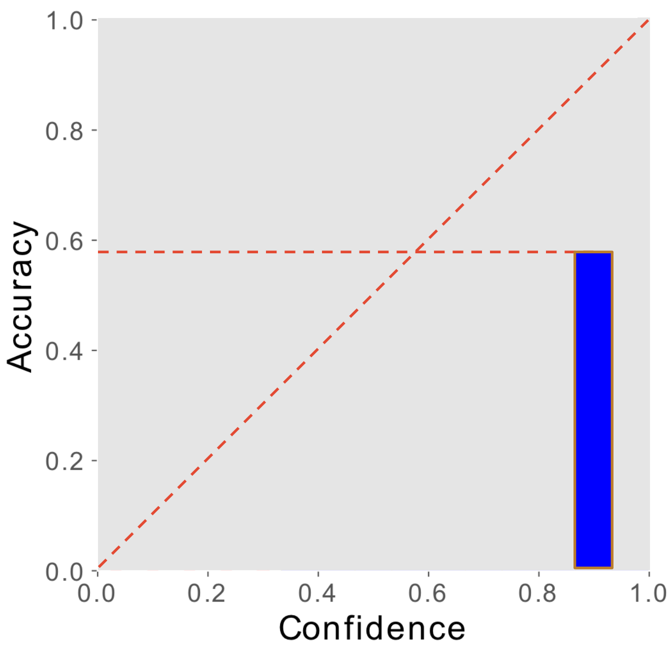


UNIVERSITY OF TARTU · University of BRISTOL · Universidade Federal da Paraíba ESTATÍSTICA · sphere · The Alan Turing Institute

Deep nets are usually over-confident

Confidence-calibrated:

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



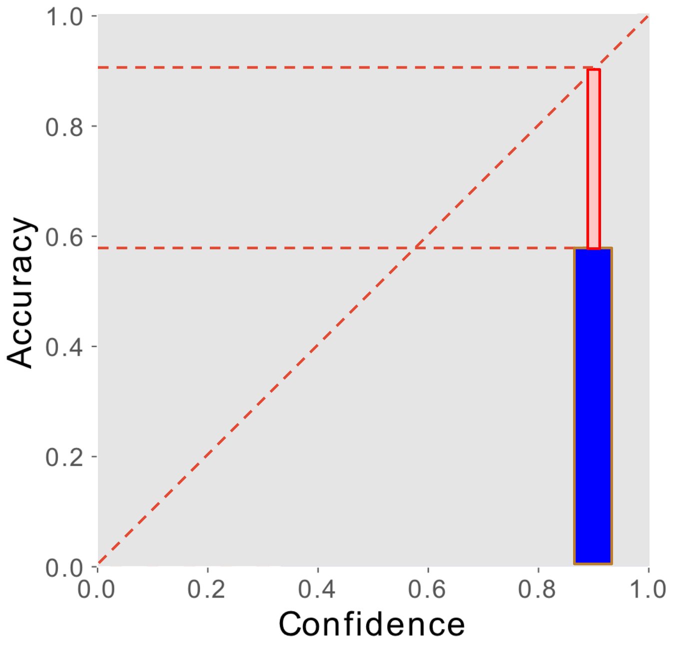
<u>Experimental setup:</u>
CIFAR-10
ResNet Wide 32

<u>Accuracy:</u>
Overall: 94%
At 90% confidence: 58%

Deep nets are usually over-confident

Confidence-calibrated:

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



Experimental setup:
CIFAR-10
ResNet Wide 32

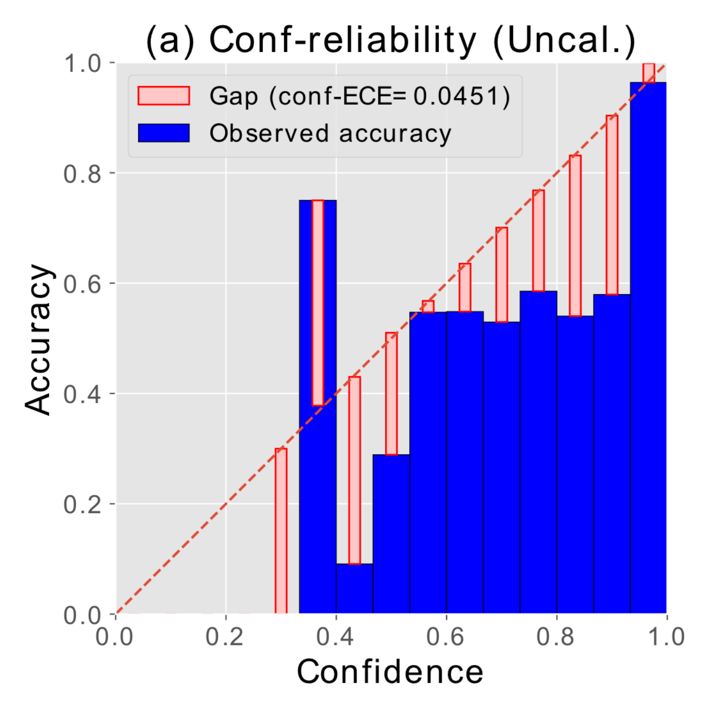
Accuracy:
Overall: 94%
At 90% confidence: 58%

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

Example: uncalibrated predictions

Confidence-calibrated:

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



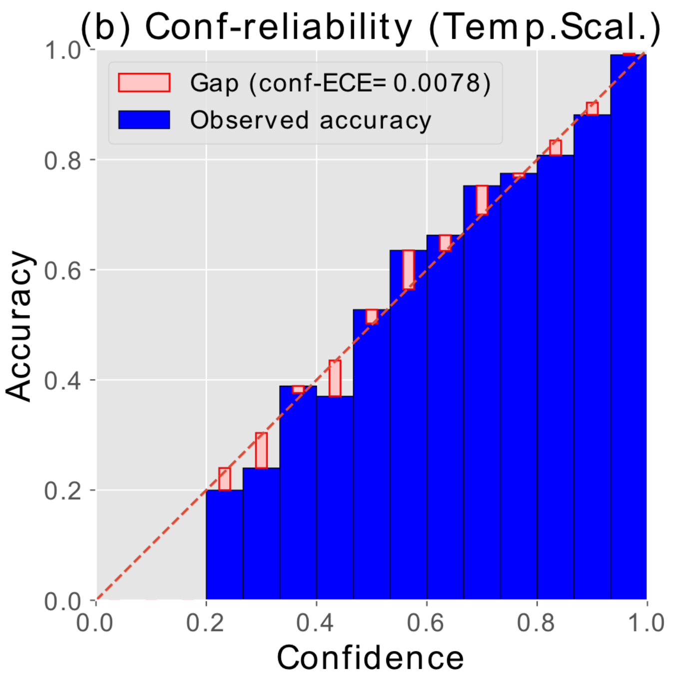
<u>Experimental setup:</u>
CIFAR-10
ResNet Wide 32

<u>Accuracy:</u>
Overall: 94%
At 90% confidence: 58%

Example: after calibration with temperature scaling

Confidence-calibrated:

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



<u>Experimental setup:</u>
CIFAR-10
ResNet Wide 32

<u>Accuracy:</u>
Overall: 94%
At 90% confidence: 58%

<u>Accuracy after Temp.Scal:</u>
Overall: 94%
At 90% confidence: 88%

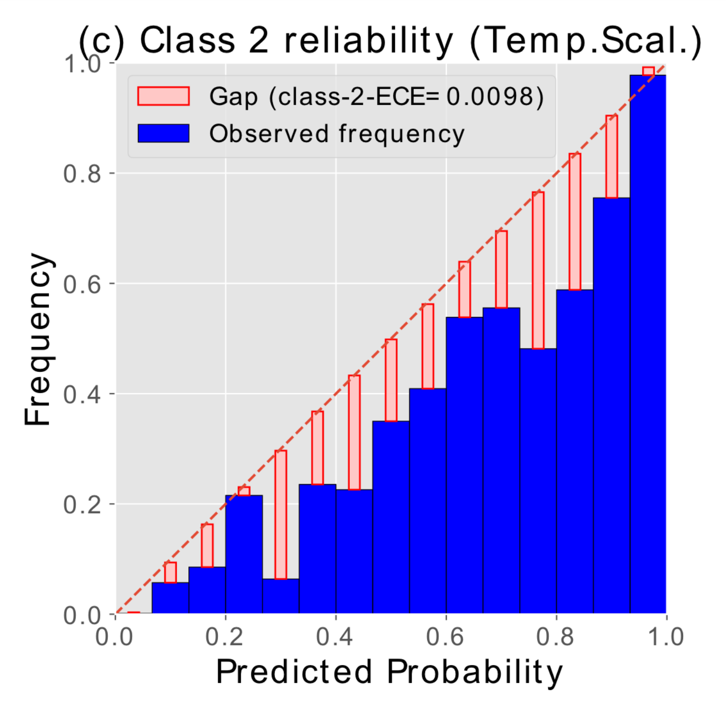
Example: after calibration with temperature scaling

Confidence-calibrated:

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

Classwise-calibrated:

$$P(Y = i \mid \hat{p}_i(X) = c) = c$$



<u>Experimental setup:</u>
CIFAR-10
ResNet Wide 32

<u>Accuracy:</u>
Overall: 94%
At 90% confidence: 58%

<u>Accuracy after Temp.Scal:</u>
Overall: 94%
At 90% confidence: 88%
At 90% class 2 prob: 76%

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

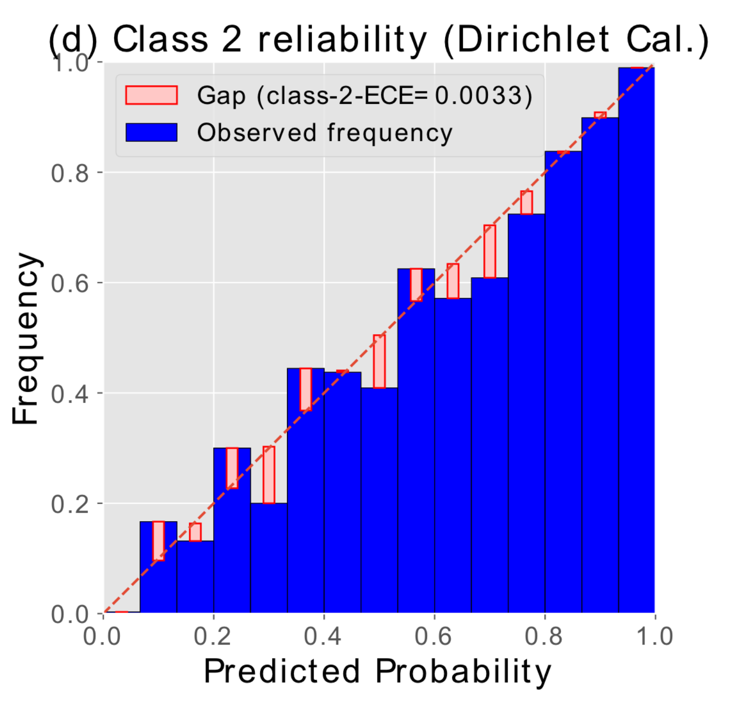
Example: after calibration with Dirichlet calibration

Confidence-calibrated:

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

Classwise-calibrated:

$$P(Y = i \mid \hat{p}_i(X) = c) = c$$



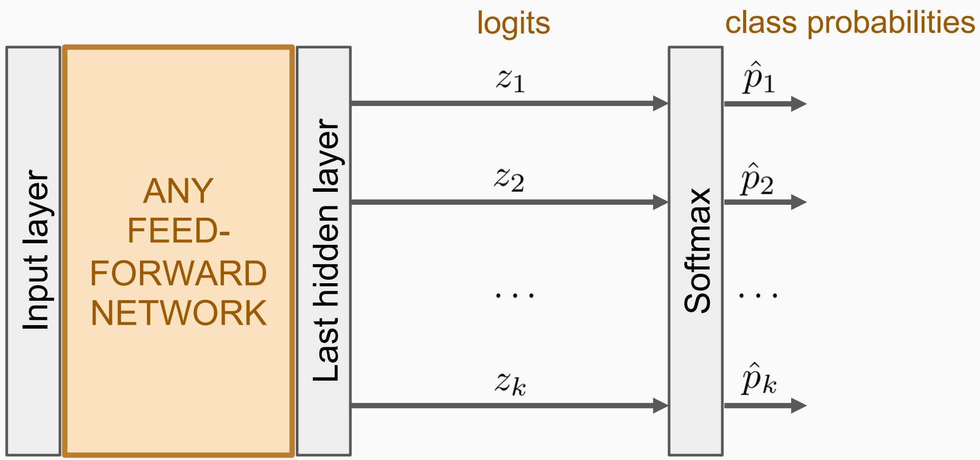
Experimental setup:
CIFAR-10
ResNet Wide 32

Accuracy:
Overall: 94%
At 90% confidence: 58%

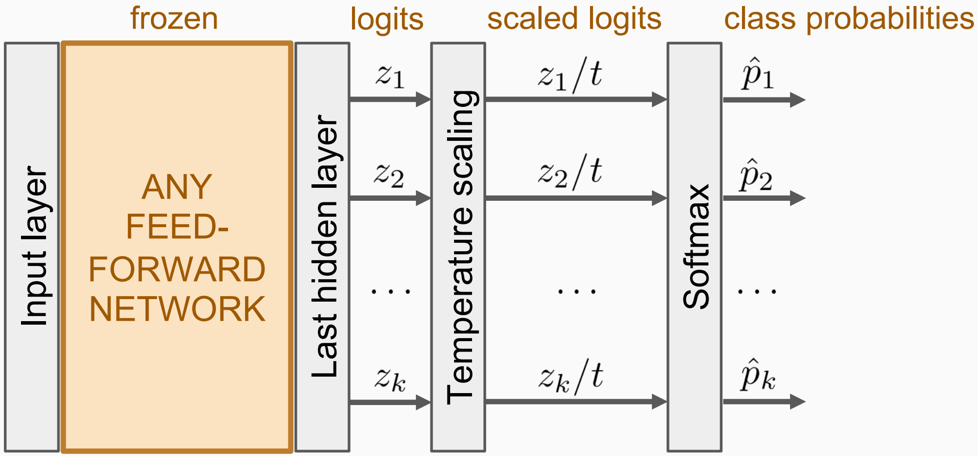
Accuracy after Temp.Scal:
Overall: 94%
At 90% confidence: 88%
At 90% class 2 prob: 76%

Accuracy after Dir.Calib:
At 90% class 2 prob: 90%

How to calibrate a multi-class classifier?



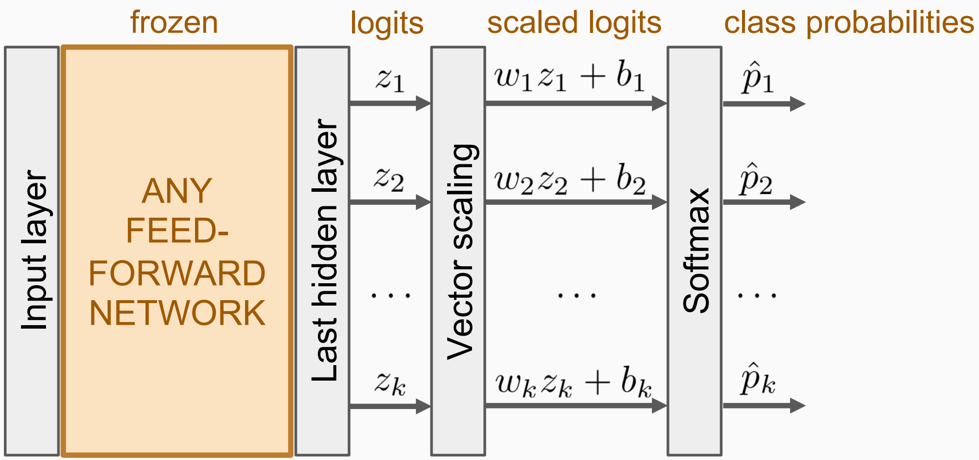
Temperature scaling



Parameters: $t \in \mathbb{R}$

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017

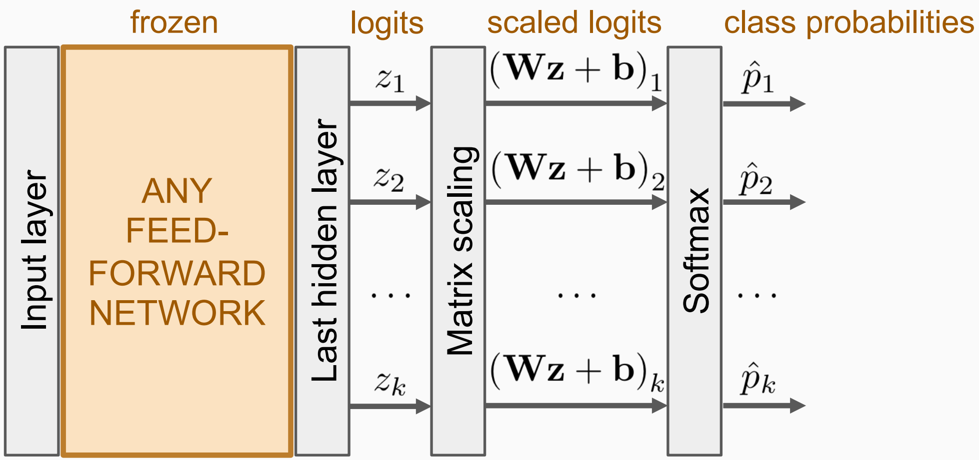
Vector scaling



Parameters: $(\mathbf{w}, \mathbf{b}) \in \mathbb{R}^{k+k}$

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017

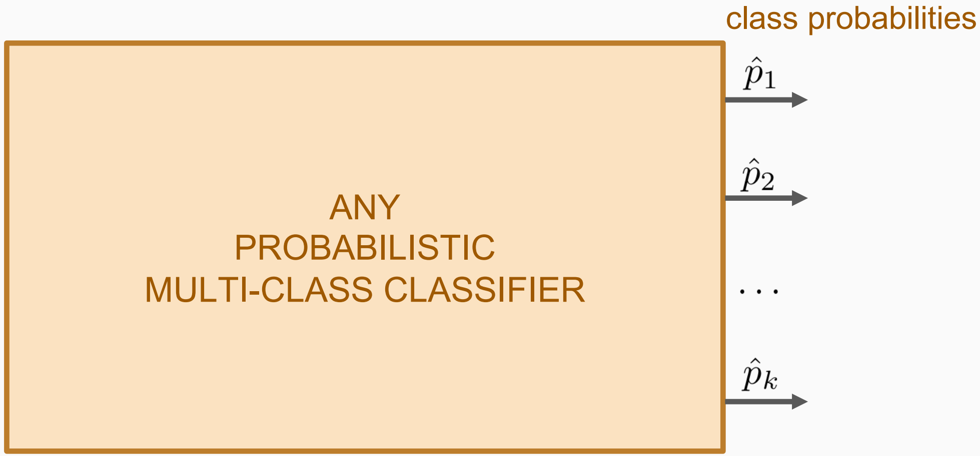
Matrix scaling



Parameters: $(\mathbf{W}, \mathbf{b}) \in \mathbb{R}^{k \times k + k}$

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017

Dirichlet calibration can calibrate any classifiers



Parametric calibration methods

	Logit space	Class probability space
Binary classification	Derived from Gaussian distribution Platt scaling[1]	Derived from Beta distribution Beta calibration[2] (+ constrained variants)
Multi-class classification		

[1] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers, pages 61–74, MIT Press,* 2000.
[2] M. Kull, T. Silva Filho, P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. AISTATS 2017

Parametric calibration methods

	Logit space	Class probability space
Binary classification	Derived from Gaussian distribution Platt scaling[1]	Derived from Beta distribution Beta calibration[2] (+ constrained variants)
Multi-class classification		Derived from Dirichlet distribution Dirichlet calibration (+ constrained variants)

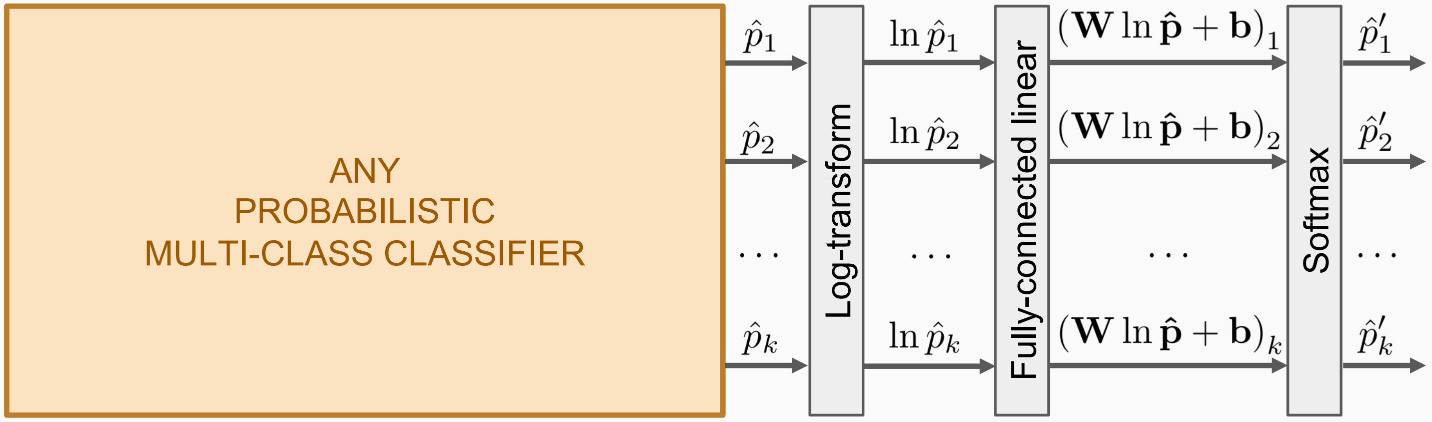
[1] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers, pages 61–74, MIT Press,* 2000.
[2] M. Kull, T. Silva Filho, P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. AISTATS 2017

Parametric calibration methods

	Logit space	Class probability space
Binary classification	Derived from Gaussian distribution Platt scaling[1]	Derived from Beta distribution Beta calibration[2] (+ constrained variants)
Multi-class classification	Matrix scaling[3] (+ vector scaling, temperature scaling)	Derived from Dirichlet distribution Dirichlet calibration (+ constrained variants)

[1] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers, pages 61–74, MIT Press,* 2000.
[2] M. Kull, T. Silva Filho, P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. AISTATS 2017
[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017

Dirichlet calibration



Parameters: $(\mathbf{W}, \mathbf{b}) \in \mathbb{R}^{k \times k + k}$

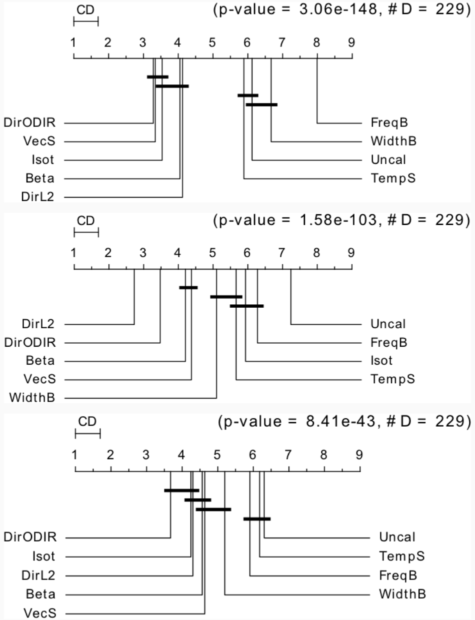
Regularisation:

- L2
- ODIR (Off-Diagonal and Intercept Regularisation)

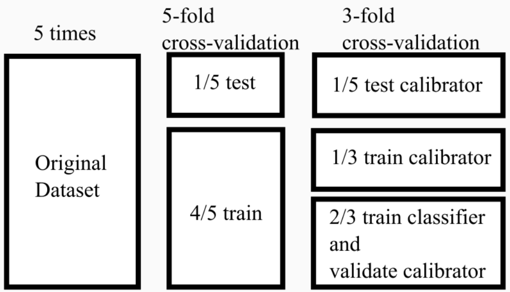
UNIVERSITY OF TARTU University of BRISTOL Universidade Federal da Paraíba ESTATÍSTICA sphere The Alan Turing Institute

Non-neural experiments

- 21 datasets x 11 classifiers = 231 settings
- Average rank
  - Classwise-ECE
  - Log-loss
  - Error rate

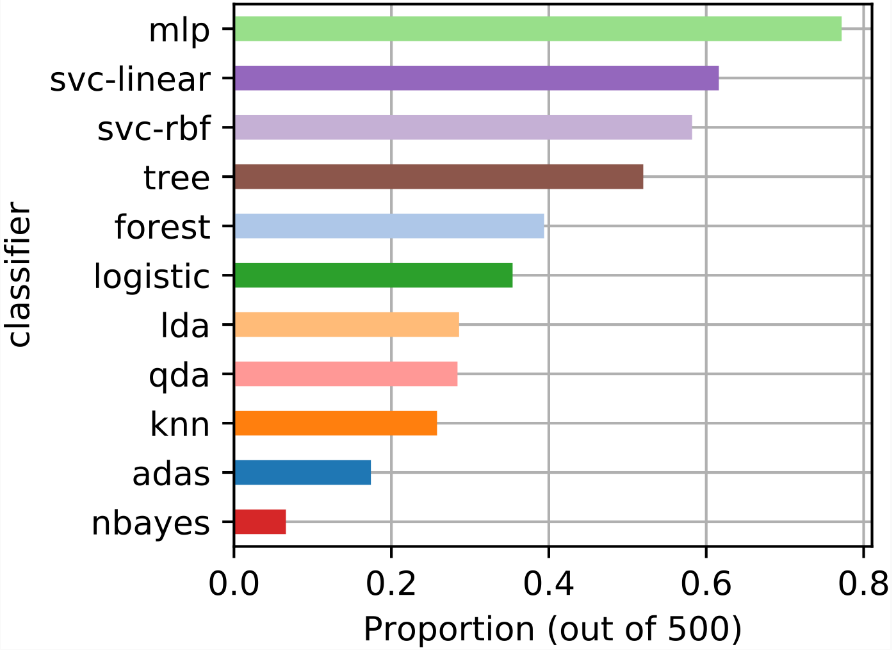


dataset	n_samples	n_features	n_classes
abalone	4177	8	3
balance-scale	625	4	3
car	1728	6	4
cleveland	297	13	5
dermatology	358	34	6
glass	214	9	6
iris	150	4	3
landsat-satellite	6435	36	6
libras-movement	360	90	15
mfeat-karhunen	2000	64	10
mfeat-morphological	2000	6	10
mfeat-zernike	2000	47	10
optdigits	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

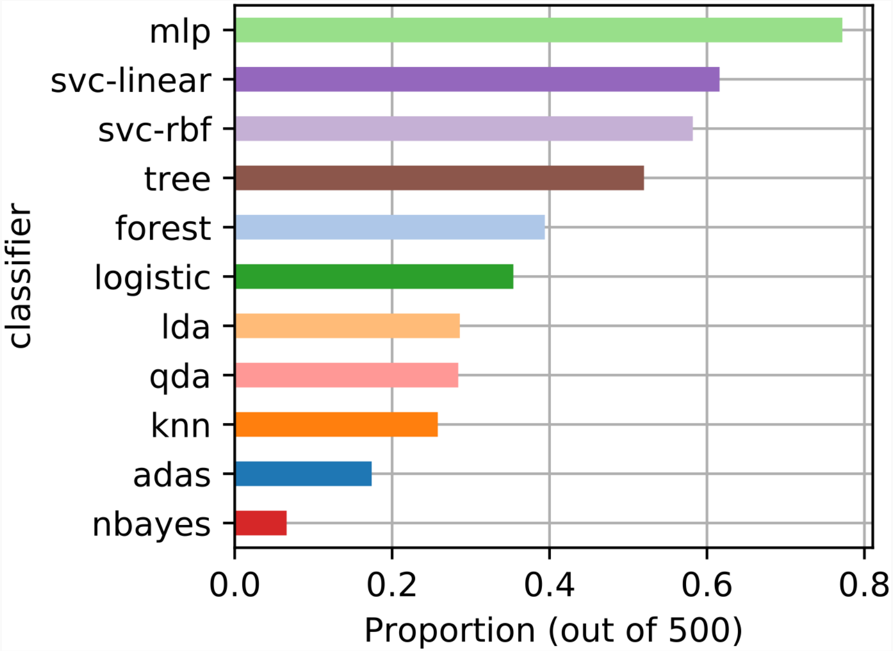


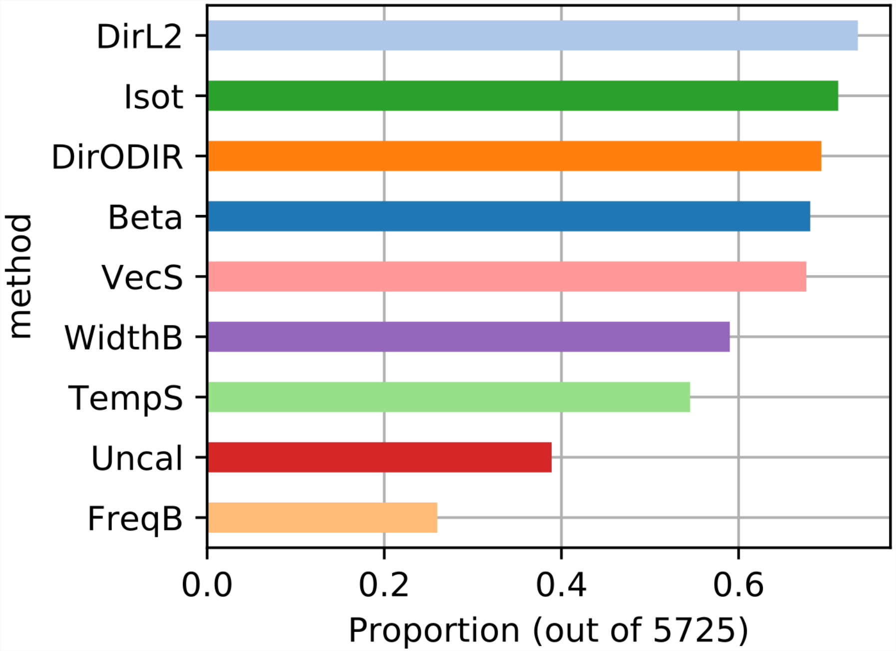
UNIVERSITY OF TARTU
University of BRISTOL
Universidade Federal da Paraíba ESTATÍSTICA
sphere
The Alan Turing Institute

Which classifiers are calibrated?



Which classifiers are calibrated?





UNIVERSITY OF TARTU

University of BRISTOL

Universidade Federal da Paraíba ESTATÍSTICA

sphere

The Alan Turing Institute

Deep Neural Networks Experiments: Settings

- 3 datasets: CIFAR-10, CIFAR-100, SVHN
- 11 convolutional NNs + 3 pretrained

Neural experiments

- Datasets: CIFAR-10, CIFAR-100, SVHN
- 11 CNNs trained as in Guo et al + 3 pretrained

Classwise-ECE

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.104_6	0.044_4	0.043_2	0.045_5	**0.043_1**	0.044_3
c10_densenet40	0.114_6	0.040_5	**0.034_1**	0.037_4	0.036_2	0.037_3
c10_lenet5	0.198_6	0.171_5	**0.052_1**	0.059_4	0.057_2	0.059_3
c10_resnet110	0.098_6	0.043_5	**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	**0.027_1**
c10_resnet_wide32	0.095_6	0.048_5	0.032_3	0.029_2	0.032_4	**0.029_1**
c100_convnet	0.424_6	**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	**0.186_1**	0.189_3	0.191_4
c100_lenet5	0.473_6	0.385_5	0.219_4	0.213_2	**0.203_1**	0.214_3
c100_resnet110	0.416_6	0.201_3	0.359_5	**0.186_1**	0.194_2	0.203_4
c100_resnet110_SD	0.375_6	0.203_4	0.373_5	0.189_3	**0.170_1**	0.186_2
c100_resnet_wide32	0.420_6	0.186_4	0.333_5	0.180_2	**0.171_1**	0.180_3
SVHN_convnet	0.159_6	0.038_4	0.043_5	0.026_2	**0.025_1**	0.027_3
SVHN_resnet152_SD	0.019_2	**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

Log-loss

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.391_6	**0.195_1**	0.197_4	0.195_2	0.197_5	0.196_3
c10_densenet40	0.428_6	0.225_5	**0.220_1**	0.224_4	0.223_3	0.222_2
c10_lenet5	0.823_6	0.800_5	0.744_2	0.744_3	0.747_4	**0.743_1**
c10_resnet110	0.358_6	0.209_5	**0.203_1**	0.205_3	0.206_4	0.204_2
c10_resnet110_SD	0.303_6	0.178_5	0.177_4	0.176_3	0.175_2	**0.175_1**
c10_resnet_wide32	0.382_6	0.191_5	0.185_4	0.182_2	0.183_3	**0.182_1**
c100_convnet	1.641_6	**0.942_1**	1.189_5	0.961_2	0.964_4	0.961_3
c100_densenet40	2.017_6	1.057_2	1.253_5	1.059_4	1.058_3	**1.051_1**
c100_lenet5	2.784_6	2.650_5	2.595_4	2.490_2	2.516_3	**2.487_1**
c100_resnet110	1.694_6	1.092_3	1.212_5	1.096_4	1.089_2	**1.074_1**
c100_resnet110_SD	1.353_6	0.942_3	1.198_5	0.945_4	**0.923_1**	0.927_2
c100_resnet_wide32	1.802_6	0.945_3	1.087_5	0.953_4	0.937_2	**0.933_1**
SVHN_convnet	0.205_6	0.151_5	0.142_3	0.138_2	0.144_4	**0.138_1**
SVHN_resnet152_SD	0.085_6	**0.079_1**	0.085_5	0.080_2	0.081_4	0.081_3
Average rank	6.0	3.5	3.79	2.93	3.14	1.64

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

Conclusion

1. Dirichlet calibration: New parametric general-purpose multiclass calibration method
    a. Natural extension of two-class Beta calibration
    b. Easy to implement with multinomial logistic regression on log-transformed class probabilities
2. Best or tied best performance with 21 datasets x 11 classifiers
3. Advances state-of-the-art on Neural Networks by introducing ODIR regularisation

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

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

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

NeurIPS 2019