Expectation consistency for calibration of neural networks (Supplementary Material)

Lucas Clarté[1] **Bruno Loureiro**[2] **Florent Krzakala**[3] **Lenka Zdeborová**[1]

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL) , Statistical Physics of Computation lab. , Lausanne, Switzerland
[2]Département d'Informatique , École Normale Supérieure - PSL & CNRS , 45 rue d'Ulm, Paris, France
[3]École Polytechnique Fédérale de Lausanne (EPFL), Information, Learning and Physics lab., Lausanne, Switzerland

A DETAILS ON TRAINING PROCEDURE

SVHN For the SVHN dataset Netzer et al. [2011], the Resnet20 model of depth 20 and containing 0.27M parameters was trained for 50 epochs, using SGD with a learning rate $\eta = 0.1$, weight decay $1e-4$ and momentum 0.9. 90% of data points were used for training and the rest was used for validation.

CIFAR10 ResNet models (of depth 20, 56 with Resnet56 having 0.85M parameters) were trained for 50 epochs, using SGD with a learning rate $\eta = 0.1$, weight decay $1e-4$ and momentum 0.9. The DenseNet 121 (containing 7.9 parameters) was trained with the same parameters as the ResNets, except for the learning rate $\eta = 0.01$. As in He et al. [2016], images in the training set were randomly cropped and flipped horizontally.

CIFAR100 On CIFAR100, we used pre-trained models from the Github repository `https://github.com/chenyaofo/pytorch-cifar-models`. These models were trained on the entirety of the training set, so the test set containing 10000 images was split in half into a validation and test set, containing 5000 images each.

A.1 ADDITIONAL PLOTS

In Figure 1, we plot the reliability diagram of Resnet20 and Resnet56 on SVHN and CIFAR10 respectively. We observe that the uncalibrated models are overconfident (as the confidence is higher than the corresponding accuracy), and both TS and EC mitigate this overconfidence.

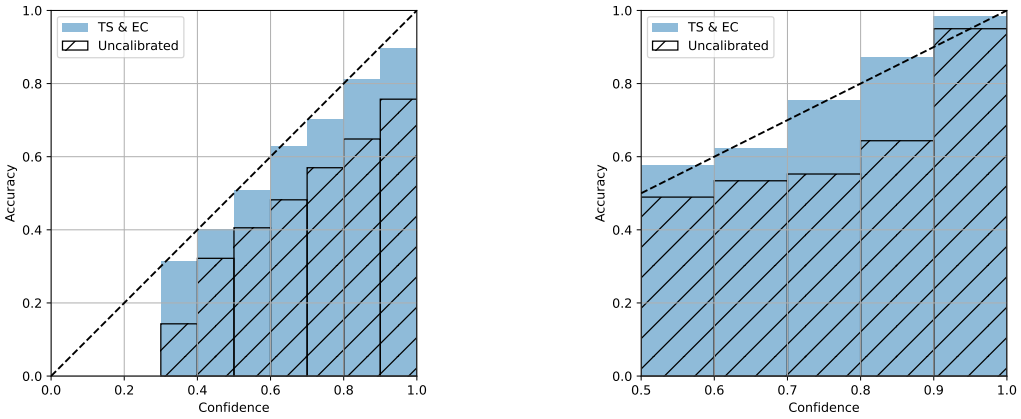

Figure 1: Reliability diagram of Resnet20 on the SVHN dataset (Left) and Resnet56 on the CIFAR10 dataset (Right). Before calibration, both methods are overconfident. TS and EC improve calibration and mitigate overconfidence.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

B STATE EVOLUTION EQUATION

In this section, we focus on the data model introduced in Section 5. Recall that we consider a dataset of n samples $\mathcal{D} = (x^\mu, y^\mu)_{\mu=1}^n$ generated by

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_d/d), \boldsymbol{w}_* \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_d), \mathbb{P}(y = 1|\boldsymbol{w}_*^\top \boldsymbol{x}) = \sigma_\star(\boldsymbol{w}_*^\top \boldsymbol{x}) \tag{1}$$

and we fit the following logistic regression model, with σ the sigmoid function:

$$\hat{f}_{\text{erm}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}) \tag{2}$$

by minimizing the following empirical risk

$$\mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda) = \sum_{\mu=1}^{n} \log \sigma(y^\mu \boldsymbol{w}^\top \boldsymbol{x}) + \lambda/2 \|\boldsymbol{w}\|^2 \tag{3}$$

we thus have $\boldsymbol{w}_{\text{erm}} = \arg\min_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda)$. For a new sample $\boldsymbol{x}$, we are interested in the joint distribution of $f_*(\boldsymbol{x})$ and $\hat{f}_{\text{erm}}(\boldsymbol{x})$. As these two functions only depend on the scalar products $\boldsymbol{w}_*^\top \boldsymbol{x}$, $\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}$ it suffices to compute the joint distribution of these scalar products. By the Gaussianity of $\boldsymbol{x}$, we just need to compute the *overlaps* $m = \boldsymbol{w}_*^\top \boldsymbol{w}_{\text{erm}}$ and $q = \|\boldsymbol{w}_{\text{erm}}\|^2$. In the high-dimensional limit where $n, d \to \infty$ but where we keep the *sampling ratio* constant $n/d = \alpha$, it is possible to compute the value of m and q. The idea is to introduce the distribution

$$\mu_{\beta,\mathcal{D},\lambda}(\boldsymbol{w}) = \frac{1}{\mathcal{Z}_\beta} \exp\left(-\beta \mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda)\right) \tag{4}$$

where $\mathcal{Z}_\beta$ is a normalization constant. In the limit $\beta \to \infty$, $\mu_{\beta,\mathcal{D},\lambda}$ converges to a Dirac distribution peaked at $\boldsymbol{w}_{\text{erm}} = \arg\min \mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda)$. To compute m, q, one needs to compute the expression of $\log \mathcal{Z}_\beta$ and its limit when $\beta \to \infty$. In the high-dimensional regime where both the dimension and number of samples diverge with a fixed ratio, this can be done using the *replica method* from statistical physics Zdeborová and Krzakala [2016]. As these computations are not the focus of the present paper, we refer to Loureiro et al. [2021], Clarté et al. [2022b] for the detailed computations. In the end, if we define

$$\mathcal{Z}_*(y, \omega, v_*) = \int \mathrm{d}z \sigma_*(y \times z) \mathcal{N}(z|\omega, v_*) \tag{5}$$

$$f(y, \omega, v) = \arg\min_z \left[\frac{(z-\omega)^2}{2v} - \log \sigma(z)\right] \tag{6}$$

then m, q are the solution of the following self-consistent equations:

$$\begin{cases} m &= \frac{\hat{m}}{\lambda + \hat{v}} \\ q &= \frac{\hat{q} + \hat{m}^2}{(\lambda+\hat{v})^2} \\ v &= \frac{1}{\lambda+\hat{v}} \end{cases}, \quad \begin{cases} \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \partial_\omega \mathcal{Z}_*(y, m/q\xi, v_*) f(y, \xi, v)\right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \mathcal{Z}_*(y, m/q\xi, v_*) f^2(y, \xi, v)\right] \\ \hat{v} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \mathcal{Z}_*(y, m/q\xi, v_*) \partial_\omega f(y, \xi, v)\right] \end{cases} \tag{7}$$

with $v_* = \rho - m^2/q$.

Calibration in the high-dimensional regime Once we obtained the overlaps m, q, we can derive the expression the calibration Δ_ℓ:

$$\Delta_\ell = \mathbb{E}\left[f_*(\boldsymbol{x})|\hat{f}_{\text{erm}}(\boldsymbol{x})\right] = \mathbb{P}\left[y = 1|\hat{f}_{\text{erm}}(\boldsymbol{x})\right] = \int \mathrm{d}z \sigma_\star(z) \mathcal{N}(z|\frac{m}{q}\hat{f}_{\text{erm}}^{-1}(\boldsymbol{x}), \rho - m^2/q) \tag{8}$$

The second line comes from the fact that the scalar product $\boldsymbol{w}_*^\top \boldsymbol{x}$ conditioned on $\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x} = \sigma^{-1}(\ell)$ follows a Gaussian distribution with mean $m/q\xi$ and variance $\rho - m^2/q$. As a consequence, the expression of ECE is

$$ECE = \mathbb{E}_{\boldsymbol{x}}\left[|\Delta_{\hat{f}_{\text{erm}}(\boldsymbol{x})}|\right] = \mathbb{E}_{\xi=\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}}\left[|\Delta_{\sigma(\xi)}|\right] = \int \mathrm{d}\xi |\Delta_{\sigma(\xi)}| \mathcal{N}(\xi|0, q) \tag{9}$$

Dataset	Model	$\mathcal{E}_g$	T_{TS}	T_{EC}	ECE	ECE_{TS}	ECE_{EC}	BS	BS_{TS}	BS_{EC}
SVHN	Resnet20	12.5	2.69	2.23	8.3	10.7	7.5	21.9	23.4	22.1
CIFAR10	Resnet20	20.9	2.4	2.0	12.8	4.6	4.2	34.2	32.2	32.1
CIFAR10	Resnet56	21	2.58	2.15	13.8	5.4	4.9	35.2	32.9	32.8
CIFAR10	Densenet121	20.4	2.76	2.54	15.8	3.6	5.0	35.9	31.8	31.9
CIFAR100	Resnet20	38.1	2.04	1.70	16.5	9.6	5.9	57.0	54.9	53.9
CIFAR100	Resnet56	34.8	2.27	2.10	21.7	7.6	7.3	56.0	50.6	50.4
CIFAR100	VGG19	35.5	2.6	2.1	28.34	5.2	5.1	61.8	50.1	50.1
CIFAR100	RepVGG-A2	30.5	1.44	1.40	13.7	11.6	11.7	47.2	47.1	47.0

Table 1: Comparison of expected calibration error (ECE) and Brier score (BS) of temperature scaling (TS) and expectation consistency (EC) when part of the validation and test data has been corrupted

C EXPERIMENTS ON CORRUPTED DATASET

We describe below an experiment where EC can significantly improve over TS for real data: we train different architectures on several image classification tasks, as in Figure 1. However, here for the validation and test set some classes are replaced with random labels. For SVHN and CIFAR10, the labels $y = 0$ are replaced by random labels. For CIFAR100, the labels $y = 0, ..., 9$ are replaced by random labels. By doing so, around 10% of validation/test data is corrupted, with a noise that depends on the class. Note that the training data is left unchanged: the goal of this experiment is to model a distribution shift between training and test data, similarly as what is done Hendrycks and Dietterich [2019].

In the table below, we compare the performance (in ECE and Brier score) of EC and TS with these corrupted datasets. We observe that in this setting, EC outperforms TS by a significant margin on several datasets and architectures.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 566–576. PMLR, 18–24 Jul 2021.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. A study of uncertainty quantification in over-parametrized high-dimensional models, 2022a.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification, 2022b.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021.

Lior Frenkel and Jacob Goldberger. Network calibration by class-based temperature scaling. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1486–1490, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks, 2021.

Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: Prediction sets, confidence intervals and calibration. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

Liam Hodgkinson, Chris van der Heide, Fred Roosta, and Michael W. Mahoney. Monotonicity and double descent in uncertainty estimation with gaussian processes, 2022.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, 1999.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in ReLU networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5436–5446. PMLR, 13–18 Jul 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Technical report, University of Toronto*, 2009.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017a. Curran Associates Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151. Curran Associates, Inc., 2021.

Wesley J. Maddox, T. Garipov, Pavel Izmailov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS*, 2019.

Pierre-Alexandre Mattei. A parsimonious tour of bayesian model uncertainty, 2019.

Cyril Méasson, Andrea Montanari, Thomas J Richardson, and Rüdiger Urbanke. The generalized area theorem and some of its consequences. *IEEE Transactions on Information Theory*, 55(11):4793–4821, 2009.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the Calibration of Modern Neural Networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press, 2015.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In *Neural Information Processing Systems*, 2021.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. KDD '02, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.