Certified Calibration: Bounding Worst-Case Calibration under Adversarial Attacks

Cornelius Emde [1] **Francesco Pinto** [2] **Thomas Lukasiewicz** [3 1] **Philip H.S. Torr** [2] **Adel Bibi** [2]

Abstract

Since neural classifiers are known to be sensitive to adversarial perturbations that alter their accuracy, *certification methods* have been developed to provide provable guarantees on the insensitivity of their predictions to such perturbations. However, in safety-critical applications, the frequentist interpretation of the confidence of a classifier (also known as model calibration) can be of utmost importance. This property can be measured via the Brier Score or the Expected Calibration Error. We show that attacks can significantly harm calibration, and thus propose certified calibration providing worst-case bounds on calibration under adversarial perturbations. Specifically, we produce analytic bounds for the Brier score and approximate bounds via the solution of a mixed-integer program on the Expected Calibration Error.

1. Introduction

Deep neural networks have shown remarkable performance in computer vision tasks such as image classification. However, the black-box nature of neural networks and the unreliability of their predictions under several forms of data shift complicates their deployment in safety critical applications (Abdar et al., 2021; Linardatos et al., 2021). It is well known that neural network classifiers are very sensitive to perturbations γ on image x that are small enough to be imperceptible to the human eye, yet $x + \gamma$ yields a different prediction than x (Goodfellow et al., 2015; Szegedy et al., 2016). This problem has been well studied in recent years with the goal of improving adversarial robustness. While a variety of methods has been proposed to improve the empirical robustness of neural networks, *certification methods* have recently gained traction, since they provide provable guarantees on the invariance of predictions under adversarial attacks (Cohen et al., 2019). Following these works, Kumar et al. (2020) have extended the notion of certification from the predicted label to the predictive confidence of the classifier (that, for the scope of this paper, is measured as the maximum softmax output).

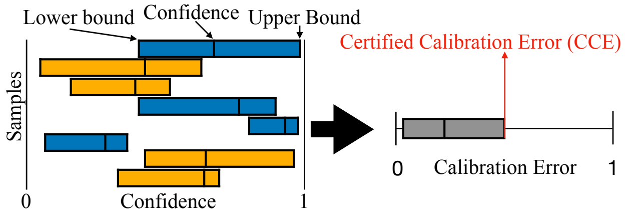


Figure 1: In this diagram, we demonstrate the goal of our work: We translate certificates on the predictive confidences into certificates on the calibration error. Each box on the left represents one prediction from a *certified model*, correct ones in blue, incorrect ones in orange. The certificates on the confidence, are represented by each box's boundaries.

While these certificates provide bounds on the confidence of the certified model as a function of the perturbation applied to each individual input, they do not inform us about the average mismatch between accuracy and confidence. The reliability of a classifier in these regards is generally quantified through both the *Expected Calibration Error* (ECE) (Naeini et al., 2015) and the *Brier Score* (BS) (Brier, 1950; Bröcker, 2009). Both describe the calibration of confidences across a set of predictions. We argue these metrics can play a strategic role in applications that leverage confidence scores in their decision-making process. For instance, one might utilise reliable confidence scores in medical diagnostics to decide whether to trust the machine learning classifier output without further human intervention or whether to defer the decision to a human expert. However, we empirically show it is possible to produce adversaries that severely impact the reliability of confidence scores while leaving the accuracy unchanged, even when the classifier is explicitly trained to be robust to adversarial attacks on the accuracy. This degradation in confidence reliability is clearly signaled by the ECE.

[1] Department of Computer Science, University of Oxford, Oxford, UK. [2] Department of Engineering Science, University of Oxford, Oxford, UK. [3] Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria.. Correspondence to: Cornelius Emde <cornelius.emde@cs.ox.ac.uk>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

For this reason, we propose the *Certified Calibration Error* (CCE) and *Certified Brier Scores* (CBS) to provide guarantees on the calibration of certified models under adversarial attacks. The CCE and CBS are set-level metrics providing worst-case bounds as set-level certificates. While certification of individual predictions enables guarantees on the accuracy of a classifier under adversaries, the additional certificates on the confidences enable worst-case bounds on the *reliability* of confidence scores as represented by the CCE and CBS (see Figure 1).

Our contributions are the following:

- We describe practical motivations to attack calibration metrics. Further, we demonstrate various attacks that severely impact them while remaining unnoticed when solely measuring the robust model accuracy.
- We introduce *certified calibration* quantified through the *certified Brier score* (CBS) and the *certified calibration error* (CCE). For the former, we present a closed-form bound as certificate. For the latter, we interpret the CCE as the solution of a *mixed-integer non-linear program* and propose an effective numerical optimisation framework to estimate it, i.e. the *approximate certified calibration error* (ACCE).

2. Confidence Calibration

We first introduce notation and formalize calibration, before explaining how attacks on confidence scores and reliability metrics may be beneficial to an attacker and showing such attacks are realisable (2.2).

2.1. Quantifying the Confidence-Accuracy Mismatch

Let $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ be a dataset of size N with $\mathbf{x}_n \in \mathbb{R}^D$, $y_n \in \mathcal{Y} = \{1, 2, ..., K\}$, and $f : \mathbb{R}^D \to \Delta_K$ be a neural network, where Δ_K is a probability simplex over K classes. We denote the kth component of f as f_k. We further, define $F : \mathbb{R}^D \to \mathcal{Y}$ to be *hard classifier* predicting a class label, usually obtained by $F(x) := \arg\max_{k \in \mathcal{Y}} f_k(x)$. This prediction is done with the *confidence* provided by the *soft classifier* $z : \mathbb{R}^D \to [0, 1]$, obtained through $\max_{k \in \mathcal{Y}} f_k(x)$. With slight abuse of notation, we refer to functions and their outputs simultaneously, i.e. $z \in [0, 1]$ is a output of $z(x)$.

Expected Calibration Error

For classification tasks, calibration describes a match between the model's confidence and its empirical performance (DeGroot & Fienberg, 1983; Naeini et al., 2015). A well-calibrated model predicts with confidence z when the fraction of correct predictions is exactly z, i.e. $\mathbb{P}(F = y|Z = z) = z$. This enables us to interpret z as a probability in the frequentist sense. We formally state:

$$\mathrm{ECE} = \mathbb{E}_Z \left[|\mathbb{P}(F = y|Z = z) - z| \right], \qquad (1)$$

which is the expected difference between confidence and accuracy over the distribution of Z.

Several estimators for (1) have been proposed. A typical approach is to discretize the empirical distribution over Z through binning (Guo et al., 2017). For each bin B_m, the average confidence is compared with the accuracy:

$$\hat{\mathrm{ECE}} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{n \in B_m} c_n - \frac{1}{|B_m|} \sum_{n \in B_m} z_n \right| \qquad (2)$$

where $c_n = \mathbb{I}\{F_n = y_n\}$ and $\mathbb{I}\{x\}$ is the indicator function, that is 1 if x, and 0 otherwise. Multiple variants of the calibration error and its estimators exist (Kumar et al., 2019). As commonly done in literature, in this paper we will focus on top label calibration that ignores the calibration of confidences of lower ranked predictions. When using an equal-width binning scheme for the estimator in (2) we will refer to it as ECE and when using an equal-count binning scheme we will refer to it as $AdaECE$ (Nguyen & O'Connor, 2015; Nixon et al., 2019).

Brier Score Accuracy and calibration represent different concepts and one may not infer model accuracy from calibration unambiguously, or vice versa (see Appendix A.1). The two concepts are unified under *proper scoring rules*, such as the *Brier Score* (BS) (Brier, 1950), which is commonly used in the calibration literature. It has been shown that these metrics can be decomposed into a calibration and a refinement term (Murphy, 1972; 1973; Bröcker, 2009). An optimal score can only be achieved by predicting accurately and with appropriate confidence. The Brier is mathematically defined as the mean squared error between the confidence vector f and a one-hot encoded label vector. Here, we will focus on the *top-label Brier score* $\mathrm{TLBS} = N^{-1} \|\mathbf{c} - \mathbf{z}\|_2^2$, which is the mean squared error between the confidences $\mathbf{z} \in [0, 1]^N$ and the correctness of each prediction $\mathbf{c} \in \{0, 1\}^N$.

2.2. Calibration under Attack

Motivation Practitioners often rely on confidence scores for decision making processes with the assumption that confidence scores can be interpreted as frequentist probabilities. Therefore, we argue it is important to certify calibration by providing worst-case bounds to its variation under adversarial perturbations. Indeed, machine learning systems deployed in safety critical applications are monitored regularly for their predictive performance and for their calibration. When abnormalities in the system are detected, the model might be pulled from deployment for further investigation. Often, this results in major operational cost, as the model might be replaced with less precise but more robust models, human labour, or pulled without replacement while being vital element of the revenue stream. To this end, an attacker might coordinate a *denial-of-service* attack.

Table 1: AdaECE↓ (%) when performing our (η, ω)-ACE attacks compared to the case in which the attack is not performed for PreAct-ResNet18 and ResNet50 on CIFAR-10 and ImageNet for both Empirical Risk Minimization (ERM) and Adversarial Risk Minimization (ARM)

			CIFAR 10	ImageNet	
ERM	Unattacked:		3.06	3.70	
	η	ω	ϵ = 8/255	ϵ = 2/255	ϵ = 3/255
	-1	y	2.49	1.06	10.62
	$+1$	y	18.79	47.23	46.92
	-1	$\hat{y}$	5.19	23.72	23.73
	$+1$	$\hat{y}$	11.87	25.17	27.84
ARM	Unattacked:		20.69	9.03	
	η	ω	ϵ = 8/255	ϵ = 2/255	ϵ = 3/255
	-1	y	21.84	7.36	8.51
	$+1$	y	23.51	11.62	12.21
	-1	$\hat{y}$	11.92	0.91	3.42
	$+1$	$\hat{y}$	25.59	13.54	14.28

We remark that, despite the extensiveness of the countermeasures to adversarial attacks literature, none of the existing techniques addresses this important scenario. For instance, defenses against adversarial attacks on labels and the notion of certified accuracy have been introduced to improve label robustness. While these bounds also indirectly bound confidence scores and calibration (e.g. because in order to prevent the prediction from changing the confidence cannot fall below a specific threshold), these are not tight enough to protect against adversaries explicitly targeting calibration.

Beyond label attacks, attacks directly targeting confidence scores have been developed. Galil & El-Yaniv (2021) extensively demonstrate the existence of attacks on confidence scores and discuss how detrimental they can be. While their discussion revolves around the importance of preserving confidences on a *per-sample* level, our analysis revolves around bounding *set-level* metrics of confidence, as these metrics are also relevant in the decision making process.

Feasibility of Attacking Reliability Galil & El-Yaniv (2021) show it is possible to produce attacks that leave the accuracy unchanged while degrading the Brier Score. In a similar fashion, expanding on their notion of Attacks on Confidence Estimation (ACE)[1], we introduce a family of parameterised ACE attacks we call (η, ω)-ACE attacks. These attacks solve the following objective:

$$\max_{\|\gamma\|_p \le \epsilon, F(x+\gamma)=F(x)} \eta \mathcal{L}_{CE}(f(x+\gamma), \omega), \tag{3}$$

where $\eta \in \{1, -1\}$ and $\omega \in \{y, \hat{y}\}$ and $\hat{y}$ is the classifier's prediction. Solving the problem above means either maximising ($\eta = 1$) or minimising ($\eta = -1$) the cross-entropy loss $\mathcal{L}_{CE}$ of the output computed on the perturbed input (i.e. to increase or decrease the confidence of the classifier) with respect to either the true label or the prediction. This is done without altering the label (i.e. $F(x+\gamma) = F(x)$). In Table 1, we show that all four possible configurations of our (η, ω)-ACE can be effective at significantly altering the ECE on the validation set of CIFAR-10 (Krizhevsky, 2009) and ImageNet-1K (Deng et al., 2009) using a PreActResNet18 (He et al., 2016b) and ResNet50 (He et al., 2016a), respectively. We show the attacks can be effective both when the model is trained with standard Expected Risk Minimization (ERM) and Adversarial Risk Minimization (ARM). As it can be seen, on an ERM trained ResNet50 on ImageNet, an $(1, y)$-ACE with radius $\epsilon = 2/555$ can increase the ECE from 3.70 to 47.23. While performing adversarial training can partly alleviate this issue, an $(1, \hat{y})$-ACE attack can still increase its ECE from 9.03 to 13.54. This clearly indicates it is possible to significantly manipulate the calibration of a model while preserving its accuracy.

3. Certifying Calibration

3.1. Prerequisites for Certified Calibration

For our definition of certified calibration, we require classifiers with predictions that are certifiably robust against label flips and require bounds on confidence scores. A state-of-the-art method to obtain such certificates on large models and datasets is *Gaussian Smoothing* which constructs a smooth classifier by adding Gaussian perturbations $\delta \sim N(0, \sigma \mathbf{I}_D)$ to its input and aggregating the predictions (Cohen et al., 2019). Let $\bar{F} : \mathbb{R}^D \to \mathcal{Y}$ be the *smoothed hard classifier*. For a radius R, we can state that, for all perturbations $\|\gamma\|_2 \le R$, the prediction will remain constant, i.e. $\bar{F}(\mathbf{x} + \gamma) = \bar{F}(\mathbf{x})$. The certifiable radius R can be computed in closed form. If the evidence to certify with $R > 0$ is insufficient, the classifier abstains. The certificate on the prediction has been extended by Kumar et al. (2020) showing a certificate on the confidence. Let $\bar{z} : \mathbb{R}^D \to [0, 1]$ be the *smoothed soft classifier* indicating the confidence of prediction $\bar{F}$. For $\bar{z}$ it is true, that $\forall \|\gamma\|_2 \le \epsilon$:

$$\Phi_\sigma(\Phi_\sigma^{-1}(\underline{p_A}) - \epsilon) \le \bar{z}(x+\gamma) \le \Phi_\sigma(\Phi_\sigma^{-1}(\overline{p_A}) + \epsilon), \tag{4}$$

where $\underline{p_A}$ and $\overline{p_A}$ are bounds on $\bar{z}(\mathbf{x})$ and Φ is the Gaussian CDF. Below, assume the bound provided in (4).

We observe that *certified accuracy* counts abstained predictions as incorrect. As we define *certified calibration* on top of models operating in a certified regime for predictions and confidences, we suggest computing it only on non-rejected samples because the confidence on abstained predictions is not well-defined and we wish to inform the reliability of confidences, when the model does not abstain.

[1] Remark: our implementation deviates from Galil & El-Yaniv (2021), e.g. we use PGD instead of FGSM to find adversaries.

3.2. Certifying Brier Score

While the Brier score generally is a unified assessment of model performance and calibration, its interpretation is slightly different for certified models. Changes in the Brier score as function of confidence perturbations solely reflect changes in the calibration as the accuracy is constant.

Consider a dataset of N samples on which we want to compute calibration. Let $\mathbf{l}, \mathbf{u} \in \mathbb{R}^N$ be the lower and upper bound on the top confidence $\mathbf{z} \in \mathbb{R}^N$, respectively, as provided by the certificate on the confidences. We can state the following upper bound on the Brier score.

Theorem 3.1. *Let* $\mathbf{l}$*,* $\mathbf{u}$ *be the certificates on* $\mathbf{z}$ *and* $\mathbf{z}$ *be the output certified classifier as defined above. Further, let* $\mathbf{c} \in \mathbb{R}^N$ *be the indicator that predictions are correct. The upper bound on the Brier score is given by:*

$$\max_{\mathbf{l} \leq \mathbf{z} \leq \mathbf{u}} \mathit{TLBS}(\mathbf{z}, \mathbf{c}) = \frac{1}{N} \|\mathbf{c} - \mathbf{l}\mathbb{I}\{\mathbf{c} = 1\} + \mathbf{u}\mathbb{I}\{\mathbf{c} = 0\}\|_2^2. \tag{5}$$

Proof. See Appendix C. □

This bound is tight and relies on the fact that shifting the confidences leaves $\mathbf{c}$ unchanged while remaining inside the certified regime. Therefore, an adversary cannot flip the prediction to increase the *confidence gap*, the sample-level distance between correctness and confidence. The Brier score is maximised when the confidence gap is large; the opposite of what a good classifier intuitively should do. Plugging the certificates on the confidence (such as (4)) into the equation above as $\mathbf{l}$ and $\mathbf{u}$, provides a certificate as function of the perturbation on the input data.

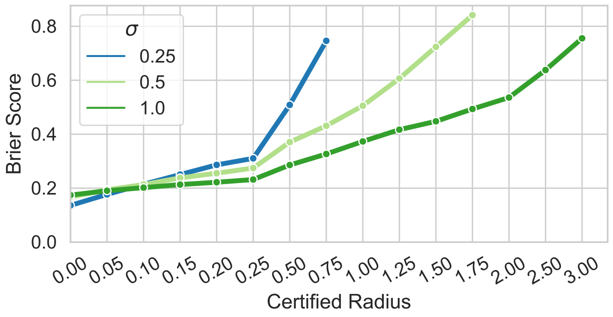


Figure 2: Certified Brier scores on ImageNet. For small radii, small smoothing σ outperforms larger ones, but as radii increase, large σ outperform smaller σ.

3.3. Certifying Calibration Error

While both, the Brier Score and the Expected Calibration Error, capture some notion of calibration, the confidence scores bounding the Brier score (5) do not bound the ECE, which can be increased even further. Therefore, we will introduce the notion of the *certified calibration error* and provide a method to approximate it.

Definition 3.2. The *certified calibration error* (*CCE*) on dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ at radius ϵ is defined as the maximum ECE, that can be observed as a result of perturbations on the inputs within an ℓ_2 ball of radius ϵ. Let, γ_n be the perturbation on input $\mathbf{x}_n$. The certified calibration error is:

$$\mathit{CCE} = \max_{\forall n: \|\gamma_n\|_2 \leq \epsilon} E\hat{C}E\left([\bar{z}(\mathbf{x}_n + \gamma_n)]_{n=1}^N, \mathbf{y}\right). \tag{6}$$

This aims to find the largest estimated calibration error on a dataset, if every sample is perturbed by at most ϵ. Finding such a bound is not trivial, as (2) is neither convex nor differentiable. Therefore, we propose a numerical method to estimate (6) and provide an empirical, approximate certificate, the *approximate certified calibration error* (ACCE).

3.4. CCE as Mixed-Integer Program

We show in this section that we may solve (6) by interpreting the calibration error as the objective of a mixed-integer problem. Since the ECE estimator (2) uses bins to estimate average confidence and accuracy, one can reformulate the problem as a bin assignment problem, where N confidence scores are assigned to M bins. We seek to jointly solve the assignment problem and find the values of z_n maximising the calibration estimator across bins.

More precisely, the estimator bins each confidence score z_n into M bins. While perturbing the data, this bin assignment might change: $\bar{z}(\mathbf{x})$ might belong to bin m, but $\bar{z}(\mathbf{x} + \gamma)$ to $m' \neq m$. While the assignment is naturally determined by the confidence score, it is key to our reformulation to split these into separate variables. The motivation is that very small changes in z_n might lead to a shift in bin assignment and thus contribute differently to the calibration error. We define the integer-valued assignment $a_{n,m}$ of confidence z_n to bin m and accordingly define $\mathbf{a} = [a_{1,1}, ..., a_{1,M}, a_{2,1}, ..., a_{N,M}]^\top \in \{0, 1\}^{NM}$, where N is the number of samples and M the number of bins. The confidence score maximising the calibration error might be different across bins, i.e. it is possible that the worst-case for bin m is different than for bin $m' \neq m$: $z^*_{n,m} \neq z^*_{n,m'}$. Therefore, we model the confidence independently for each bin introducing bin-specific confidences $z_{n,m}$ and with it $\mathbf{z} = [z_{1,1}, ..., z_{1,M}, z_{2,1}, ..., z_{N,M}]^\top \in [0, 1]^{NM}$. Further, let c_n be the indicator whether prediction n is correct, i.e. $c_n = \mathbb{I}\{\bar{F}(\mathbf{x}_n) = y_n\}$ and let $e_{n,m} = c_n - z_{n,m}$, the sample confidence gap. Note that c_n is independent of the bin assignment as a result of certification (i.e. while the confidence may shift, the prediction will remain unchanged). Analog to $\mathbf{a}$ and $\mathbf{z}$, we define $\mathbf{e} = [e_{1,1}, ..., e_{1,M}, e_{2,1}, ..., e_{N,M}]^\top \in \mathbb{R}^{NM}$. Now let $\mathbf{B}$ be a stack of N identity matrices of size M, i.e., $\mathbf{B} = [\mathbf{I}_M, ..., \mathbf{I}_M]^\top$. We define $\mathbf{E}(\mathbf{z}) = \left(\mathbf{e}\mathbf{1}_M^\top\right) \odot \mathbf{B} \in \mathbb{R}^{NM \times M}$ where $\odot$ is the Hadamard product. We can now rewrite the calibration error.

Theorem 3.3. *Let* $\mathbf{a}$ *and* $\mathbf{E}$ *be the output of a certified classifier as defined above. The calibration error estimator in (2) can be expressed as:*

$$E\hat{C}E = \frac{1}{N}\|\mathbf{E}(\mathbf{z})^\top \mathbf{a}\|_1 \tag{7}$$

where $\mathbf{a}$ *and* $\mathbf{E}$ *are subject to the* unique assignment, confidence *and* valid assignment *constraints below. Thus, maximising* (7) *is equivalent to solving* (6).

Proof. See Appendix D.1. □

Unique Assignment Constraint The assignment variable $\mathbf{a}$ has to be constrained such that each data point is assigned to exactly one bin, i.e. $\sum_m \mathbf{a}_{n,m} = 1$. To this end, we define $\mathbf{C} = \mathbf{I}_N \otimes \mathbf{1}_M \in \mathbb{R}^{NM \times N}$, where $\otimes$ is the Kronecker product. $\mathbf{C}$ sums up all assignments per data point and hence our constraint becomes $\mathbf{C}^\top \mathbf{a} = \mathbf{1}_N$.

Confidence Constraint The problem in (6) is defined over perturbations $\|\gamma_n\|_2 \le \epsilon$ for each input $\mathbf{x}_n$, which can be propagated into bounds on confidences z_n using certificates on the confidences, such as (4) following Kumar et al. (2020). Let $l_n^z \le z_{n,m} \le u_n^z$ be the lower and upper bound on the confidence as provided by the certificate. In addition, any confidence assigned to bin m has to adhere to the boundaries of this bin, i.e. $l_m^B \le z_{n,m} \le u_m^B$. We can combine these two conditions to unify the bounds: $\max(l_n^z, l_m^B) = l_{n,m} \le z_{n,m} \le u_{n,m} = \min(u_n^z, u_m^B)$. With this, we define $\mathbf{l} = [l_{1,1}, ..., l_{1,M}, l_{2,1}, ..., l_{N,M}]^\top$ and $\mathbf{u} = [u_{1,1}, ..., u_{1,M}, u_{2,1}, ..., u_{N,M}]^\top$ and state the full constraint: $\mathbf{l} \le \mathbf{z} \le \mathbf{u}$.[2]

It is possible that the bounds due to the binning and the certificate on z_n do not intersect, i.e. for some n, m it might be that $[l_n^z, u_n^z) \cap [l_m^B, u_m^B) = \emptyset$. This is expected for narrow certificates on z or a large number of bins. For these instances, we will set $\mathbf{z} \leftarrow 0$ and define $\mathbf{l}'$ and $\mathbf{u}'$ to be $\mathbf{l}$ and $\mathbf{u}$ with the same elements set to 0. We define $S_z = \{\mathbf{z} : \forall n, m : l'_{n,m} \le z_{n,m} \le u'_{n,m}\}$ to be the feasible set on $\mathbf{z}$.

Valid Assignment Constraint Above we have identified that some confidences can never be assigned to some bins. For those instances, we constrain $a_{n,m} = 0$. Let $k_{n,m} = \mathbb{I}\{l_{n,m} \ge u_{n,m}\}$ be the indicator that bin m is inaccessible to data point n. We define matrix $\mathbf{K}$ to be a $NM \times N$ matrix summing all inaccessible bin assignments. Formally, letting $\mathbf{k} = [k_{1,1}, ..., k_{1,M}, k_{2,1}, ..., k_{N,M}]^\top$, and $\mathbf{K} = \mathbf{k}\mathbf{1}_N^\top \odot (\mathbf{I}_N \otimes \mathbf{1}_M)$, the constraint is: $\mathbf{K}^\top \mathbf{a} = \mathbf{0}_N$.

Formal Program Statement We summarize the constraints above and state the program in its canonical form for clarity. The mixed-integer program over $(\mathbf{a}, \mathbf{z})$ is given by:

$$\text{maximise} \quad \frac{1}{N}\|\mathbf{E}(\mathbf{z})^\top \mathbf{a}\|_1 \tag{8}$$

subject to:

$$\mathbf{a} \in \{0,1\}^{NM}, \quad \mathbf{C}^\top \mathbf{a} = \mathbf{1}_N, \quad \mathbf{K}^\top \mathbf{a} = \mathbf{0}_N, \quad \mathbf{z} \in S_z$$

While this expression comes a the cost of increasing the number of variables compared to (6), this provides us with a useful framework to run a numerical solver.

3.5. ADMM Solver

We propose to use the ADMM algorithm (Boyd et al., 2011) to solve minimise (8). While ADMM has proofs for convergence on convex problems, it is well known that it enjoys good convergence properties even on non-convex problems. ADMM minimises the augmented Lagrangian of the constrained problem by sequentially solving sub-problems alternating between minimizing the primal variables and maximizing the dual variables.

We follow Wu & Ghanem (2019) and Bibi et al. (2023) to relax the binary-constraints on $\mathbf{a}$. Note that, $\mathbf{a} \in \{0,1\}^{NM} \Leftrightarrow \mathbf{a} \in S_b \cap S_2$, where S_b is the unit hypercube and S_2 is the ℓ_2-sphere, both centered at $\frac{\mathbf{1}}{\mathbf{2}}$. We introduce auxiliary variables $\mathbf{q}_1 \in S_b$ and $\mathbf{q}_2 \in S_2$ and add constraints $\mathbf{a} = \mathbf{q}_1$ and $\mathbf{a} = \mathbf{q}_2$. Similarly, we replace the constraint $\mathbf{z} \in S_z$ by enforcing it on $\mathbf{g}$ and adding $\mathbf{z} = \mathbf{g}$. Updates on the primal variables $(\mathbf{a}, \mathbf{z}, \mathbf{q}_1, \mathbf{q}_2, \mathbf{g})$ are performed via gradient descent (see Appendix D).

In our experiments, ADMM always converges in under 3000 steps, and runs in a matter of minutes. At convergence, we observe that all constraints are sufficiently met. Exploring various factors on the convergence, we conducted a large hyperparameter search (see Appendix E.2.1) and base our final hyperparameter on these results. We recommend running ADMM on two different initialisation of $\mathbf{z}$, the observed adversary-free confidences, as well as those achieving the CBS and subsequently picking the larger ACCE. We find that the maximum ACCE is achieved well before ADMM has converged. Therefore, we recommend projecting a copy of $\mathbf{a}$ and $\mathbf{z}$ into their feasible set after each step and calculate the ACCE.

4. Experiments

4.1. Experimental setup and details

We follow the work on certifying confidences (Kumar et al., 2020) in our experimental setup. We use a ResNet-110

[2] In Section 3.2 on the CBS, $\mathbf{l}, \mathbf{u} \in [0,1]^N$ are the immediate bounds on $\mathbf{z} \in [0,1]^N$ provided by the certificate on the confidences. Here, $\mathbf{l}, \mathbf{u} \in [0,1]^{NM}$ provide bounds on the expanded $\mathbf{z} \in \mathbb{R}^{NM}$ as defined in this section. Be aware that these definitions are overloaded and these bounds now combine binning and confidence certificates.

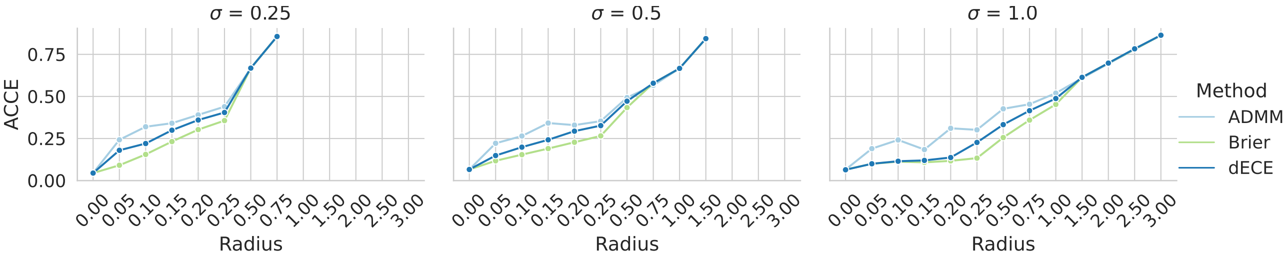


Figure 3: The ACCE returned by ADMM, dECE and the Brier confidences are shown here for ImageNet. ADMM is the most effective method as it uniformly yields the largest bounds.

model for CIFAR-10 experiments and a ResNet-50 for ImageNet trained by Cohen et al. (2019). We rely on the certificates provided by Cohen et al. (2019) and (Kumar et al., 2020) as they provide us with closed form certificates. As commonly done, we sample 500 images from the test set of ImageNet to conduct our experiments. For CIFAR-10, we use the entire test-set to compute the CBS, but sample 2000 certified images for the ACCE. Gaussian Smoothing is performed on $100,000$ samples and we certify at $\alpha = .001$. For our work, we use the certifiable radius provided by Cohen et al. (2019) and rely on the confidence bounds in (4). As mentioned, we only certify calibration at ϵ when the prediction can be certified at ϵ. We compute certified metrics on $\epsilon \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1.00, 1.50, 2, 2.5\}$ for CIFAR-10 and additionally 3.0 for ImageNet.

4.2. Certified Brier Score

We use Theorem 3.1 to obtain the certificate on the Brier Score in closed form. We report the CBSs for smooth models with different smoothing σs and for a range of certified radii. The results for CIFAR-10 are shown in Figure 8 in Appendix E.1 and in Figure 2 for ImageNet. For both datasets, we observe that the CBSs *increases* with larger certified radii. Models with small σ suffer from about a 100% increase in Brier score at $\epsilon = 0.25$, while stronger smoothed models only increase by $< 50\%$. As strongly smoothed models yield tighter certificates on confidences (see Figure 7 in Appendix B), we find that those models are more robust for larger radii at the cost of performance on smaller radii.

4.3. Certified Calibration Error

We compare the ACCE from ADMM with two other techniques and find that ADMM outperforms both. First, we obtain the confidences bounding the Brier score (the "Brier confidences") and compute the resulting ECE as baseline. Second, we utilise the *differentiable calibration error* (dECE) (Bohdal et al., 2021) and perform gradient ascent to maximise it (see Appendix F.1). We compare these methods for a range of certified radii and different smoothing σ. For CIFAR-10 the results are shown in Figure 12 in Appendix E.2.2 and for ImageNet in Figure 3. As with for the CBS, large certified radii are associated with worse calibration. We find that ADMM uniformly yields *higher* ACCE than the dECE with differences up to approximately 0.2, which is a strong qualitative difference in calibration. While all method yield very similar bounds on large radii, we may conclude the ADMM is by far more effective than the other methods in approximating the CCE.

4.4. Discussion

Across all experiments, we observe that even small perturbations on input data can harm calibration significantly and thus the certificates take on large values. This finding is in line with Galil & El-Yaniv (2021), who report similar results on non-certified models. Strongly smoothed models have worse calibration for small radii and better calibration for large radii suggesting a *calibration-robustness-trade-off*. Interestingly, we observe that all three methods to find the ACCE are approximately equal at large radii. We believe that this is not an insufficiency of the ACCE to approximate the CCE, but rather believe that the CCE and CBS solutions converge. For accuracies 0 and 1, and unbounded confidences, it is trivial to see that the CCE is achieved by the same confidences as the CBS. We conjecture this is the case for other accuracies as well.

5. Related Work

Only few papers discuss the confidence scores on certified models. Jeong et al. (2021) propose a variant of *mixup* (Zhang et al., 2018) for certified models to reduce overconfidence in runner-up classes with the goal of increasing the certified radius, but do not examine the confidence as uncertainty estimator. A wider body of literature has been published relating adversarial robustness to calibration on non-certified models. Grabinski et al. (2022) show that robust models are better calibrated while other work shows that poorly calibrated data points are easier to attack (Qin et al., 2021). This is used by the latter to improve calibration through adversarial training. Stutz et al. (2020), utilise confidence scores and calibration techniques to improve adversarial robustness. Few works investigate the calibration of uncertainty calibration under adversarial attack (Sensoy et al., 2018; Tomani & Buettner, 2021; Kopetzki et al., 2021), however their work is not very applicable as these use more elaborate uncertainty scores than softmax output.

While some works provide bounds on calibration in various contexts, none of them are applicable to our setup (Kumar et al., 2019; Qiao & Valiant, 2021; Wenger et al., 2020).

Acknowledgements

We thank the reviewers for the efforts and helpful comments. Further, we thank Mohsen Pourpouneh for his support.

This work is supported by a UKRI grant Turing AI Fellowship (EP/W002981/1). C. Emde is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1) and Cancer Research UK (CRUK). A. Bibi has received an Amazon Research Award. F. Pinto's PhD is funded by the European Space Agency (ESA). T. Lukasiewicz is supported by the AXA Research Fund. We also thank the Royal Academy of Engineering and FiveAI.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535.

Bibi, A., Alqahtani, A., and Ghanem, B. Constrained Clustering: General Pairwise and Cardinality Constraints. *IEEE Access*, 11:5824–5836, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3236608.

Bohdal, O., Yang, Y., and Hospedales, T. Meta-Calibration: Meta-Learning of Model Calibration Using Differentiable Expected Calibration Error. In *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*, 2021. _eprint: 2106.09613.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. ISSN 1935-8237. doi: 10.1561/2200000016.

Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1, January 1950. doi: 10.1175/1520-0493(1950)078\textless0001:VOFEIT\textgreater2.0.CO;2.

Bröcker, J. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009. ISSN 0035-9009. doi: 10.1002/qj.456. Place: Chichester, UK.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, June 2019.

DeGroot, M. H. and Fienberg, S. E. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983. ISSN 00390526, 14679884. Publisher: [Royal Statistical Society, Wiley].

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *cvpr*, 2009.

Galil, I. and El-Yaniv, R. Disrupting Deep Uncertainty Estimation Without Harming Accuracy. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21285–21296. Curran Associates, Inc., 2021.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

Grabinski, J., Gavrikov, P., Keuper, J., and Keuper, M. Robust Models are less Over-Confident. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143, June 2017. ISSN 9781510855144.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*, 2016b.

Jeong, J., Park, S., Kim, M., Lee, H.-C., Kim, D.-G., and Shin, J. SmoothMix: Training Confidence-calibrated Smoothed Classifiers for Certified Robustness. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30153–30168. Curran Associates, Inc., 2021.

Kopetzki, A.-K., Charpentier, B., Zügner, D., Giri, S., and Günnemann, S. Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5707–5718. PMLR, July 2021.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. 2009.

Kumar, A., Liang, P. S., and Ma, T. Verified Uncertainty Calibration. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kumar, A., Levine, A., Feizi, S., and Goldstein, T. Certifying Confidence via Randomized Smoothing. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5165–5177. Curran Associates, Inc., 2020.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018.

Murphy, A. H. Scalar and Vector Partitions of the Probability Score : Part II. N-State Situation. *Journal of Applied Meteorology (1962-1982)*, 11(8):1183–1192, January 1972. Publisher: American Meteorological Society.

Murphy, A. H. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015: 2901–2907, January 2015. ISSN 2159-5399.

Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models, 2015. _eprint: 1508.05154.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Qiao, M. and Valiant, G. Stronger Calibration Lower Bounds via Sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pp. 456–466, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8053-9. doi: 10.1145/3406325.3451050. event-place: Virtual, Italy.

Qin, Y., Wang, X., Beutel, A., and Chi, E. Improving Calibration through the Relationship with Adversarial Robustness. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14358–14369. Curran Associates, Inc., 2021.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pp. 3179–3189, June 2018.

Stutz, D., Hein, M., and Schiele, B. Confidence-Calibrated Adversarial Training: Generalizing to Unseen Attacks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9155–9166. PMLR, July 2020.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pp. 2818–2826, December 2016. ISBN 978-1-4673-8850-4. doi: 10.1109/CVPR.2016.308.

Tomani, C. and Buettner, F. Towards Trustworthy Predictions from Deep Neural Networks with Fast Adversarial Calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9886–9896, May 2021. doi: 10.1609/aaai.v35i11.17188.

Wenger, J., Kjellström, H., and Triebel), R. Non-Parametric Calibration for Classification. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 178–190. PMLR, August 2020.

Wu, B. and Ghanem, B. \ell _p-Box ADMM: A Versatile Framework for Integer Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1695–1708, July 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2845842.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018.

A. Motivation

A.1. Calibration $\neq$ Accuracy

It is important to note, that accuracy and calibration measure different concepts and one may not infer model performance from calibration with certainty or vice versa. Consider the following examples, where we fix one quantity and construct datasets resulting in other quantity taking on opposing values. These are illustrated in Figure 4. First, we fix the calibration error to $ECE = 0.5$ and for *Case 1*, we construct N data points with label $y_n = 1$, confidence $z_n = 0.5$ and thus prediction $\hat{y}_n = 1$. The calibration error here is 0.5 and the accuracy is 1. For *Case 2*, our predictions remain the same, but we change the labels to $y_n = 0$ resulting in an accuracy of 0 while keeping the calibration error of 0.5. Next, we fix the accuracy to 1 and construct examples with $ECE = 0.5$ and $ECE = 0$. The former is given by *Case 1*. The latter (*Case 3*) can be constructed using $y_n = 1$ and $z_n = 1$. Thus, we can construct a distribution over Z and Y such knowing the accuracy tells us nothing about the calibration error and vice versa. Clearly, when evaluating the quality of predictions, it is insufficient to only assess the accuracy. Hence, we argue that certifying accuracy in safety relevant applications is insufficient and calibration should be considered.

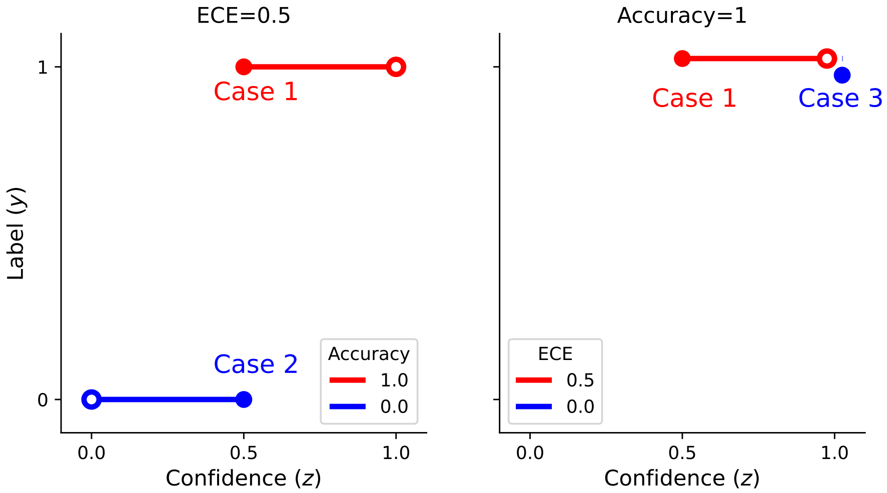
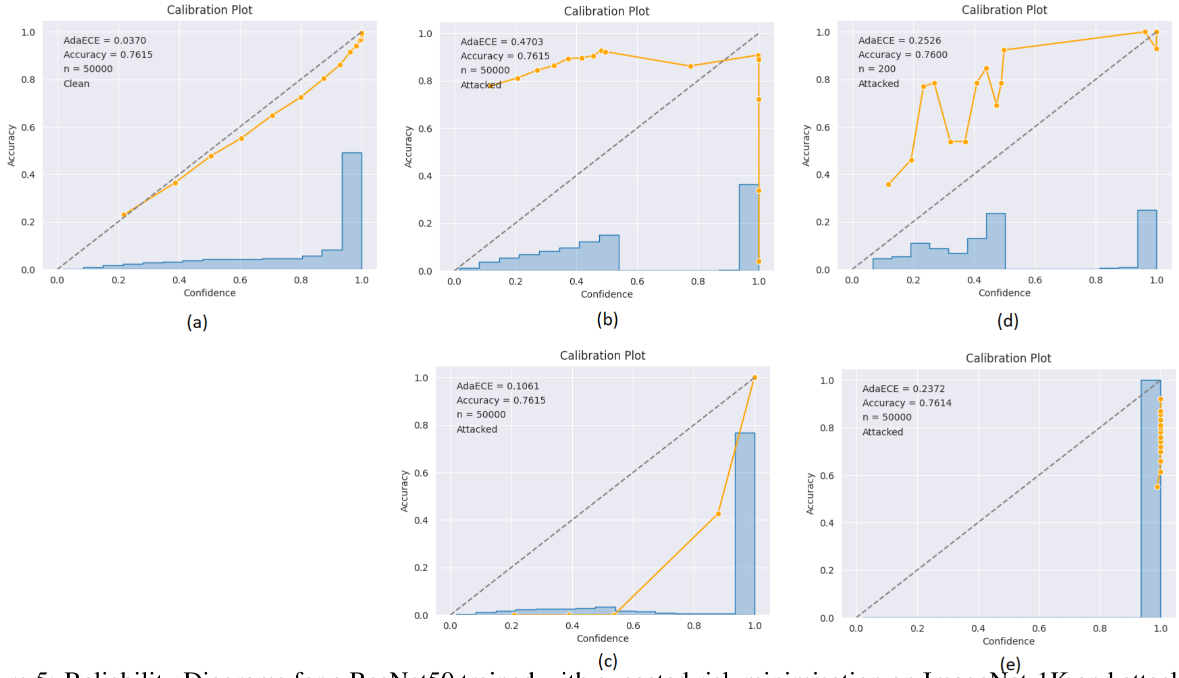


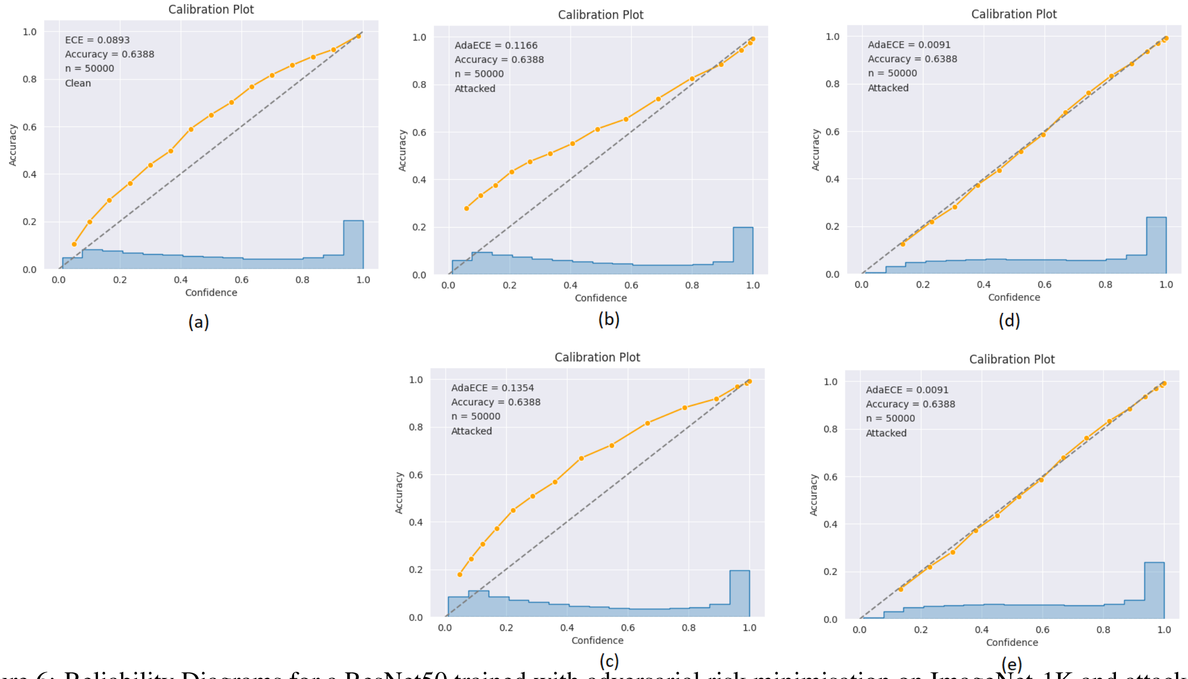
Figure 4: This visualisation shows that we can fix either the accuracy or the calibration and construct a dataset to obtain the other quantity with opposite values. The example looks at a binary classification problem. The empty circle displays the point of perfect calibration and the full circle is the calibration on the data. The distance of the line in-between is the calibration error.

A.2. Empirical Attacks



(a) (b) (d) (c) (e)

Figure 5: Reliability Diagrams for a ResNet50 trained with expected risk minimisation on ImageNet-1K and attack radius $\epsilon = 2/255$. (a) No Attack, (b) $(1, y)$-ACE attack, (c) $(-1, y)$-ACE attack, (d) $(1, \hat{y})$-ACE attack, (e) $(-1, \hat{y})$-ACE attack. The histogram on the bottom represents the distribution of confidence scores.



(a) (b) (d) (c) (e)

Figure 6: Reliability Diagrams for a ResNet50 trained with adversarial risk minimisation on ImageNet-1K and attack radius $\epsilon = 2/255$. (a) No Attack, (b) $(1, y)$-ACE attack, (c) $(-1, y)$-ACE attack, (d) $(1, \hat{y})$-ACE attack, (e) $(-1, \hat{y})$-ACE attack. The histogram on the bottom represents the distribution of confidence scores.

B. Confidence Bounds

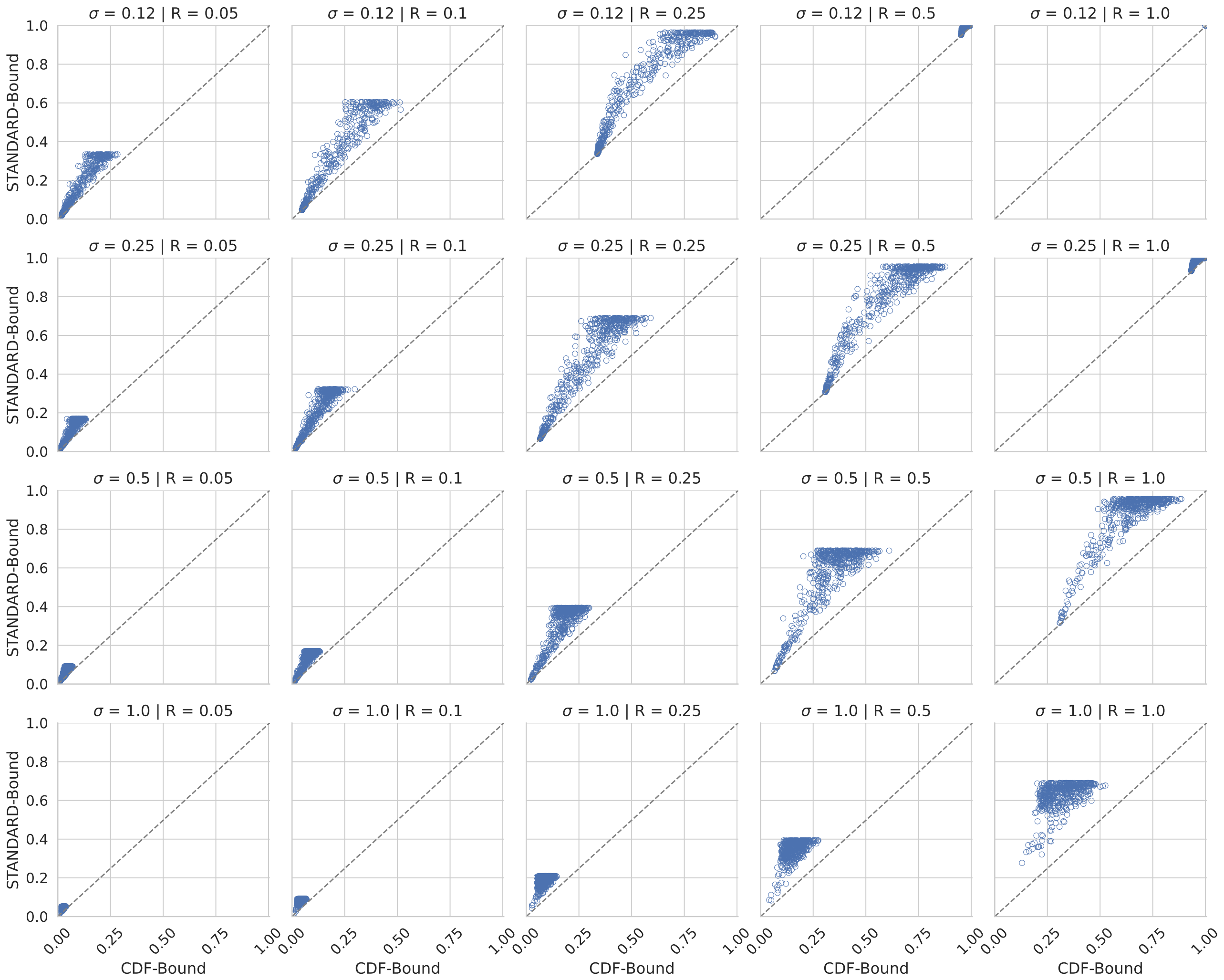


Figure 7: Distance between upper and lower bound of the confidence certificates provided by Kumar et al. (2020). Subsampled to 500 samples from the test set of CIFAR10.

.

Kumar et al. (2020) introduce the bounds on the confidence scores, but their work focuses on issuing a certificate given a lower bound on the confidence. Therefore, they do not investigate the upper bounds on the confidence. Here, we compute the interval of certified confidences and compare the two certificates on the confidence: The standard STANDARD bound as given in (4) and the more advanced CDF bound. In Figure 7, the distance between the upper and lower bound is plotted for a range of smoothing σ and certification radii R on a random subset of 500 samples from the CIFAR-10 test set. We may observe that the CDF method yields uniformly tighter bounds in practice.

C. Brier Bound

Here we provide a proof to Theorem 3.1 as stated above.

Proof. Assume $\mathbf{z}^* = \mathbf{l}\mathbb{I}\{\mathbf{c} = 1\} + \mathbf{u}\mathbb{I}\{\mathbf{c} = 0\}$ is not the maximum and we want to change $\mathbf{z}^*$ to maximise the TLBS. For some data point with $c_n = 1$, the bound is $z^*_n = l_n$. Reducing z^*_n is not possible without leaving its feasible set ($l_n \leq z^*_n \leq u_n$), and thus the only way to find the maximum is to increase it. However, increasing z^*_n would reduce $c_n - z^*_n$

which in turn reduces the error, as $\|\cdot\|_2$ is strictly increasing in $c_n - z_n$. Thus, we have a contradiction for $c_n = 1$. The other case, $c_n = 0$ is analog and both are true for all n and thus, the maximum is proven.

□

D. Calibration as Mixed-Integer Program

D.1. Restating the Calibration Error

Here we provide a proof of Theorem 3.3.

Proof. Let $\mathbf{E}(\mathbf{z})$ and $\mathbf{a}$ be as defined as above. We start with the expression in (7) and show equality to (2). Note that $a_{n,m}$ is 1 if data point n is in bin B_m and 0 otherwise. Also note that by the definition of $\mathbf{E}$ that $e_{n,m} = e_n$ when $a_{n,m} = 1$:

$$\begin{aligned}
\|\mathbf{E}^\top \mathbf{a}\|_1 &= \sum_{m=1}^{M} \left|\mathbf{e}_m(\mathbf{z})^\top \mathbf{a}\right| && (9)\\
&= \sum_{m=1}^{M} \left|\sum_{n=1}^{N} e_{n,m} a_{n,m}\right| && (10)\\
&= \sum_{m=1}^{M} \left|\sum_{n \in B_m} e_n\right| && (11)\\
&= \sum_{m=1}^{M} \left|\sum_{n \in B_m} c_n - z_n\right| && (12)\\
&= \sum_{m=1}^{M} |B_m| \left|\frac{1}{|B_m|} \sum_{n \in B_m} c_n - \frac{1}{|B_m|} \sum_{n \in B_m} z_n\right| && (13)
\end{aligned}$$

Dividing both sides by N yields the result. □

D.2. Lagrangian

We formally state the Augmented Lagrangian. The variables $\mathbf{a}$, $\mathbf{z}$, $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{g}$ are described as above, each of dimension NM.

$$\begin{aligned}
\mathcal{L}(\mathbf{a}, \mathbf{z}, \mathbf{g}, \mathbf{q}_{1,2}, \boldsymbol{\lambda}_{1,2,3,4,5}) = &-|\mathbf{E}(\mathbf{z})^\top \mathbf{a}| \mathbf{1}_M \\
&+ \mathbb{I}_\infty\{\mathbf{q}_1 \in S_b\} + \boldsymbol{\lambda}_1^\top [\mathbf{a} - \mathbf{q}_1] + \frac{\rho_1}{2}\|\mathbf{a} - \mathbf{q}_1\|_2^2 \\
&+ \mathbb{I}_\infty\{\mathbf{q}_2 \in S_2\} + \boldsymbol{\lambda}_2^\top [\mathbf{a} - \mathbf{q}_2] + \frac{\rho_2}{2}\|\mathbf{a} - \mathbf{q}_2\|_2^2 \\
&+ \boldsymbol{\lambda}_3^\top [\mathbf{C}^\top \mathbf{a} - \mathbf{1}_N] + \frac{\rho_3}{2}\|\mathbf{C}^\top \mathbf{a} - \mathbf{1}_N\|_2^2 \\
&+ \boldsymbol{\lambda}_4^\top \mathbf{K}^\top \mathbf{a} + \frac{\rho_4}{2}\|\mathbf{K}^\top \mathbf{a}\|_2^2 \\
&+ \mathbb{I}_\infty\{\mathbf{g} \in S_z\} + \boldsymbol{\lambda}_5^\top [\mathbf{z} - \mathbf{g}] + \frac{\rho_5}{2}\|\mathbf{z} - \mathbf{g}\|_2^2
\end{aligned} \qquad (14)$$

with dual variables $\boldsymbol{\lambda}_{1,2,5} \in \mathbb{R}^{NM}$, $\boldsymbol{\lambda}_{3,4} \in \mathbb{R}^N$. Here, $\mathbb{I}_\infty$ is 0 if the statement is true and ∞ otherwise. The values of $\rho_i > 0$ are hyperparameters to be tuned.

D.3. ADMM Updates

We perform T ADMM steps. At each ADMM step we cycle through the primal variables $\mathbf{a}$ and $\mathbf{z}$ and perform gradient descent. For the variables $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{g}$, we obtain an analytic solution by equating the gradient to 0, solving the equation and projecting the variables into their feasible set. For $\mathbf{q}_1$, the update $\mathcal{U}_{\mathbf{q}_1}$ is given by

$$\mathbf{q}_1 \leftarrow \text{clamp}_{[0,1]}\left(\frac{\boldsymbol{\lambda}_1}{\rho_1} + \mathbf{a}\right), \tag{15}$$

for $\mathbf{q}_2$ the update $\mathcal{U}_{\mathbf{q}_2}$ is given by

$$\mathbf{q}_2 \leftarrow \frac{1}{2}\mathbf{1} + \frac{\sqrt{NM}}{2}\frac{\frac{\boldsymbol{\lambda}_2}{\rho_2} + \mathbf{a} - \frac{1}{2}\mathbf{1}}{\|\frac{\boldsymbol{\lambda}_2}{\rho_2} + \mathbf{a} - \frac{1}{2}\mathbf{1}\|_2}, \tag{16}$$

and finally, the update $\mathcal{U}_{\mathbf{g}}$ is given by

$$\mathbf{g} \leftarrow \text{clamp}_{[\mathbf{l}',\mathbf{u}']}\left(\frac{\boldsymbol{\lambda}_5}{\rho_5} + \mathbf{z}\right). \tag{17}$$

The updates on the dual variables $\boldsymbol{\lambda}_i$ are performed through a single gradient ascent step with step size ρ_i.

With these definitions, we can formalise the algorithm for the ADMM updates.

Algorithm 1 The ADMM Updates

```
input ADMM parameters $\{\alpha_{\mathbf{a}}, \alpha_{\mathbf{z}}, \rho_{1,2,3,4,5}\}$, primal variables $\left\{\mathbf{a}^{(0)}, \mathbf{z}^{(0)}, \mathbf{q}_1^{(0)}, \mathbf{q}_2^{(0)}, \mathbf{g}^{(0)}\right\}$ and dual variables $\{\boldsymbol{\lambda_i}\}_{i=1}^{5}$
output ACCE
  for $t = 1$ to $T$ do
    $\mathbf{a}^{(t)} \leftarrow \mathbf{a}^{(t-1)} - \alpha_{\mathbf{a}}\nabla_{\mathbf{a}}\mathcal{L}\left(\mathbf{a}^{(t-1)}, \mathbf{z}^{(t-1)}, \mathbf{g}^{(t-1)}, \mathbf{q}_{1,2}^{(t-1)}, \boldsymbol{\lambda}_{1,2,3,4,5}^{(t-1)}\right)$
    $\mathbf{z}^{(t)} \leftarrow \mathbf{z}^{(t-1)} - \alpha_{\mathbf{z}}\nabla_{\mathbf{z}}\mathcal{L}\left(\mathbf{a}^{(t)}, \mathbf{z}^{(t-1)}, \mathbf{g}^{(t-1)}, \mathbf{q}_{1,2}^{(t-1)}, \boldsymbol{\lambda}_{1,2,3,4,5}^{(t-1)}\right)$
    $\mathbf{g}^{(t)} \leftarrow \mathcal{U}_{\mathbf{g}}\left(\mathbf{z}^{(t)}, \boldsymbol{\lambda}_5^{(t-1)}\right)$
    $\mathbf{q}_1^{(t)} \leftarrow \mathcal{U}_{\mathbf{q}_1}\left(\mathbf{a}^{(t)}, \boldsymbol{\lambda}_1^{(t-1)}\right)$
    $\mathbf{q}_2^{(t)} \leftarrow \mathcal{U}_{\mathbf{q}_2}\left(\mathbf{a}^{(t)}, \boldsymbol{\lambda}_2^{(t-1)}\right)$
    $\boldsymbol{\lambda}_i^{(t)} \leftarrow \boldsymbol{\lambda}_i^{(t-1)} + \rho_i\nabla_{\boldsymbol{\lambda}_i}\mathcal{L}\left(\mathbf{a}^{(t)}, \mathbf{z}^{(t)}, \mathbf{g}^{(t)}, \mathbf{q}_{1,2}^{(t)}, \boldsymbol{\lambda}_{1,2,3,4,5}^{(t-1)}\right)$ for $i = 1, 2, 3, 4, 5$
  end for
```

E. Experimental Results

E.1. Brier Score

E.2. ADMM for Certified Calibration Error

E.2.1. ADMM Hyperparameter Search

We performed a random search hyperparameter search for the ADMM solver to find reasonable hyperparameters that are efficient across experiments.

- The learning rate for $\mathbf{z}$, $\alpha_{\mathbf{z}}$: We test values from 1×10^{-5} to 1×10^{-2} and find it has little influence.
- The learning rate for $\mathbf{a}$, $\alpha_{\mathbf{a}}$: We test values from 5×10^{-4} to 5×10^{-2} for ImageNet and from 1×10^{-5} to 1×10^{-2} for CIFAR10. Larger learning rates are preferred. In our experiments 0.001 to 0.05 worked best.
- The Lagrangian smoothing variable ρ_i: We tested starting values from 0.001 to 0.5. We find little influence but values around 0.01 to 0.05 works best. We test multiplicative schedules to increase ρ_i over time with increases starting at $1‰$ to 4%. While larger values of ρ improve constraint convergence, they can dominate gradients when the constraints have not sufficiently converged yet, and thus ADMM easily diverges. Schedules around 1.02% to 1.05% per step are effective and we stop increasing ρ at 10, which is completely sufficient to meet the constraints. We find, that ρ_i scheduling is important. But applying the same schedule described above for all $\rho = \rho_1, ..., \rho_5$ works reasonably well.
- We test performing 1 to 3 updates for $\mathbf{a}$ and $\mathbf{z}$ per ADMM step and find that more steps do not aid results while slowing down ADMM. We thus, recommend 1 step per ADMM step.

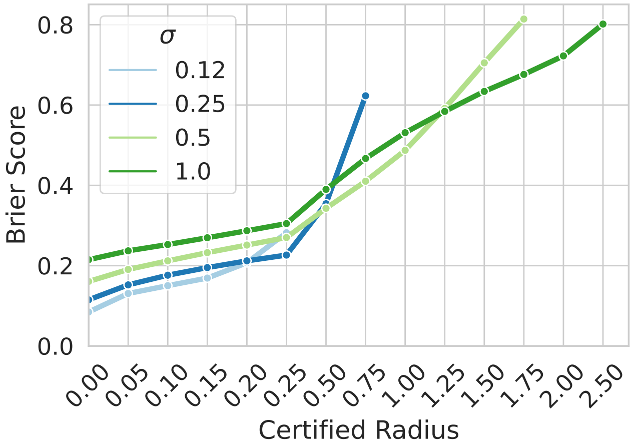


Figure 8: Certified Brier scores (CBS) on CIFAR-10 across a range of certified radii. For small radii, models with small smoothing σ outperforms those with larger σ, but as radii increase, large σ outperform smaller σ. As noted in section 3.2, changes in the Brier score as a function of confidence perturbations are solely reflecting calibration on certified models. However, in this and Figure 2 the dataset changes across certified radii, and thus increases in the CBS cannot solely be attributed to calibration.

- We test clipping values of $\mathbf{a}$. As we constrain $\mathbf{a} = \mathbf{q}_2$, we clip $\|\mathbf{a} - 1/2\|_\infty \leq 1.2\|\mathbf{q}_2 - 1/2\|_\infty$. Note that when $\mathbf{a}$ converges, it will be significantly smaller than this constraint. We have never observed any decline in performance with this approach but seen that it stabilises the optimisation problem in rare instances.

- We clip the gradients of $\mathbf{a}$ to 5 in infinity norm aiding stability.

- The initialisation of $\mathbf{a}$ has a major effect on the performance of ADMM, more so than any other hyperparameter. While it is an obvious solution to initialise $\mathbf{a}^{(0)}$ such that it is a valid assignment for $\mathbf{z}^{(0)}$, we find that this is majorly outperformed by uniform initialisations. We tested 0, $1/M$ and 1 and find that $1/M$ works best.

- The initialisation of $\mathbf{z}$ has a mild effect on ADMM performance given that sometimes, we might have prior knowledge on how and in which direction the model currently is miscalibrated. We recommend initialising it with adversary-free confidences and with the confidences achieving the Brier bound and subsequently picking the larger one.

E.2.2. ADMM vs dECE vs Brier Bound

To further assess the efficacy of the ADMM algorithm, we compare it to an alternative methods of approximating bounds: the *differentiable calibration error* (dECE) (Bohdal et al., 2021) and the Bounds obtained by the confidences that maximise the Brier score (*Brier confidences)*. We find that ADMM outperforms both other techniques. We performed hyperparameter searches for ADMM and dECE with a wide range, but carefully selected hyperparameters. For ADMM we run between 259 and 1560 trials (depending on the runtime) and for dECE always 2000 trials. We explore a subset of radii and smoothing σ (as visible from our plots). Our first observation is, that the dECE is very sensitive to the initialisation of confidence scores indicating that gradient ascent on dECE (even with very large learning rates) does not sufficiently explore the loss surface. While differences are also observable for ADMM, these are very small. Figure 9 show these differences for ImageNet.

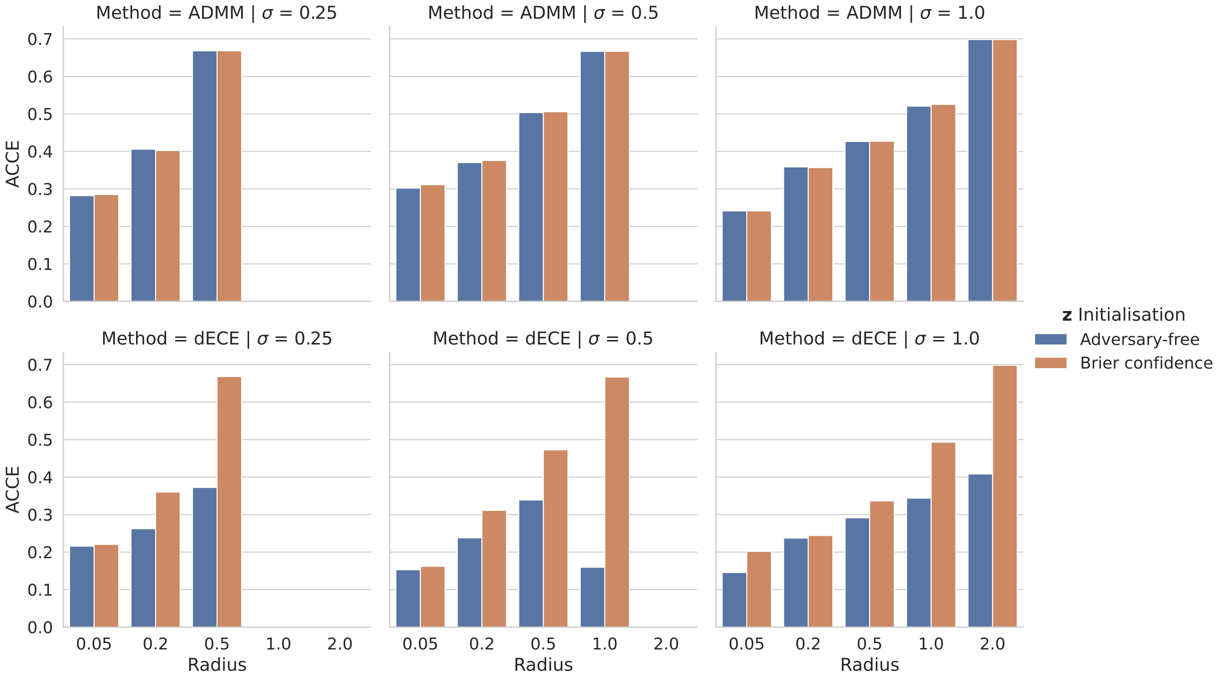


Figure 9: The results of the hyperparameter search are shown here. For each combination of Dataset, σ and Radius, the maximum achieved by the three methods, ADMM, dECE and Brier are shown here. While dECE is able to be uniformly better than the Brier confidences, ADMM outperforms both by a significant margin.

Second, as a result of our hyperparameter search, we note that the best ADMM results outperform the best dECE by a significant margin. To demonstrate this, we compare the maxima achieved across trials by the dECE, the Brier bound and ADMM in Figure 10.

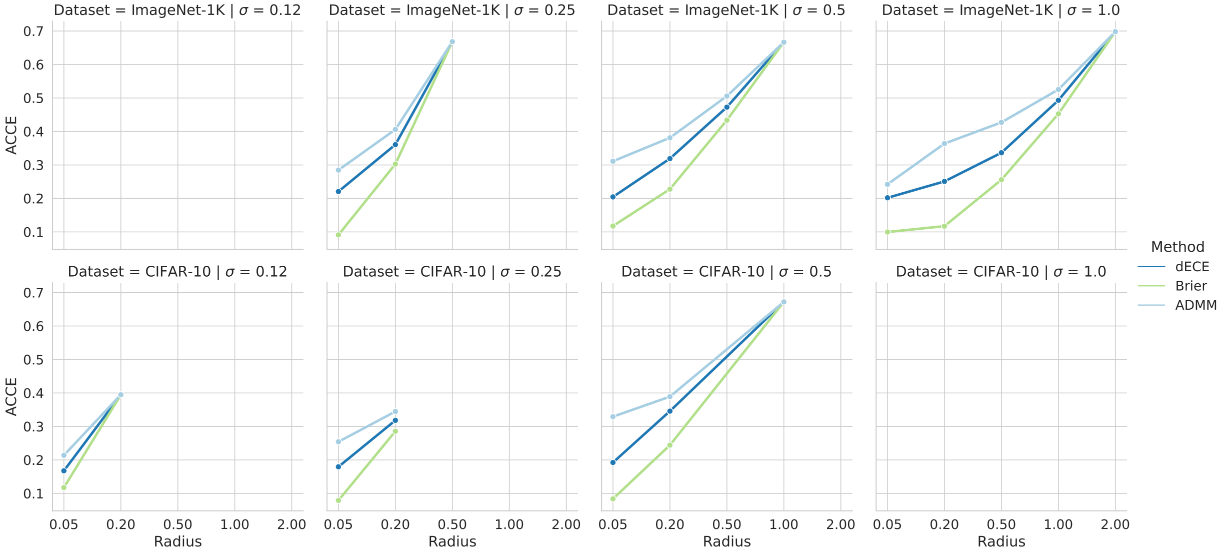


Figure 10: The results of the hyperparameter search are shown here. For each combination of Dataset, σ and Radius, the maximum achieved by the three methods, ADMM, dECE and Brier are shown here. While dECE is able to be uniformly better than the Brier confidences, ADMM outperforms both by a significant margin.

As noted in the main paper, for large radii, the methods converge to each other. Beyond, the comparison of maximum values above, we find that even a single *one-size-fits-all* set of hyperparameters for ADMM outperforms the maximum achieved by the dECE hyperparameter search in most cases as shown in Figure 11.

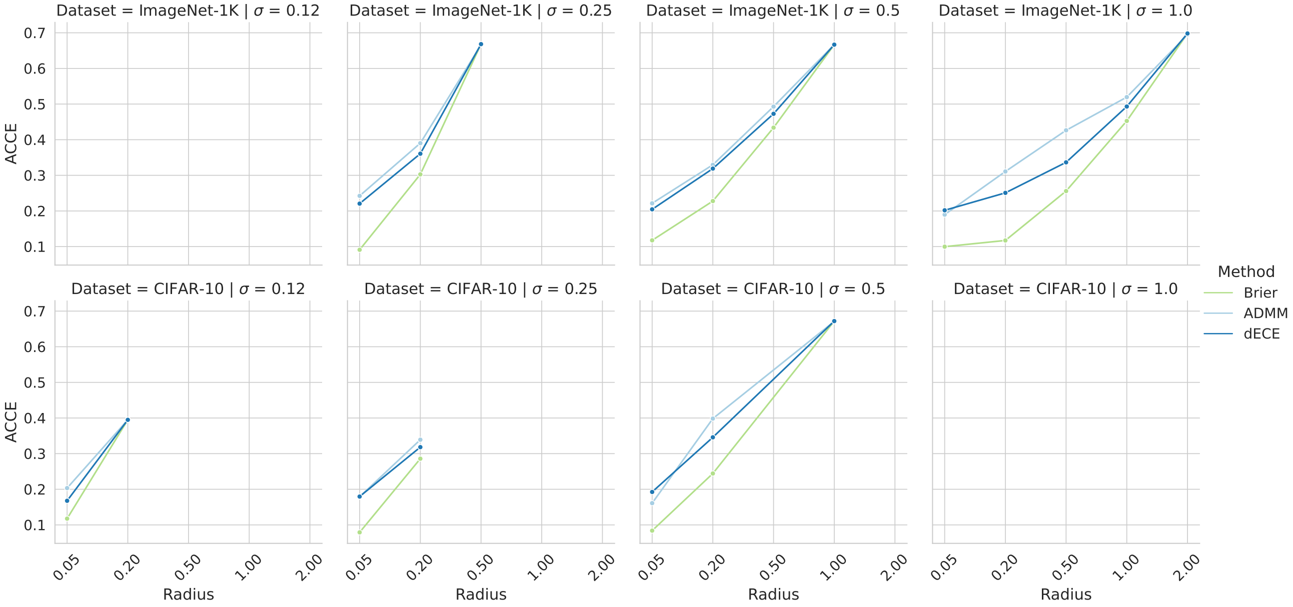


Figure 11: We compare a single set of parameters for ADMM across all Dataset, Smoothing and Epsilon combinations and find that it outperforms the *maximum* dECE from a hyperparameter search of 2000 samples in 17/19 instances.

The hyperparameters used for ADMM are as follows: All ρ_is are initialised to 0.01 with increases every step by a factor of 0.4%. Both learning rates are set to 0.001. The assignments are initialised to $1/M$, the confidence is initialised to adversary-free and Brier confidence, a measure benefiting the dECE much more than the ADMM. The gradients of the assignment are clipped to 1.

We use our results from the hyperparameter searches to run ADMM and dECE on a finer grid of certified confidences as referenced in section 4.1. Here, we present the results for the finer grid for CIFAR-10.

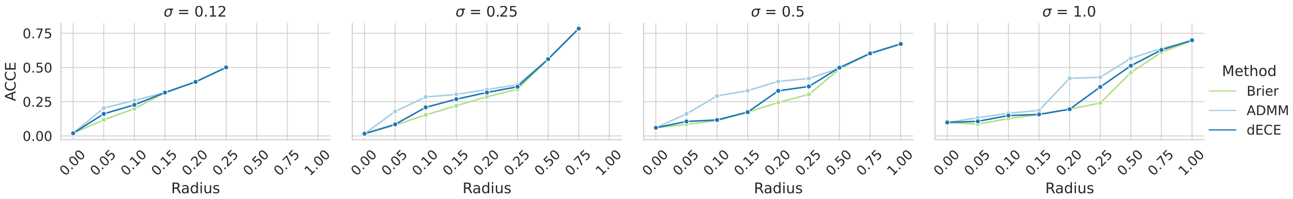


Figure 12: The ACCE is shown here for ADMM, dECE and the Brier confidences. One set of carefully selected hyperparameter is used. For a wider overview see Figure 10

F. Differentiable Calibration Error

F.1. Definition

Bohdal et al. (2021) note that the standard calibration error estimator is non-differentiable in two operations: the accuracy and the hard binning. As we are assuming certified predictions, their correctness c_n (and thus the accuracy per bin) is constant with respect to the confidence scores z_n and thus does not need to be differentiable. We therefore simplify their dECE and apply only one differentiable approximation. We restate the calibration error estimator from (2) and simplify:

$$\hat{\text{ECE}} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{n \in B_m} c_n - \frac{1}{|B_m|} \sum_{n \in B_m} z_n \right| \tag{18}$$

$$= \frac{1}{N} \sum_{m=1}^{M} \left| \sum_{n=1}^{N} a_{n,m}(c_n - z_n) \right| \tag{19}$$

where $a_{n,m} \in \{0, 1\}$ is the hard indicator whether data point n is assigned to bin m. This is replaced with a soft indicator $s_{n,m} \in [0, 1]$ with $\sum_m s_{n,m} = 1$. Define matrix $\mathbf{S} \in [0, 1]^{N \times M}$ with elements $s_{n,m}$ and let $\mathbf{e} = [c_1 - z_1, ..., c_N - z_N]$. We can write the differentiable calibration error d$\hat{\text{ECE}}$ as:

$$\text{d}\hat{\text{ECE}} = \frac{1}{N} \|\mathbf{S}^\top \mathbf{e}\|_1 \tag{20}$$

The soft assignment $s_{n,m}$ is obtained the following way. For M equal width bins with cut-offs $\beta_1 < ... < \beta_{M-1}$, we define $\mathbf{b}$ with elements $b_i = -\sum_{m=1}^{i-1} \beta_m$. Further, let $\mathbf{w} = [1, 2, ..., M]$ we obtain the soft assignments $\mathbf{s}_n$ through a tempered softmax function: $\mathbf{s}_n = \boldsymbol{\sigma}((\mathbf{w} z_n + \mathbf{b})/\tau)$. For $\tau \to 0$, the vector $\mathbf{s}_n$ approximates a one-hot encoded vector and (20) recovers the original ECE.

F.2. dECE Parameters

As for ADMM, we perform an extensive hyperparameter search for the dECE. Our key insight is to start the optimisation with high values of τ, i.e. 0.01 as this increases the smoothness of the objective function and slowly decrease τ to about 1×10^{-6} at which point we usually observe a difference between the ECE and dECE of less than 1×10^{-6}.

As for the ADMM, we test a wide range of hyperparameters for the dECE. The one most significant hyperparameter is the initialization of $\mathbf{z}$: We test the same initialisation as for ADMM: adversary-free and Brier confidences. In addition, we test random Gaussian and random uniform initialisation. We do not find a single optimal strategy. It is possible that the adversary-free ECE is *higher* than the ECE obtained by using the Brier confidences. As dECE shows poor performance in exploring the loss surface, we recommend initialisation to whatever initialisation yields the largest error to begin with. All other hyperparameters are insignificant in comparison and usually over 95% of variation in results explained by the initialisation of $\mathbf{z}$. We test various learning rates, schedulers for τ (as mentioned above), learning rate schedulers (Cosine Annealing, Constant, ReduceOnPleateau) and optimizer momentum and find none of these hyperparameters to be significant.