Soft Calibration Objectives for Neural Networks

Archit Karandikar[*]
Google Research
archk@google.com

Nicholas Cain[*]
Google Research
nicholascain@google.com

Dustin Tran
Google Research
trandustin@google.com

Balaji Lakshminarayanan
Google Research
balajiln@google.com

Jonathon Shlens
Google Research
shlens@google.com

Michael C. Mozer
Google Research
mcmozer@google.com

Becca Roelofs
Google Research
rolfs@google.com

Abstract

Optimal decision making requires that classifiers produce uncertainty estimates consistent with their empirical accuracy. However, deep neural networks are often under- or over-confident in their predictions. Consequently, methods have been developed to improve the calibration of their predictive uncertainty, both during training and post-hoc. In this work, we propose differentiable losses to improve calibration based on a soft (continuous) version of the binning operation underlying popular calibration-error estimators. When incorporated into training, these soft calibration losses achieve state-of-the-art single-model ECE across multiple datasets with less than 1% decrease in accuracy. For instance, we observe an 82% reduction in ECE (70% relative to the post-hoc rescaled ECE) in exchange for a 0.7% relative decrease in accuracy relative to the cross-entropy baseline on CIFAR-100. When incorporated post-training, the soft-binning-based calibration error objective improves upon temperature scaling, a popular recalibration method. Overall, experiments across losses and datasets demonstrate that using calibration-sensitive procedures yield better uncertainty estimates under dataset shift than the standard practice of using a cross-entropy loss and post-hoc recalibration methods.[2]

1 Introduction

Despite the success of deep neural networks across a variety of domains, they are still susceptible to miscalibrated predictions. Both over- and under-confidence contribute to miscalibration, and empirically, deep neural networks empirically exhibit significant miscalibration [Guo et al., 2017]. Calibration error (*CE*) quantifies a model's miscalibration by measuring how much its confidence, i.e. the predicted probability of correctness, diverges from its accuracy, i.e. the empirical probability of correctness. Models with low CE are critical in domains where satisfactory outcomes depend on well-modeled uncertainty, such as autonomous vehicle navigation [Bojarski et al., 2016] and medical diagnostics [Jiang et al., 2012, Caruana et al., 2015, Kocbek et al., 2020]. Calibration has also been shown to be useful for improving model fairness [Pleiss et al., 2017] and detecting out-of-distribution data [Kuleshov and Ermon, 2017, Devries and Taylor, 2018, Shao et al., 2020]. More generally, low

[*]co-first author
[2]Code available on GitHub: https://github.com/google/uncertainty-baselines/tree/main/experimental/caltrain

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

CE is desirable in any setting in which thresholds are applied to the predicted confidence of a neural network in order to make a decision.

Methods for quantifying CE typically involve binning model predictions based on their confidence. CE is then computed empirically as a weighted average of the absolute difference in average prediction confidence and average accuracy across different bins [Naeini et al., 2015]. Oftentimes these bins are selected heuristically such as *equal-width* (uniformly binning the score interval) and *equal-mass* (with equal numbers of samples per-bin) [Nixon et al., 2019].

However, these commonly used measures of CE are not trainable with gradient-based methods because the binning operation is discrete and has zero derivatives. As a result, neural network parameters are not directly trained to minimize CE, either during training or during post-hoc recalibration. In this paper, we introduce new objectives based on a differentiable binning scheme that can be used to efficiently and directly optimize for calibration.

Contributions. We propose estimating CE with soft (i.e., overlapping, continuous) bins rather than the conventional hard (i.e., nonoverlapping, all-or-none) bins. With this formulation, the CE estimate is differentiable, allowing us to use it as (1) a secondary (i.e., auxiliary) loss to incentivize model calibration during training, and (2) a primary loss for optimizing post-hoc recalibration methods such as temperature scaling. In the same spirit, we soften the AvUC loss [Krishnan and Tickoo, 2020], allowing us to use it as an effective secondary loss during training for non-Bayesian neural networks where the AvUC loss originally proposed for Stochastic Variational Inference (SVI) typically does not work. Even when training with the cross-entropy loss results in training set memorization (perfect train accuracy and calibration), Soft Calibration Objectives are still useful as secondary training losses for reducing test ECE using a procedure we call *interleaved training*.

In an extensive empirical evaluation, we compare Soft Calibration Objectives as secondary losses to existing calibration-incentivizing losses. In the process, we find that soft-calibration losses outperform prior work on in-distribution test sets. Under distribution shift, we find that calibration-sensitive training objectives as a whole (not always the ones we propose) result in better uncertainty estimates compared to the standard cross-entropy loss coupled with temperature scaling.

Our contributions can be summarized as follows:

- We propose simple Soft Calibration Objectives S-AvUC, SB-ECE as secondary losses which optimize for CE *throughout training*. We show that across datasets and choice of primary losses, the S-AvUC secondary loss results in the largest improvement in ECE as per the Cohen's d effect-size metric (Figure 1). We also show that such composite losses obtain *state-of-the-art single-model ECE* in exchange for less than 1% reduction in accuracy (Figure 5) for CIFAR-10, CIFAR-100, and Imagenet.
- We improve upon temperature scaling - a popular post-hoc recalibration method - by directly optimizing the temperature parameter for soft calibration error instead of the typical log-likelihood. Our extension (TS-SB-ECE) consistently beats original temperature scaling (TS) across different datasets, loss functions and calibration error measures, and we find that the performance is better under dataset shift (Figure 2).
- Overall, our work demonstrates a fundamental advantage of objectives which better incentivize calibration over the standard practice of training with cross-entropy loss and then applying post-hoc methods such as temperature scaling. Uncertainty estimates of neural networks trained using these methods generalize better in and out-of-distribution.

2 Related Work

Many techniques have been proposed to train neural networks for calibration. These can be organized into three categories. One category augments or replaces the primary training loss with a term to explicitly incentivize calibration. These include the AvUC loss [Krishnan and Tickoo, 2020], MMCE loss [Kumar et al., 2018] and Focal loss [Mukhoti et al., 2020, Lin et al., 2018]. The Mean Squared Error loss also compares favourably [Hui and Belkin, 2021] to cross-entropy loss. We show that our methods outperform all these calibration-incentivizing training objectives, applied across multiple primary losses. Label smoothing [Müller et al., 2020] has been shown to improve calibration and can be interpreted as a modified primary loss.

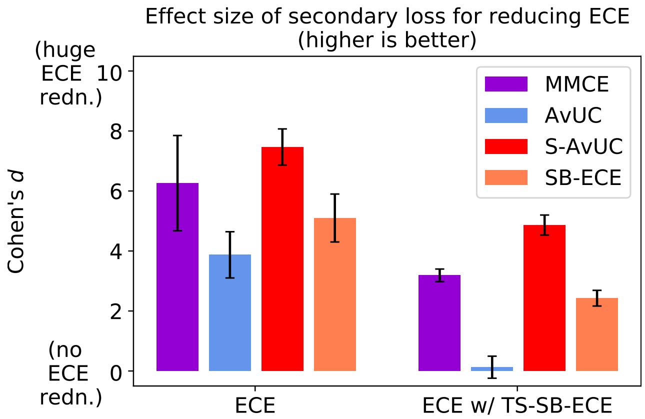


Figure 1: We compare the effect size of various secondary losses on ECE (equal-mass binning, ℓ_2 norm) across datasets and primary losses, both with and without post-hoc temperature scaling. The S-AvUC secondary loss we propose shows the strongest positive effect, followed by MMCE and SB-ECE. Note that a d-value of 0.8 (resp., 2.0) is considered a large (resp., huge) positive effect and d-values obtained here are much larger. Secondary losses which incentivize calibration show strong positive effect for reducing ECE even after temperature scaling.

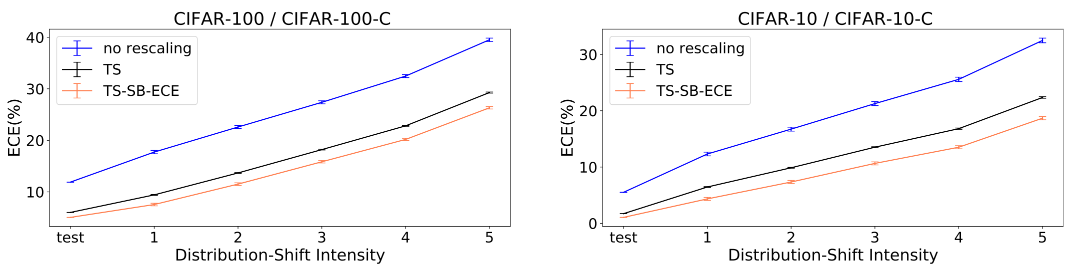


Figure 2: Post-hoc temperature scaling with the soft calibration error objective (TS-SB-ECE) reduces ECE more than standard post-hoc temperature scaling (TS), particularly under distribution shift. This result holds across datasets (left and right panels), distribution shift intensities (along abscissa) and training objectives (not shown). The training objective shown here for both datasets is the most popular one: NLL. The ECE value (equal-mass binning, ℓ_2 norm) shown is the mean ECE across the corruption types that constitute CIFAR-10-C and CIFAR-100-C. Error bars are ± 1 standard error of mean (SEM), corrected for intrinsic variability due to type of corruption [Masson and Loftus, 2003].

A second category of methods are post-hoc calibration methods, which rescale model predictions after training. These methods optimize additional parameters on a held-out validation set [Platt, 1999, Zadrozny and Elkan, 2002, Kull et al., 2019, Zadrozny and Elkan, 2001, Naeini and Cooper, 2016, Allikivi and Kull, 2019, Kull et al., 2017, Naeini et al., 2015, Wenger et al., 2020, Gupta et al., 2020]. The most popular technique is temperature scaling [Guo et al., 2017], which maximizes a single temperature parameter on held-out NLL. We examine temperature scaling and propose an improvisation that directly optimizes temperature for a soft calibration objective instead of NLL. Temperature scaling has shown to be ineffective under distribution shift in certain scenarios [Ovadia et al., 2019]. We show that uncertainty estimates of methods which train for calibration generalize better than temperature scaling under distribution shift.

A third category of methods examines model changes such as ensembling multiple predictions [Lakshminarayanan et al., 2017, Wen et al., 2020] or priors [Dusenberry et al., 2020]. Similar to previous work [Kumar et al., 2018, Lin et al., 2018], we focus on the choice of loss functions for improving calibration of a single neural network—whether during training or post-hoc—and do not compare against ensemble models or Bayesian neural networks. These techniques are complementary to ours and can be combined with our techniques to further improve performance.

Recent works have investigated issues [Nixon et al., 2019, Kumar et al., 2019, Roelofs et al., 2020, Gupta et al., 2020] with the originally proposed ECE [Guo et al., 2017] and suggested new ones. Debiased CE [Kumar et al., 2019] and mean-sweep CE [Roelofs et al., 2020] have been shown to have lesser bias and more consistency across the number of bins parameter than ECE whereas KS-error [Gupta et al., 2020] avoids binning altogether. We report these metrics in the Appendix.

3 Background

3.1 The Task and the Model

Consider a classification task over K classes with a dataset of N samples $D = \langle(\boldsymbol{x}_i, y_i)\rangle_{i=1}^{N}$ drawn from the joint probability distribution $\mathcal{D}(\mathcal{X}, \mathcal{Y})$ over the input space $\mathcal{X}$ and label space $\mathcal{Y} = \{1, 2, \dots K\}$. The task is modelled using a deep neural network with parameters $\boldsymbol{\theta}$ whose top layer is interpreted as a softmax layer. The top layer consists of K neurons which produce logits $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle g_{\boldsymbol{\theta}}(y|\boldsymbol{x})\rangle_{y\in\mathcal{Y}}$. The predictive probabilities for a given input are:

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle f_{\boldsymbol{\theta}}(y|\boldsymbol{x})\rangle_{y\in\mathcal{Y}} = \mathrm{softmax}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{x})).$$

The parameters $\boldsymbol{\theta}$ of the neural network are trained to minimize $\mathbb{E}_{(\boldsymbol{x},y)}\mathcal{L}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$ where $(\boldsymbol{x}, y)$ is sampled from $\mathcal{D}(\mathcal{X}, \mathcal{Y})$. Here $\mathcal{L}$ is a trainable loss function which incentivizes the predictive distribution $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ to fit the label y. The model's prediction on datapoint $(\boldsymbol{x}, y)$ is denoted by $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$. We denote by $c_{\boldsymbol{\theta}}(\boldsymbol{x}) = \max(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$ the confidence of this prediction and by boolean quantity $a_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \mathbf{1}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})=y}$ the accuracy of this prediction.

3.2 Expected Calibration Error (ECE)

Given a distribution $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$ on datapoints, there are two standard notions of Ideal Calibration Error that we refer to as Ideal Binned Expected Calibration Error (of order p), $\mathrm{IECE}_{\mathrm{bin},p}(\hat{\mathcal{D}}, \boldsymbol{\theta}) = \mathrm{IECE}_{\mathrm{bin},p}$, and Ideal Expected Label-Binned Calibration Error (of order p), $\mathrm{IECE}_{\mathrm{lb},p}(\hat{\mathcal{D}}, \boldsymbol{\theta}) = \mathrm{IECE}_{\mathrm{lb},p}$. This nomenclature is consistent with that introduced by Roelofs et al. [2020]. Both these measure, in slightly different ways, the p^{th} root of the p^{th} moment of the absolute difference between model confidence and the empirical accuracy given that confidence. They are defined as follows:

$$\mathrm{IECE}_{\mathrm{bin},p}(\hat{\mathcal{D}}, \boldsymbol{\theta}) = \left(\mathbb{E}_{c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_0})}\left[|\mathbb{E}[a_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_1}, \hat{y_1})|c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_1}) = c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_0})] - c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_0})|^p\right]\right)^{1/p} \tag{1}$$

$$\mathrm{IECE}_{\mathrm{lb},p}(\hat{\mathcal{D}}, \boldsymbol{\theta}) = \left(\mathbb{E}_{(\hat{\boldsymbol{x}_0}, \hat{y_0})}\left[|\mathbb{E}[a_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_1}, \hat{y_1})|c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_1}) = c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_0})] - c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}_0})|^p\right]\right)^{1/p}. \tag{2}$$

Note that the critical dependence on $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$ is implicit in both definitions since the datapoint $(\hat{\boldsymbol{x}_0}, \hat{y_0})$ from the outer expectation and the datapoint $(\hat{\boldsymbol{x}_1}, \hat{y_1})$ from the inner expectation are both sampled from $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$.

We cannot compute $\mathrm{IECE}_{\mathrm{bin},p}$ and $\mathrm{IECE}_{\mathrm{lb},p}$ in practice since the number of datapoints are finite. Instead we consider a dataset $\hat{D} = \langle(\hat{\boldsymbol{x}}_i, \hat{y}_i)\rangle_{i=1}^{\hat{N}}$ drawn from $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$ and partition the confidence interval $[0, 1]$ into bins $\mathcal{B} = \langle B_i\rangle_{i\in\{1,2,\dots M\}}$, each of which also corresponds to a confidence interval. We will use c_i as a shorthand for $c_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_i)$ and a_i as a shorthand for $a_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_i, \hat{y}_i)$. We denote by $b_i(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = b_i$ the bin to which c_i belongs. We define the size of bin j as $S_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = S_j$, the average confidence of bin j as $C_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = C_j$ and the average accuracy of bin j as $A_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = A_j$. These are expressed as follows:

$$S_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = |\{i|b_i = j\}| \tag{3}$$

$$C_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \tfrac{1}{S_j}\Sigma_{i|b_i=j}c_i \tag{4}$$

$$A_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \tfrac{1}{S_j}\Sigma_{i|b_i=j}a_i. \tag{5}$$

We are now in a position to define the Expected Binned Calibration Error of order p which we denote by $\mathrm{ECE}_{bin,p}$ and Expected Label-Binned Calibration Error of order p we denote by $\mathrm{ECE}_{lb,p}$. These serve as empirical approximations to the corresponding intractable ideal notions from equations 1 and 2. They are defined as follows:

$$\mathrm{ECE}_{bin,p}(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \left(\Sigma_{i=1}^{M}\tfrac{S_j}{\hat{N}}\cdot|A_j - C_j|^p\right)^{1/p} \tag{6}$$

$$\mathrm{ECE}_{lb,p}(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \left(\tfrac{1}{\hat{N}}\Sigma_{i=1}^{\hat{N}}|A_{b_i} - c_i|^p\right)^{1/p}. \tag{7}$$

It follows from Jensen's inequality that $\mathrm{ECE}_{lb,p}(\mathcal{B}, D, \boldsymbol{\theta}) \geq \mathrm{ECE}_{bin,p}(\mathcal{B}, D, \boldsymbol{\theta})$ [Roelofs et al., 2020].

4 Soft Calibration Objectives

In this section, we define quantities that can be used to better incentivize calibration during training.

4.1 Soft-Binned ECE (SB-ECE)

The quantities in the definitions of $\mathrm{ECE}_{bin,p}$ and $\mathrm{ECE}_{lb,p}$ can be written in terms of a formal definition of the bin membership function. Let us denote the bin-membership function for a given binning $\mathcal{B} = \langle B_i \rangle_{i \in \{1,2,\dots M\}}$ by $\boldsymbol{u}_{\mathcal{B}} : [0,1] \to \mathcal{U}_M$, where $\mathcal{U}_M = \{\boldsymbol{v} \in [0,1]^M | \Sigma_j v_j = 1\}$ is the set of possible bin membership vectors over M bins. The membership function for bin j is denoted by $u_{\mathcal{B},j} : [0,1] \to [0,1]$ and is defined by $u_{\mathcal{B},j}(c) = \boldsymbol{u}_{\mathcal{B}}(c)_j$. The size, average accuracy, and average confidence of bin j from equations 3, 4, and 5 can now be written in terms of $\boldsymbol{u}_{\mathcal{B}}$ as follows:

$$S_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \Sigma_{i=1}^{\hat{N}} u_{\mathcal{B},j}(c_i) \tag{8}$$

$$C_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \tfrac{1}{S_j} \Sigma_{i=1}^{\hat{N}} (u_{\mathcal{B},j}(c_i) \cdot c_i) \tag{9}$$

$$A_j(\mathcal{B}, \hat{D}, \boldsymbol{\theta}) = \tfrac{1}{S_j} \Sigma_{i=1}^{\hat{N}} (u_{\mathcal{B},j}(c_i) \cdot a_i). \tag{10}$$

The quantities $\mathrm{ECE}_{bin,p}$ and $\mathrm{ECE}_{lb,p}$ can further be written in terms of the quantities S_j, C_j and A_j using equation 6 and a modification of equation 7 (see equation 12 below). We know that the differentials $\partial\mathrm{ECE}_{bin,p}/\partial\boldsymbol{\theta}$ and $\partial\mathrm{ECE}_{lb,p}/\partial\boldsymbol{\theta}$ are non-trainable. The formulation above makes it clear that this is precisely because $\partial u_{\mathcal{B},j}/\partial c$ is zero within bin boundaries and undefined at bin boundaries. Moreover, this observation implies that if we could come up with a trainable soft bin-membership function $\boldsymbol{u}^*_{\mathcal{B}}$ then we could use it in place of the usual hard bin-membership function $\boldsymbol{u}_{\mathcal{B}}$ to obtain a trainable version of $\mathrm{ECE}_{bin,p}$ and $\mathrm{ECE}_{lb,p}$.

With this motivation, we define the soft bin-membership function that has a well-defined non-zero gradient in $(0,1)$. It is parameterized by the number of bins M and a temperature parameter T. We consider equal-width binning for simplicity and so we represent it as $\boldsymbol{u}^*_{M,T}$ rather than $\boldsymbol{u}^*_{\mathcal{B}}$. We desire unimodality over confidence: if ξ_j denotes the center of bin j then we want $\partial u^*_{M,T,j}/\partial c$ to be positive for $c < \xi_j$ negative for $c > \xi_j$. Similarly, we also desire unimodality over bins: if $c < \xi_i < \xi_j$ or $c > \xi_i > \xi_j$, then we want that $u^*_{M,T,i}(c) > u^*_{M,T,j}(c)$. Finally, we also want the aforementioned temperature parameter T to control how close the binning is to hard binning (i.e. how steeply membership drops off). This would give us the nice property of hard-binning being a limiting condition of soft-binning. With this motivation, we define the soft bin-membership function as

$$\begin{aligned} \boldsymbol{u}^*_{M,T}(c) &= \mathrm{softmax}(\boldsymbol{g}_{M,T}(c)), \\ \text{where } g_{M,T,i}(c) &= -(c - \xi_i)^2/T \qquad \forall i \in \{1, 2, \dots M\}. \end{aligned}$$

Figure 3 visualizes soft bin-membership. We can now formulate trainable calibration error measures. We define the Expected Soft-Binned Calibration Error $\text{SB-ECE}_{\text{bin},p}(M, T, \hat{D}, \boldsymbol{\theta})$ and Expected Soft-Label-Binned Calibration Error $\text{SB-ECE}_{\text{lb},p}(M, T, \hat{D}, \boldsymbol{\theta})$:

$$\text{SB-ECE}_{\text{bin},p}(M, T, \hat{D}, \boldsymbol{\theta}) = \left(\Sigma_{i=1}^{M} (\tfrac{S_j}{\hat{N}} |A_j - C_j|^p) \right)^{1/p}, \tag{11}$$

$$\text{SB-ECE}_{\text{lb},p}(M, T, \hat{D}, \boldsymbol{\theta}) = \left(\tfrac{1}{\hat{N}} \Sigma_{i=1}^{\hat{N}} \Sigma_{j=1}^{M} (u^*_{\mathcal{M},T,j}(c_i) \cdot |A_j - c_i|^p) \right)^{1/p}. \tag{12}$$

The quantities S_j, C_j and A_j in these expressions are obtained by using the soft bin membership function $\boldsymbol{u}^*_{M,T}$ in place of the hard bin membership function $\boldsymbol{u}_{\mathcal{B}}$ in equations 8, 9 and 10 respectively.

We use these trainable calibration error measures as: (1) part of the training loss function and (2) the objective that is minimized to tune the temperature scaling parameter T_{ts} during post-hoc calibration.

4.2 Soft AvUC (S-AvUC)

The accuracy versus uncertainty calibration (AvUC) loss [Krishnan and Tickoo, 2020] categorizes each prediction that a model with parameters $\boldsymbol{\theta}$ makes for a labelled datapoint $d_i = (\boldsymbol{x}_i, y_i)$ from dataset $\hat{D}$ according to two axes: (1) accurate [A] versus inaccurate [I], based on the value of the boolean quantity $a_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i) = a_i$ (2) certain [C] versus uncertain [U], based on whether the entropy

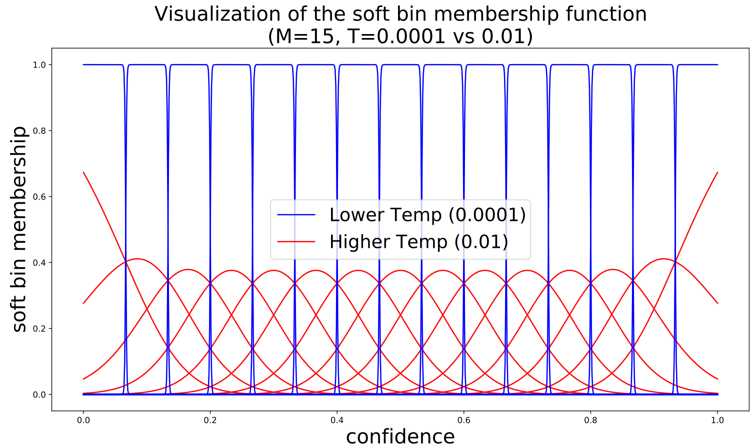


Figure 3: Visualization of the soft bin membership function which shows that the temperature parameter determines the sharpness of the binning. Soft binning limits to hard binning as temperature tends to zero.

$h(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = h_i$ of the predictive distribution is above or below a threshold κ. Denote the number of elements from $\hat{D}$ that fall in each of these 4 categories by $\hat{n}_{\text{AC}}$, $\hat{n}_{\text{AU}}$, $\hat{n}_{\text{IC}}$ and $\hat{n}_{\text{IU}}$ respectively. The AvUC loss incentivizes the model to be certain when accurate and uncertain when inaccurate:

$$\text{AvUC}(\kappa, \hat{D}, \boldsymbol{\theta}) = \log\left(1 + \frac{n_{\text{AU}} + n_{\text{IC}}}{n_{\text{AC}} + n_{\text{IU}}}\right), \tag{13}$$

where the discrete quantities are relaxed to be differentiable:

$$\begin{aligned} n_{\text{AU}} &= \Sigma_{i|(\boldsymbol{x}_i,y_i)\in S_{\text{AU}}}(c_i \tanh h_i) & n_{\text{IC}} &= \Sigma_{i|(\boldsymbol{x}_i,y_i)\in S_{\text{IC}}}((1-c_i)(1-\tanh h_i)) \\ n_{\text{AC}} &= \Sigma_{i|(\boldsymbol{x}_i,y_i)\in S_{\text{AC}}}(c_i(1-\tanh h_i)) & n_{\text{IU}} &= \Sigma_{i|(\boldsymbol{x}_i,y_i)\in S_{\text{IU}}}((1-c_i)\tanh h_i). \end{aligned} \tag{14}$$

Krishnan and Tickoo [2020] have showed good calibration results using the AvUC loss in SVI settings. However, in our experiments we found that the addition of the AvUC loss term resulted in poorly calibrated models in non-SVI neural network settings (see Appendix A). One reason for this seems to be that minimizing the AvUC loss results in the model being incentivized to be even more confident in its inaccurate and certain predictions (via minimizing n_{IC}) and even less confident in its accurate and uncertain predictions (via minimizing n_{AU}). This conjecture is validated by experimental observations: when we stopped the gradients flowing through the c_i terms in equation 14, we were able to obtain calibrated models (see Appendix F). Fixing this incentivization issue in a more principled manner than stopping gradients is desirable. Another desirable improvisation is replacing the hard categorization into the certain/uncertain bins with a soft partitioning scheme. We meet both these objectives by defining a notion of soft uncertainty.

We want a limiting case of the soft uncertainty function to be the hard uncertainty function based on an entropy threshold κ. This implies that we will continue to have a parameter κ despite getting rid of the hard threshold. As before, we desire a temperature parameter T that will determine how close the function is to the hard uncertainty function. The soft uncertainty function $t_{\kappa,T} : [0,1] \to [0,1]$ takes as input the $[0,1]$-normalized entropy $h_i^* = h_i/\log(K)$ of the predicted posterior where K is the number of classes. We also need $\partial t_{\kappa,T}/\partial h^*$ to be positive in $[0,1]$ and would like $t_{\kappa,T}$ to satisfy $\lim_{h^*\to 0} t_{\kappa,T}(h^*) = 0$ and $\lim_{h^*\to 1} t_{\kappa,T}(h^*) = 1$. Finally, it would be good to have the $[0,1]$ identity mapping as a special case of $t_{\kappa,T}$-family for some value of κ and T. We now define the soft-uncertainty function in the following way so that it meets all stated desiderata:

$$t_{\kappa,T}(h^*) = \text{logistic}\left(\frac{1}{T}\log\frac{h^*(1-\kappa)}{(1-h^*)\kappa}\right).$$

Finally, we define Soft AvUC in terms of soft uncertainty by modifying equations 13 and 14. In our experiments, we use Soft AvUC as part of the loss function to obtain calibrated models.

$$\text{S-AvUC}(\kappa, T, \hat{D}, \boldsymbol{\theta}) = \log\left(1 + \frac{n'_{\text{AU}} + n'_{\text{IC}}}{n'_{\text{AC}} + n'_{\text{IU}}}\right), \tag{15}$$

where

$$\begin{aligned} n'_{\text{AU}} &= \Sigma_{i|a_i=1}(t_{\kappa,T}(h_i^*)\tanh h_i) & n'_{\text{IC}} &= \Sigma_{i|a_i=0}((1-t_{\kappa,T}(h_i^*))(1-\tanh h_i)) \\ n'_{\text{AC}} &= \Sigma_{i|a_i=1}((1-t_{\kappa,T}(h_i^*))(1-\tanh h_i)) & n'_{\text{IU}} &= \Sigma_{i|a_i=0}(t_{\kappa,T}(h_i^*)\tanh h_i). \end{aligned} \tag{16}$$

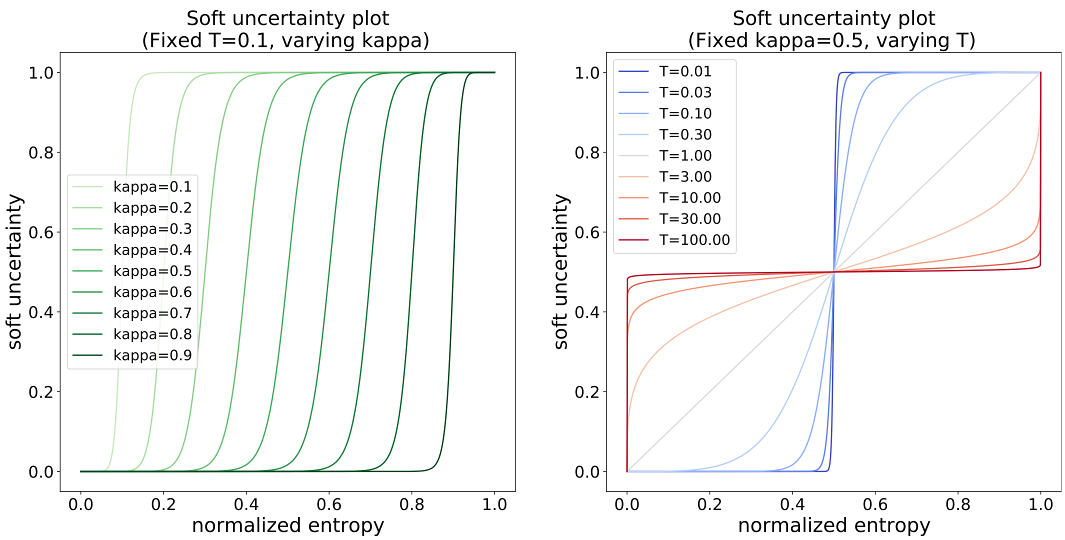


Figure 4: Visualization of the soft uncertainty function $t_{\kappa,T}(h^*)$ which shows that the parameter κ captures the soft-threshold whereas the parameter T captures the sharpness of the thresholding.

5 Results

We compare our Soft Calibration Objectives to recently proposed calibration-incentivizing training objectives MMCE, focal loss, and AvUC on the CIFAR-10, CIFAR-100, and ImageNet datasets. We evaluate the full cross-product of primary and secondary losses: the options for primary loss are cross-entropy (NLL), focal or mean squared error (MSE) loss; and the options for secondary loss are MMCE, AvUC, SB-ECE or S-AvUC. Results for the MSE primary loss and the AvUC secondary loss are in Appendix A. Our experiments build on the Uncertainty Baselines and Robustness Metrics libraries [Nado et al., 2021, Djolonga et al., 2020].

5.1 Soft Calibration Objectives for End-to-End Training

Our results demonstrate that training losses which include Soft Calibration Objectives obtain state-of-the-art single-model ECE on the test set in exchange for less than 1% reduction in accuracy for all three datasets that we experiment with. In fact, our methods (especially S-AvUC) without post-hoc temperature scaling are better than or as good as other methods with or without post-hoc temperature scaling on all three datasets.

The primary losses we work with for CIFAR-10/100 are the cross-entropy (NLL) loss, the focal loss and mean squared error (MSE) loss. Focal loss [Mukhoti et al., 2020] and MSE loss [Hui and Belkin, 2021] have recently shown to outperform the NLL loss in certain settings. The cross-entropy loss outperforms the other two losses on Imagenet, and is thus our sole focus for this dataset.

The primary loss even by itself (especially NLL) can overfit to the train ECE [Mukhoti et al., 2020], without help from the soft calibration losses. Even in such settings, we show that soft calibration losses yield reduction of test ECE using a technique we call 'interleaved training' (Appendix B).

We use the Wide-Resnet-28-10 architecture [Zagoruyko and Komodakis, 2017] trained for 200 epochs on CIFAR-100 and CIFAR-10. For Imagenet, we use the Resnet-50 [He et al., 2015] architecture training for 90 epochs. All our experiments use the SGD with momentum optimizer with momentum fixed to 0.9 and learning rate fixed to 0.1. The loss function we use in our experiments is PL $+ \beta \cdot$ SL $+ \lambda \cdot$ L2 where PL and SL denote the primary and secondary losses respectively and L2 denotes the weight normalization term with ℓ_2 norm. We tune the β and λ parameters along with the parameters κ and T relevant to the secondary losses SB-ECE$_{lb,p}(M, T, \hat{D}, \boldsymbol{\theta})$ and S-AvUC$(\kappa, T, \hat{D}, \boldsymbol{\theta})$. We tune these hyperparameters sequentially. We fix the learning rate schedule and the number of bins M to keep the search space manageable. Appendix G has more details of our hyperparameter search.

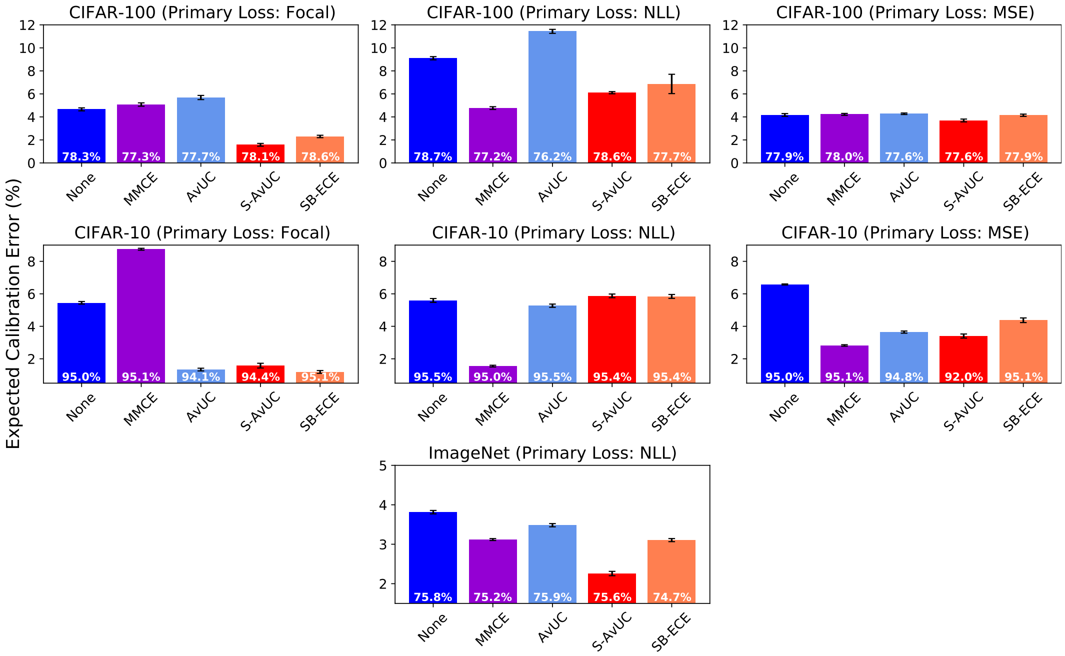


Figure 5: Soft Calibration Objectives (S-AvUC, SB-ECE), when used as secondary losses with a primary loss, achieve lower ECE (equal-mass binning, ℓ_2 norm) than the corresponding primary loss for CIFAR-10, CIFAR-100, and ImageNet. These statistically significant wins come at the cost of less than 1% accuracy (reported at bottom of the bar). Values reported are mean over 10 runs, and the error bars indicate ± 1 standard error of the mean (SEM). For each dataset (each row) the best (across primary losses) ECE obtained using the S-AvUC and SB-ECE secondary losses is lower than the best ECE obtained using existing techniques.

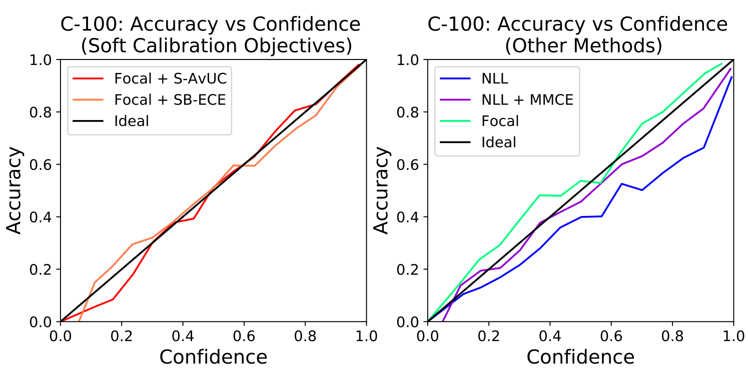


Figure 6: Accuracy vs Confidence plots for various methods on CIFAR-100. NLL is significantly overconfident and NLL + MMCE is somewhat overconfident. While Focal loss is underconfident, augmenting it with Soft Calibration Objectives fixes this issue, resulting in curves closest to the ideal.

In Table 1, we report the runs with the best ECE values that also came within 1% of the primary loss run with the highest accuracy. Figure 6 is the accuracy-confidence plot corresponding to Table 1a. In Figure 5, we visualize the ECE on the test set for all combinations of primary loss, secondary loss and dataset. The best ECE for each dataset is attained using Soft Calibration Objectives as secondary losses. More such figures are in Appendix D and the complete table can be found in Appendix A.

5.2 Soft Calibration Objectives for Post-Hoc Calibration

Standard temperature scaling (TS) uses a cross-entropy objective to optimize the temperature. However, using our differentiable soft binning calibration objective (SB-ECE), we can optimize the temperature using a loss function designed to directly minimize calibration error. In Figure 2 (and Figure 9 in Appendix C), we compare temperature scaling with a soft binning calibration objective (TS-SB-ECE) to standard temperature scaling with a cross-entropy objective (TS) on out-of-distribution shifts of increasing magnitude on both CIFAR-10 and CIFAR-100. The distribution

Table 1: We report average accuracy (with standard error across 10 trials), ECE, and ECE obtained after post-hoc temperature scaling (TS) for models trained with different objectives on the CIFAR-10, CIFAR-100, and ImageNet datasets. ECE is computed with the ℓ_2 norm and equal-mass binning. We find Soft Calibration Objectives (SB-ECE, S-AvUC) result in better or equivalent ECEs compared to previous methods with or without TS. We also find that TS does not always improve ECE. The best ECE value is highlighted for each dataset. The best ECE with TS value is highlighted if it improves over the best value from the ECE column.

(a) CIFAR-100

Loss Fn.	Accuracy	ECE	ECE with TS
NLL	78.7±0.122	9.10±0.139	5.36±0.091
NLL + MMCE	77.2±0.072	4.77±0.121	4.06±0.138
Focal	78.3±0.086	4.66±0.130	6.47±0.140
Focal + SB-ECE	78.6±0.062	2.30±0.105	5.16±0.108
Focal + S-AvUC	78.1±0.084	**1.57**±0.122	4.15±0.090

(b) CIFAR-10

Loss Fn.	Accuracy	ECE	ECE with TS
NLL	95.5±0.040	5.59±0.119	1.95±0.127
NLL + MMCE	95.0±0.031	1.55±0.053	**1.09**± 0.098
Focal	95.0± 0.085	5.45±0.079	2.69±0.190
Focal + SB-ECE	95.1± 0.056	**1.19**± 0.088	2.08± 0.143
Focal + S-AvUC	94.4± 0.145	1.58± 0.146	1.34± 0.172

(c) ImageNet

Loss Fn.	Accuracy	ECE	ECE with TS
NLL	75.8± 0.036	3.81± 0.043	2.17± 0.045
NLL + MMCE	75.2± 0.048	3.12± 0.025	2.18± 0.029
NLL + SB-ECE	74.7 ± 0.028	3.11 ± 0.039	**1.92**± 0.024
NLL + S-AvUC	75.6 ± 0.053	**2.26** ± 0.055	2.02 ± 0.041

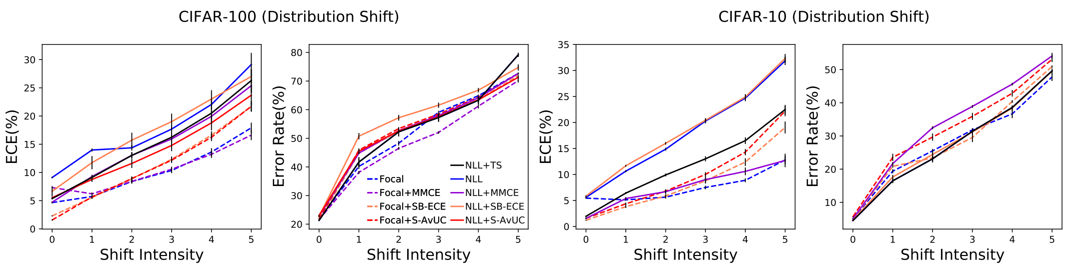


Figure 7: For the CIFAR-10/100 datasets, methods which train for calibration outperform the popular methods - NLL and NLL + TS - under distribution shift. Focal primary loss and MMCE secondary loss result in the lowest ECEs under shift. We note that these methods start off with worse ECEs than our SB-ECE and S-AvUC methods on the test set but end up with better ECE under increasing levels of skew. The OOD datasets that we have used here are CIFAR-10/100-C with skew levels 1-5.

shifts come from either the CIFAR-10-C or CIFAR-100-C datasets. Whereas Figure 2 focuses on cross-entropy loss, Figure 9 also has the plots for other primary losses. Table 2 contains comparisons between the two methods based on test ECE for all combinations of dataset, primary loss and secondary loss. We find that TS-SB-ECE outperforms TS under shift in most cases, and the performance increase is similar across shifts of varying magnitude. Note that temperature scaling (TS) does not always improve ECE, especially when the training loss is different from NLL. In such cases TS-SB-ECE still outperforms TS but may or may not result in ECE improvement.

5.3 Training for Calibration Under Distribution Shift

In previous sections we have shown that training for calibration outperforms the popular cross-entropy loss coupled with post-hoc TS on the in-distribution test set. We find that methods which train for calibration (not always our proposed methods) also outperform the cross-entropy loss with TS under dataset shift. Prior work has shown that temperature scaling can perform poorly under distribution shift [Ovadia et al., 2019], and our experiments reproduce this issue. Moreover, we show that training for calibration makes progress towards fixing this problem. However, different methods perform best under distribution shift on different datasets. Whereas S-AvUC does well on ImageNet OOD (see figure 8), Focal loss does better than our methods on CIFAR-10-C and CIFAR-100-C (see figure 7). We cannot prescribe one method in particular under distribution shift given these results but we have shown a crucial benefit of using methods to train for calibration as a whole.

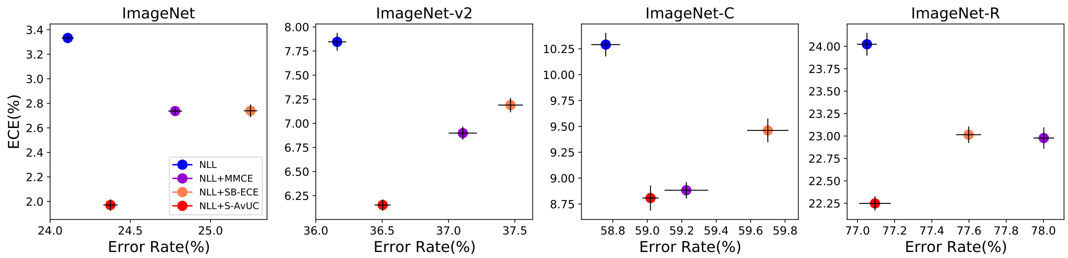


Figure 8: A comparison of methods for ImageNet (primary loss: cross-entropy) shows that the S-AvUC secondary loss yields lowest ECEs under dataset shift. Error bars are ± 1 SEM over 10 runs.

6 Conclusions

We proposed Soft Calibration Objectives motivated by the goal of directly training models for calibration. These objectives - SB-ECE and S-AvUC - are softened versions of the usual ECE measure and the recently-proposed AvUC loss, respectively. They are easy-to-implement augmentations to the popular cross-entropy loss. We performed a thorough comparison of existing methods of training for calibration. Our experiments show that methods based on soft-calibration objectives can be used to obtain the best ECE among such methods in exchange for less than 1% drop in accuracy. We note that a model being better calibrated overall does not necessarily mean that it is better calibrated for every group and hence the fairness of our methods as well as related methods must be studied. However, our methods of training-for-calibration can be adapted to encourage fairness by applying the methods separately to each protected group.

Even when one does not wish to incorporate secondary losses to train for calibration, we showed that post-hoc temperature scaling works better when tuned using the SB-ECE objective instead of the standard cross-entropy loss. Practitioners can easily replace the cross-entropy loss with our SB-ECE loss when performing post-hoc temperature scaling.

Finally, we demonstrated that the uncertainty estimates of methods which train for calibration generalize better under dataset shift as compared to post-hoc calibration, which is a fundamental motivation for transitioning to training for calibration.

Acknowledgements

The authors thank Brennan McConnell and Mohammad Khajah who conducted initial explorations of soft binning calibration loss. The authors also thank Zack Nado, D. Sculley and Jeremiah Liu for help with implementation and suggestions for the writeup.

References

Mari-Liis Allikivi and Meelis Kull. Non-parametric Bayesian isotonic calibration: Fighting over-confidence in binary classification. In *ECML/PKDD*, 2019.

Mariusz Bojarski, D. Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, L. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, X. Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *ArXiv*, abs/1604.07316, 2016.

R. Caruana, Yin Lou, J. Gehrke, Paul Koch, M. Sturm, and Noémie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Terrance Devries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *ArXiv*, abs/1802.04865, 2018.

Josip Djolonga, Minderer Matthias, Zack Nado, Jeremy Nixon, Rob Romijnders, Dustin Tran, and Mario Lucic. Robustness Metrics, 2020. URL https://github.com/google-research/robustness_metrics.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017.

Kartik Gupta, Amir M. Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, C. Sminchisescu, and R. Hartley. Calibration of neural networks using splines. *ArXiv*, abs/2006.12800, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks, 2021.

Xiaoqian Jiang, M. Osl, J. Kim, and L. Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 19:263 – 274, 2012.

Simon Kocbek, Primoz Kocbek, Leona Cilar, and Gregor Stiglic. Local interpretability of calibrated prediction models: A case of type 2 diabetes mellitus screening test. *arXiv preprint arXiv:2006.13815*, 2020.

R. Krishnan and O. Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *ArXiv*, abs/2012.07923, 2020.

Volodymyr Kuleshov and S. Ermon. Estimating uncertainty online against an adversary. In *AAAI*, 2017.

Meelis Kull, Telmo de Menezes e Silva Filho, and Peter A. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, 2017.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *NeurIPS*, 2019.

A. Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, 2018.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *NeurIPS*, 2019.

Balaji Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

M. E. J. Masson and G. R. Loftus. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57:203–220, 2003. URL http://web.uvic.ca/psyc/masson/ML03.pdf.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, S. Golodetz, P. Torr, and P. Dokania. Calibrating deep neural networks using focal loss. *ArXiv*, abs/2002.09437, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim Rudner, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning, 2021. URL https://github.com/google/uncertainty-baselines.

M. Naeini and G. Cooper. Binary classifier calibration using an ensemble of near isotonic regression models. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 360–369, 2016.

M. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 2015.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

Yaniv Ovadia, E. Fertig, J. Ren, Zachary Nado, D. Sculley, S. Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration, 2017.

R. Roelofs, N. Cain, Jonathon Shlens, and M. Mozer. Mitigating bias in calibration error estimation. *ArXiv*, abs/2012.08668, 2020.

Zhihui Shao, Jianyi Yang, and Shaolei Ren. Calibrating deep neural network classifiers on out-of-distribution datasets. *ArXiv*, abs/2006.08914, 2020.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

Jonathan Wenger, H. Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *AISTATS*, 2020.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML*, 2001.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017.

A Complete Table of Results

We perform thorough experimentation for the full cross-product of primary and secondary losses for each dataset. We tune several hyperparameters (section G) for each of these settings. For the CIFAR-100 and CIFAR-10 datasets we consider three primary loss functions: cross-entropy, MSE, and Focal. For ImageNet, we consider only the cross-entropy primary loss since the other two did not match its performance in our experiments. We consider five secondary losses: MMCE [Kumar et al., 2018], AvUC [Krishnan and Tickoo, 2020], AvUC-GS (section F), SB-ECE (section 4.1), and S-AvUC (section 4.2). We also consider the setting with no secondary loss. This leads to 18 configurations (3×6) for CIFAR-100, 18 configurations (3×6) for CIFAR-10 and 6 configurations (1×6) for ImageNet for a total of 42 hyperparameter tunings. The results for the "best" (section G) hyperparameter configuration for each of these settings are listed in Table 2.

In order to measure calibration of a model we consider (1) the ECE of the trained model, (2) the ECE with standard post-hoc temperature scaling, and (3) the ECE with post-hoc temperature scaling for the SB-ECE objective. The ECE measured here is ℓ_2 norm equal-mass ECE with 15 bins. We showed in Figure 1 that the S-AvUC loss performs best aggregated across settings of datasets and primary losses. However, there isn't one secondary loss that is always the best for every setting. Even so, we do see that for all three datasets, the best ECE values for the trained model result from Soft Calibration Objectives used as secondary losses. These wins come at the cost of less than 1% accuracy. Soft Calibration Objectives are also either the best or equivalent to other methods for all datasets when we consider the numbers after temperature scaling. Note that post-hoc temperature scaling doesn't always help and should be used only if applying it results in better calibration than the trained model on some held out dataset.

B Train Set Memorization and Interleaved Training

The NLL primary loss has recently shown to heavily overfit the train ECE [Mukhoti et al., 2020] in some settings. In these cases, it essentially memorizes the train set, achieving near-perfect accuracy and calibration on it without help from any calibration-incentivizing losses. This raises a question about the effectiveness of using soft calibration during training for reducing test ECE in such settings.

We saw this happen on 2 of our 7 dataset + primary loss settings (CIFAR-100/10 datasets with the NLL primary loss; see table 3). As expected, using Soft Calibration Objectives as secondary training losses did not help reduce train ECE here. To fix this issue, we modified the training procedure. The train set was split into two - the 'majority' train set and the 'held-out' train set. Each epoch was also correspondingly split into two - the first part optimized NLL on the majority train set and the second part optimized Soft Calibration Objectives on the held-out train set. This way we avoided incentivizing something that was already overfit. The second part of each epoch incentivized the predicted distribution to have higher entropy. We call this 'interleaved training'.

We observed that Soft Calibration Objectives with interleaved training helped reduce test ECE relative to baseline on CIFAR-100 + NLL, as can be seen in table 2. For a fixed primary loss, the recommendation to practitioners is to either (1) have a calibration dataset and use interleaving if it yields better ECE on it or (2) in absence of a calibration dataset, use interleaving if train ECE is suspiciously low (e.g. less than 1%). Another approach is to replace the primary loss which overfits (e.g. NLL) with one that doesn't (e.g. Focal) and then use Soft Calibration Objectives for further gains.

C Soft Calibration Objectives for Post-Hoc Temperature Scaling

We have seen in Figure 2 that temperature scaling for the SB-ECE objective (TS-SB-ECE) outperforms standard temperature scaling (TS) both in- and out-of-distribution for models trained with the popular NLL loss on the CIFAR-100 and CIFAR-10 datasets. We see in figure 9 that this also holds true for the Focal and MSE primary losses. Whereas temperature scaling does not always help to improve calibration, particularly out-of-distribution, we do see that TS-SB-ECE outperforms standard TS in most cases. We plot comparisons for only 6 of our 42 settings in figure 9 for the sake of conciseness - these are for the CIFAR-100 and CIFAR-10 datasets for models trained using each of the three primary losses. The plots demonstrate that TS-SB-ECE outperforms TS on the i.i.d. test

Table 2: Results for the full cross-product of experimental settings between primary and secondary losses for CIFAR-100, CIFAR-10 and ImageNet. We report average accuracy, ECE and ECE obtained after post-hoc temperature scaling (TS, TS-SB-ECE) for each of the 42 configurations after hyperparameter tuning. These metrics corresponding to the best hyperparameter configuration for that setting. ECE is computed with the ℓ_2 norm and equal-mass binning. Each cell reports average $\pm$ SEM across 10 runs. The best ECE numbers for each dataset in each of the three ECE columns are highlighted. We see that the best ECE values for all three datasets are obtained using Soft Calibration Objectives as secondary losses.

			Accuracy	ECE	ECE+TS	ECE+TS-SB-ECE
Dataset	Primary	Secondary				
CIFAR-100	NLL	<none>	78.7 $\pm$ 0.122	9.10 $\pm$ 0.139	5.36 $\pm$ 0.091	4.69 $\pm$ 0.069
		MMCE	77.2 $\pm$ 0.072	4.77 $\pm$ 0.121	4.06 $\pm$ 0.138	4.11 $\pm$ 0.173
		AvUC	76.2 $\pm$ 0.146	11.4 $\pm$ 0.175	4.40 $\pm$ 0.190	7.64 $\pm$ 0.187
		AvUC-GS	78.3 $\pm$ 0.126	6.21 $\pm$ 0.084	8.36 $\pm$ 0.166	5.40 $\pm$ 0.241
		S-AvUC	78.6 $\pm$ 0.079	6.10 $\pm$ 0.095	9.13 $\pm$ 0.085	5.88 $\pm$ 0.243
		SB-ECE	77.7 $\pm$ 0.167	6.86 $\pm$ 0.839	**3.18** $\pm$ 0.093	3.51 $\pm$ 0.314
	Focal	<none>	78.3 $\pm$ 0.086	4.66 $\pm$ 0.130	6.47 $\pm$ 0.140	4.84 $\pm$ 0.115
		MMCE	77.3 $\pm$ 0.104	5.07 $\pm$ 0.147	3.58 $\pm$ 0.098	**2.82** $\pm$ 0.110
		AvUC	77.7 $\pm$ 0.145	5.68 $\pm$ 0.181	5.87 $\pm$ 0.182	3.93 $\pm$ 0.209
		AvUC-GS	78.1 $\pm$ 0.056	3.38 $\pm$ 0.109	4.89 $\pm$ 0.118	2.91 $\pm$ 0.122
		S-AvUC	78.1 $\pm$ 0.084	**1.57** $\pm$ 0.122	4.15 $\pm$ 0.090	2.89 $\pm$ 0.077
		SB-ECE	78.6 $\pm$ 0.062	2.30 $\pm$ 0.105	5.16 $\pm$ 0.108	4.10 $\pm$ 0.095
	MSE	<none>	77.9 $\pm$ 0.089	4.17 $\pm$ 0.122	5.22 $\pm$ 0.126	4.79 $\pm$ 0.281
		MMCE	78.0 $\pm$ 0.071	4.22 $\pm$ 0.085	5.06 $\pm$ 0.108	4.90 $\pm$ 0.111
		AvUC	77.6 $\pm$ 0.134	4.27 $\pm$ 0.067	5.29 $\pm$ 0.125	4.42 $\pm$ 0.224
		AvUC-GS	77.8 $\pm$ 0.041	4.31 $\pm$ 0.114	5.03 $\pm$ 0.091	4.15 $\pm$ 0.135
		S-AvUC	77.6 $\pm$ 0.104	3.69 $\pm$ 0.122	4.60 $\pm$ 0.144	4.26 $\pm$ 0.322
		SB-ECE	77.9 $\pm$ 0.071	4.14 $\pm$ 0.100	5.00 $\pm$ 0.092	4.27 $\pm$ 0.184
CIFAR-10	NLL	<none>	95.5 $\pm$ 0.040	5.59 $\pm$ 0.119	1.95 $\pm$ 0.127	1.16 $\pm$ 0.106
		MMCE	95.0 $\pm$ 0.031	1.55 $\pm$ 0.053	**1.09** $\pm$ 0.098	1.45 $\pm$ 0.114
		AvUC	95.5 $\pm$ 0.053	5.27 $\pm$ 0.101	2.39 $\pm$ 0.110	1.30 $\pm$ 0.106
		AvUC-GS	95.6 $\pm$ 0.036	4.94 $\pm$ 0.109	2.07 $\pm$ 0.129	1.31 $\pm$ 0.072
		S-AvUC	95.4 $\pm$ 0.027	5.87 $\pm$ 0.113	2.24 $\pm$ 0.130	1.20 $\pm$ 0.115
		SB-ECE	95.4 $\pm$ 0.043	5.84 $\pm$ 0.121	2.28 $\pm$ 0.091	1.27 $\pm$ 0.139
	Focal	<none>	95.0 $\pm$ 0.085	5.45 $\pm$ 0.079	2.69 $\pm$ 0.190	1.77 $\pm$ 0.115
		MMCE	95.1 $\pm$ 0.068	8.74 $\pm$ 0.059	3.13 $\pm$ 0.110	2.16 $\pm$ 0.210
		AvUC	94.1 $\pm$ 0.128	1.34 $\pm$ 0.084	2.53 $\pm$ 0.124	1.12 $\pm$ 0.133
		AvUC-GS	95.2 $\pm$ 0.063	1.39 $\pm$ 0.081	2.30 $\pm$ 0.125	1.46 $\pm$ 0.108
		S-AvUC	94.4 $\pm$ 0.145	1.58 $\pm$ 0.146	1.34 $\pm$ 0.172	**1.05** $\pm$ 0.197
		SB-ECE	95.1 $\pm$ 0.056	**1.19** $\pm$ 0.088	2.08 $\pm$ 0.143	1.38 $\pm$ 0.187
	MSE	<none>	95.0 $\pm$ 0.041	6.58 $\pm$ 0.034	5.33 $\pm$ 0.122	4.43 $\pm$ 0.106
		MMCE	95.1 $\pm$ 0.034	2.82 $\pm$ 0.046	3.23 $\pm$ 0.137	2.34 $\pm$ 0.141
		AvUC	94.8 $\pm$ 0.050	3.64 $\pm$ 0.066	4.72 $\pm$ 0.139	4.01 $\pm$ 0.178
		AvUC-GS	94.8 $\pm$ 0.031	5.44 $\pm$ 0.146	5.56 $\pm$ 0.151	5.20 $\pm$ 0.173
		S-AvUC	92.0 $\pm$ 0.117	3.40 $\pm$ 0.126	1.87 $\pm$ 0.080	1.50 $\pm$ 0.134
		SB-ECE	95.1 $\pm$ 0.056	4.37 $\pm$ 0.143	5.08 $\pm$ 0.159	4.26 $\pm$ 0.136
ImageNet	NLL	<none>	75.8 $\pm$ 0.036	3.81 $\pm$ 0.043	2.17 $\pm$ 0.045	3.32 $\pm$ 0.043
		MMCE	75.2 $\pm$ 0.048	3.12 $\pm$ 0.025	2.18 $\pm$ 0.029	2.68 $\pm$ 0.022
		AvUC	75.9 $\pm$ 0.035	3.48 $\pm$ 0.041	3.37 $\pm$ 0.037	3.18 $\pm$ 0.036
		AvUC-GS	75.8 $\pm$ 0.035	3.84 $\pm$ 0.030	3.15 $\pm$ 0.028	3.44 $\pm$ 0.034
		S-AvUC	75.6 $\pm$ 0.053	**2.26** $\pm$ 0.055	2.02 $\pm$ 0.041	**1.92** $\pm$ 0.046
		SB-ECE	74.7 $\pm$ 0.028	3.11 $\pm$ 0.039	**1.92** $\pm$ 0.024	2.62 $\pm$ 0.039

Table 3: Overconfident training set memorization happens in 2 of our 7 dataset + primary loss settings. This is characterized by a very low train ECE (in bold) and a high ratio of train ECE to test ECE (in bold). ECE is computed with the ℓ_1 norm and equal-width binning. We use interleaved training with Soft Calibration Objectives as secondary losses in these 2 cases to reduce train ECE.

Dataset	Primary Loss	Test ECE	Train ECE	Test ECE / Train ECE
CIFAR-100	NLL	6.88%	**0.45%**	**15.14**
	Focal	4.21%	9.18%	0.46
	MSE	3.51%	2.49%	1.41
CIFAR-10	NLL	2.66%	**0.05%**	**49.82**
	Focal	5.28%	7.13%	0.74
	MSE	6.53%	9.82%	0.66
Imagenet	NLL	3.35%	4.34%	0.77

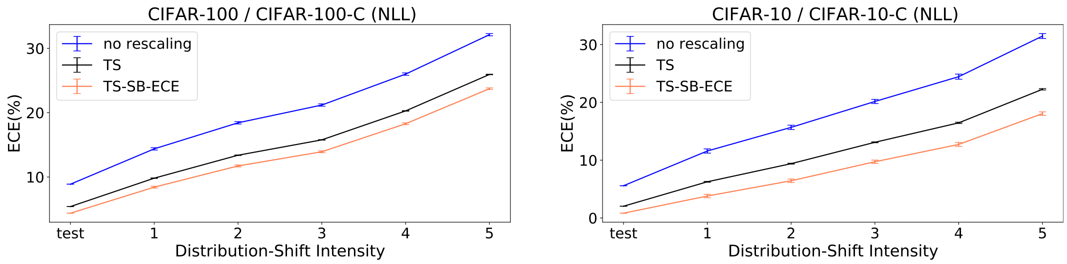
set and under all levels of skew in the CIFAR-100-C and CIFAR-10-C datasets. Table 2 shows that TS-SB-ECE outperforms TS in 35 of the 42 settings that we experiment on. We conclude that Soft Calibration Objectives can be used to improve upon TS - the popular post-hoc calibration method.

A pertinent follow-up question for temperature scaling is whether we can take this approach of directly optimizing temperature for SB-ECE a step further and directly optimize temperature for (hard-binned) ECE instead. ECE is not trainable, but we might still be able to do a search for temperature. Indeed, multi-resolution search to optimize temperature for hard-binned ECE (hereby, TS-HB-ECE) is an alternative to gradient-based training of temperature for soft-binned ECE (i.e. TS-SB-ECE). Our results suggest that this might be promising. This approach is worth investigating further.

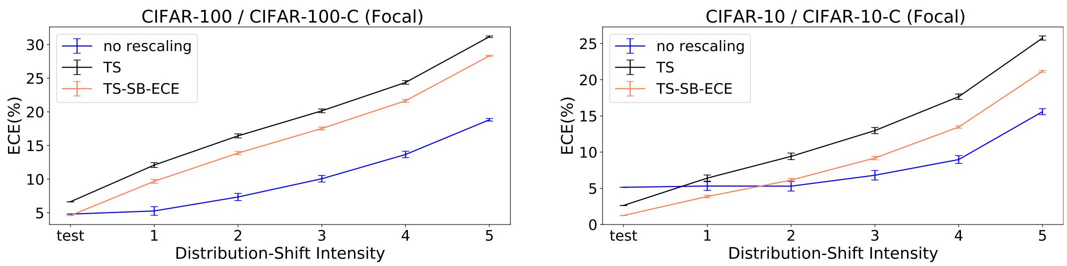
The two methods have potential advantages and disadvantages. TS-HB-ECE has higher variance than TS-SB-ECE for a given sample size, which risks a suboptimal solution even if TS-HB-ECE is the quantity we wish to optimize. TS-HB-ECE is also less computationally efficient than TS-SB-ECE. To formally compare complexities, assume that we want to compute top-label equal-width ECE and that we save logits in the last training epoch. Let K be the number of classes, N_c be size of the recalibration dataset and V be the number of temperature values inspected for multi-resolution search. Assuming that the gradient-based method trains for a small constant number of epochs, the post-processing complexity for TS-SB-ECE is $O(N_cK)$ and for TS-HB-ECE is $O(VN_cK)$. The quantity V may not be small if we want to get close to the optimal temperature. Consequently, this cost difference can be high for language models with large vocabularies and a lot of training data.

D Reliability Plots

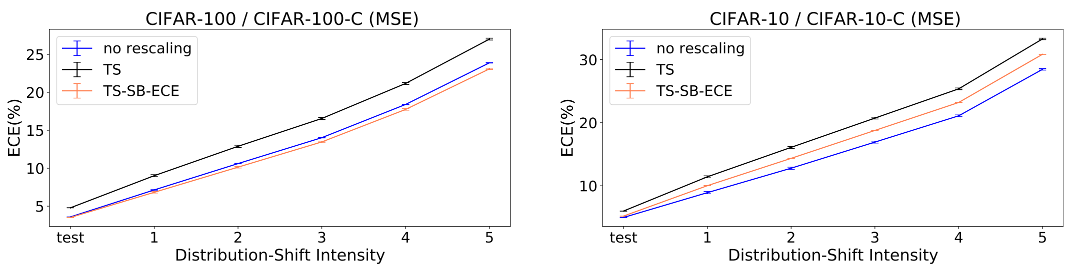
In this section, we use reliability plots (figure 10) to compare the calibration of models trained using different training objectives for CIFAR-100, CIFAR-10 and ImageNet. We compare the cross-entropy baseline with each of the proposed methods to train for calibration: Focal loss [Lin et al., 2018], MSE Loss [Hui and Belkin, 2021], MMCE loss [Kumar et al., 2018], S-AvUC and SB-ECE. We do not include the AvUC [Krishnan and Tickoo, 2020] loss since it does not help to improve calibration for non-Bayesian neural networks (section F). We find that NLL results in overconfident models and that adding MMCE as a secondary loss to the NLL primary loss reduces the amount of overconfidence. The MSE primary loss results in models that are overconfident for some confidence bins and underconfident for others, which helps to explain why temperature scaling is not as effective for MSE as compared to Focal and NLL (see table 2). Focal loss, on the other hand, results in underconfident models. This is also seen in the ECE vs average uncertainty plots (figure 10), where we find that NLL results in least average uncertainty for all datasets whereas Focal loss is amongst the highest average uncertainties for both CIFAR-100 and CIFAR-10. Obtaining both lower ECE and lower average uncertainty simultaneously as compared to the standard cross-entropy loss remains an open challenge. Finally we note that models trained using Soft Calibration Objectives as secondary losses are the most visually calibrated for all three datasets, consistent with our findings in Section 5.



(a) NLL



(b) Focal



(c) MSE

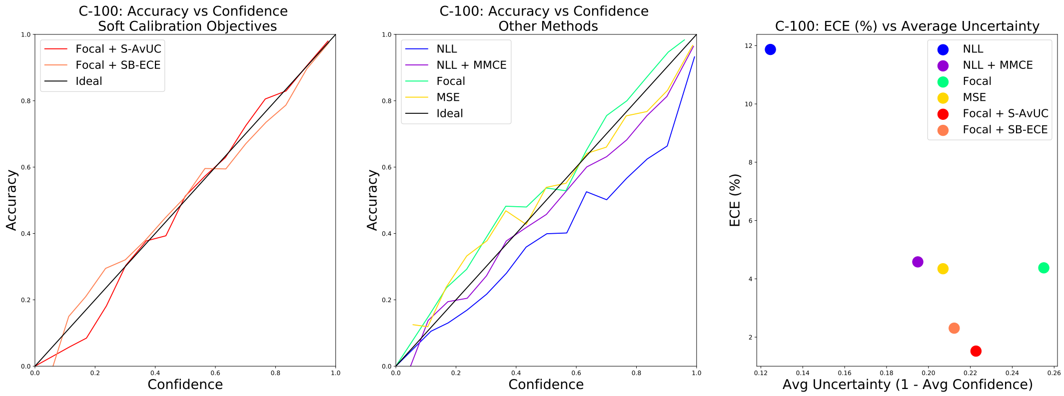
Figure 9: Post-hoc temperature scaling with the soft calibration error objective (TS-SB-ECE) outperforms standard post-hoc temperature scaling (TS), particularly under distribution shift. This result holds across datasets (left and right panels), distribution shift intensities (along abscissa) and training objectives (rows). The ECE value (equal-mass binning, ℓ_2 norm) shown is the mean ECE across all corruption types in CIFAR-10-C and CIFAR-100-C. Error bars are ± 1 standard error of mean (SEM), corrected for intrinsic variability due to type of corruption [Masson and Loftus, 2003]

.

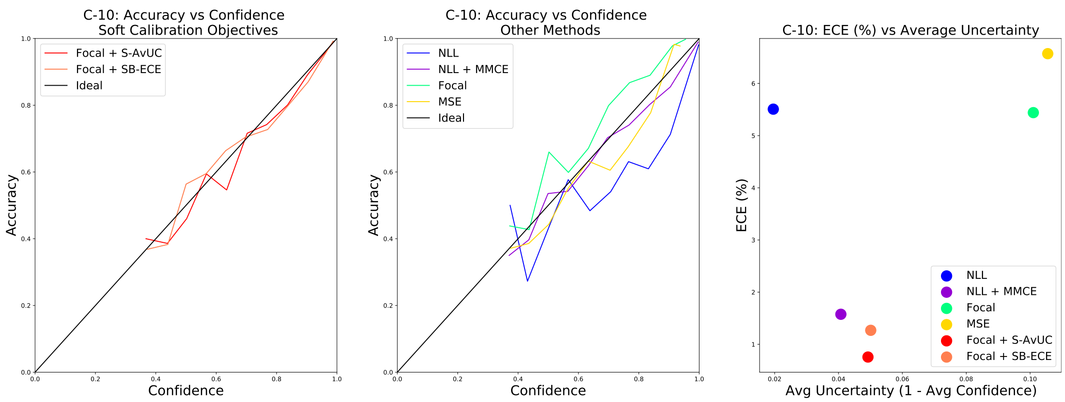
E Other Calibration Measures

Even though ECE is the most popular metric for measuring calibration, it has issues related to consistency and bias [Nixon et al., 2019, Kumar et al., 2019, Roelofs et al., 2020, Gupta et al., 2020]. Debiased CE [Kumar et al., 2019] and mean-sweep CE [Roelofs et al., 2020] have been shown to have lesser bias and more consistency across the number of bins parameter than ECE whereas KS-error [Gupta et al., 2020] avoids binning altogether.

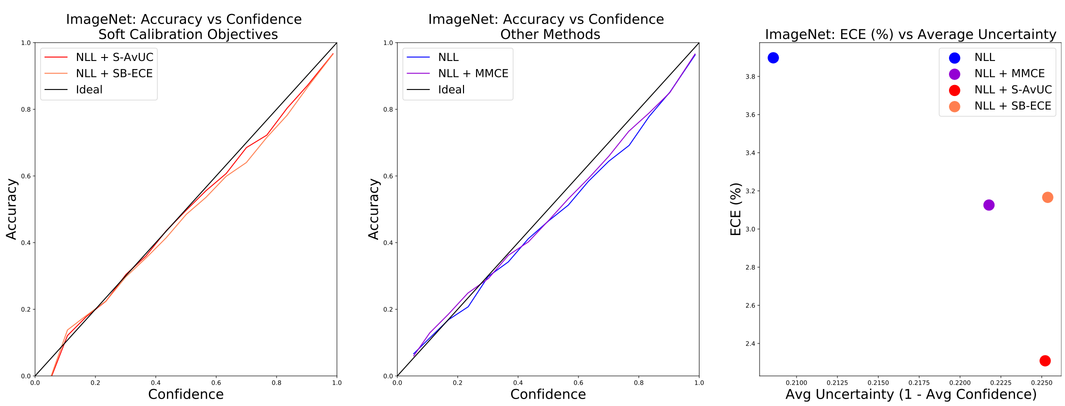
We have validated our findings on CIFAR-100 and CIFAR-10 using KS-error and mean-sweep CE. The findings based on these measures (table 4) are consistent with those based on ECE i.e. soft-calibration objectives outperform all other methods. We have reported results corresponding only to tables 1a and 1b for conciseness rather than those correponding to the full table 2. As stated before, these account for the best-performing experiments on these datasets and in particular the best results for the cross-entropy baseline, Focal loss [Mukhoti et al., 2020] and MMCE [Kumar et al., 2018].



(a) Reliability plots for C-100



(b) Reliability plots for C-10



(c) Reliability plots for ImageNet

Figure 10: Models trained with the NLL loss are overconfident across all three datasets. Adding MMCE as a secondary loss reduces the overconfidence. The MSE loss results in underconfidence and overconfidence for different bins. Models trained with Focal loss are underconfident. Soft Calibration Objectives (S-AvUC, SB-ECE) result in the most visually calibrated reliability plots across all datasets. Note that the high confidence regions have much higher density than the low confidence regions and are thus more critical to the ECE value.
In the ECE vs average uncertainty plots, we see that Focal and NLL losses result in the highest and lowest average uncertainties respectively. S-AvUC results in the lowest ECE in all three datasets and SB-ECE is the next best for CIFAR-10 and CIFAR-100. Obtaining lower ECE as well as lower average uncertainty as compared to the NLL loss remains an open challenge.

Table 4: We report accuracy, average ECE, KS-Error and mean-sweep CE for the CIFAR-10 and CIFAR-100 datasets corresponding to the entries in tables 1a and 1b. Mean-sweep CE is computed with the ℓ_2 norm and equal-mass binning. KS-error is computed as originally defined with the ℓ_1 norm. These additional measures further demonstrate that soft-calibration objectives outperform all other methods.

(a) CIFAR-100

Loss Fn.	Accuracy	ECE	KS-Error	mean-sweep CE
NLL	78.7	9.10	6.71	9.02
NLL + MMCE	77.2	4.77	3.27	4.77
Focal	78.3	4.66	4.20	4.57
Focal + SB-ECE	78.6	2.30	0.71	2.21
Focal + S-AvUC	78.1	**1.57**	**0.51**	**1.24**

(b) CIFAR-10

Loss Fn.	Accuracy	ECE	KS-Error	mean-sweep CE
NLL	95.5	5.59	2.64	4.78
NLL + MMCE	95.0	1.55	0.83	1.30
Focal	95.0	5.45	5.22	5.41
Focal + SB-ECE	95.1	**1.19**	**0.41**	**0.77**
Focal + S-AvUC	94.4	1.58	0.66	1.43

F AvUC with Gradient Stopping

The AvUC loss [Krishnan and Tickoo, 2020] was proposed to train for calibration in Stochastic Variational Inference (SVI) settings. It is based on the idea of giving an incentive to the model to be certain when accurate and uncertain when inaccurate via a secondary loss. The secondary loss term is described in equations 13 and 14. These are restated here for readability.

$$\text{AvUC}(\kappa, \hat{D}, \boldsymbol{\theta}) = \log\Big(1 + \frac{n_{\text{AU}} + n_{\text{IC}}}{n_{\text{AC}} + n_{\text{IU}}}\Big), \tag{13}$$

$$\begin{aligned} n_{\text{AU}} &= \Sigma_{i|(\boldsymbol{x}_i, y_i)\in S_{\text{AU}}}(c_i \tanh h_i) & n_{\text{IC}} &= \Sigma_{i|(\boldsymbol{x}_i, y_i)\in S_{\text{IC}}}((1-c_i)(1-\tanh h_i)) \\ n_{\text{AC}} &= \Sigma_{i|(\boldsymbol{x}_i, y_i)\in S_{\text{AC}}}(c_i(1-\tanh h_i)) & n_{\text{IU}} &= \Sigma_{i|(\boldsymbol{x}_i, y_i)\in S_{\text{IU}}}((1-c_i)\tanh h_i). \end{aligned} \tag{14}$$

Note that S_{AU}, S_{IC}, S_{AC} and S_{IU} form a partition of datapoints from the training batch $\hat{D}$ which fall in each of the four categories resulting from two classifications: (1) accurate [A] vs. inaccurate [I] based on whether the model's prediction is correct and (2) certain [C] vs uncertain [U] whether the model's entropy is above or below a threshold κ. In our experiments we tune κ rather than inferring it from the first few epochs as was suggested in [Krishnan and Tickoo, 2020]. Despite this additional degree of freedom, we consistently observe that the originally proposed AvUC loss does not help for calibration in non-Bayesian settings.

A closer look at equations 13 and 14 suggests that this might be because of some of the incentives provided by the secondary loss. As observed in section 4.2, minimizing the AvUC loss results in the model being incentivized to be even more confident in its inaccurate and certain predictions via minimizing n_{IC}, specifically the $(1 - c_i)$ multiplicand. Similarly, it also encourages the model to be even less confident in its accurate and uncertain predictions via minimizing n_{AU}, specifically the c_i multiplicand.

To test whether these misincentives are really the cause of our observations, we conduct experiments with a variant of the loss where we stop gradients flowing through the $(1 - c_i)$ and c_i multiplicands in each of the four expressions in equation 14. We denote this modified version of the AvUC loss as the AvUC-GS loss, where "GS" denotes gradient stopping. The experiments confirm our hypothesis - we observe that the AvUC-GS secondary loss helps in improving calibration in situations where the original AvUC secondary loss did not. This observation holds across datasets and primary losses. This can be seen in table 2, where the rows corresponding to AvUC-GS in the secondary loss column have lower ECE on an average than the respective rows corresponding to the AvUC secondary loss. We conclude that AvUC-GS is also an effective secondary loss. However, the soft calibration objective S-AvUC which is inspired from AvUC-GS generalizes, outperforms and supersedes it.

G Hyperparameter Tuning and the Accuracy-Calibration Tradeoff

We look at both the accuracy and the ECE of competing hyperparameter configurations. Some comparisons yield a clear winner but often there is a tradeoff between these. In such cases, we choose the lowest ECE whilst giving up less than 1% accuracy relative to the hyperparameter configuration with the highest accuracy.

As stated in section 5.1, we use the Wide-Resnet-28-10 architecture [Zagoruyko and Komodakis, 2017] trained for 200 epochs on CIFAR-100 and CIFAR-10. For Imagenet, we use the Resnet-50 [He et al., 2015] architecture trained for 90 epochs. The loss function we use in our experiments is PL $+ \beta \cdot$ SL $+ \lambda \cdot$ L2 where PL and SL denote the primary and secondary losses respectively and L2 denotes the weight normalization term with ℓ_2 norm.

All our experiments use the SGD with momentum optimizer with momentum fixed to 0.9 and base learning rate fixed to 0.1. We follow a learning rate schedule which is fixed for each dataset across training losses. The number of bins (M, if applicable) for the soft binning secondary loss is fixed to 15. We used a per-core-batch-size of 64 on a 2x2 TPU topology with 8 cores for an effective batch size of 512. Both the baseline runs and experimental runs (all runs from Table 2) used this batch size. In general, larger batch size is better for the computation of soft calibration losses, but we did not see significant ECE gains with a further increase in batch size. These form the set of hyperparameters we fixed rather than tuned. Fixing some hyperparamers allows us to keep the search space manageable. The values we use for these are those which work well for the cross-entropy baseline.

The β (if applicable) and λ parameters along with parameters relevant to the secondary loss are the set of hyperparameters that we tune. In our experiments with the SB-ECE secondary loss, the secondary loss function we use is SB-ECE$_{lb,p}(M, T, \hat{D}, \boldsymbol{\theta})$ from equation 12, for which we tune the T parameter. In our experiments with the Soft-AvUC loss the secondary loss function we use is S-AvUC$(\kappa, T, \hat{D}, \boldsymbol{\theta})$ from equation 15, for which we tune the κ and T parameters. As stated above, these combined with β and λ form our set of tuned parameters. These hyperparameters are the most critical ones for demonstrating the effectiveness of our techniques.

We tune these parameters one-at-a-time starting with the threshold parameter for the secondary loss (κ, if applicable), followed by temperature parameter for the secondary loss (T, if applicable), the β parameter (if applicable) and the L2-normalization coefficient λ. We retain the best value from the tuning experiments for one parameter while tuning a subsequent parameter.

We show in table 2 that Soft Calibration Objectives result in lower ECEs than previous methods in exchange for a small reduction in accuracy relative to the cross-entropy baseline. In our experiments, we found that other methods to train for calibration (MMCE, AvUC, Focal loss, MSE loss) also have to sacrifice a small amount of accuracy relative to the cross-entropy baseline in order to attain better calibrated models. This fundamental tradeoff can be summarized by the pareto-optimal curve between accuracy and calibration. Our methods result in points on this curve which are better calibrated than previously proposed methods, whilst trading off less than 1% accuracy.

H Is SB-ECE as secondary loss a proper scoring rule?

We start this discussion by asking the following question: what happens to SB-ECE for perfectly calibrated models as the dataset size goes to infinity? We know that ECE tends to zero in this case. Proposition 1 shows that the same holds for SB-ECE. We will use terminology introduced in sections 3 and 4 in this section.

Proposition 1. *Consider a dataset $D = \langle (\boldsymbol{x}_i, y_i) \rangle_{i=1}^{N}$ drawn from the joint probability distribution $\mathcal{D}(\mathcal{X}, \mathcal{Y})$. Say we have a perfectly calibrated model for it such that $E[a|c] = c$, where a and c denote accuracy and confidence respectively. If we consider SB-ECE$_{bin,p}$ with M bins, temperature T and norm p as defined in equation 11, we have*

$$\textit{SB-ECE}_{bin,p}(M, T, \hat{D}, \boldsymbol{\theta}) \to 0 \textit{ as } N \to \infty$$

Proof. Let p_c denote the p.d.f. of the confidence c viewed as a random variable. Let us denote the membership function for bin j by $u_j(c)$, a shorthand for our earlier notation $u_{M,T,j}(c)$ from section 4.1. The size, confidence and accuracy of bin j, as defined in equations 9 and 10 are denoted by C_j and A_j respectively. We observe that as $N \to \infty$, these quantities satisfy the following:

$$C_j = \frac{\Sigma_{i=1}^{N} u_j(c_i) \cdot c_i}{\Sigma_{i=1}^{N} u_j(c_i)} \to \frac{\int_0^1 x u_j(x)\, p_c(x) dx}{\int_0^1 u_j(x)\, p_c(x) dx}$$

$$A_j = \frac{\Sigma_{i=1}^{N} u_j(c_i) \cdot a_i}{\Sigma_{i=1}^{N} u_j(c_i)} \to \frac{\int_0^1 E[a|c=x] u_j(x)\, p_c(x) dx}{\int_0^1 u_j(x)\, p_c(x) dx}$$

Since $E[a|c = x] = x$ for perfectly calibrated models, we infer that as $N \to \infty$:

$$C_j - A_j \to 0$$

$$\text{SB-ECE}_{\text{bin},p}(M, T, \hat{D}, \boldsymbol{\theta}) \to 0$$

□

Note that this does not necessarily hold for datasets of a given finite size. If we measure SB-ECE with $N = 2$ datapoints for a perfectly-calibrated model which always has a confidence of 30% and is correct 30% of the time, we will always end up with ECE and SB-ECE both greater than zero.

If we consider an optimal classifier which always outputs the true probabilities, then it minimizes NLL since NLL is a proper scoring rule and it minimizes SB-ECE (note that SB-ECE ≥ 0 by definition) as dataset size goes to infinity as per proposition 1 since it is perfectly calibrated. Hence, it minimizes any positive linear combination of NLL and SB-ECE which implies that these linear combinations are proper scoring rules in the limit of infinite data.

I Post-hoc Dirichlet Calibration

In Tables 1 and 2, we have compared our methods to existing calibration-incentivizing losses that operate during training, with and without post-hoc temperature scaling. There are a several post-hoc recalibration techniques which can be used complementary to our methods. For this reason, comparing to these has not been the focus of our work, similar to [Mukhoti et al., 2020] and [Kumar et al., 2018]. Nevertheless, in this section we show that our methods do significantly better than training with the cross-entropy, focal or MSE primary losses and using such post-hoc recalibration methods. In particular, we compare against Dirichlet calibration [Kull et al., 2019]. Table 5 records ECE numbers for the best-performing training objectives for CIFAR-10 corresponding to table 1b, both with and without post-hoc Dirichlet calibration.

Table 5: We report accuracy and average ECE across 10 runs for the best performing training objectives for the CIFAR-10 dataset corresponding to the entries in table 1b, both with and without post-hoc dirichlet calibration. ECE is computed with the ℓ_2 norm and equal-mass binning. Our methods outperform post-hoc dirichlet calibration applied to the NLL, Focal and MSE primary losses.

(a) CIFAR-10

Loss Fn.	Accuracy	ECE	ECE w/ Dirichlet
NLL	95.5	5.59	2.45
NLL + MMCE	95.0	1.55	**1.28**
Focal	95.0	5.45	2.48
Focal + SB-ECE	95.1	**1.19**	2.24
Focal + S-AvUC	94.4	1.58	1.70
MSE	95.0	6.58	4.69

J Multiclass Calibration Error

Reducing top-label calibration error using calibration-incentivizing training loss functions is the focus of our work, similar to [Kumar et al., 2018], [Krishnan and Tickoo, 2020] and [Mukhoti et al., 2020]. In this section, we additionally evaluate marginal ECE [Kumar et al., 2019] which measures the calibration of the predicted probability distribution over all classes rather than just the top class. Table 6 shows that our methods which are trained to minimize top-label ECE yield the best marginal ECE numbers as well.

The terms in the S-AvUC loss are tied to the top-label and this does readily lend itself to a multiclass extension. The soft binning approach however immediately yields a trainable soft version of marginal ECE analogous to the top-label case. Soft-binned marginal ECE is beyond the scope of this paper but can be investigated further.

Table 6: We report accuracy, average top-label ECE and average marginal ECE for the CIFAR-10 and CIFAR-100 datasets corresponding to the entries in tables 1a and 1b. Marginal ECE and top-label ECE are computed with the ℓ_2 norm and equal-mass binning. Soft-calibration objectives outperform other methods on multiclass CE in addition to the top-label CE they are designed to optimize.

(a) CIFAR-100

Loss Fn.	Accuracy	ECE	Marginal ECE
NLL	78.7	9.10	3.78
NLL + MMCE	77.2	4.77	3.05
Focal	78.3	4.66	4.35
Focal + SB-ECE	78.6	2.30	3.05
Focal + S-AvUC	78.1	**1.57**	**2.72**

(b) CIFAR-10

Loss Fn.	Accuracy	ECE	Marginal ECE
NLL	95.5	5.59	2.31
NLL + MMCE	95.0	1.55	2.06
Focal	95.0	5.45	5.64
Focal + SB-ECE	95.1	**1.19**	2.24
Focal + S-AvUC	94.4	1.58	**1.65**