Chair of Computer Vision and Artifical Intelligence
Department of Informatics
Technical University of Munich

Calibration via Mix-n-Match

Anika Apel
20. December 2020

Abstract — Deep learning has become a growing research field and has been applied to various practical problems. However, an issue encountered with neural networks is that they tend to be over-confident with their decisions due to overfitting on the data and that the predicted output probability does not represent the true certainty of the decision. Especially when applying learned algorithms to decision-making in real-world problems such as object detection in automated driving or medical diagnosis a system has to provide an uncertainty measurement which indicates how likely the decision is wrong. This report explores different post-hoc calibration methods which can be directly applied to every probabilistic classifier and compares parametric, non-parametric methods and the recently proposed mix-n-match approach towards calibration.

1 Introduction

Machine learning and especially deep learning have been applied to many classification tasks with grat success. To use these methods in real-world decision-making systems and for safety-critical applications such as medical diagnosis or object detection in autonomous driving, an uncertainty estimate of the decision is necessary to ensure reliability of the system as well as trustworthiness for the end user. With these given circumstances, the requirement for a system is that it is both accurate predicting the correct outcome and can also indicate how confident it is in this decision. For a probabilistic classifier this implies that the classification scores obtained by the model should reflect the true correctness likelihood of the output. This problem is referred to as calibration of a model. Simple but not very accurate methods such as logistic regression and decision trees are well-calibrated. However, it has been shown in empirical studies that these requirements are not fulfilled by neural networks which are more expressive in general. Model capacity, batch normalization and weight decay as a regularization technique are examples for factors to decrease calibration [1].

Different approaches for calibration have been studied in the literature and can be clustered into three categories. A visualized conceptual map is given in Figure 1. The first category which we refer to as ab-initio calibration comprises all methods which train ab-initio well-calibrated methods. The second group includes Bayesian frameworks to represent prediction uncertainty. Examples to be mentioned here are MC Dropout [2] or Stochastic Variational Inference [3]. The last group of categories which we will study in depth in this report is post-hoc calibration. Post-hoc calibration covers all methods which calibrate the models after training by applying a transformation of a classifier's predictions in a post-processing step. A major advantage of this approach is that it is build on top of a pretrained classifier and can therefore be applied to any classification method with a probabilistic output prediction.

To explore the topic of post-hoc calibration, we introduce in Section 2 what calibration in multi-class classification means, what the goal of calibration is and which properties a calibration method needs to fulfill. In Section 3, we review different evaluation methods for calibration, explain the different metrics and show the differences between them. In Section 4, we then explore different calibration methods including two basic methods as well as novel mix-n-match extensions of these methods. In Section 5 we compare the different approaches in multiple setups to emphasize their advantages and disadvantages. Finally, in section 6 we summarize the main post-hoc calibration methods and their usages.

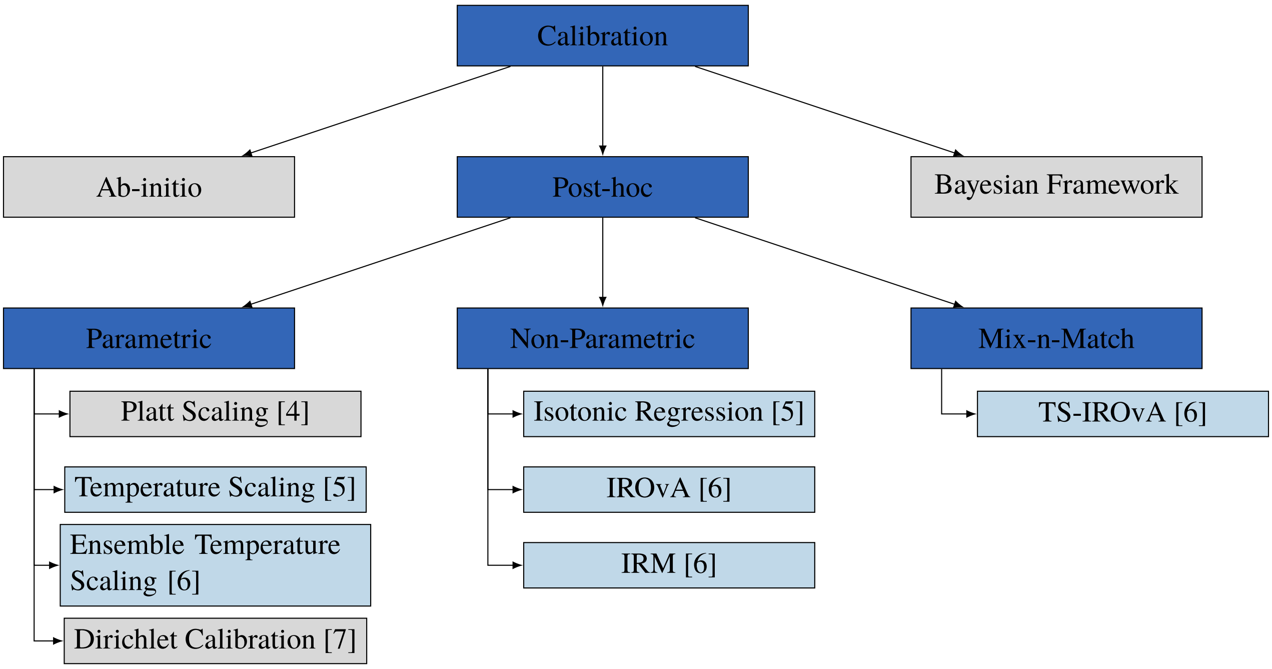


Figure 1 Conceptual map of calibration methods with focus on post-hoc calibration methods. In blue, methods that will be explained in depth in Section 4. Methods marked in gray, on the other hand, are not covered in this report.

2 Problem Setup

In this report we will study the task of calibrating classification methods and will therefore first define the classification setup as well as the calibration task.

Lets denote $\mathcal{X}$ as the input space and $\mathcal{Y}$ as the label space where X and Y denote random variables with an unknown joint probability distribution $P(X, Y)$. X represents the input features and Y represents the L-class one hot encoded label vector. Given a probabilistic classifier $f : \mathcal{X} \longrightarrow \mathcal{Z} \subseteq \Delta^L$ the output $z = f(X)$ represents the predictive distribution over the class labels for a given input X. Thus, the vector $z = (z_1, \ldots, z_L)$ defines a probability distribution of the probability simplex Δ^L with $z_1, \ldots, z_L \in [0, 1]$ and $\sum_{i=1}^{L} z_i = 1$. In machine learning, the goal is to find the optimal probabilistic classifier such that the negative log likelihood on a training data set $\{x_i, y_i\}_{i=1}^n$ is minimized.

The information about calibration of a probabilistic classifier is contained in the canonical calibration function $\pi(z)$ defined as

$$\pi(z) = P(Y \in \cdot | f(X) = z) \tag{1}$$

where $\cdot$ represents the placeholder for any label $Y \in \{1, \ldots, L\}$ [8]. A model is called reliable or class-wise calibrated if and only if the following condition is fulfilled

$$\pi(z)^{(l)} = P(Y^l = 1 | f(X) = z) = z \forall X \in \mathcal{X} \tag{2}$$

where $pi(z)^{(l)}$ represent the actual likelihood of X being labeled as class l. A deviation in the confidence from the actual likelihood is called miscalibration of a model.

In some literature [1] a weaker condition for calibration can be found where only the predicted label has to be calibrated resulting in the following condition for a well-calibrated system

$$argmax_{l \in \{1,\ldots,L\}} \pi(z) \tag{3}$$

This weaker condition should not be used for defining calibration as the model should be calibrated as well on less likely predictions.

To improve the calibration of an already existing trained classifier such as a deep neural network the output can be transformed in a post-processing step referred to as post-hoc calibration. By relying on already existing classifiers, it is especially useful for modern deep learning classifiers due to its generality, flexibility and freedom in designing the classifier [8]. The pipeline to perform post-hoc calibration consists of two steps. In a first step, a calibration map $T : \Delta^L \rightarrow \Delta^L$ is learned on a validation set with n_c data samples to transform the outputs to estimated probabilities. To avoid unwanted bias in the calibration a training-independent hold-out validation set should be used here. In the second step, the learned calibration map is evaluated on another evaluation set with n_e data samples.

Depending on the design of the calibration map, post-hoc calibration methods can be further distinguished in parametric and non-parametric methods. Whereas parametric methods comprise all calibration maps defined by a finite-dimensional parametric space, non-parametric methods rely on calibration maps described by infinite-dimensional parameters.

Zhang et al. [6] propose three desideratas for post-hoc calibration which should be carefully considered when designing such methods. The first desiderata is accuracy preservation indicating that calibration should not come at cost of accuracy. The underlying problem is that by applying the post-hoc transformation the predicted labels can change causing a decrease of a classifier's accuracy. To achieve accuracy preservation the calibration map should not change the order of the classification scores. If this property is not fulfilled the design scheme in Equation 4 can be applied. This design scheme learns a strictly isotonic function g to transform each classification score z_l and then normalizing the transformed vector to obtain a probabilistic prediction $T(z)$.

$$\begin{gathered} g : [0, 1] \rightarrow \mathbb{R}_{\geq 0} \\ T(z) = (g(z_1), g(z_2), \ldots, g(z_L)) / \sum_{l=1}^{L} g(z_l) \end{gathered} \tag{4}$$

In addition, data efficiency is a desirable property of a calibration method. Whereas data efficiency means the ability to achieve well calibration without requiring large amount of calibration data for learning the calibration map. The last desiderata is high expressive power of the calibration method. High expressive power is understood as sufficient representation power to estimate the class probabilities given enough calibration data.

3 Calibration Evaluation

In the following section, we present different evaluation methods for calibration. In general, the task of evaluating the calibration of a model is to check whether the probabilities predicted by a model - a classifier in this case - match the distribution of realized outcomes [8].

3.1 Reliability Diagram

A visual tool for evaluating the calibration of a model is the so called Reliability Diagram which have been first proposed for evaluating the reliability of meteor forecasts [9]. By plotting the sample accuracy as a function of the model confidence the miscalibration of the model can be visualized. For a model with perfect calibration the accuracy equals exactly the confidence of the model and the plotted function corresponds exactly the identity function.

If data points in the reliability diagram lay above the identity line the model is over-confident and if data points lay below the identity line the model is under-confident. An exemplary reliability diagram can be seen in Figure 2 where the left diagram is showing an overconfident model and the right diagram is showing an under-confident model.

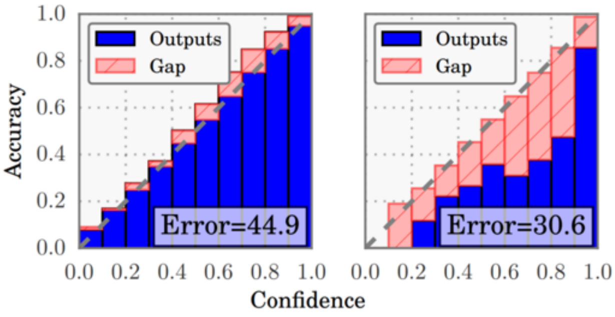


Figure 2 Exemplary reliability diagram from [1]

3.2 Expected Calibration Error

An intuitive statistic to measure calibration relative to the ideal reliability diagram was introduced with Expected Calibration Error (ECE). The metrics measures the expected difference between confidence and accuracy of the classifier [10].

Mathematically written ECE is defined as

$$\mathop{\mathbb{E}}_{\hat{P}}[|\mathbb{P}(\hat{Y} = Y|\hat{P} = p) - p|] \tag{5}$$

where p stands for the accuracy and $\mathbb{P}(\hat{Y} = Y|\hat{P} = p)$ for the confidence of the evaluated classifier. There exist different versions for calculating the calibration error. We will present in this report two different ones, the most common one, histogram-based ECE, and the more recently proposed kernel-based ECE. Further extensions of ECE are defined in [11] which we won't explore in detail here.

To use the ECE metric for multi-class classification it has to be reduced to one effective dimension to avoid a curse of dimensionality in convergence of the metric. In this setting, top-label ECE^d can be examined measuring the expected calibration error for the model's top d predictions and the corresponding true probabilities [12]. A second option if the calibration error among all classes is of interest, the class-wise ECE can be measured to measure the average gap across all class-wise predictions [7].

3.2.1 Histogram-based ECE

The initial calculation of ECE as proposed in [10] uses binning for estimating confidence and accuracy of the classifier. Using this histogram-based approach toward estimation ECE is defined as the weighted average gap among these bins mathematically written as

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{6}$$

where M corresponds to the number of bins, $|B_m|$ the number of predictions in bin m, n is the total number of predictions over all bins and $\text{acc}(B_m)$ and $\text{conf}(B_m)$ correspond to the accuracy and respectively confidence in bin m.

In recent research several issues of this approach has been identified. First, using a finite set of bins for estimating the calibration error can underestimate the actual calibration error[12]. Second, due to the used binning scheme for estimating confidence and accuracy histogram-based ECE suffers from the same drawbacks as histogram. These drawbacks include a bias-variance trade-off due to the sensitivity to binning schemes as well as data-inefficiency [6].

3.2.2 Kernel-based ECE

Due to the above mentioned issues of the histogram-based ECE Zhang et al. propose a kernel-based ECE using a kernel density estimator for estimating confidence and accuracy of the classifier [6]. Therefore, a kernel smoothing function K is introduced for smoothing the estimated probabilities

$$\begin{aligned} &K : \mathbb{R} \to \mathbb{R}_{\geq 0} \\ &K_h(a) = h^{-1} K(a/h) \end{aligned} \tag{7}$$

where $h > 0$ corresponds to the kernel bandwidth.

Using this kernel function, the probabilities can be estimated as

$$\begin{aligned} \tilde{p}(z) &= \frac{h^{-L}}{n_e} \sum_{i=1}^{n_e} \prod_{l=1}^{L} K_h(z_l - z_l^{(i)}) \\ \tilde{\pi}(z) &= \frac{\sum_{i=1}^{n_e} y^{(i)} \prod_{l=1}^{L} K_h(z_l - z_l^{(i)})}{\sum_{i=1}^{n_e} \prod_{l=1}^{L} K_h(z_l - z_l^{(i)})} \end{aligned} \tag{8}$$

resulting in the following definition of kernel-based ECE:

$$\widetilde{ECE}^{d}(f) = \int \|z - \tilde{\pi}(z)\|_d^d \tilde{p}(z) dz \tag{9}$$

Especially for a small evaluation calibration dataset this method is less unbiased and represents the true calibration error more accurate compared to histogram-based ECE[6]. The kernel-based ECE can thus be seen as a data-efficient variant of ECE with high potential for an evaluation metric.

3.3 Maximal Calibration Error

In safety-critical applications the worst-case deviation between confidence and accuracy is more meaningful than ECE. Therefore the Maximal Calibration Error (MCE) is defined as seen in Equation 10.

$$\max_{m \in \{1,\ldots,M\}} |\mathbb{P}(\hat{Y} = Y | \hat{P} = p) - p| \tag{10}$$

To give an intuitive explanation of the MCE we refer to the reliability diagram where the maximum calibration error corresponds to the maximum confidence gap of the classifier visualized in red in Figure 2.

3.4 Calibration gain

To directly evaluate the enhancement caused by applying a calibration map can be measured by the calibration gain. The calibration gain measures the reduction of ECE achieved by the calibration map. In mathematical terms the calibration gain is defined as follows

$$\Delta ECE^2(T) = ECE^2(f) - ECE^2(T \circ f) \tag{11}$$

where f represents the classifier mapping and T the calibration map. To note here is that a higher calibration gain indicates a better performance of an analyzed calibration method [6].

In comparison to ECE calibration gains reflects the calibration capabilities of calibration methods making it a very powerful ranking method. Furthermore, calibration gain is a dimensionality-independent method and avoids the curse of dimensionality which affects.

4 Post-hoc Calibration Methods

After introducing the general framework of calibration as well as methods to evaluate calibration we will now give an overview over different post-hoc calibration methods. Therefore, we will first explore Isotonic Regression (IR) and Temperature Scaling (TS) as common basic methods of post-hoc calibration. In a second step we will explore recent mix-n-match extensions of IR and TS using ensemble techniques to enforce the three properties elucidated earlier. As a last category we will elaborate a mix-n-match approach by composing basic methods.

4.1 Calibration Basic Methods

4.1.1 Temperature Scaling

Temperature Scaling [1] is a parametric calibration method emerged as a single-parameter variant of Platt scaling [4]. It can correct a monotonic disortion in the logits by multiplying the logits by a scalar before applying softmax. This approach is summarized in the following calibration map:

$$T(z_i)^{(l)} = \sigma(s_i/t)(l) = (z_i^{1/t})^{(l)}/|z_i^{1/t}| \tag{12}$$

where s_i are the logits of trained classifier, z_i the predictive labels and t is the learnable parameter referred to as the temperature of the calibration map. The optimal temperature for the calibration map is found by minimizing negative log likelihood on the validation data set.

The introduced temperature parameter t can be interpreted as a measurement of the confidence level of a classifier. A temperature of $t = 1$ corresponds to no possible calibration correction in the present scenario indicating a well confidence. A temperature below 1 increases the estimated probabilities indicating that the model was under-confident whereas a temperature above 1 shrinks the estimated probabilities indicating that the model was over-confident. In the extreme case of $t \to \infty$ every prediction has the same estimated probability and the classifier has maximum uncertainty. In empirical studies Guo et al. [1] have shown that despite the simplicity of the method it shows surprising effectiveness in various setups.

4.1.2 Isotonic Regression

IR is a non-parametric approach proposed by [5] as one of the first approaches for calibrating classifiers. The basic implementation of IR is only applicable for binary classification. Therefore, we present later an extension to multi-class classification.

In IR a calibration map is chosen from family of isotonic functions $\mathcal{G}$. This function is determined by learning a stepwise constant function T to transform uncalibrated outputs using pair-adjacent violators algorithm [13]. The idea is based on the intuition that the ordering of the classification scores implies information about the certainty of a system. If enough calibration data is given IR is very powerful [14].

Algorithm 1 Pair-Adjacent Violators Algorithm to find optimal isotonic regression $T^*(x)$

```
Input: $\{x_i\}$ sorted by their classification scores $z_i = f(x_i)$
Output:stepwise constant function $T^*(z)$
Initialization: $T(z_i) = 0 for x_i = 0$ and $T(z_i) = 1 for x_i = 1$
while $T(z)$ is not isotonic do
    find pair-adjacent violators $(z_{i-1}, z_i)$ with $T(z_{i-1}) > T(z_i)$
    Replace with average $T(z_{i-1}) = T(z_i) = T(z_{i-1}) + T(z_i)/2$
end while
$T^*(z) = T(z)$
```

The algorithm to determine the isotonic regression function as the used calibration map is described in Algorithm 1. In the worst-case N pair-adjacent violators (z_{i-1}, z_i) can be found resulting in a constant calibration map representing maximum uncertainty of the classifier. In the best case the ordering of the classification scores contains already all information and calibration can't be improved with IR.

As mentioned earlier, this method is only applicable to binary classification. Therefore a multi-class extension was proposed by [6] referred to in the paper as IROvA. To make IR applicable to multi-class classification with L labels, the task of multi-class classification is decomposed into L binary one-versus-all classifications. Therefore we learn a set of calibration maps $\{T_l\}_{1,\dots,L}$ differing for each class. Problematic with this approach is the lack of accuracy preservation as well as the lack of data efficiency as the required amount of data scales linearly with number of classes in the best case of a balanced dataset.

4.2 Mix-n-Match via Ensemble Techniques

4.2.1 Ensemble Temperature Scaling

Ensemble Temperature Scaling (ETS) was introduced as an extension of TS to enforce accuracy preservation, data efficiency and high expressive power of the methods. The basic method TS is due to its amount of parameters already very data-efficient and by construction accuracy-preserving but shows limited expressive power compared to non-parametric methods. To improve the expressive power an ensemble of calibration maps of the same type (see Equation 13) is used.

$$T(z) = w_1 T(z;\Theta_1) + \cdots + w_M T(z;\Theta_M) \tag{13}$$

Precisely, ETS uses a three-component ensemble weighting three TS calibration maps with different temperature parameters (see Equation 14). The ensemble is build of the original TS calibration map $T(z;t) = (z_1^{1/t},\dots,z_L)/\sum_{l=1}^{L} z_l^{1/t}$, a calibration map with $t = 1$ to increase stability if original classifier is well calibrated and a calibration map with $t \longrightarrow \infty$ to smooth the predictions.

$$T(z;w,t) = w_1 T(z;t) + w_2 z + w_3 \frac{1}{L} \tag{14}$$

As with the basic method TS, this method is accuracy preserving by construction for $w_i \geq 0, i \in \{1,2,3\}$. The data efficiency should be obtained when just adding three additional parameters but with this small amount of additional parameters large expressivity is added to the calibration map as we will see in Section 5.

4.2.2 Multi-class Isotonic Regression

IRM [6] is an extension to IROvA enforcing accuracy preservation, data efficiency and high expressive power. The data-efficiency issues of IROvA, the convential one-versus-all extension for IR, lies on the necessity to learn L calibration maps. Especially for a large number of classes this approach needs a lot of calibration data to represent a good fit. To address this problem IRM was proposed with the approach to learn one calibration map over all classes and define the calibration map using the design scheme for accuracy preservation described in Equation 4.

The following procedure is performed for learning a single strictly isotonic function g used for transforming the classification scores. In a first step, the predictions and labels are ensembled to obtain a total of $n_c L$ data samples denoted as the predictions $a^j{}_{j=1}^{n_c L}$ and the corresponding labels $b^j{}_{j=1}^{n_c L}$. As in the initialization step of IR, the set of predictions and labels are now sorted by their label-wise classification scores $\{a^j\}_{j=1}^{n_c L}$. In a second step, the isotonic function g^* is learned by minimizing the squared error loss between $g(a)$ and b over all $n_c L$ data entries. We use here the same approach as in basic IR and apply the pair-adjacent violator algorithm to find the best stepwise constant isotonic function. To apply the design scheme for accuracy preservation a strictly isotonic function is required and thus, the learned function g^* has to be modified to impose strict isotonicity. Therefore a very small ϵ is used to add a linear term ϵa to the isotonic regression g^* whenever strict isotonicity is violated to defuse the violations. To summarize, this results in the following calibration map

$$\begin{aligned} \hat{g}(z) &= g^*(z) + \epsilon z \\ T(z) &= (\hat{g}(z_1),\dots,\hat{g}(z_L)/\sum_{l=1}^{L} \hat{g}(z_l) \end{aligned} \tag{15}$$

where z are predictions of the classifier, ϵ a small positive number to enforce strict isotonicity and g^* the learned isotonic regression based on the ensembled predictions a and labels b.

Reducing the problem to one calibration map instead of L calibration maps as it is handled in IROvA gives a less expressive calibration but significantly improves data efficiency. This trade-off is referred to as the efficiency-expressivity trade-off of IRM [6].

4.3 Mix-n-Match via Composition

The lack of expressivity in the parametric models and the lack of data-efficiency in non-parametric models yield to the introduction of a mix-n-match via composition approach by Zhang et al. [6]. The general idea within this approach is to combine the best of both worlds by composition of parametric and non-parametric methods.

The calibration map is defined as the composition

$$T(z) = T_{np}(T_p(z)) \tag{16}$$

where T_{np} is the calibration map of a chosen non-parametric method and T_p is the calibration map of a chosen parametric method.

Considering the three properties for calibrations this approach can benefit from the high expressive power of non-parametric methods as well as the data efficiency of the parametric method. Accuracy preservation is maintained if each of the composed methods are accuracy-preserving as the composition of two isotonic functions remains isotonic.

5 Comparison

In the following section we will compare the methods presented in Section 4 based on the results of [6]. Therefore, we will evaluate the different methods in two setups. In a first which is referred to in the literature as the classical calibration evaluation setup the calibration performance is compared on a fixed sized dataset. The second setup should evaluate the data-amount dependent behaviour referred to in the literature as learning curve analysis [14].

Moreover, the methods will be evaluated on calibration tasks with varying complexity. The complexity of a calibration depends on the complexity of the canonical calibration function $\pi(z)$ and without deeper knowledge about the calibration complexity it is assumed that the complexity depends on the model and data complexity [6]. Therefore, the following combinations are defined as low, moderately and highly complex calibration tasks:

1. **low complex calibration:** low model complexity and low data complexity
2. **moderately complex calibration:** low model complexity but high data complexity and high model complexity but low data complexity
3. **highly complex calibration:** high model complexity and high data complexity

In order to cover the aforementioned spectrums of complexity, the models were evaluated on CIFAR-10 with 10 class labels, CIFAR-100 with 100 class labels and ImageNet with 1000 class labels indicating the different levels of data complexity. Further, the calibration methods were applied to different classification neural networks such as DenseNet40, LeNet 5, ResNet 100, WideResnet 28-10 with varying model capacities to include the aspect model complexity.

A general summary of the comparison based on the three properties accuracy-preserving, data efficiency and high expressive power which we will explain in detail in the following sections is given in Table 1.

Method	Accuracy-preserving	Data efficiency	Expressive power
TS	yes	good	low
ETS	yes	good	improved compared to TS
IR	yes	low	high
IROvA	no	low	high for large n_c
IRM	yes	improved compared to IROvA	lower than IROvA
IROVA-TS	no	better than non-parametric methods, lower than parametric-methods	high

Table 1 General overview over reviewed models evaluated on properties accuracy-preserving, data efficiency and expressive power grouped by type of calibration method (parametric, non-parametric, composition).

5.1 TS and ETS

The main motivation for introducing ETS as an extension of TS was to improve the expressive power of the parametric method.

In Table 2 it is shown that ETS achieved for every combination of dataset and model at least the same top-label ECE[1] and outperformed it in most of the cases. On a fixed calibration dataset size ETS indeed achieves higher expressive power compared to TS. The added expressive power to ETS by using an ensemble of different temperature scalings sufficiently increased the performance and let ETS perform best over mostly all setups presented in Table 2.

To further explore whether this is also the case for data-dependent behaviour Zhang et al. [6] evaluated the methods as well on different amounts of calibration data. The results of this experiment visualized in Figure 3a

Dataset	Model	Uncalibrated	TS	ETS	IRM	IROvA	IROvA-TS
CIFAR-10	DenseNet 40	3.30	**1.04**	**1.04**	1.18	1.16	1.11
CIFAR-10	LeNet 5	1.42	1.16	**1.13**	1.19	1.26	1.26
CIFAR-10	ResNet 110	4.25	2.05	2.05	1.53	1.45	**1.39**
CIFAR-10	WRN 28-10	2.53	1.61	1.61	1.02	0.994	**0.967**
CIFAR-100	DenseNet 40	12.22	1.55	**1.54**	3.32	4.48	2.22
CIFAR-100	LeNet 5	2.76	1.11	**1.05**	1.33	3.67	3.18
CIFAR-100	ResNet 110	13.61	2.75	**1.93**	4.78	5.27	3.00
CIFAR-100	WRN 28-10	4.41	3.24	**2.80**	3.16	3.45	2.92
ImageNet	DenseNet 161	5.09	1.72	**1.33**	2.13	3.97	3.01
ImageNet	ResNeXt 101	7.44	3.03	**2.02**	3.51	4.64	3.09
ImageNet	VGG19	3.31	1.64	**1.36**	1.85	3.77	3.03
ImageNet	WRN 50-2	4.83	2.52	**1.81**	2.54	3.91	3.03

Table 2 Calibration errors of different calibration methods in classical calibration setup with fixed calibration dataset size evaluated on top-label ECE[1] as presented by [6].

show that the used ensemble of different temperature scalings added expressive power to ETS reducing the final ECE by 16%. Additionally, adding three more parameters to the model did not decrease the data efficiency of the method significantly which can be seen by the fast convergence of both methods.

5.2 IROvA and IRM

IRM as an extension to IROvA was introduced to improve the data-efficiency by learning one transforming function g for all class instead of L different transforming functions. Regarding the design of the calibration maps it can be said that IRM is accuracy-preserving whereas IROvA doesn't fulfill this important property. We will further explore in this section how this affected the other two properties data efficiency and expressive power.

The expressive power can be evaluated in the standard setup as well as in the learning curve analysis reflecting the data-dependent behaviour. As we can see in Table 2 IRM does not outperform IROvA in all setups and the gap between the obtained ECEs varies over the different calibration taks. Especially it can be seen that for tasks with increasing number of classes IRM performs better than IROvA indicating a better data efficiency.

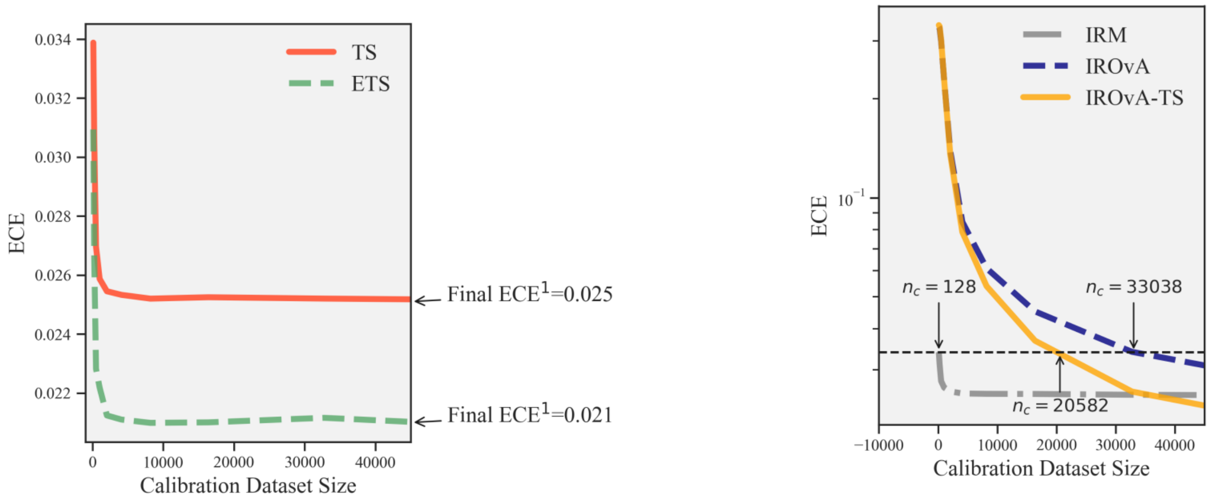
The improvement in data efficiency of IRM over IROvA is visualized using learning curve analysis in Figure 3b where the ECE is plotted against the calibration dataset size. Additionally, the dataset size needed for reaching a specific level of calibration error is emphasized in the plot. This plot especially highlights the high boost in data efficiency and shows the fast convergence of IRM in ECE which is aspired.

By learning only one calibration map in IRM instead of L maps in IROvA a decrease in expressive power was assumed. This decrease was visible in some of the experiments presented in [6] but this loss in expressive power is small compared to the achieved improvement in data efficiency.

5.3 Basic Methods and Mix-n-Match via Composition

The intuition of introducing Mix-n-Match was to combine the best of both worlds. We will therefore study in this section which effect composing IROvA and TS has on the overall performance of the calibration map. Additionally, we will compare the performance to the extended methods ETS and IRM.

As mentioned before, accuracy preservation of a mix-n-match approach is only given if both of the composed models are accuracy-preserving. Therefore, the mix-n-match approach TS-IROvA [6] is not accuracy-preserving and it requires a large calibration dataset to achieve the same accuracy as the uncalibrated classifier (see Figure 4.



(a) Expressive power of TS and ETS based on final ECE

(b) Data efficiency of IRM, IROvA and IROvA-TS based on needed calibration dataset size to reach equal calibration error

Figure 3 Comparison of learning curves of mix-n-match approaches and basic approaches as given in [6]

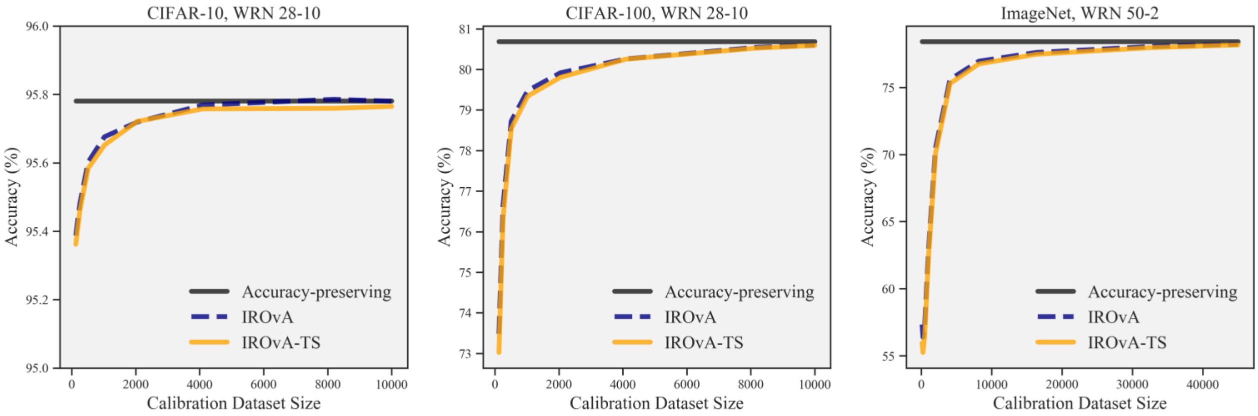


Figure 4 Learning curve analysis plotting accuracy towards calibration dataset for IROvA and IROvA-TS and as a reference the accuracy-preserving baseline classifier [6].

Indeed, it can be said that the composition does not affect the accuracy-preserving property of the non-parametric approach.

In a next step, we will evaluate whether the composition improved the data-efficiency properties compared to the non-parametric methods and how it affected the expressive power of the calibration. To assess this, the learning curves as presented in [6] are analyzed. The data-amount dependent behaviour covers data efficiency of a method as well as expressive power of a method. Data efficiency is reflected in the learning curves as fast convergence in ECE and can be compared over methods for a fixed ECE. As shown in 3b the data efficiency of TS-IROvA has improved compared to the basic approach IROvA but the data-efficiency of IRM cannot be reached. Overall, the convergence of IROvA-TS compared to IROvA stays the same but it converges at a lower ECE indicating better calibration.

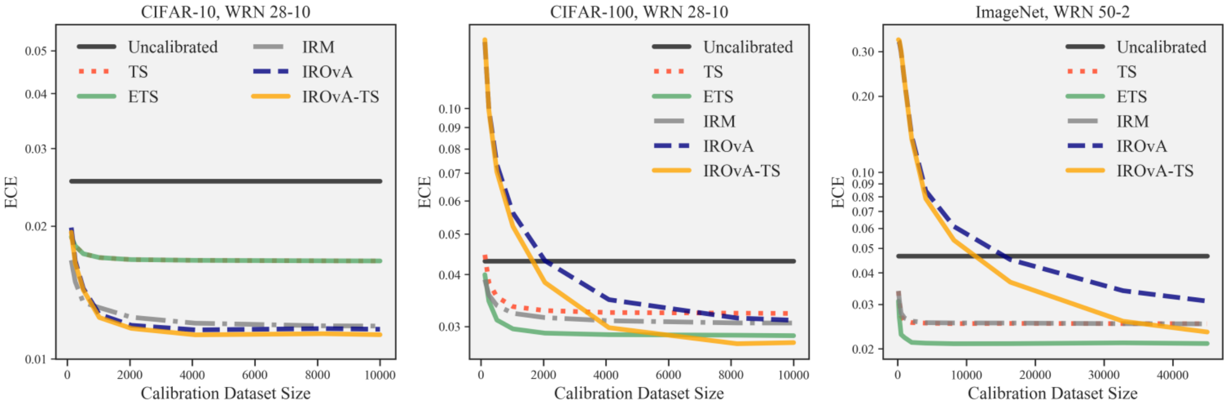


Figure 5 Learning cure analysis plotting ECE towards calibration dataset for all presented methods to assess data-amount dependent behaviour as presented in [6]. Data complexity increases from left to right.

Finally, expressive power of the mix-n-match approach will be analyzed. As shown in Figure 5 IROvA-TS outperforms all other methods in the final ECE score after convergence. The gap between the different methods increases with increasing data complexity resulting in the conclusion that mix-n-match is especially powerful for calibration tasks with high data complexity.

6 Conclusion

In this report we examined different post-hoc calibration methods. Based on the proposed properties by [6] and the necessity of these we compared the post-hoc calibration methods in different scenarios. It has been shown that the lack of expressivity in TS and lack of data efficiency in IROvA can be compensated by different ensemble techniques in calibration. Additionally, composition of non-parametric and parametric methods have shown improvement over using a single non-parametric method but was not as promising as the ensemble techniques.

In general, the choice of the calibration method depends on the model and data complexity and parametric and non-parametric are suitable for different purposes. If calibration is applied in a data-limited regime where high data-efficiency is important parametric methods are preferred over non-parametric methods. However, in data-rich regimes where the focus is especially on achieving very good calibration non-parametric calibrations can be applied. The mix-n-match approach via composition outperformed the other methods only in regard to expressive power for a very large calibration dataset. However, this approach seems promising if a calibration task requires high expressive power due to its complexity and data amount is limited.

References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks.”

[2] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*, 2016, pp. 1050–1059.

[3] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[4] J. Platt and others, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” vol. 10, no. 3, pp. 61–74, publisher: Cambridge, MA.

[5] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’02*. ACM Press, p. 694.

[6] J. Zhang, B. Kailkhura, and T. Y.-J. Han, “Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning.”

[7] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, “Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration,” vol. 32, pp. 12 316–12 326.

[8] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. B. Schön, “Evaluating model calibration in classification.”

[9] A. H. Murphy and R. L. Winkler, “Reliability of subjective probability forecasts of precipitation and temperature,” vol. 26, no. 1, pp. 41–47, publisher: Wiley Online Library.

[10] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI’15. AAAI Press, pp. 2901–2907.

[11] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, “Measuring calibration in deep learning.” in *CVPR Workshops*, pp. 38–41.

[12] A. Kumar, P. S. Liang, and T. Ma, “Verified uncertainty calibration,” vol. 32, pp. 3792–3803.

[13] R. E. Barlow, “Statistical inference under order restrictions; the theory and application of isotonic regression.”

[14] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML ’05. Association for Computing Machinery, pp. 625–632.