Abstract

Deep Neural Networks (DNNs), despite their tremendous success in recent years, could still cast doubts on their predictions due to the intrinsic uncertainty associated with their learning process. Ensemble techniques and post-hoc calibrations are two types of approaches that have individually shown promise in improving the uncertainty calibration of DNNs. However, the synergistic effect of the two types of methods has not been well explored. In this paper, we propose a truth discovery framework to integrate ensemble-based and post-hoc calibration methods. Using the geometric variance of the ensemble candidates as a good indicator for sample uncertainty, we design an accuracy-preserving truth estimator with provably no accuracy drop. Furthermore, we show that post-hoc calibration can also be enhanced by truth discovery-regularized optimization. On large-scale datasets including CIFAR and ImageNet, our method shows consistent improvement against state-of-the-art calibration approaches on both histogram-based and kernel density-based evaluation metrics. Our code is available at https://github.com/horsepurve/truly-uncertain.

1 INTRODUCTION

We live in an uncertain world. With the increasing use of deep learning in the real world, quantitative estimation of the predictions from deep neural networks (DNNs) must not be neglected, especially when it comes to medical imaging [Esteva et al., 2017] [Ma et al., 2019], disease diagnosis [De Fauw et al., 2018] [Ma et al., 2018], and autonomous driving [Kendall et al., 2017]. Uncertainty also plays an important role in differentially private data analysis [Bassily et al., 2013].

Modern deep neural networks, despite their extraordinary performance, are oft-criticized as being poorly calibrated and prone to be overconfident, thus leading to unsatisfied uncertainty estimation. The process of adapting deep learning’s output to be consistent with the actual probability is called uncertainty calibration [Guo et al., 2017], and has drawn a growing attention in recent years.

For a better calibration of the uncertainty of DNNs, the efforts to date have been concentrated on developing more effective calibration and evaluation methods. Existing calibration methods roughly fall into two categories, depending on whether an additional hold-out calibration dataset is used. (1) Post-hoc calibration methods use a calibration dataset to learn a parameterized transformation that maps from classifiers’ raw outputs to their expected probabilities. Quite a few techniques in this category can be used to learn the mapping, such as Temperature Scaling (TS) [Guo et al., 2017] [Kull et al., 2019], Ensemble Temperature Scaling (ETS) [Zhang et al., 2020], and cubic spline [Gupta et al., 2021], etc. However, the expressivity of the learnable mapping could still be limited in all of them. This is evidenced by the fact that in TS a single temperature parameter $T$ is tuned, while ETS brings in three additional ensemble parameters. Thus, it is desirable to explore a more sophisticated form of the mapping function. (2) Another line of methods adapt the training process so that the predictions are better calibrated. Techniques in this category include mixup training [Thulasidasan et al., 2019], pre-training [Hendrycks et al., 2019a], label-smoothing [Müller et al., 2019], data augmentation [Ashukha et al., 2020], self-supervised learning [Hendrycks et al., 2019b], Bayesian approximation [Gal and Ghahramani, 2016] [Gal et al., 2017], and Deep Ensemble (DE) [Lakshminarayanan et al., 2017], with its variants (Snapshot Ensemble [Huang et al., 2017a], Fast Geometric Ensemble (FGE) [Garipov et al., 2018], SWA-Gaussian (SWAG) [Maddox et al., 2019]). Methods from these two categories thrive in recent years, and a natural idea is to combine them together. Recently, Ashukha et al. [2020] points out the
Accordingly, in this paper, we propose truth discovery as an ideal tool for improving uncertainty calibration of deep learning, and make several contributions as follows:

1. We propose Truth Discovery Ensemble (TDE) that improves Deep Ensemble, and show that model uncertainty can be easily derived from truth discovery.

2. Considering that uncertainty calibration approaches may potentially cause a diminished accuracy, we further develop a provably accuracy-preserving Truth Discovery Ensemble (aTDE) via geometric optimization.

3. We propose an optimization approach that directly minimizes ECEs, works for both histogram-based and KDE-based metrics, and integrates multiple metrics via compositional training.

4. We further incorporate the discovered information (i.e. Entropy based Geometric Variance) into the post-hoc calibration pipeline (pTDE) and elevate the performance to a higher level.

To summarize, we show how truth discovery can benefit both ensemble-based and post-hoc uncertainty calibrations, and validate our proposed methods via experiments upon large-scale datasets, using both binning-based and binning-free metrics, along with comprehensive ablation studies.

2 PRELIMINARIES OF UNCERTAINTY CALIBRATION

For an arbitrary multi-class classifier (not necessarily neural network) \( f_\theta : \mathcal{D} \subseteq \mathbb{R}^d \rightarrow \mathcal{Z} \subseteq \Delta^L \) that can make \( L \) predictions for \( L \) classes, its outputs (in any scale) can be transformed into a "probability vector" \( \mathbf{z} \in \mathcal{Z} \) such that:

\[
\sum_{l=1}^{L} z_l = 1, 0 \leq z_l \leq 1.
\]

This can be done by the softmax function, which usually tails the last layer of a deep neural network. Here, \( \Delta^L \) is the probability simplex in \( L \) dimensional space. Note that the classifier parameters can also be drawn from a distribution \( \theta \sim q(\theta) \), e.g., ResNets with random initialization being the only difference.

Although \( \mathbf{z} \) is in the probability simplex \( \Delta^L \), its components may not necessarily have anything to do with, but sometimes are misinterpreted as, the probability of each class. Similarly, the maximum value of the \( L \) outputs, \( \max_l z_l \), was used to represent the "confidence" that the classifier has on its prediction. To avoid possible misleading, \( \max_l z_l \) is referred to as winning score \( v \) (i.e., \( v = \max_l z_l \)\) [Thulasidasan et al., 2019] hereinafter.

For both ensemble-based and post-hoc calibration methods, the model is trained based on a set of \( N_t \) training samples \( \{x^{(i)}, y^{(i)}\}_{i=1}^{N_t}, x^{(i)} \in \mathcal{D}, y^{(i)} \in \{1, ..., L\} \). Let random variables \( X, Y \) represent input data and label, respectively. Then,
another random variable $Z = f_0(X)$ stands for the probability vector. If $z_l$ indeed represents the actual probability of class $l$ (which usually not), then, the following should hold:

$$P(Y = l | Z = z) = z_l.$$  \hfill (2)

At this time, we also call the classifier $f_0$ to be perfectly calibrated. It is well known that the probabilities $P(Y = l | Z = z)$ are hard to evaluate, since there is no ground-truth for the probability of an input $x^{(i)}$ being misclassified as class $l \neq y^{(i)}$. In this paper, we focus on a variant of Eq. \textcolor{red}{\ref{eq:2}}, which only measures the probability of the sample being correctly classified:

$$P(Y = y^{(i)} | Z = z^{(i)}) = v^{(i)}$$  \hfill (3)

where $v^{(i)}$ is the winning score which is also the only value taken into consideration when evaluate top-1 accuracy (ACC).

**Ensemble of Deep Neural Networks.** Although it is computationally demanding, ensemble (Deep Ensemble [Lakshminarayanan et al., 2017], Snapshot Ensemble [Huang et al., 2017a], etc.) remains as a popular approach for uncertainty calibration. Formally, for a classifier $f_0$ with parameter distribution $q(\theta)$, the prediction of sample $x^{(i)}$ is given by:

$$z^{(i)}_{ens} = \int f_0(x^{(i)})q(\theta)d\theta$$  \hfill (4)

which can be approximated by $S$ independently trained classifiers as $\frac{1}{S} \sum_{s=1}^{S} f_{\theta^{(s)}}(x^{(i)}), \theta^{(s)} \sim q(\theta)$. The $S$ classifiers can be obtained either by independent random initialization (Deep Ensemble) or periodically convergence into local minimum via learning rate decay (Snapshot Ensemble). Since each $z^{(i)}_{y} \in \Delta^L$, $z^{(i)}_{ens}$ is also on the probability simplex.

**Post-hoc Calibration of Deep Neural Networks.** Until now, we have not yet introduced the concept of confidence, since all non-post-hoc approaches take the winning score as a representation of the confidence, based on which the calibration error is subsequently measured. In post-hoc calibration, on the other hand, a set of hold-out calibration samples $(x^{(i)}, y^{(i)})_{i=1}^{N_c}$ is required to learn a mapping from $z$ to another probability vector $\pi = \mathcal{T}(z)$ with a learnable function $\mathcal{T} : \Delta^L \rightarrow \Delta^L$, and max\_confidence is referred to as confidence $w$ in this paper, i.e. $w = \max_{i} \pi_i$. Note that $\text{arg max}_i \pi_i$ is not necessarily equal to $\text{arg max}_i z_i$, and at this time the calibration may potentially decrease the accuracy, if the latter is the correct class. With this, our goal Eq. \textcolor{red}{\ref{eq:3}} now becomes $P(Y = y^{(i)} | Z = z^{(i)}) = v^{(i)}$. Since we tackle both post-hoc and non-post-hoc calibrations in this paper, we distinguish winning score and confidence explicitly in that the former directly comes from the classifier $f_0$ while the latter is derived from a mapping deliberately learned for confidence modeling.

**Calibration Error Evaluation.** For the evaluation of a calibration algorithm, the calibration function $\mathcal{T}$ is applied on another evaluation dataset $(x^{(i)}, y^{(i)})_{i=1}^{N_e}$ of size $N_e$ which has no overlapping with neither the training nor the calibration datasets. In histogram-based evaluation metric, the $N_e$ samples are split into $B$ predefined bins. Formally, we define $B$ pairs of endpoints $\{(\mu_b, \nu_b)\}_{b=1}^{B}$, $\nu_b = \mu_{b+1}$, and $B$ point sets $\{\{w_i\}_{i=1}^{N_e}\}_{b=1}^{B}$ such that $\mu_b \leq w < \nu_b, \forall w \in P_b$. Then, the Expected Calibration Error (ECE) is defined as:

$$\text{ECE}(f_0) = \sum_{b=1}^{B} \frac{|P_b|}{K} \left[ \text{ACC}(P_b) - \text{conf}(P_b) \right],$$  \hfill (5)

which measures the empirical deviation of the sample accuracy in the $b^{th}$ bin: $\text{ACC}(P_b) = \frac{1}{|P_b|} \sum_{j=1}^{|P_b|} \text{arg max}_i z^{(i)} = y^{(i)}$ and the average confidence in it: $\text{conf}(P_b) = \frac{1}{|P_b|} \sum_{j=1}^{|P_b|} w^{(j)}$. The indicator function $1 : B \rightarrow \{0, 1\}$ returns 1 if the Boolean expression is true and otherwise 0.

The ECE metric can be easily affected by the number of bins $B$ and the positions of the endpoints. Without the use of binning, $\text{ECE}^{KDE}$ estimates the calibration error by a kernel function $K : \mathbb{R} \rightarrow \mathbb{R}_{>0}$ with bandwidth $h > 0$. Based on Bayesian rule, $\text{ECE}^{KDE}$ is given as:

$$\text{ECE}^{KDE}(f_0) = \int |w - \tilde{\text{ACC}}(w)| \tilde{P}(w)dw$$  \hfill (6)

in which $\tilde{\text{ACC}}(w)$ is the expected accuracy if the sample confidence is $w$. $\tilde{P}(w)$ and $\tilde{\text{ACC}}(w)$ are determined by kernel density estimation as:

$$\tilde{P}(w) = \frac{1}{N_e} \sum_{i=1}^{N_e} K_h(w - w^{(i)}),$$

$$\tilde{\text{ACC}}(w) = \frac{1}{N_e} \sum_{i=1}^{N_e} \frac{\text{arg max}_i z^{(i)} = y^{(i)}}{\sum_{i=1}^{N_e} K_h(w - w^{(i)})}.$$  

### 3 TRUTH DISCOVERY ENSEMBLE

Existing ensemble techniques in deep learning take the average of predictions made by multiple classifiers derived either from random initialization (Deep Ensemble [Lakshminarayanan et al., 2017]), periodically learning rate decay (Snapshot Ensemble [Huang et al., 2017a]), or from connected optima on the loss functions (Fast Geometric Ensembling [Garipov et al., 2018]), but scarcely utilize the sample level variance among the members of an ensemble. To make use of such information, we first introduce a vanilla truth discovery algorithm in Deep Ensemble context, and then extend it to one with accuracy-preserving guarantee.
3.1 TRUTH DISCOVERY WITHIN PROBABILITY SIMPLEX

To be consistent with truth discovery literature, we use sources to denote the $S$ independently trained models. For every sample $(x^{(i)}, y^{(i)})$ in the evaluation dataset, $S$ independent predictions $z^{(i,s)} = f_{θ^{(s)}}(x^{(i)})$ are made from all $S$ sources (denoted by $z_s$ henceforth for brevity). Since the classifiers were trained with stochastic gradient descent (SGD), they may make wrong decisions on every $x^{(i)}$. To model such a behavior, we assign a reliability value $ω_s$ to each classifier $f_{θ^{(s)}}$.

Definition 3.1 (Truth discovery [Li et al.] [2014]). Given the set of points $\{z_s\}_{s=1}^S \subseteq \triangle^L$ from $S$ classifiers, truth discovery aims at finding the truth probability vector $z^* \in \triangle^L$ and meanwhile the reliability $ω_s$ for the $s^{th}$ classifier, such that the following objective function is minimized:

$$\begin{align*}
\min_{z^*, \{ω_s\}_{s=1}^S} & \sum_{s=1}^S ω_s||z^* - z_s||^2 \\
\text{s.t.} & \sum_{s=1}^S e^{-ω_s} = 1.
\end{align*}$$

(7)

With this definition, interestingly, we can show a direct relationship between truth discovery and model uncertainty. We additionally define uncertainty of source, $υ_s$, as the opposite of source reliability, i.e. $υ_s = e^{-ω_s}$. Then, (7) can be written as:

$$\begin{align*}
\min_{z^*, \{υ_s\}_{s=1}^S} & \sum_{s=1}^S -\frac{1}{2}||z^* - z_s||^2 \ln υ_s \\
\text{s.t.} & \sum_{s=1}^S υ_s = 1.
\end{align*}$$

(8)

This is essentially the Cross Entropy (CE) of source uncertainty $υ_s$ and $||z^* - z_s||^2/2$, which is the similarity between the optimum probability vector to each source vector $(0 ≤ υ_s ≤ 1, 0 ≤ ||z^* - z_s||^2/2 ≤ 1)$. Thus, the minimization process is to ensure that the solution resolves the ambiguity of the system as much as possible. Hence, truth discovery can ideally benefit uncertainty calibration through finding the truth vector.

Algorithms for approximating the global optimum exist [Ding and Xu] [2020], [Huang et al.] [2019]. But here with the assumption Eq. 1 that all the possible truth vectors fall on the probability simplex $\triangle^L$, we adopt a simpler solution. Since both the truth vector $z^*$ and source reliabilities are unknown, we can alternatively update the reliability/uncertainty and the truth vector. Specifically, if $z^*$ is temporarily fixed, the optimum reliability values can be found through Lemma 3.1.

Lemma 3.1 (Li et al.] [2014]). If $z^*$ is fixed, the following reliability value for each source $ω_s$ minimizes the objective function (7).

$$ω_s = \ln(\frac{\sum_{s=1}^S ||z^* - z_s||^2}{||z^* - z_s||^2}).$$

(9)

After the reliabilities have been fixed, the new truth vector can be updated by simply taking the average of source vectors weighted by the found reliabilities, i.e., $\sum_{s=1}^S ω_s z_s$. It can be easily justified that the updated vector is still on the probability simplex. Initially, the ensemble vector $z_{ens}$ can be an educated guess of $z^*$. The iterative updating of the truth vector and the source reliability can be terminated if the position of the truth vector changed by less than $ε$ within maximum $I$ iterations. The process is summarized in Algorithm 1 (ignore line #5 at this time).

Algorithm 1: Optimization of the truth vector.

Data: $\{z_s\}_{s=1}^S$

Result: $z^*$

1. $z^{*(0)} \leftarrow z_{ens}$;
2. for $i \leftarrow 1, ..., I$
3. $ω_s^{(i)} \leftarrow \frac{\ln(\sum_{s=1}^S ||z^{*(i-1)} - z_s||^2)}{||z^{*(i-1)} - z_s||^2}$;
4. $z^{*(i)} \leftarrow \sum_{s=1}^S ω_s^{(i)} z_s$;  \hspace{1em} \triangleright update truth vector
5. $z^{*(i)} \leftarrow Algorithm 1[z^{*(i)}]$;  \hspace{1em} \triangleright preserve accuracy
6. if $||z^{*(i)} - z^{*(i-1)}||^2 < ε$ then
7. \hspace{2em} return $z^{*(i)}$

3.2 ACCURACY-PRESERVING TRUTH DISCOVERY

Post-hoc calibration methods may potentially cause a decrease of the prediction accuracy, if the ranks of the scores for each class cannot be maintained. Hence, we usually anticipate the calibration algorithm to be accuracy-preserving [Zhang et al.] [2020]. In the vanilla truth discovery algorithm, however, this requirement may not be satisfied, since $z^*$ could change the predictions. Suppose that $c$ is the predicted class derived from ensemble $z_{ens}$, i.e., $c = \arg\max_i(z_{ens})_i$. 
After the truth vector $z^*$ is found, we want to at least maintain the accuracy of the ensemble. Thus, Eq. (7) can be retrofitted to be an accuracy-preserving version.

**Definition 3.2** (Accuracy-preserving truth discovery). Given the set of points $\{z_s\}_{s=1}^S \subseteq \Delta^L$ and the ensemble vector $z_{ens}$, find the truth vector $z^* \in \Delta^L$ and reliabilities $\{\omega_s\}_{s=1}^S$ such that the following objective function is minimized:

$$\min_{z^*, \{\omega_s\}_{s=1}^S} \sum_{s=1}^S \omega_s ||z^* - z_s||^2$$

s.t. $\sum_{s=1}^S e^{-\omega_s} = 1$

$$\arg \max_l z^*_l = \arg \max_l (z_{ens})_l.$$  

It can be formulated as a geometric optimization problem. See Figure 1 for an illustration when $L = 3$. The constraint $\arg \max_l z^*_l = \arg \max_l (z_{ens})_l$ introduces a subspace $\Omega : z_l > z_m, \forall m \neq l$. The discovery of the truth vector must be performed within the accuracy-preserving simplex $\Delta_a = \Delta_L \cap \Omega$. Intuitively, when $z^*$ falls outside of the accuracy-preserving simplex $\Delta_a$ in one iteration, we can find the projection of $z^*$ onto $\Delta_a$ to pull it back to $\Delta_a$. This projection can be found by Algorithm 2 which is proved in Theorem 3.1. When the projection is done, Algorithm 1 continues until the desired truth vector is found.

**Theorem 3.1.** Algorithm 2 preserves the accuracy of the prediction.

**Proof.** The theorem can be proved using Lagrange Multipliers with Karush-Kuhn-Tucker (KKT) Conditions, see supplementary material Sec. A.1 for details. \hfill $\square$

**Algorithm 2:** Projection from truth vector $z$ onto accuracy-preserving simplex $\Delta_a$.

<table>
<thead>
<tr>
<th>Data: $z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result: $\tilde{z}$</td>
</tr>
<tr>
<td>1 $c \leftarrow \arg \max_l (z_{ens})_l$; $\tilde{z} \leftarrow z$;</td>
</tr>
<tr>
<td>2 $M \leftarrow \text{ARGSORT}((z_1, \ldots, z_L))$;</td>
</tr>
<tr>
<td>3 if $M[1] = c$ then</td>
</tr>
<tr>
<td>4 return $\tilde{z}$</td>
</tr>
<tr>
<td>5 else</td>
</tr>
<tr>
<td>6 for $l \leftarrow 1, \ldots, L$ do</td>
</tr>
<tr>
<td>7 $\tilde{z} \leftarrow \frac{1}{t+1}(\tilde{z} + z_{M[1]} + \ldots + z_{M[l]})$;</td>
</tr>
<tr>
<td>8 if $\tilde{z} &gt; z_{M[l+1]}$ then</td>
</tr>
<tr>
<td>9 $\tilde{z}<em>e \leftarrow \tilde{z}; \tilde{z}</em>{M[n]} \leftarrow \tilde{z}, \forall n \leq l$;</td>
</tr>
<tr>
<td>10 return $\tilde{z}$</td>
</tr>
</tbody>
</table>

## 4 TRUTH DISCOVERY-REGULARIZED POST-HOC CALIBRATION

Although ECE-like scores are difficult to be optimized directly, recent works have attempted to minimize ECE either by using maximum mean calibration error (MMCE), a kernelized version of the ECE [Kumar et al., 2018], or by rank preserving transforms [Bai et al., 2021]. To provide a better solution, we formulate the minimization of ECE (and also ECE\(_{KDE}\)) as an optimization problem in high dimensions, and show how it can be easily extended to incorporate the information gained from truth discovery.

**Optimization of ECEs.** For simplicity, we only consider the confidence for the ground-truth (namely, top-1) class, which is essentially the probability of a sample being correctly predicted. Then, our learnable mapping becomes $w = T(z) : \Delta_L \rightarrow \mathbb{R}$. One step further, if only the winning score is considered, then $w = T(a) : \mathbb{R} \rightarrow \mathbb{R}$. Next, we find the specific form of $T$. It has been shown that deep neural networks tend to be overconfident on most of the predictions [Guo et al., 2017]. Inspired by this, we impose an attenuation factor $\varphi(v)$ on every sample, which is a function of the winning score $v$ so that the adjusted confidence becomes $w = v - \varphi(v)$. The simplest form of $\varphi(v)$ is to use a constant within a bin $P_b$. Thus, we define an attenuation weight $\psi_b$ for the bin $P_b$. All the attenuation weights $\{\psi_b\}_{b=1}^B$ can be viewed as a point $\psi \in \mathbb{R}^B$. Now we have the definition of the mapping function:

$$w = v - \varphi(v) = v - \psi_K$$

s.t. $\mu_K \leq v < \nu_K.$

Notice that for consistency we call $\varphi(v)$ the attenuation factor, but it can also enhance the confidence $w$ if $\varphi(v) < 0$ somewhere.

Now our goal is to find the location of $\psi$ in $\mathbb{R}^B$ such that the expected calibration error is minimized:

$$\min_{\psi_b} \text{ECE}(\{P_b\}_{b=1}^B),$$

where ECE is shown in Eq. (5). Since all the computations in ECE (and ECE\(_{KDE}\)) are differentiable, the minimization of ECEs can be done by gradient descent methods. Here, a mini-batch Stochastic Gradient Descent (SGD) approach is used. In each epoch, a subset of calibration data is sampled, based on which the attenuation weights are updated. The optimization process is encapsulated in Algorithm 3, in which ECE and ECE\(_{KDE}\) can be used interchangeably. The algorithm can be easily implemented by virtue of automatic differentiation libraries e.g. PyTorch [Paszke et al., 2019].

### Compositional Approach for the Optimization of ECEs.

Even an accelerated method for KDE computation is used [O’Brien et al., 2016]. KDE-based metric is still much more...
Table 1: Comparison of TDE/aTDE with Deep Ensemble (DE) before post-hoc calibration.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>ECE$^{KDE}$</th>
<th>ECE$^{↓}$</th>
<th>ACC$^\uparrow$</th>
<th>ECE$^{KDE}$</th>
<th>ECE$^{↓}$</th>
<th>ACC$^\uparrow$</th>
<th>ECE$^{KDE}$</th>
<th>ECE$^{↓}$</th>
<th>ACC$^\uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PreResNet110</td>
<td>DE</td>
<td>2.88</td>
<td>3.06</td>
<td>82.83</td>
<td>1.31</td>
<td>0.48</td>
<td>96.38</td>
<td>1.15</td>
<td>0.43</td>
<td>96.41</td>
</tr>
<tr>
<td></td>
<td>TDE</td>
<td>1.60</td>
<td>1.81</td>
<td>82.89</td>
<td>1.05</td>
<td>0.75</td>
<td>96.38</td>
<td>1.01</td>
<td>0.59</td>
<td>96.42</td>
</tr>
<tr>
<td></td>
<td>aTDE</td>
<td>1.55</td>
<td>1.78</td>
<td>82.83</td>
<td>1.07</td>
<td>0.75</td>
<td>96.38</td>
<td>1.02</td>
<td>0.60</td>
<td>96.41</td>
</tr>
<tr>
<td>PreResNet164</td>
<td>DE</td>
<td>2.39</td>
<td>2.46</td>
<td>83.52</td>
<td>1.32</td>
<td>0.34</td>
<td>96.68</td>
<td>1.31</td>
<td>0.35</td>
<td>96.67</td>
</tr>
<tr>
<td></td>
<td>TDE</td>
<td>1.31</td>
<td>1.44</td>
<td>83.49</td>
<td>1.14</td>
<td>0.65</td>
<td>96.66</td>
<td>1.08</td>
<td>0.52</td>
<td>96.68</td>
</tr>
<tr>
<td></td>
<td>aTDE</td>
<td>1.33</td>
<td>1.42</td>
<td>83.52</td>
<td>1.13</td>
<td>0.63</td>
<td>96.68</td>
<td>1.08</td>
<td>0.53</td>
<td>96.67</td>
</tr>
<tr>
<td>WideResNet</td>
<td>DE</td>
<td>6.45</td>
<td>6.40</td>
<td>84.38</td>
<td>1.06</td>
<td>0.35</td>
<td>97.17</td>
<td>1.20</td>
<td>0.41</td>
<td>97.20</td>
</tr>
<tr>
<td></td>
<td>TDE</td>
<td>5.48</td>
<td>5.58</td>
<td>84.37</td>
<td>1.14</td>
<td>0.49</td>
<td>97.15</td>
<td>1.10</td>
<td>0.40</td>
<td>97.17</td>
</tr>
<tr>
<td></td>
<td>aTDE</td>
<td>5.49</td>
<td>5.57</td>
<td>84.38</td>
<td>1.13</td>
<td>0.47</td>
<td>97.17</td>
<td>1.12</td>
<td>0.41</td>
<td>97.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ImageNet</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50</td>
<td>DE/SE</td>
<td>3.11</td>
<td>3.24</td>
<td>79.25</td>
<td>1.85</td>
<td>2.11</td>
<td>78.46</td>
<td>1.73</td>
<td>2.03</td>
<td>78.52</td>
</tr>
<tr>
<td></td>
<td>TDE</td>
<td>2.16</td>
<td>2.41</td>
<td>79.22</td>
<td>1.69</td>
<td>2.08</td>
<td>78.45</td>
<td>1.66</td>
<td>2.10</td>
<td>78.50</td>
</tr>
<tr>
<td></td>
<td>aTDE</td>
<td>2.19</td>
<td>2.43</td>
<td>79.25</td>
<td>1.70</td>
<td>2.10</td>
<td>78.46</td>
<td>1.69</td>
<td>2.13</td>
<td>78.52</td>
</tr>
</tbody>
</table>

Algorithm 3: Optimization of ECEs.

Data: \( \{(v^{(i)}, y^{(i)})\}_{i=1}^{N_c} \)

Result: \( \{ \psi_b \}_{b=1}^{B} \)

for epoch ← 1, ..., #epoch do

1. sample \( \{(v^{(j)}, y^{(j)})\}_{j=1}^{N_c} \) \( \sim \) \( \{(v^{(i)}, y^{(i)})\}_{i=1}^{N_c} \);
2. \( w = v - \varphi(v) \); \( \triangleright \) apply attenuation factor
3. loss ← ECE or ECE$^{KDE}$; \( \triangleright \) forward propagation
4. update \( \{ \psi_b \}_{b=1}^{B} \); \( \triangleright \) backward propagation

return \( \{ \psi_b \}_{b=1}^{B} \)

time-consuming compared to histogram-based metrics. A natural question that arises here is whether the minimization of ECE$^{KDE}$ can be sped up by the minimization of ECE. To answer this, we first find the attention weights by using ECE as the loss function, and then use the obtained \( \{ \psi_b \}_{b=1}^{B} \) as an initial guess for the minimization of ECE$^{KDE}$. This approach enables the compositional optimization of histogram-based and KDE-based calibration errors.

ECE Optimization Regularized by Truth Discovery. Given a discovered truth vector \( z^* \), let \( V \) denote the total squared distance to \( z^* \) (i.e., \( V = \sum_{s=1}^{S} ||z^* - z_s||^2 \)) and \( q_s \) denote the contribution of each \( z_s \) to \( V \) (i.e., \( q_s = ||z^* - z_s||^2 / V \)). Then, the entropy induced by \( \{ q_s \}_{s=1}^{S} \) is:

\[
H = - \sum_{s=1}^{S} q_s \log q_s = \frac{1}{V} \sum_{s=1}^{S} ||z^* - z_s||^2 \log \frac{V}{||z^* - z_s||^2}.
\]

Based on these, we can define the **Entropy based Geometric Variance** (HV):

**Definition 4.1** (Entropy based Geometric Variance [Ding and Xu 2020]). Given the point set \( \{ z_s \}_{s=1}^{S} \) and a point \( z^* \), the entropy based Geometric Variance \( (HV) \) is \( H \times V \) where \( H \) and \( V \) are defined as shown above.

With this definition, it is easy to see that the objective function of truth discovery \( (7) \) is exactly the entropy based geometric variance \( (HV) \) and the optimization problem \( (7) \) is equivalent to finding a point \( z^* \) to minimize \( HV \).

If the truth vector \( z^* \) has been determined, then \( HV \) is an indicator of the ambiguity of the system, and can be borrowed as an external information for our ECE optimization. Despite being in the same bin and overconfident, the sample confidences should not be attenuated at the same scale. Instead, the sample with higher \( HV \) (i.e., higher variance and uncertainty) is to be attenuated by a larger magnitude. Consequently, the mapping function can be reshaped as:

\[
w^{(i)} = v^{(i)} - \varphi(v^{(i)}) = v^{(i)} - \alpha_1 \psi_i - \alpha_2 HV^{(i)},
\]

where \( \alpha_1, \alpha_2 \) are hyperparameters, and \( HV^{(i)} \) is essentially the value of the objective function of truth discovery as computed in Section 3.2 for each sample \( (x^{(i)}, y^{(i)}) \) in a total of \( N_c + N_e \) calibration and evaluation samples. The learning of the mapping function \( \mathcal{T} \) from the calibration data is a supervised learning problem. Hence, \( \mathcal{T} \) is expected to be overwritten to calibration data. By incorporating orthogonal information (i.e. \( HV \)) acquired from the truth discovery of multiple ensemble classifiers, we can learn a mapping \( \mathcal{T} \) that generalizes better on the evaluation datasets.

**5 EXPERIMENTS**

The main goals of our experiments are to: (1) compare Truth Discovery Ensemble (TDE), especially the accuracy-preserving version (aTDE), with ensemble-based calibration
Figure 2: Results of DE/TDE/aTDE on PreResNet164 trained upon CIFAR100 in an unsupervised manner (i.e. no hold-out calibration set). The number of sources is increased from 10 to 100. Our method works favorably with various metrics for the evaluation of calibration.

Table 2: Comparison of post-hoc calibration methods. The best results are highlighted in bold. We also underline the best results excluding our method. ECEs are reported in terms of mean±standard deviation obtained from 5 random replications.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>DE</th>
<th>TS</th>
<th>ETS</th>
<th>IRM</th>
<th>pTDE</th>
<th>DE</th>
<th>TS</th>
<th>ETS</th>
<th>IRM</th>
<th>pTDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>ResNet18</td>
<td>2.86±0.42</td>
<td>2.70±0.36</td>
<td>2.31±0.30</td>
<td>5.25±0.35</td>
<td>1.64±0.37</td>
<td>2.41±0.35</td>
<td>2.37±0.37</td>
<td>2.25±0.33</td>
<td>2.09±0.36</td>
<td>1.63±0.37</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet121</td>
<td>1.69±0.15</td>
<td>1.68±0.14</td>
<td>1.45±0.15</td>
<td>1.70±0.19</td>
<td>1.29±0.19</td>
<td>1.67±0.07</td>
<td>1.66±0.08</td>
<td>1.65±0.12</td>
<td>1.60±0.15</td>
<td>1.49±0.12</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNetX29</td>
<td>1.92±0.26</td>
<td>1.93±0.27</td>
<td>1.54±0.22</td>
<td>1.88±0.36</td>
<td>1.20±0.34</td>
<td>1.89±0.12</td>
<td>1.86±0.07</td>
<td>1.87±0.11</td>
<td>1.74±0.24</td>
<td>1.63±0.12</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>ResNet18</td>
<td>0.94±0.14</td>
<td>0.65±0.13</td>
<td>0.65±0.13</td>
<td>0.67±0.24</td>
<td>0.75±0.10</td>
<td>1.50±0.11</td>
<td>1.48±0.06</td>
<td>1.48±0.06</td>
<td>1.35±0.08</td>
<td>1.21±0.18</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>DenseNet121</td>
<td>0.88±0.11</td>
<td>0.65±0.13</td>
<td>0.65±0.13</td>
<td>0.55±0.06</td>
<td>0.50±0.14</td>
<td>1.78±0.17</td>
<td>1.76±0.16</td>
<td>1.76±0.16</td>
<td>1.56±0.15</td>
<td>1.45±0.20</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>ResNetX29</td>
<td>0.46±0.09</td>
<td>0.33±0.10</td>
<td>0.34±0.10</td>
<td>0.45±0.13</td>
<td>0.39±0.15</td>
<td>1.38±0.06</td>
<td>1.43±0.09</td>
<td>1.43±0.09</td>
<td>1.42±0.10</td>
<td>1.36±0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#sources</th>
<th>Dataset</th>
<th>Model</th>
<th>DE</th>
<th>TS</th>
<th>ETS</th>
<th>IRM</th>
<th>pTDE</th>
<th>DE</th>
<th>TS</th>
<th>ETS</th>
<th>IRM</th>
<th>pTDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>10 (DE)</td>
<td>3.06±0.12</td>
<td>1.59±0.09</td>
<td>0.96±0.07</td>
<td>1.93±0.16</td>
<td>0.88±0.11</td>
<td>3.13±0.11</td>
<td>1.16±0.14</td>
<td>1.06±0.11</td>
<td>1.79±0.11</td>
<td>0.93±0.08</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>30 (DE)</td>
<td>3.19±0.06</td>
<td>1.47±0.15</td>
<td>0.87±0.10</td>
<td>1.89±0.12</td>
<td>0.81±0.16</td>
<td>3.26±0.05</td>
<td>1.10±0.12</td>
<td>0.97±0.09</td>
<td>1.75±0.07</td>
<td>0.95±0.11</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>10 (SE)</td>
<td>2.32±0.09</td>
<td>1.56±0.08</td>
<td>1.04±0.10</td>
<td>1.74±0.04</td>
<td>0.71±0.17</td>
<td>2.18±0.09</td>
<td>1.14±0.06</td>
<td>1.07±0.08</td>
<td>1.51±0.07</td>
<td>0.95±0.07</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>30 (SE)</td>
<td>2.13±0.07</td>
<td>1.57±0.07</td>
<td>0.95±0.15</td>
<td>1.70±0.07</td>
<td>0.76±0.22</td>
<td>1.89±0.09</td>
<td>1.16±0.08</td>
<td>0.95±0.17</td>
<td>1.43±0.11</td>
<td>0.87±0.22</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Ablation Study of the proposed truth discovery-regularized post-hoc calibration (pTDE). For the four variants of pTDE, the blue/red color denotes if compositional training (Comp./)truth-discovery regularization (Truth. Reg.) is utilized.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>optbw</th>
<th>optKDE</th>
<th>pTDEbw</th>
<th>pTDEKDE</th>
<th>optbw</th>
<th>optKDE</th>
<th>pTDEbw</th>
<th>pTDEKDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>ResNet18</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet121</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNetX29</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>ResNet18</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>DenseNet121</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>ResNetX29</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>ImageNet</td>
<td>10 (DE)</td>
<td>0.76±0.13</td>
<td>0.83±0.13</td>
<td>0.81±0.09</td>
<td>0.88±0.11</td>
<td>0.99±0.08</td>
<td>0.92±0.07</td>
<td>1.01±0.08</td>
<td>0.93±0.08</td>
</tr>
<tr>
<td>ImageNet</td>
<td>30 (DE)</td>
<td>0.88±0.11</td>
<td>0.85±0.11</td>
<td>0.87±0.16</td>
<td>0.81±0.16</td>
<td>1.09±0.12</td>
<td>1.02±0.09</td>
<td>1.03±0.10</td>
<td>0.95±0.11</td>
</tr>
<tr>
<td>ImageNet</td>
<td>10 (SE)</td>
<td>0.80±0.19</td>
<td>0.73±0.19</td>
<td>0.75±0.19</td>
<td>0.71±0.17</td>
<td>1.04±0.12</td>
<td>0.96±0.09</td>
<td>1.01±0.11</td>
<td>0.95±0.07</td>
</tr>
<tr>
<td>ImageNet</td>
<td>30 (SE)</td>
<td>0.81±0.16</td>
<td>0.76±0.20</td>
<td>0.78±0.17</td>
<td>0.76±0.22</td>
<td>0.98±0.19</td>
<td>0.90±0.22</td>
<td>0.95±0.22</td>
<td>0.87±0.22</td>
</tr>
</tbody>
</table>
within typically 5 iterations. Were generated from the \( S \) (TDE) and its accuracy-preserving counterpart (aTDE), we further investigate the stability of TDE/aTDE with different number of sources by changing \( S \) by an interval of 10 for CIFARs and 5 for ImageNet, as illustrated in Figure 2 and Tables A1, A2, A3, and A4. Interestingly, Deep Ensemble tends to be overconfident with larger number of sources, i.e., ensemble members, while TDE/aTDE works favorably with even larger amount of available sources, and this is when a high accuracy is usually reached (Figure 2c), suggesting TDE and aTDE’s superior ability in utilizing information from multiple sources than Deep Ensemble.

5.1 IMPROVED DEEP ENSEMBLE BY TRUTH DISCOVERY

Experimental Setup. For a fair comparison, we downloaded the trained models of A Ashukha et al. [2020] including PreResNet110/164 [He et al. 2016] and WideResNet28x10 [Zagoruyko and Komodakis 2016] trained on CIFAR10/100 [Krizhevsky 2009] (10/100 classes), and ResNet50 trained on ImageNet [Deng et al. 2009] (10000 classes). All 3 network architectures on CIFAR10/100 were trained 100 times (i.e. \( S = 100 \)) following the Deep Ensemble (DE) workflow, while ResNet50 was trained on ImageNet resulting in 50 models either by Deep Ensemble or by Snapshot Ensemble (i.e. \( S = 50 \)). For every sample in the standard testing dataset, \( S \) ensemble members were generated from the \( S \) models. While looking for the truth vector, for both the vanilla Truth Discovery Ensemble (TDE) and its accuracy-preserving counterpart (aTDE), we set \( \epsilon = e^{-8} \) in all experiments, and observed a convergence within typically 5 iterations.

Results. The comparison of truth discovery ensemble methods with Deep Ensemble on CIFARs (\( S = 50 \) or 100), and with Deep Ensemble/Snapshot Ensemble on ImageNet (\( S = 25 \) or 50) is shown in Table 1. Clearly, TDE ameliorates either ECE or \( \text{ECE}^\text{KDE} \) by a large margin in most of the experimental settings, especially on datasets with higher complexity (i.e., CIFAR100 and ImageNet), but fails at maintain accuracy. The accuracy-preserving version aTDE, on the other hand, successfully preserves the accuracy, with nearly the same capability of lowering the ECEs, which validates the correctness of our accuracy-preserving Algorithm 2. It can also be concluded from Table 1 that higher ACC and lower ECEs are hard to be reached simultaneously, but the metrics contributed to by the two variants TDE/aTDE are usually very similar. The KS metric recently proposed by Gupta et al. [2021] which measures the maximal distance between the accumulated output probability to the actual probability, is a binning-free calibration evaluator that different from ECEs. The KS error is also measured for all the experiments. By leveraging truth discovery, our TDE method lowers the KS to as low as 0.6% (100 sources) as shown in Figure 2(e), even without any hold-out calibration sample.

Further, we investigate the stability of TDE/aTDE with different number of sources by downloading from https://github.com/bayesgroup/pytorch-ensembles

5.2 IMPROVED POST-HOC CALIBRATION BY TRUTH DISCOVERY-REGULARIZED OPTIMIZATION

Experimental Setup. In this section, we evaluate the performance of our truth discovery-regularized post-hoc calibration methods (pTDE), to which the information elicited from ensemble-based methods is incorporated. To this end, we first train the vanilla optimization method using histogram-based ECE as the loss function (opt\( ^\text{hist} \)) as described in Section 4 with batch size at 1000 for CIFARs and 10000 for ImageNet for 70 epochs. Then, we apply compositional training by switching to \( \text{KDE} \)-based loss function (opt\( ^\text{KDE} \)) for 5 additional epochs. To leverage the information gained from truth discovery, the entropy based geometric variance (HV) values are computed for all \((N_c + N_r)\) samples. By taking HV into the training process, opt\( ^\text{hist} \) is promoted to truth discovery-regularized post-hoc calibration pTDE\( ^\text{hist} \), which is subsequently optimized for 5 additional epochs using ECE\( ^\text{KDE} \) as the loss function to be pTDE\( ^\text{KDE} \). Finally, pTDE\( ^\text{KDE} \) (or pTDE for short) is compared with several state-of-the-art post-hoc calibration methods, namely, Temperature Scaling (TS) [Guo et al. 2017], Ensemble Temperature Scaling (ETS) [Zhang et al. 2020], and multi-class isotonic regression (IRM) [Zhang et al. 2020].

Results. The attenuation/enhancement factor we apply in Eq. 11 enables an iterative rearrangement of samples across bins, until a small discrepancy between confidence and actual accuracy is achieved within every bin. Figure 3 shows how this recalibration (vanilla opt\( ^\text{hist} \)) affects confidence distribution. From Table 2 we can see that our method (pTDE) significantly outperforms all competing methods (except on CIFAR10 with ECE). It is worth noting that when pTDE is excluded, there is no sweeping method under every circumstances, but pTDE overall shows better consistency especially with the ECE\( ^\text{KDE} \) metric which is not susceptible to the binning strategy, e.g., the number and positions of the bins. To inspect the individual contributions
Table 4: KS error (in %) on ImageNet evaluation dataset by various post-hoc calibration methods including four variants of our proposed method. The best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Deep Ensemble</th>
<th>Snapshot Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S = 10$</td>
<td>$S = 30$</td>
</tr>
<tr>
<td>DE/SE</td>
<td>2.71±0.10</td>
<td>2.86±0.03</td>
</tr>
<tr>
<td>TS</td>
<td>0.88±0.13</td>
<td>1.03±0.09</td>
</tr>
<tr>
<td>ETS</td>
<td>0.59±0.13</td>
<td>0.42±0.10</td>
</tr>
<tr>
<td>IRM</td>
<td>0.97±0.13</td>
<td>0.93±0.10</td>
</tr>
<tr>
<td>Spline</td>
<td>0.38±0.09</td>
<td>0.34±0.07</td>
</tr>
<tr>
<td>optHist</td>
<td>0.43±0.20</td>
<td>0.39±0.16</td>
</tr>
<tr>
<td>optKDE</td>
<td>0.37±0.15</td>
<td>0.33±0.11</td>
</tr>
<tr>
<td>pTDEHist</td>
<td>0.42±0.18</td>
<td>0.36±0.19</td>
</tr>
<tr>
<td>pTDEKDE</td>
<td><strong>0.37±0.14</strong></td>
<td><strong>0.27±0.12</strong></td>
</tr>
</tbody>
</table>

6 CONCLUSION

In this work, we first present Truth Discovery Ensemble (TDE) that neither requires hold-out calibration data nor alters any training process, but significantly surpasses the original ensemble result, and in the meanwhile preserves the accuracy (aTDE). For post-hoc calibration, the superiority of our final methods (pTDE) is attributed not only to truth discovery, but also to the compositional training strategy. In conclusion, truth discovery is well positioned to assist both ensemble-based and post-hoc calibration. We hope that the proposed calibrators augmented by truth discovery can enlarge the arsenal of uncertainty calibration methods for deep learning. Our source code is available at https://github.com/horsepurve/truly-uncertain.

Acknowledgements

This research was supported in part by NSF through grants CCF-1716400 and IIS-1910492.

References


Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In International Conference on Management of Data, SIGMOD, pages 1187–1198, 2014.


