Diagnostic Uncertainty Calibration: Towards Reliable Machine Predictions in Medical Domain (Appendix)

A Background for proper loss decomposition

We describe the proofs for proper loss decompositions introduced in section 2.

A.1 Decomposition of proper losses and calibration

As we have described in section 2.2, the expected loss L can be decomposed as follows:

Theorem 2 (DeGroot and Fienberg (1983))**.** *The expectation of proper loss ℓ is decomposed into non-negative terms as follows:*

$$L = \mathrm{CL} + \mathrm{RL}, \quad \textit{where} \quad \begin{cases} \mathrm{CL} := \mathbb{E}[d(C, Z)], & (\textit{Calibration Loss}) \\ \mathrm{RL} := \mathbb{E}[d(Y, C)], & (\textit{Refinement Loss}) \end{cases} \tag{20}$$

where a calibration map $C := \mathbb{E}[Y|Z] \in \Delta^{K-1}$ is defined as in Def. 1.

Proof.

$$\begin{aligned} L &= \mathbb{E}[d(Y, Z)] \\ &= \mathbb{E}[\ell(Y, Z) - \ell(Y, Y)] \\ &= \mathbb{E}[\ell(Y, Z) - \ell(Y, C)] + \mathbb{E}[\ell(Y, C) - \ell(Y, Y)] \\ &= \mathbb{E}[\mathbb{E}[\ell(Y, Z) - \ell(Y, C)|Z]] + \mathbb{E}[d(Y, C)], \end{aligned}$$

where the second term equals to the RL term. For the first term, as we have defined $\ell(q, Z) := \mathbb{E}_{Y \sim \mathrm{Cat}(q)}[\ell(Y, Z)]$ when $q \in \Delta^{K-1}$, the subterms can be rewritten as follows:

$$\begin{aligned} \mathbb{E}[\ell(Y, Z)|Z] &= \mathbb{E}_{Y \sim \mathrm{Cat}(\mathbb{E}[Y|Z])}[\ell(Y, Z)] = \ell(\mathbb{E}[Y|Z], Z) = \ell(C, Z), \\ \mathbb{E}[\ell(Y, C)|Z] &= \mathbb{E}_{Y \sim \mathrm{Cat}(\mathbb{E}[Y|Z])}[\ell(Y, C)] = \ell(\mathbb{E}[Y|Z], C) = \ell(C, C). \end{aligned}$$

Hence, the first term equals to the CL term as follows:

$$\mathbb{E}[\mathbb{E}[\ell(Y, Z) - \ell(Y, C)|Z]] = \mathbb{E}[\mathbb{E}[\ell(C, Z) - \ell(C, C)|Z]] = \mathbb{E}[d(C, Z)].$$

□

Note that we have followed the terminology used in Kull and Flach (2015). The terms CL and RL are also referred to as reliability (Bröcker, 2012; Ferro and Fricker, 2012) and sharpness (DeGroot and Fienberg, 1983).

A.2 Decomposition of proper losses under label uncertainty

As we have described in section 2.2, if Y follows an instance-wise categorical distribution with a probability vector, *i.e.*, $Y|X \sim \mathrm{Cat}(Q)$, where $Q(X) \in \Delta^{K-1}$, L can be further decomposed as follows:

Theorem 3 (Kull and Flach (2015))**.** *The expectation of proper loss ℓ is decomposed into non-negative terms as follows:*

$$L = \mathrm{EL} + \mathrm{IL} = \underbrace{\mathrm{CL} + \mathrm{GL}}_{\mathrm{EL}} + \mathrm{IL}, \tag{21}$$

$$\textit{where} \quad \begin{cases} \mathrm{EL} = \mathbb{E}[d(Q,Z)], & (\textit{Epistemic Loss}) \\ \mathrm{IL} = \mathbb{E}[d(Y,Q)], & (\textit{Irreducible Loss}) \\ \mathrm{GL} = \mathbb{E}[d(Q,C)]. & (\textit{Grouping Loss}) \end{cases} \tag{22}$$

Note that the CL term is the same form as in equation (5).

Proof. We first prove the first equality.

$$\begin{aligned} L &= \mathbb{E}[d(Y,Z)] \\ &= \mathbb{E}[\ell(Y,Z) - \ell(Y,Y)] \\ &= \mathbb{E}[\ell(Y,Z) - \ell(Y,Q)] + \mathbb{E}[\ell(Y,Q) - \ell(Y,Y)] \\ &= \mathbb{E}[\mathbb{E}[\ell(Y,Z) - \ell(Y,Q)|Q]] + \mathbb{E}[d(Y,Q)], \end{aligned}$$

where the second term is IL. As similar to the proof of Theorem 2, the following relations hold:

$$\begin{aligned} \mathbb{E}[\ell(Y,Z)|Q] &= \mathbb{E}_{Y\sim Q}[\ell(Y,Z)|Q] = \mathbb{E}[\ell(Q,Z)|Q], \\ \mathbb{E}[\ell(Y,Q)|Q] &= \mathbb{E}_{Y\sim Q}[\ell(Y,Q)|Q] = \mathbb{E}[\ell(Q,Q)|Q]. \end{aligned}$$

Therefore, the first term turns out to be EL as follows:

$$\mathbb{E}[\mathbb{E}[\ell(Y,Z) - \ell(Y,Q)|Q]] = \mathbb{E}[\mathbb{E}[\ell(Q,Z) - \ell(Q,Q)|Q]] = \mathbb{E}[d(Q,Z)].$$

This term is further decomposed as follows:

$$\begin{aligned} \mathbb{E}[d(Q,Z)] &= \mathbb{E}[\ell(Q,Z) - \ell(Q,C)] + \mathbb{E}[\ell(Q,C) - \ell(Q,Q)] \\ &= \mathbb{E}[\mathbb{E}[\ell(Q,Z) - \ell(Q,C)|Z]] + \mathbb{E}[d(Q,C)], \end{aligned}$$

where the second term is GL. To show that the first term is CL, we have to prove the following results:

$$\begin{aligned} \mathbb{E}[\ell(Q,Z)|Z] &= \mathbb{E}[\ell(C,Z)|Z], \\ \mathbb{E}[\ell(Q,C)|Z] &= \mathbb{E}[\ell(C,C)|Z]. \end{aligned}$$

As these are proven with the same procedure, we only show the proof for the first equality.

$$\begin{aligned} \mathbb{E}[\ell(Q,Z)|Z] &= \mathbb{E}[\mathbb{E}_{Y\sim Q}\,\ell(Y,Z)|Z] \\ &= \mathbb{E}[\sum_k \ell(Y_k,Z)Q_k|Z] \\ &= \mathbb{E}[\sum_k \ell(Y_k,Z)\,\mathbb{E}[Y|Z]_k|Z] \\ &= \mathbb{E}[\sum_k \ell(Y_k,Z)C_k|Z] \\ &= \mathbb{E}[\mathbb{E}_{Y\sim C}[\ell(Y,Z)]|Z] = \mathbb{E}[\ell(C,Z)|Z]. \end{aligned}$$

□

Theorems and proofs for more generalized decompositions are found in Kull and Flach (2015).

B Details on CPE evaluation metrics with label histograms

We describe supplementary information for Section 3: proofs for propositions, additional discussion, and experimental setup.

B.1 Unbiased Estimators of L_{sq}

We give a proof for Prop. 1.

Restatement of Proposition 1 (Unbiased estimator of expected squared loss)**.** *The following estimator of L_{sq} is unbiased.*

$$\widehat{L}_{\mathrm{sq}} := \frac{1}{W}\sum_{i=1}^{N} w_i \sum_{k=1}^{K}\left[(\widehat{\mu}_{ik} - z_{ik})^2 + \widehat{\mu}_{ik}(1-\widehat{\mu}_{ik})\right], \tag{23}$$

where $\widehat{\mu}_{ik} := y_{ik}/n_i$, $w_i \geq 0$, and $W := \sum_{i=1}^{N} w_i$.

Proof. We begin with the following plugin estimator of L_{sq} with an instance i:

$$\widetilde{L}_{\mathrm{sq},i} := \sum_k \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ik}^{(j)} - z_{ik})^2. \tag{24}$$

By taking an expectation with respect to y_{ik} and z_{ik},

$$\mathbb{E}[\widetilde{L}_{\mathrm{sq},i}] = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_k \mathbb{E}[(y_{ik}^{(j)} - z_{ik})^2] = \sum_k \mathbb{E}[(Y_k - Z_k)^2] = \mathbb{E}[\|Y - Z\|^2] = L_{\mathrm{sq}}\,.$$

Therefore, $\widetilde{L}_{\mathrm{sq},i}$ is an unbiased estimator of L_{sq}. Intuitively, an estimator combined with N instances is expected to have a lower variance than that with a single instance. A linear combination of $\widetilde{L}_{\mathrm{sq},1}, \ldots, \widetilde{L}_{\mathrm{sq},N}$ is also an unbiased estimator as follows:

$$\frac{1}{W}\,\mathbb{E}[\sum_i w_i \widetilde{L}_{\mathrm{sq},i}] = \frac{1}{W}\sum_i w_i\,\mathbb{E}[\widetilde{L}_{\mathrm{sq},i}] = L_{\mathrm{sq}},$$

where $\sum_i w_i \geq 0, W := \sum_i w_i$. The proof completes by transforming $\widetilde{L}_{\mathrm{sq},i}$ as follows:

$$\widetilde{L}_{\mathrm{sq},i} = \sum_k \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ik}^{(j)} - 2y_{ik}^{(j)} z_{ik} + z_{ik}^2\right) = \sum_k \left(\widehat{\mu}_{ik} - 2\,\widehat{\mu}_{ik}\,z_{ik} + z_{ik}^2\right) = \sum_k \widehat{\mu}_{ik}(1-\widehat{\mu}_{ik}) + (\widehat{\mu}_{ik} - z_{ik})^2\,.$$

□

Determination of weights $\boldsymbol{w}$ For the undetermined weights $w_1, \ldots, w_N$, we have argued that the optimal weights would be constant when the numbers of annotators $n_1, \ldots, n_N$ were constant. As we assume that each of an instance i follows an independent categorical distribution with a parameter $Q_i \in \Delta^{K-1}$, the variance of $\widehat{L}_{\mathrm{sq},k}$ is decomposed as follows:

$$\mathbb{V}[\widehat{L}_{\mathrm{sq},k}] = \sum_{i=1}^{N}\left(\frac{w_i}{W}\right)^2 \mathbb{V}[\widehat{L}_{\mathrm{sq},ik}]. \tag{25}$$

Thus, if n_i is constant for all the instance, the optimal weights are found as follows:

$$\min_{\boldsymbol{w}'}\sum_{i=1}^{N} w_i'^2 \quad s.t. \quad \sum_{i=1}^{N} w_i' = 1, \quad \forall i,\, w_i' \geq 0, \tag{26}$$

By taking a derivative with respect to $\boldsymbol{w}'$ of $\sum_{i=1} w_i'^2 + \lambda(\sum_{i=1}^{N} w_i' - 1)$, the solution is $\forall i,\, w_i' = 1/N$.

For cases with varying numbers of annotators per instance, it is not straightforward to determine the optimal weights. From a standard result of variance formulas, the variance of $\widehat{L}_{\mathrm{sq}}$ is further decomposed as follows:

$$\begin{aligned}\mathbb{V}[\widehat{L}_{\mathrm{sq},ik}] &:= \mathbb{E}[\mathbb{V}[\widehat{L}_{\mathrm{sq},ik}\,|X_i]] + \mathbb{V}[\mathbb{E}[\widehat{L}_{\mathrm{sq},ik}\,|X_i]] \\ &= \frac{1}{n_i}\,\mathbb{E}[\sigma^2_{\mathrm{sq},k}(X)] + \mathbb{V}[\mu_{\mathrm{sq},k}(X)],\end{aligned}$$

where $\mu_{\mathrm{sq},k}(X) := \mathbb{E}[(Y_k - Z_k)^2|X]$ and $\sigma^2_{\mathrm{sq},k}(X) := \mathbb{V}[(Y_k - Z_k)^2|X]$. Therefore, the optimal weights depend on the ratio of the first and the second terms. If the first term is negligible compared to the second term, using the constant weights regardless of n_i would be optimal. In contrast, if the first term is dominant, $w_i \propto n_i$ would be optimal. However, the ratio of the two terms depend on the dataset and is not determined a priori. In this work, we have used $w_i = 1$.

B.2 Unbiased Estimators of EL

In this section, we give a proof for Prop. 2.

Definition 4 (Plugin estimator of EL)**.**

$$\widetilde{\mathrm{EL}} := \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\widehat{\mu}_{ik} - z_{ik})^2. \tag{27}$$

Restatement of Proposition 2 (Unbiased estimator of EL)**.** *The following estimator of EL is unbiased.*

$$\widehat{\mathrm{EL}} := \widetilde{\mathrm{EL}} - \frac{1}{N}\sum_{i}\sum_{k}\frac{1}{n_i - 1}\,\widehat{\mu}_{ik}(1 - \widehat{\mu}_{ik}). \tag{28}$$

Proof. The term EL is decomposed as $\mathrm{EL} = \sum_k \mathrm{EL}_k$, where

$$\mathrm{EL}_k = \mathbb{E}[(Q_k - Z_k)^2] = \mathbb{E}[Q_k^2] - 2\,\mathbb{E}[Q_k Z_k] + \mathbb{E}[Z_k^2].$$

As for the terms in the plugin estimator $\widetilde{\mathrm{EL}} = \sum_k \widetilde{\mathrm{EL}}_k$, we can show that

$$\begin{aligned}\mathbb{E}[\widehat{\mu}_{ik}\,z_{ik}] &= \frac{1}{n_i} n_i\,\mathbb{E}[Q_k Z_k] = \mathbb{E}[Q_k Z_k], \\ \mathbb{E}[z_{ik}^2] &= \mathbb{E}[Z_k^2].\end{aligned}$$

The bias of $\widetilde{\mathrm{EL}}_k$ comes from $\widehat{\mu}_{ik}^2$, that corresponds to $\mathbb{E}[Q_k^2]$ term. We can replace $\widehat{\mu}_{ik}^2$ by an unbiased estimator as follows:

$$\widehat{\mu}_{ik}^2 \to \frac{1}{n_i(n_i - 1)}\sum_{j=1}^{n_i}\sum_{j'=1:j'\neq j}^{n_i} y_{ik}^{(j)} y_{ik}^{(j')}, \tag{29}$$

where an expectation of each of the summand of *r.h.s.* is $\mathbb{E}[Q_k^2]$, hence that of *r.h.s.* is also be $\mathbb{E}[Q_k^2]$. Consequently, the difference of the plugin estimator $\widetilde{\mathrm{EL}}_k$ and the unbiased estimator $\widehat{\mathrm{EL}}_k$ is calculated as follows:

$$\begin{aligned}\widetilde{\mathrm{EL}}_k - \widehat{\mathrm{EL}}_k &= \frac{1}{N}\sum_i \left[\widehat{\mu}_i^2 - \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\sum_{j'=1:j'\neq j}^{n_i} y_{ik}^{(j)} y_{ik}^{(j')}\right] \\ &= \frac{1}{N}\sum_i \left\{\widehat{\mu}_i^2 - \frac{1}{n_i(n_i-1)}\left[\left(\sum_{j=1}^{n_i} y_{ik}^{(j)}\right)^2 - \sum_{j=1}^{n_i} y_{ik}^{(j)}\right]\right\} \\ &= \frac{1}{N}\sum_i \left\{\widehat{\mu}_i^2 - \frac{1}{n_i(n_i-1)}\left(n_i^2\,\widehat{\mu}_{ik}^2 - n_i\,\widehat{\mu}_{ik}\right)\right\} \\ &= \frac{1}{N}\sum_i \frac{1}{n_i - 1}\,\widehat{\mu}_{ik}\,(1 - \widehat{\mu}_{ik})\,.\end{aligned}$$

□

B.3 Debiased Estimators of CL

In this section, we give a proof for Prop. 3.

Definition 5 (Plugin estimator of CL)**.**

$$\widetilde{\mathrm{CL}}_{kb}(\mathcal{B}_k) := \frac{|I_{kb}|}{N}(\bar{c}_{kb} - \bar{z}_{kb})^2, \quad \textit{where} \quad \bar{c}_{kb} := \frac{\sum_{i\in I_{kb}} \widehat{\mu}_{ik}}{|I_{kb}|}, \quad \bar{z}_{kb} := \frac{\sum_{i\in I_{kb}} z_{ik}}{|I_{kb}|}, \tag{30}$$

where $I_{kb} := \{i \,|\, z_{ik} \in \mathcal{B}_{kb}\}$ *denotes an index set of b-th bin and* $\mathcal{B}_{kb} := [\zeta_{kb}, \zeta_{kb+1})$ *is a b-th interval of the binning scheme* $\mathcal{B}_k$*.*

Restatement of Proposition 3 (Debiased estimator of CL_{kb})**.** *The plugin estimator of* CL_{kb} *is debiased with the following estimator:*

$$\widehat{\mathrm{CL}}_{kb}(\mathcal{B}_k) := \widetilde{\mathrm{CL}}_{kb}(\mathcal{B}_k) - \frac{|I_{kb}|}{N}\frac{\bar{\sigma}^2_{kb}}{|I_{kb}|-1}, \quad \textit{where} \quad \bar{\sigma}^2_{kb} := \frac{1}{|I_{kb}|}\sum_{i\in I_{kb}} \widehat{\mu}^2_{ik} - \left(\frac{1}{|I_{kb}|}\sum_{i\in I_{kb}} \widehat{\mu}_{ik}\right)^2. \tag{31}$$

Proof. The bias of the plugin estimator $\widetilde{\mathrm{CL}}_{kb}$ is explained in a similar manner as in the case of $\widetilde{\mathrm{EL}}_k$. Concretely, a bias of the term $\bar{c}^2_{kb}$ for an estimation of $\bar{C}^2_{kb}$ can be reduced with a following replacement:

$$\bar{c}^2_{kb} \to \frac{1}{|I_{kb}|\,(|I_{kb}|-1)} \sum_{i\in I_{kb}} \sum_{i'\in I_{kb}: i'\neq i} \widehat{\mu}_{ik}\,\widehat{\mu}_{i'k}\,. \tag{32}$$

Note that the *r.h.s.* term is only defined for the bin with $|I_{kb}| > 1$. In this case, a conditional expectation of the term is as follows:

$$\begin{aligned}
\mathbb{E}[\frac{|I_{kb}|}{N} \cdot r.h.s., |I_{kb}| > 1] &= \sum_{m=2}^{N} \frac{\mathbb{E}[|I_{kb}| = m]m}{N} \frac{1}{m(m-1)} \sum_{i\in I_{kb}} \sum_{i'\in I_{kb}: i'\neq i} \mathbb{E}[\widehat{\mu}_{ik}\,\widehat{\mu}_{i'k} \,|\, |I_{kb}| = m] \\
&= \sum_{m=2}^{N} \frac{\mathbb{E}[|I_{kb}| = m]m}{N} \bar{C}^2_{kb} = \frac{\mathbb{E}[|I_{kb}| \,|\, |I_{kb}| > 1]}{N} \bar{C}^2_{kb} \\
&= (\mathbb{E}[Z_k \in \mathcal{B}_{kb}] - \eta_{kb})\,\bar{C}^2_{kb},
\end{aligned}$$

where $\eta_{kb} = \mathbb{E}[I_{kb} \leq 1]/N$, which can be reduced by increasing N relative to the bin size. When we use the $r.h.s. = 0$ for $|I_{kb}| \leq 1$, $\eta_{kb}\bar{C}^2_{kb}$ is a remained bias term after the replacement in equation (32). When we define an estimator $\widehat{\mathrm{CL}}_{kb}$ as a modified $\widetilde{\mathrm{CL}}_{kb}$ that has been applied the replacement (32), a debiasing amount of the bias with the modification is calculated as follows:

$$\begin{aligned}
\widetilde{\mathrm{CL}}_{kb} - \widehat{\mathrm{CL}}_{kb} &= \frac{|I_{kb}|}{N}\left\{\bar{c}^2_{kb} - \frac{1}{|I_{kb}|\,(|I_{kb}|-1)} \sum_{i\in I_{kb}} \sum_{i'\in I_{kb}: i'\neq i} \widehat{\mu}_{ik}\,\widehat{\mu}_{i'k}\right\} \\
&= \frac{|I_{kb}|}{N}\left\{\bar{c}^2_{kb} - \frac{1}{|I_{kb}|\,(|I_{kb}|-1)}\left[(|I_{kb}|\,\bar{c}_{kb})^2 - \sum_{i\in I_{kb}} \widehat{\mu}^2_{ik}\right]\right\} \\
&= \frac{|I_{kb}|}{N}\left\{\frac{1}{|I_{kb}|-1}\left[\frac{1}{|I_{kb}|}\sum_{i\in I_{kb}} \widehat{\mu}^2_{ik} - \bar{c}^2_{kb}\right]\right\} = \frac{|I_{kb}|}{N}\frac{\bar{\sigma}^2_{kb}}{|I_{kb}|-1}.
\end{aligned}$$

□

Note that as we mentioned in the proof, the bin-wise debiasing cannot be applied for the bins with $|I_{kb}| \leq 1$. We use 0 for the estimators with such bins. For single-labeled data, the remaining bias from this limitation is also analyzed in the literature (Ferro and Fricker, 2012).

B.4 Definition and estimators of dispersion loss

We consider to estimate the remainder term $\mathrm{EL}-\mathrm{CL}$. As we present in equation (6), EL is decomposed into $\mathrm{CL}+\mathrm{GL}$, in which GL is a loss relating to the lack of predictive sharpness. However, the approximate calibration loss $\mathrm{CL}(\mathcal{B})$ is known to be underestimated (Vaicenavicius et al., 2019; Kumar et al., 2019) in relation to the coarseness of the selected binning scheme $\mathcal{B}$. On the other hand, EL does not suffer from a resolution of $\mathcal{B}$. Instead of estimating the GL term for binned predictions with $\mathcal{B}$, we use the difference term $\mathrm{DL}(\mathcal{B}) := \mathrm{EL}-\mathrm{CL}(\mathcal{B})$, which we call dispersion loss. The non-negativity of $\mathrm{DL}(\mathcal{B})$ is shown as follows.

Proposition 4 (Non-negativity of dispersion loss)**.** *Given a binning scheme $\mathcal{B}$, a dispersion loss for class k is decomposed into bin-wise components, where each term takes a non-negative value:*

$$\mathrm{DL}_k(\mathcal{B}) := \mathrm{EL}_k - \mathrm{CL}_k(\mathcal{B}) = \sum_b \mathrm{DL}_{kb}(\mathcal{B}), \tag{33}$$

$$\mathrm{DL}_{kb}(\mathcal{B}) := \mathbb{E}[\mathbb{E}[\{(Q_k - \bar{C}_{kb}) - (Z_k - \bar{Z}_{kb})\}^2 | Z_k \in \mathcal{B}_{kb}]] \geq 0. \tag{34}$$

Proof. From the definition of $\mathrm{DL}_k(\mathcal{B})$,

$$\begin{aligned}
\mathrm{DL}_k(\mathcal{B}) &:= \mathrm{EL}_k - \mathrm{CL}_k(\mathcal{B}) \\
&= \mathbb{E}[(Q_k - Z_k)^2] - \sum_b \mathbb{E}[\mathbb{E}[(\bar{C}_{kb} - \bar{Z}_{kb})^2 \mid Z_k \in \mathcal{B}_{kb}]] \\
&= \sum_b \mathbb{E}[\mathbb{E}[(Q_k - Z_k)^2 - (\bar{C}_{kb} - \bar{Z}_{kb})^2 \mid Z_k \in \mathcal{B}_{kb}]] \\
&= \sum_b \mathrm{DL}_{kb}(\mathcal{B}), \\
&\quad \text{where} \quad \mathrm{DL}_{kb}(\mathcal{B}) := \mathbb{E}[\mathbb{E}[(Q_k - Z_k)^2 - (\bar{C}_{kb} - \bar{Z}_{kb})^2 \mid Z_k \in \mathcal{B}_{kb}]].
\end{aligned}$$

By noting that $\bar{C}_{kb} - \bar{Z}_{kb} = \mathbb{E}[Q_k - Z_k | Z_k \in \mathcal{B}_{kb}]$, the last term is further transformed as follows:

$$\begin{aligned}
\mathrm{DL}_{kb}(\mathcal{B}) &= \mathbb{E}[\mathbb{E}[(Q_k - Z_k)^2 - (\bar{C}_{kb} - \bar{Z}_{kb})^2 \mid Z_k \in \mathcal{B}_{kb}]] \\
&= \mathbb{E}[\mathbb{E}[(Q_k - Z_k)^2 - 2(Q_k - Z_k)(\bar{C}_{kb} - \bar{Z}_{kb}) + (\bar{C}_{kb} - \bar{Z}_{kb})^2 \mid Z_k \in \mathcal{B}_{kb}]] \\
&= \mathbb{E}[\mathbb{E}[\{(Q_k - Z_k) - (\bar{C}_{kb} - \bar{Z}_{kb})\}^2 \mid Z_k \in \mathcal{B}_{kb}]] \\
&= \mathbb{E}[\mathbb{E}[\{(Q_k - \bar{C}_{kb}) - (Z_k - \bar{Z}_{kb})\}^2 \mid Z_k \in \mathcal{B}_{kb}]].
\end{aligned}$$

Then, the last term is apparently ≥ 0. □

From equation (34), the DL term can be interpreted as the average of the bin-wise overdispersion of the true class probability Q_k, which is unaccounted for by the deviation of Z_k. For single-labeled cases, similar argument is found in (Stephenson et al., 2008). The plugin and debiased estimators of DL are derived from those of EL and CL, respectively.

By using the plugin and the debiased estimators of EL and CL, those estimators of DL are defined as follows:

Definition 6 (Plugin / debiased estimators of dispersion loss)**.**

$$\widetilde{\mathrm{DL}}_{kb}(\mathcal{B}) = \frac{1}{N} \sum_{i \in I_{kb}} \{(\hat{\mu}_{ik} - \bar{c}_{kb}) - (z_{ik} - \bar{z}_{kb})\}^2, \tag{35}$$

$$\widehat{\mathrm{DL}}_{kb}(\mathcal{B}) = \widetilde{\mathrm{DL}}_{kb}(\mathcal{B}) - \frac{1}{N} \sum_{i \in I_{kb}} \left(\frac{1}{n_i - 1} \hat{\mu}_{ik}(1 - \hat{\mu}_{ik}) - \frac{\bar{\sigma}^2_{kb}}{|I_{kb}| - 1} \right). \tag{36}$$

B.5 Experimental setup for debiasing effects of EL and CL terms

The details of the experimental setup for Section 3.3 are described. We experimented on evalutions of a perfect predictor that indicated correct instance-wise CPEs, using synthetic binary labels with varying instance sizes: from 100 to 10,000. For each instance, the positive probability for label generation was drawn from a uniform

distribution $U(0,1)$, and two or five labels were generated in an i.i.d. manner following a Binomial distribution with the corresponding probability. Since the predictor indicated the correct probability, EL and CL would be zero in expectation. For a binning scheme $\mathcal{B}$ of the estimators, we adopted 15 equally-spaced binning, which was regularly used to evaluate calibration errors (Guo et al., 2017).

C Details on higher-order statistics evaluation

The details and proofs for the statements in section 4 are described. Let $X \in \mathcal{X}$ be an input feature and $\{Y^{(j)} \in e^K\}_{j=1}^n$ be n distinct labels for the same instance. We define a symmetric categorical statistics $\phi : e^{K\times n} \to e^M (M \geq 2)$ for the n labels. For the case of $M = 2$, ϕ can be equivalently represented as $\phi : e^{K\times n} \to \{0, 1\}$, and we use this definition for the successive discussion. In our experiments, we particularly focus on a disagreement between paired labels $\phi^D = \mathbb{I}[Y^{(1)} \neq Y^{(2)}]$ as predictive target.

Consider a probability prediction $\varphi : \mathcal{X} \to [0, 1]$ for statistics $\phi : e^{K\times n} \to \{0, 1\}$, a strictly proper loss $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}$ encourages $\varphi(X)$ to approach the right probability $P(\phi(Y^{(1)}, \ldots, Y^{(n)})|X)$ in expectation. We use (one dimensional) squared loss $\ell(\phi, \varphi) = (\phi - \varphi)^2$ in our evaluation. The expected loss is as follows:

Definition 7 (Expected squared loss for ϕ and φ)**.**

$$L_\phi := \mathbb{E}[(\phi - \varphi)^2], \tag{37}$$

where the expectation is taken over the random variables X and $Y^{(1)}, \ldots, Y^{(n)}$.

Note that L_ϕ for an empirical distribution is equivalent to Brier score of ϕ and φ. A decomposition of L_ϕ into CL_ϕ and RL_ϕ is readily available by applying Theorem 2.

$$L_\phi := \mathbb{E}[(\phi - \varphi)^2] = \underbrace{\mathbb{E}[(\mathbb{E}[\phi|\varphi] - \varphi)^2]}_{\mathrm{CL}_\phi} + \underbrace{\mathbb{E}[(\phi - \mathbb{E}[\phi|\varphi])^2]}_{\mathrm{RL}_\phi}. \tag{38}$$

We will derive the estimators of L_ϕ and CL_ϕ as evaluation metrics. However, the number of labels per instance is $n_i \geq n$ in general[5], which results in multiple inconsistent statistics ϕ for the same instance. The problem can be solved with similar treatments as in the evaluation of CPEs.

As we stated in section 4, an unbiased estimator of the mean statistics for each instance $\mu_{\phi,i} := \mathbb{E}[\phi|X = x_i]$ is a useful building block in the estimation of L_ϕ and CL_ϕ. Recall that we assume a conditional independence of an arbitrary number of labels given an input feature, *i.e.*, $Y^{(1)}, \ldots, Y^{(n_i)}|X \underset{i.i.d.}{\sim} \mathrm{Cat}(Q(X))$, $\mu_{\phi,i}$ is estimated as follows:

Theorem 4 (Unbiased estimator of $\mu_{\phi,i}$)**.** *For an instance i with n_i labels obtained in a conditional i.i.d. manner, an unbiased estimator of the conditional mean $\mu_{\phi,i}$ is given as follows:*

$$\widehat{\mu}_{\phi,i} := \binom{n_i}{n}^{-1} \sum_{j \in \mathrm{Comb}(n_i,n)} \phi(y_i^{(j_1)}, \ldots, y_i^{(j_n)}), \tag{39}$$

where $\mathrm{Comb}(n_i, n)$ denotes the distinct subset of size n drawn from $\{1, \ldots, n_i\}$ without replacement.

Proof. This is directly followed from the fact that $\widehat{\mu}_{\phi,i}$ is a U-statistic of n-sample symmetric kernel function ϕ (Hoeffding et al., 1948). □

C.1 Unbiased estimator of L_ϕ

We give an unbiased of L_ϕ as follows:

Theorem 5 (Unbiased estimator of L_ϕ)**.** *The following estimator is an unbiased estimator of L_ϕ.*

$$\widehat{L}_\phi := \frac{1}{N} \sum_i \widehat{L}_{\phi,i}, \quad \textit{where} \quad \widehat{L}_{\phi,i} := \binom{n_i}{n}^{-1} \sum_{j \in \mathrm{Comb}(n_i,n)} \left(\phi(y_i^{(j_1)}, \ldots, y_i^{(j_n)}) - \varphi_i\right)^2. \tag{40}$$

[5]We omit an instance with $n_i < n$ where ϕ cannot be calculated with distinct n labels.

Proof. We first confirm that, for each random variables $\phi(y_i^{(1)}, \dots, y_i^{(n)})$ and φ_i of sample i,

$$\mathbb{E}[f(y_i^{(1)}, \dots y_i^{(n)}; \varphi_i)] = L_\phi, \quad \text{where} \quad f(y_i^{(1)}, \dots y_i^{(n)}; \varphi_i) := \left(\phi(y_i^{(1)}, \dots, y_i^{(n)}) - \varphi_i\right)^2,$$

is satisfied by definition. As f is an n-sample symmetric kernel of variables $y_i^{(1)}, \dots, y_i^{(n)}$, $\widehat{L}_{\phi,i}$ is a U-statistic (Hoeffding et al., 1948) of the kernel given φ_i and also an unbiased estimator of L_ϕ as follows:

$$\mathbb{E}[\widehat{L}_{\phi,i}] = \mathbb{E}[\mathbb{E}[\widehat{L}_{\phi,i} \,|\varphi_i]] = \mathbb{E}[\mathbb{E}[f(y_i^{(1)}, \dots y_i^{(n)}; \varphi_i)|\varphi_i]] = L_\phi. \tag{41}$$

Hence

$$\mathbb{E}[\widehat{L}_\phi] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\widehat{L}_{\phi,i}] = L_\phi.$$

□

C.2 Debiased estimator of $\mathrm{CL}_\phi(\mathcal{B})$

Following the same discussion as the $\mathrm{CL}(\mathcal{B})$ term of CPEs, we also consider a binning based approximation of CL_ϕ stratified with a binning scheme $\mathcal{B}$ for predictive probability $\varphi \in [0, 1]$. Then, the plugin estimator of $\mathrm{CL}_\phi(\mathcal{B})$ is defined as follows:

Definition 8 (Plugin estimator of CL_ϕ)**.**

$$\widetilde{\mathrm{CL}}_\phi(\mathcal{B}) := \sum_{b=1}^{B} \widetilde{\mathrm{CL}}_{\phi,b}(\mathcal{B}), \tag{42}$$

$$\textit{where} \quad \widetilde{\mathrm{CL}}_{\phi,b}(\mathcal{B}) := \frac{|I_{\phi,b}|}{N}(\bar{c}_{\phi,b} - \bar{\varphi}_b)^2, \quad \bar{c}_{\phi,b} := \frac{\sum_{i\in I_{\phi,b}} \widehat{\mu}_{\phi,i}}{|I_{\phi,b}|}, \quad \bar{\varphi}_{\phi,b} := \frac{\sum_{i\in I_{\phi,b}} \varphi_i}{|I_{\phi,b}|}. \tag{43}$$

We again improve the plugin estimator with the following debiased estimator $\widehat{\mathrm{CL}}_{\phi,b}(\mathcal{B})$:

Corollary 6 (Debiased estimator of $\mathrm{CL}_{\phi,b}$)**.** *A plugin estimator $\widetilde{\mathrm{CL}}_{\phi,b}(\mathcal{B})$ of $\mathrm{CL}_{\phi,b}(\mathcal{B})$ is debiased to $\widehat{\mathrm{CL}}_{\phi,b}(\mathcal{B})$ with a correction term as follows:*

$$\widehat{\mathrm{CL}}_{\phi,b}(\mathcal{B}) := \widetilde{\mathrm{CL}}_{\phi,b}(\mathcal{B}) - \frac{|I_{\phi,b}|}{N}\frac{\bar{\sigma}^2_{\phi,b}}{|I_{\phi,b}| - 1}, \tag{44}$$

$$\textit{where} \quad \bar{\sigma}^2_{\phi,b} := \frac{1}{|I_{\phi,b}|}\sum_{i\in I_{\phi,b}} \widehat{\mu}^2_{\phi,i} - \left(\frac{1}{|I_{\phi,b}|}\sum_{i\in I_{\phi,b}} \widehat{\mu}_{\phi,i}\right)^2. \tag{45}$$

Note that the estimator is only available for bins with $|I_{\phi,b}| \geq 2$.

Proof. The proof follows a similar reasoning to Prop. B.3. We reduce the bias introduced with the term $\bar{c}^2_{\phi,b}$ by replacing the term with unbiased one for $\bar{C}^2_{\phi,b} = \mathbb{E}[\phi|\varphi \in \mathcal{B}_b]^2$ as follows:

$$\bar{c}^2_{\phi,b} \to \frac{1}{|I_{\phi,b}|\,(|I_{\phi,b}| - 1)}\sum_{i\in I_{\phi,b}}\sum_{i'\in I_{\phi,b}:i'\neq i} \widehat{\mu}_{\phi,i}\,\widehat{\mu}_{\phi,i'},$$

where $\widehat{\mu}_{\phi,i}$ is defined in equation (39). An improvement with the debiased estimator $\widetilde{\mathrm{CL}}_{\phi,b} - \widehat{\mathrm{CL}}_{\phi,b}$ is also calculated with the same manner as in Prop. B.3. □

C.3 Summary of evaluation metrics introduced for label histograms

In Table 3, we summarize evaluation metrics introduced for label histograms, where *order* shows the required numbers of labels for each instance to define the metrics, and *rater* represents those for estimating the metrics.

Table 3: Summary of evaluation metrics introduced for label histograms

Order	Signature	Description	Rater
1	$L_{\mathrm{sq}} = \mathrm{EL} + \mathrm{IL}$	Expected squared loss of CPEs	≥ 1
	$\mathrm{EL} = \mathrm{CL} + \mathrm{DL}$	Epistemic loss of CPEs	≥ 2
	$\mathrm{CL} = \mathrm{CE}^2$	Calibration loss of CPEs	≥ 1
	DL	Dispersion loss of CPEs	≥ 2
2	$L_{\phi^{\mathrm{D}}}$	Expected squared loss of DPEs	≥ 2
	$\mathrm{CL}_{\phi^{\mathrm{D}}}$	Calibration loss of DPEs	≥ 2

D Details on post-hoc uncertainty calibration methods

D.1 CPE calibration methods based on linear transformations

To complement Section 2.3, we summarize the formulation of CPE (class probability estimation) calibration that is based on linear transformations. Let $x \in \mathcal{X}$ denotes an input data, $u : \mathcal{X} \to \mathbb{R}^K$ denotes a DNN function that outputs a logit vector, and $f(x) = \mathrm{softmax}(u(x)) \in \Delta^{K-1}$ denotes CPEs. A common form of CPE calibration with linear transformations is given as follows:

$$\widetilde{u}(x) = Wu(x) + b, \tag{46}$$

$$\widetilde{f}(x) = \mathrm{softmax}(\widetilde{u}(x)), \tag{47}$$

where $\widetilde{u}(x)$ denotes transformed logits with parameters $W \in \mathbb{R}^{K \times K}$ and $b \in \mathbb{R}^K$, and $\tilde{f} : \mathcal{X} \to \Delta^{K-1}$ denotes CPEs after calibration.

The most general form of equation (46) is referred to as matrix scaling (Guo et al., 2017; Kull et al., 2019). A version of that with a constraint $W = \mathrm{diag}(v)$, $v \in \mathbb{R}^K$ and that with a further constraint $v = 1/t$, $t \in \mathbb{R}, b = 0$ are called vector and temperature scaling, respectively. In particular, temperature scaling has a favorable property; it does not change the maximum predictive class of each instance, and hence neither the overall accuracy, as the order of vector elements between u and $\tilde{u}$ for each x is unchanged.

For vector and matrix scaling, regularization terms are required to prevent over-fitting; L2 regularization of b:

$$\Omega_{\mathrm{L2}}(b) := \lambda_b \frac{1}{K} \sum_k b_k^2 \tag{48}$$

is commonly used for vector scaling, and off-diagonal and intercept regularization (ODIR):

$$\Omega_{\mathrm{ODIR}}(W, b) := \lambda_w \frac{1}{K(K-1)} \sum_{k \neq k'} W_{kk'}^2 + \lambda_b \frac{1}{K} \sum_k b_k^2 \tag{49}$$

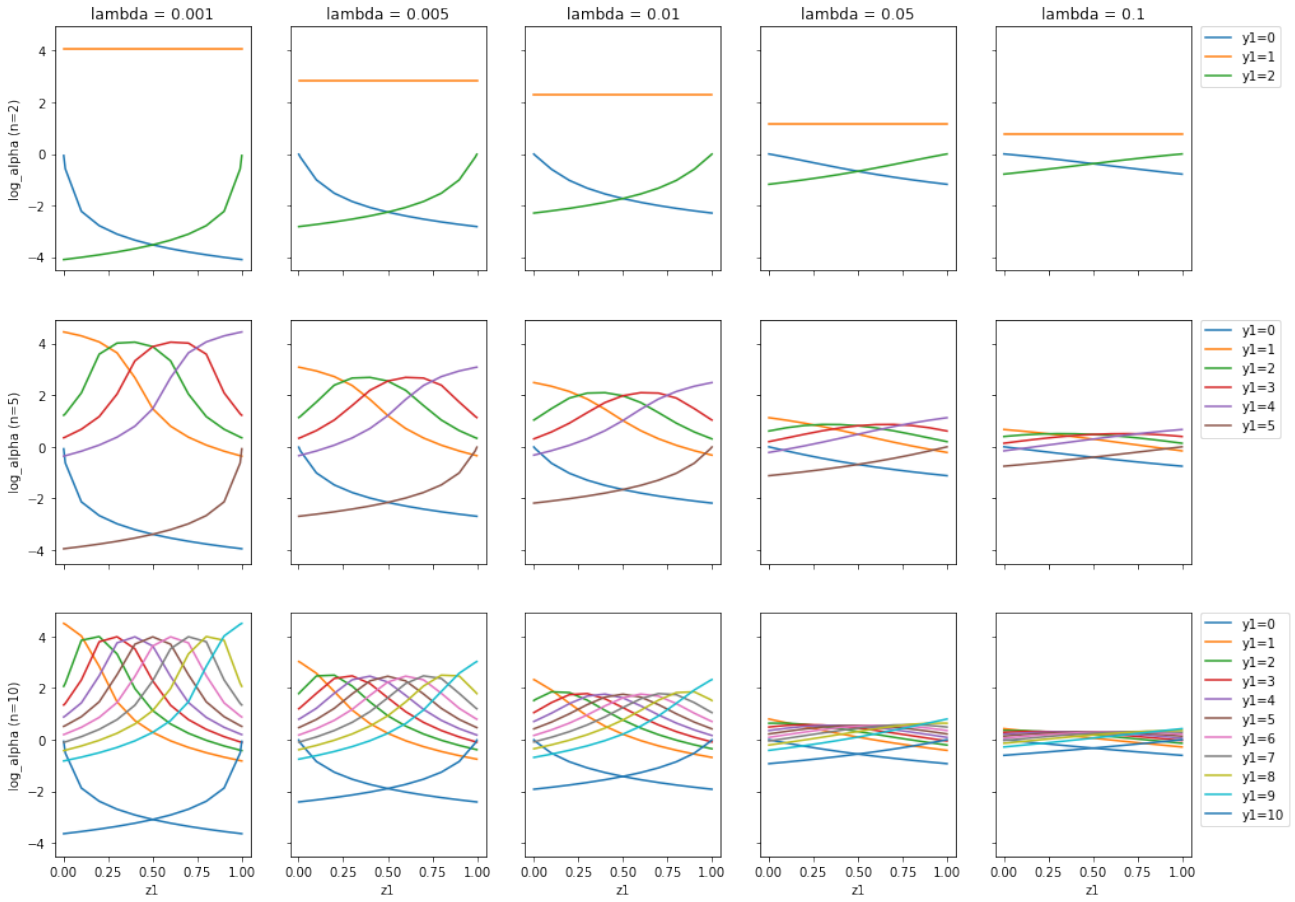
is proposed for matrix scaling, which is used for improving class-wise calibration (Kull et al., 2019).

D.2 Details on α-calibration

Loss function For the loss function for α-calibration, we use a variant of NLL in equation (7) as follows:

$$-\frac{1}{\sum_i n_i} \sum_i \log \mathrm{DirMult}(y_i | \alpha_0(x_i) f(x_i)) + \frac{\lambda_\alpha}{N} \sum_i (\log \alpha_0(x_i))^2, \tag{50}$$

where $\mathrm{DirMult}(\cdot)$ denotes the Dirichlet multinomial distribution, and a regularization term $\lambda_\alpha (\log \alpha_0)^2$ is introduced for stabilization purpose, which penalizes the deviation from $\alpha_0 = 1$ to both directions towards extreme concentrations of the mass: $P(\zeta|X) \to \delta(\zeta = f(X))$ with $\alpha_0 \to \infty$ or $P(\zeta|X) \to \sum_k \delta(\zeta = e_k)\, \mathbb{E}[\zeta_k]$ with $\alpha_0 \to 0$. We employ $\lambda_\alpha = 0.005$ throughout this study.



Optimal values of $\log \alpha_0$ for different values of the hyperparameter λ_α

Figure 3: Optimal values of $\log \alpha_0(x_i)$ are numerically evaluated for binary class problems with the number of labels $n_i \in \{2, 5, 10\}$, label histograms y_i with $0 \leq y_{i1} \leq n_i$, class probability estimations of the first class: $z1 \in \{.001, .01, .1, .2, \ldots, .8, .9, .99, .999\}$, and the hyperparameter λ_α (Appendix D.2). As expected, the range of $\log \alpha_0$ contains zero and gets narrower as λ_α increases.

Hyperparameter analysis for the optimal values of α_0 Intuitively, $\log \alpha_0$ is likely to get close to zero as the regularization coefficient λ_α increases. If we regard $\log \alpha_0(x_i)$ as a free parameter, the optimal value of $\log \alpha_0(x_i)$ only depends on the CPEs $f(x_i)$, the number of labels n_i and the observed labels y_i for each instance. We assume that the number of labels n_i is common for all the instances for simplicity. The optimality condition for α_0 is obtained by taking a derivative of equation (50) with respect to $\log \alpha_0(x_i)$ as follows:

$$
\begin{aligned}
0 &= -\frac{\alpha_0}{n_i}\left[\sum_k f_k \left(\psi(\alpha_0 f_k + y_{ik}) - \psi(\alpha_0 + n_i)\right) - \sum_k f_k \left(\psi(\alpha_0 f_k) - \psi(\alpha_0)\right)\right] + 2\lambda_\alpha \log \alpha_0 \\
&= -\frac{\alpha_0}{n_i}\left(\sum_k \sum_{l=1}^{y_{ik}} \frac{f_k}{\alpha_0 f_k + l - 1} - \sum_{l=1}^{n_i} \frac{1}{\alpha_0 + l - 1}\right) + 2\lambda_\alpha \log \alpha_0, \qquad (51)
\end{aligned}
$$

where $\psi(\cdot)$ denotes the digamma function, and a recurrence formula $\psi(s+1) = \psi(s) + \frac{1}{s}$ is used for the derivation.

One can verify that divergences of the optimal $\log \alpha_0$ occur in some special cases with $\lambda_\alpha = 0$. For example, if the labels are unanimous, i.e., $y_{i1} = n_i$, $n_i > 1$, and $f_1 < 1$, the r.h.s. of equation (51) turns out to be positive as follows:

$$
-\frac{\alpha_0}{n_i} \sum_{l=1}^{n_i} \left(\frac{f_1}{\alpha_0 f_1 + l - 1} - \frac{1}{\alpha_0 + l - 1}\right) > 0,
$$

which implies that $\log \alpha_0 \to -\infty$. In contrast, if $n_i = 2$, $K = 2$, and $y_{i1} = y_{i2} = 1$, the r.h.s of equation (51) is calculated as follows:

$$-\frac{\alpha_0}{n_i}\left(\frac{2}{\alpha_0} - \frac{1}{\alpha_0} - \frac{1}{\alpha_0 + 1}\right) = -\frac{1}{n_i(\alpha_0 + 1)} < 0,$$

which results in $\log \alpha_0 \to \infty$.

For a finite $\lambda_\alpha > 0$, the optimal values of $\log \alpha_0$ can be numerically evaluated by Newton's method. We show these values in Fig. 3 for several conditions of binary class problems. As expected, the range of the optimal $\log \alpha_0$ contains zero and gets narrower as λ_α increases.

D.3 Proof for Theorem 1

Restatement of Theorem 1. *There exist intervals for parameter $\alpha_0 \geq 0$, which improve task performances as follows.*

1. *For DPEs, $L_{\phi^{\mathrm{D}},G} \leq L^{(0)}_{\phi^{\mathrm{D}},G}$ holds when $(1 - 2u_Q + s_Z)/2(u_Q - s_Z) \leq \alpha_0$, and $L_{\phi^{\mathrm{D}},G}$ takes the minimum value when $\alpha_0 = (1 - u_Q)/(u_Q - s_Z)$, if $u_Q > s_Z$ is satisfied.*

2. *For posterior CPEs, $\mathrm{EL}'_G \leq \mathrm{EL}'^{(0)}_G$ holds when $(1 - u_Q - \mathrm{EL}_G)/2\,\mathrm{EL}_G \leq \alpha_0$, and EL'_G takes the minimum value when $\alpha_0 = (1 - u_Q)/\mathrm{EL}_G$, if $\mathrm{EL}_G > 0$ is satisfied.*

Note that we denote $s_Z := \sum_k Z_k^2$, $u_Q := \mathbb{E}[\sum_k Q_k^2|G]$, $v_Q := \mathbb{V}[Q_k|G]$, and $\mathrm{EL}_G := \mathbb{E}[\sum_k (Z_k - Q_k)^2|G]$. The optimal α_0 of both tasks coincide to be $\alpha_0 = (1 - u_Q)/v_Q$, if CPEs match the true conditional class probabilities given G, i.e., $Z = \mathbb{E}[Q|G]$.

Proof. We will omit the superscript D from ϕ^{D} and φ^{D} for brevity. First, we rewrite φ and Z' in equation (19) as follows:

$$\varphi = \gamma\left(1 - \sum_k Z_k^2\right), \quad Z'_k = \gamma Z_k + (1 - \gamma)Y_k, \tag{52}$$

where $\gamma := \alpha_0/(\alpha_0 + 1)$. Note that $\gamma \in (0, 1)$ since $\alpha_0 \in (0, +\infty)$. We also introduce the following variables:

$$s_Q := \sum_k \mathbb{E}[Q_k|G]^2, \qquad \bar{s}_Q := 1 - s_Q, \qquad v_Q := \sum_k \mathbb{V}[Q_k|G], \tag{53}$$

$$u_Q := \sum_k \mathbb{E}[Q_k^2|G] = s_Q + v_Q, \qquad \bar{u}_Q := 1 - u_Q, \tag{54}$$

$$s_Z := \sum_k Z_k^2, \qquad \bar{s}_Z := 1 - s_Z, \tag{55}$$

where, all the variables reside within $[0, 1]$ since $Z, Q \in \Delta^{K-1}$.

1. The first statement: DPE

The objective function to be minimized is as follows:

$$L_{\phi,G} = \mathbb{E}[(\phi - \varphi)^2|G] = (\mathbb{E}[\phi|G] - \varphi)^2 + \mathbb{V}[\phi|G] \tag{56}$$

$$= \mathbb{E}[(\bar{u}_Q - \gamma\bar{s}_Z)^2] + \mathbb{V}[\phi|G], \tag{57}$$

where we use the relation $\mathbb{E}[\phi|G] = \bar{u}_Q$ and $\varphi = \gamma\bar{s}_Z$. Note that only the first term is varied with α_0, and $L_{\phi,G} \to L^{(0)}_{\phi,G}$ $(\gamma \to 1)$. The condition for satisfying $L_{\phi,G} \leq L^{(0)}_{\phi,G}$ is found by solving

$$0 = L_{\phi_G} - L^{(0)}_{\phi,G} = (\gamma - 1)\bar{s}_Z\{(\gamma + 1)\bar{s}_Z - 2\bar{u}_Q\} = (\gamma - 1)\bar{s}_Z\{\gamma\bar{s}_Z + \bar{s}_Z - 2\bar{u}_Q\}. \tag{58}$$

$\bar{s}_Z = 0$ and $\gamma \to 1$ are trivial solutions that correspond to a hard label prediction (*i.e.*, $Z \in e^K$) and $\alpha_0 \to \infty$, respectively. The remaining condition for $L_{\phi,G} \leq L^{(0)}_{\phi,G}$ is

$$\gamma \in [-1 + 2\bar{u}_Q/\bar{s}_Z, 1), \tag{59}$$

which is feasible when $\bar{u}_Q < \bar{s}_Z$, *i.e.*, $u_Q > s_Z$. In this case,

$$\gamma^* = \bar{u}_Q/\bar{s}_Z \tag{60}$$

is the optimal solution for γ. By using a relation $\alpha_0 = \gamma/(1-\gamma)$ with equations (59) and (60), the first statement of the theorem is obtained as follows:

$$\alpha_0 \geq \frac{-\bar{s}_Z + 2\bar{u}_Q}{2\bar{s}_Z - 2\bar{u}_Q} = \frac{1 - 2u_Q + s_Z}{2(u_Q - s_Z)}, \quad \alpha_0^* = \frac{\bar{u}_Q}{\bar{s}_Z - \bar{u}_Q} = \frac{1 - u_Q}{u_Q - s_Z}. \tag{61}$$

If $Z = \mathbb{E}[Q|G]$ is satisfied, $s_Z = s_Q$ holds, and the above conditions become as follows:

$$\alpha_0 \geq \frac{1 - u_Q - v_Q}{2v_Q}, \quad \alpha_0^* = \frac{1 - u_Q}{v_Q}. \tag{62}$$

2. The second statement: posterior CPE

The objective for the second problem is as follows:

$$\begin{aligned}
\mathrm{EL}'_G &= \mathbb{E}[\sum_k (Z'_k - Q_k)^2 | G] \\
&= \mathbb{E}[\sum_k (\gamma Z_k + (1-\gamma) Y_k - Q_k)^2 | G] \\
&= \mathbb{E}[\sum_k ((Z_k - Q_k) + (1-\gamma)(Y_k - Z_k))^2 | G] \\
&= \sum_k \mathbb{E}[(Z_k - Q_k)^2 | G] + 2(1-\gamma)\,\mathbb{E}[(Z_k - Q_k)(Y_k - Z_k)|G] + (1-\gamma)^2\,\mathbb{E}[(Y_k - Z_k)^2|G],
\end{aligned} \tag{63}$$

where the first term equals to EL_G, and the second and third term are further transformed as follows:

$$\begin{aligned}
\sum_k \mathbb{E}[(Z_k - Q_k)(Y_k - Z_k)|G] &= \sum_k \mathbb{E}[\mathbb{E}[(Z_k - Q_k)(Y_k - Z_k)|Q]|G] \\
&= -\sum_k \mathbb{E}[\mathbb{E}[(Z_k - Q_k)^2|Q]|G] = -\,\mathrm{EL}_G,
\end{aligned} \tag{64}$$

$$\begin{aligned}
\sum_k \mathbb{E}[(Y_k - Z_k)^2|G] &= \sum_k \mathbb{E}[\mathbb{E}[(Y_k - Z_k)^2|Q]|G] \\
&= \sum_k \mathbb{E}[\mathbb{E}[(Y_k - 2Y_k Z_k + Z_k^2)|Q]|G] \\
&= \sum_k \mathbb{E}[(Q_k - 2Q_k Z_k + Z_k^2)|G] \\
&= \sum_k \mathbb{E}[(Q_k(1-Q_k) + (Q_k - Z_k)^2|G] = \bar{u}_Q + \mathrm{EL}_G\,.
\end{aligned} \tag{65}$$

Hence equation (63) can be written as

$$\mathrm{EL}'_G = \mathrm{EL}_G - 2(1-\gamma)\,\mathrm{EL}_G + (1-\gamma)^2(\mathrm{EL}_G + \bar{u}_Q). \tag{66}$$

The condition for satisfying $\mathrm{EL}'_G \leq EL_G$ is obtained by solving

$$0 = \mathrm{EL}'_G - \mathrm{EL}_G = (1-\gamma)\left\{(1-\gamma)(\mathrm{EL}_G + \bar{u}_Q) - 2\,\mathrm{EL}_G\right\}. \tag{67}$$

If $\mathrm{EL}_G = 0$, which means $Z_k = Q_k$ given G, $\gamma \to 1$ is optimal as expected. For the other case, *i.e.*, $\mathrm{EL}_G > 0$, γ that satisfying $\mathrm{EL}'_G \leq \mathrm{EL}_G$ and the optimal γ are

$$\gamma \in \left[\frac{\bar{u}_Q - \mathrm{EL}_G}{\bar{u}_Q + \mathrm{EL}_G}, 1\right), \quad \gamma^* = \frac{\bar{u}_Q}{\bar{u}_Q + \mathrm{EL}_G}, \tag{68}$$

respectively. By using $\alpha_0 = \gamma/(1-\gamma)$, the corresponding α_0 and α_0^* are

$$\alpha_0 \geq \frac{\bar{u}_Q - \mathrm{EL}_G}{2\,\mathrm{EL}_G} = \frac{1 - u_Q - \mathrm{EL}_G}{2\,\mathrm{EL}_G}, \quad \alpha_0^* = \frac{\bar{u}_Q}{\mathrm{EL}_G} = \frac{1 - u_Q}{\mathrm{EL}_G}, \tag{69}$$

respectively, which are the second statement of the theorem. If $Z = \mathbb{E}[Q|G]$ is satisfied, $\mathrm{EL}_G = v_Q$ holds, and the above conditions become as follows:

$$\alpha_0 \geq \frac{1 - u_Q - v_Q}{2v_Q}, \quad \alpha_0^* = \frac{1 - u_Q}{v_Q}. \tag{70}$$

Notably, these are the same conditions as the terms in equation (62), respectively. □

D.4 Summary of DPE computations

We use α-calibration, ensemble-based methods (MCDO and TTA), and a combination of them for predicting DPEs as follows.

α-calibration

$$\widehat{\varphi}^D = 1 - \sum_k \int \zeta_k^2 \,\mathrm{Dir}(\zeta|\alpha_0 f)\, d\zeta = \frac{\alpha_0}{\alpha_0 + 1}\left(1 - \sum_k f_k^2\right). \tag{71}$$

Ensemble-based methods

$$\widehat{\varphi}^D = \frac{1}{S}\sum_{s=1}^{S}\left(1 - \sum_k \left(f_k^{(s)}\right)^2\right), \tag{72}$$

where, $f^{(s)} : \mathcal{X} \to \Delta^{(k-1)}$ is the s-th prediction of the ensemble, and S is the size of the ensemble.

Ensemble-based methods with α-calibration Although an output α already represents a CPE distribution without ensemble-based methods: MCDO and TTA, it can be formally combined with these methods. In such cases, we calculate the predictive probability $\widehat{\varphi}$ as follows:

$$\widehat{\varphi}^{\mathrm{D}} = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{\alpha_0^{(s)}}{\alpha_0^{(s)} + 1}\left(1 - \sum_k \left(f_k^{(s)}\right)^2\right)\right], \tag{73}$$

where $\alpha_0^{(s)}, f^{(s)}$ denote the s-th ensembles of α_0 and f, respectively.

D.5 Summary of posterior CPE computations

We consider a task for updating CPE of instance $x \in \mathcal{X}$ after an expert annotation $y \in e^K$. For this task, the posterior CPE distribution $P_{\mathrm{model}}(\zeta|x, y)$ is computed from an original (prior) CPE distribution model $P_{\mathrm{model}}(\zeta|x)$ as follows:

$$P_{\mathrm{model}}(\zeta|x, y) = \frac{P_{\mathrm{model}}(y, \zeta|x)}{P_{\mathrm{model}}(y|x)}, \tag{74}$$

where

$$P_{\mathrm{model}}(y, \zeta|x) = P_{\mathrm{model}}(\zeta|x)\prod_{k=1}^{K}\zeta_k^{y_k}. \tag{75}$$

For the case with multiple test instances, we assume that a predictive model is factorized as follows:

$$P_{\mathrm{model},N}(\zeta_{1:N}|x_{1:N}) = \prod_{i=1}^{N} P_{\mathrm{model}}(\zeta_i|x_i). \tag{76}$$

In this case, the posterior of CPEs is also factorized as follows:

$$P_{\mathrm{model},N}(\zeta_{1:N}|x_{1:N}, y_{1:N}) = \frac{P_{\mathrm{model},N}(y_{1:N}, \zeta_{1:N}|x_{1:N})}{P_{\mathrm{model},N}(y_{1:N}|x_{1:N})} = \prod_{i=1}^{N} \frac{P_{\mathrm{model}}(y_i, \zeta_i|x_i)}{P_{\mathrm{model}}(y_i|x_i)} = \prod_{i=1}^{N} P_{\mathrm{model}}(\zeta_i|x_i, y_i). \tag{77}$$

α-calibration Prior and posterior CPE distributions are computed as follows:

$$P_\alpha(\zeta|x) = \mathrm{Dir}(\zeta|\alpha_0(x)f(x)), \tag{78}$$
$$P_\alpha(\zeta|x, y) = \mathrm{Dir}(\zeta|\alpha_0(x)f(x) + y). \tag{79}$$

Ensemble-based methods Prior and posterior CPE distributions are computed as follows:

$$P_{\mathrm{ens.}}(\zeta|x) = \frac{1}{S}\sum_{s=1}^{S} \delta(\zeta - f^{(s)}(x)), \tag{80}$$

$$P_{\mathrm{ens.}}(\zeta|x, y) = \frac{1}{W'}\sum_{s=1}^{S} w'_s \delta(\zeta - f^{(s)}(x)), \tag{81}$$

where S is the size of the ensemble, $f^{(s)}$ denotes the s-th CPEs of the ensemble, $w'_s := \sum_s \prod_k f^{(s)}(x)_k^{y_k}$, and $W' := \sum_{s=1}^{S} w'_s$.

We omit the cases of predictive models combining the ensemble-based methods and α-calibration, where the posterior computation requires further approximation.

D.6 Discussion on conditional i.i.d. assumption of label generations for α-calibration

At the beginning of section 3, we assume a conditional i.i.d distribution for labels Y given input data X, which is also a basis for α-calibration. We expect that the assumption roughly holds in typical scenarios, where experts are randomly assigned to each example. However, α-calibration may not be suitable for counter-examples that break the assumption. For instance, if two fixed experts with different policy annotate all examples, these two labels would be highly correlated. In such a case, the disagreement probability between them may be up to one and exceeds the maximum possible value φ^{D} allowed in equation (19), where φ^{D} always decreases from the original value, which corresponds to $\alpha_0 \to \infty$, by α-calibration.

E Experimental details

We used the Keras Framework with Tensorflow backend (Chollet et al., 2015) for implementation.

E.1 Preprocessing

Mix-MNIST and CIFAR-10 We generated two synthetic image dataset: Mix-MNIST and Mix-CIFAR-10 from MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky et al., 2009), respectively. We randomly selected half of the images to create mixed-up images from pairs and the other half were retained as original. For each of the paired images, a random ratio that followed a uniform distribution $U(0, 1)$ was used for the mix-up and a class probability of multiple labels, which were two or five in validation set. For Mix-MNIST (Mix-CIFAR-10), the numbers of generated instances were $37,500$ $(30,000)$, $7,500$ $(7,500)$, and $7,500$ $(7,500)$ for training, validation, and test set, respectively.

MDS Data We used $80,610$ blood cell images with a size of 360×363, which was a part of the dataset obtained in a study of myelodysplastic syndrome (MDS) (Sasada et al., 2018), where most of the images showed a white blood cell in the center of the image. For each image, a mean of 5.67 medical technologists annotated the cellular category from 22 subtypes, in which six were anomalous types. We partitioned the dataset into training, validation, and test set with $55,356$, $14,144$, and $11,110$ images, respectively, where each of the partition consisted of images from distinct patient groups. Considering the high labeling cost with experts in the medical domain, we focused on scenarios that training instances were singly labeled, and multiple labels were only available for validation and test set. The mean number of labels per instance for validation and test set were 5.79 and 7.58, respectively.

E.2 Deep neural network architecture

Mix-MNIST For Mix-MNIST dataset, we used a neural network architecture with three convolutional and two full connection layers. Specifically, the network had the following stack:

- Conv. layer with 32 channels, 5×5 kernel, and ReLU activation
- Max pooling with with 2×2 kernel and same padding
- Conv. layer with 64 channels, 7×7 kernel, and ReLU activation
- Max pooling with with 4×4 kernel and same padding
- Conv. layer with 128 channels, 11×11 kernel, and ReLU activation
- Global average pooling with 128 dim. output
- Dropout with 50% rate
- Full connection layer with $K = 3$ dim. output and softmax activation

Mix-CIFAR-10 and MDS We adopted a modified VGG16 architecture (Simonyan and Zisserman, 2014) as a base model, in which the full connection layers were removed, and the last layer was a max-pooling with 512 output dimensions. On top of the base model, we appended the following layers:

- Dropout with 50% rate and 512 dim. output
- Full connection layer with 128 dim. output and ReLU activation
- Dropout with 50% rate and 128 dim. output
- Full connection layer with 22 dim. output and softmax activation

E.3 Training

We used the following loss function for training:

$$\mathcal{L}(y, z) = -\frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log z_{ik}, \tag{82}$$

which was equivalent to the negative log-likelihood for instance-wise multinomial observational model except for constant. We used Adam optimizer (Kingma and Ba, 2014) with a base learning rate of 0.001. Below, we summarize conditions specfic to each dataset.

Mix-MNIST We trained for a maximum of 100 epochs with a minibatch size of 128, applying early stopping with ten epochs patience for the validation loss improvement. We used no data augmentation for Mix-MNIST.

Mix-CIFAR-10 We trained for a maximum of 250 epochs with a minibatch size of 128, applying a variant of warm-up and multi-step decay scheduling (He et al., 2016) as follows:

- A warm-up with five epochs
- A multi-step decay that multiplies the learning rate by 0.1 at the end of 100 and 150 epochs

We selected the best weights in terms of validation loss. While training, we applied data augmentation with random combinations of the following transformations:

- Rotation within -180 to 180 degrees
- Width and height shift within ± 10 pixels
- Horizontal flip

MDS We trained for a maximum of 200 epochs with a minibatch size of 128, applying a warm-up and multi-step decay scheduling as follows:

- A warm-up with five epochs
- A multi-step decay that multiplies the learning rate by 0.1 at the end of 50, 100, and 150 epochs

We recorded training weights for every five epochs, and selected the best weights in terms of validation loss. While training, we applied data augmentation with random combinations of the following transformations:

- Rotation within -20 to 20 degrees
- Width and height shift within ± 5 pixels
- Horizontal flip

For each image, the center 224×224 portion is cropped from the image after the data augmentation.

E.4 Post-hoc calibrations and predictions

We applied temperature scaling for CPE calibration and α-calibration for obtaining CPE distributions. For both calibration methods, we used validation set, which was splited into 80% calibration set for training and 20% calibration-validation (cv) set for the validation of calibration. We trained for a maximum of 50 epochs using Adam optimizer with a learning rate of 0.001, applying early stopping with ten epochs patience for the cv loss improvement. The loss functions of equation (82) and (50) were used for CPE- and α-calibration, respectively. For the feature layer that used for α-calibration, we chose the penultimate layer that corresponded to the last dropout layer in this experiment. The training scheme is the same as that of the CPE calibration, except for a loss function that we described in D.2. We also used ensemble-based methods: Monte-Calro dropout (MCDO) (Gal and Ghahramani, 2016) and Test-time augmentation (TTA) (Ayhan and Berens, 2018) for CPE distribution predictions, which were both applicable to DNNs at prediction-time, where 20 MC-samples were used for ensemble. A data augmentation applied in TTA was the same as that used in training, and we only applied TTA for Mix-CIFAR-10 and MDS data.

F Additional experiments

F.1 Evaluations of class probability estimates

We present evaluation results of class probability estimates (CPEs) for Mix-MNIST and Mix-CIFAR-10 in Table 4 and 5, respectively. Overall, CPE measures were comparable between the same datasets with different validation labels (two and five). By comparing $\widehat{L}_{\mathrm{sq}}$ and $\widehat{\mathrm{EL}}$, the relative ratio of $\widehat{\mathrm{EL}}$ against irreducible loss could be evaluated which was much higher in Mix-CIFAR-10 than in Mix-MNIST. Among Raw predictions, temperature

scaling kept accuracy and showed a consistent improvement in $\widehat{\mathrm{EL}}$ and $\widehat{\mathrm{CE}}$ as expected. While TTA showed a superior performance over MCDO and Raw predictions in accuracy, $\widehat{L}_{\mathrm{sq}}$ and $\widehat{\mathrm{EL}}$, the effect of calibration methods for CPEs with ensemble-based predictions was not consistent, which might be because calibration was not ensemble-aware.

Table 4: Evaluations of CPEs for Mix-MNIST

	Mix-MNIST(2)				Mix-MNIST(5)			
Method	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓
Raw	**.9629**	.1386	.0388	.0518	**.9629**	.1386	.0388	.0518
Raw+α	**.9629**	.1386	.0388	.0518	**.9629**	.1386	.0388	.0518
Raw+ts	**.9629**	**.1376**	**.0379**	**.0473**	**.9629**	**.1376**	**.0379**	**.0475**
Raw+ts+α	**.9629**	**.1376**	**.0379**	**.0473**	**.9629**	**.1376**	**.0379**	**.0475**
MCDO	**.9635**	.1392	.0395	**.0425**	.9635	.1392	.0395	**.0425**
MCDO+α	.9621	**.1391**	**.0394**	.0442	**.9644**	**.1387**	**.0389**	.0463
MCDO+ts	.9628	.1408	.0410	.0632	.9627	.1402	.0404	.0638
MCDO+ts+α	.9624	.1412	.0415	.0653	.9629	.1407	.0409	.0631

Table 5: Evaluations of CPEs for Mix-CIFAR-10

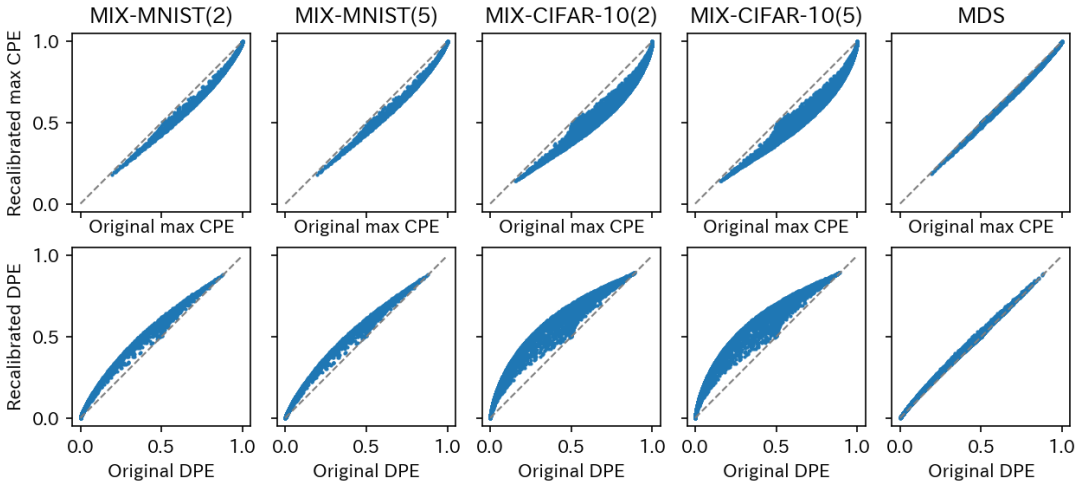
	Mix-CIFAR-10(2)				Mix-CIFAR-10(5)			
Method	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓
Raw	**.7965**	.3518	.2504	.1093	**.7965**	.3518	.2504	.1093
Raw+α	**.7965**	.3518	.2504	.1093	**.7965**	.3518	.2504	.1093
Raw+ts	**.7965**	**.3437**	**.2423**	**.0687**	**.7965**	**.3438**	**.2423**	**.0685**
Raw+ts+α	**.7965**	**.3437**	**.2423**	.0688	**.7965**	**.3438**	**.2423**	**.0685**
MCDO	**.7983**	.3488	.2474	.0955	**.7983**	.3488	.2474	.0955
MCDO+α	.7968	.3488	.2474	.0962	.7965	.3493	.2479	.0973
MCDO+ts	.7972	**.3442**	**.2428**	.0684	.7977	.3446	.2431	**.0675**
MCDO+ts+α	.7965	.3445	.2430	**.0650**	.7975	**.3444**	**.2430**	.0723
TTA	.8221	.3230	.2216	**.0877**	.8221	.3230	.2216	**.0877**
TTA+α	.8241	**.3221**	**.2206**	.0884	.8277	**.3223**	**.2209**	.0894
TTA+ts	.8257	.3467	.2452	.1688	**.8279**	.3466	.2451	.1707
TTA+ts+α	**.8277**	.3446	.2432	.1697	.8245	.3460	.2446	.1726

F.2 Discussion on the effect of temperature scaling for disagreement probability estimates

In Table 1, it is observed that disagreement probability estimates (DPEs) are consistently degraded by temperature scaling (Raw+ts) from the original scores (Raw), despite the positive effects for calibration of class probability estimates (CPEs) with ts. Though we observe that the degradation can be overcome with α-calibration (see Raw+ts+α or Raw+α in Table 1), the mechanism that causes the phenomena is worth analyzing. Since there exists well-known overconfidence in the maximum class probabilities from DNN classifiers (Guo et al., 2017), the recalibration of CPEs by ts tends to reduce the maximum class probabilities. On the other hand, the amount of change in a DPE for each instance can be written as follows:

$$\Delta^{\mathrm{D}} := \varphi'^{\mathrm{D}} - \varphi^{\mathrm{D}} = \sum_{k=1}^{K} (f_k + f'_k)(f_k - f'_k), \tag{83}$$

where f and φ^{D} denote the original CPEs and DPE, respectively, and f' and φ'^{D} denote those after ts, respectively. It is likely that Δ^{D} takes a positive value as the dominant term of Δ^{D} in equation (83) is k with the maximum $f_k + f'_k$ value, where $f_k > f'_k$ is satisfied for overconfident predictions. In fact, the averages of



Changes in the maximum CPEs and DPEs with temperature scaling

Figure 4: Changes in the maximum class probability estimates (CPEs) and disagreement probability estimates (DPEs) with temperature scaling (ts) are presented for each instance. In contrast to the maximum CPEs, which are generally decreased with ts to compensate overconfidence, DPEs are increased with ts as discussed in Appendix F.2.

Δ^{D} between Raw+ts and Raw for each of the five settings in Table 1 are all positive, which are 0.027, 0.025, 0.101, 0.103, and 0.019, respectively. Instance-wise changes in the maximum CPEs and DPEs with ts are shown in Fig. 4. Simultaneously, DPEs without α-calibration systematically overestimate the empirical disagreement probabilities, as shown in Fig. 1. Therefore, the positive Δ^{D} means that φ'^{D} is even far from a target probability $\mathbb{E}[\phi^{\mathrm{D}}|X]$ than φ^{D} is, despite the improvement in CPEs with ts.

F.3 Additional experiments for MDS data

In addition to MDS data with single training labels per instance (MDS-1) used in the main experiment, we trained and evaluated with full MDS data (MDS-full), where all the multiple training labels per example were employed. Also, we included additional CPE calibration methods: vector and matrix scaling (vs and ms, respectively), which were introduced in Section D.1, for these experiments. We adopted an L2 regularization for vs and an ODIR for ms, in which the following hyper-parameter candidates were examined:

- vs: $\lambda_b \in \{0.1, 1.0, 10\}$
- ms: $(\lambda_b, \lambda_w) \in \{0.1, 1.0, 10\} \times \{0.1, 1.0, 10\}$

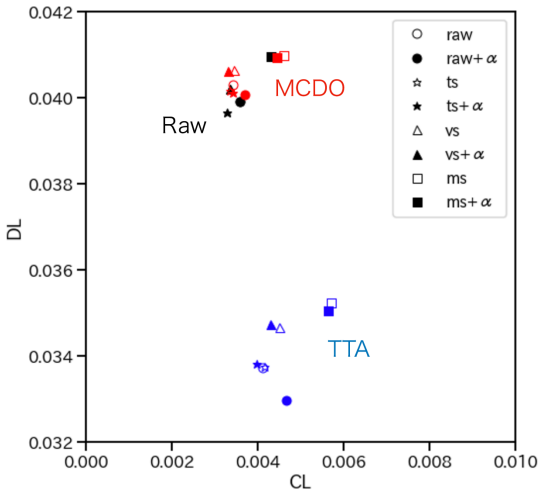
where λ_b and λ_w were defined in equation (48) and (49), respectively. The hyper-parameters were selected with respect to the best cv loss, which were $\lambda_b = 0.1$ for vs and $\lambda_b = 1.0, \lambda_w = 10$ for ms in the single-training MDS data, and $\lambda_b = 0.1$ for vs and $\lambda_b = 0.1, \lambda_w = 10$ for ms in the full MDS data.

Results We summarize the order-1 and -2 performance metrics for predictions with MDS-1 and MDS-full datasets in Table 6. For both datasets, temperature scaling consistently improved (decreased) $\widehat{\mathrm{EL}}$ and $\widehat{\mathrm{CL}}$ for each of Raw, MCDO, and TTA predictions. While vector scaling was slightly better at obtaining the highest accuracy than the other methods, the effect of vs and ms for the metrics of probability predictions were limited. As same as the results of synthetic experiments, TTA showed a superior performance over MCDO and Raw predictions in accuracy, $\widehat{L}_{\mathrm{sq}}$ and $\widehat{\mathrm{EL}}$. Since $\widehat{\mathrm{EL}}$ can be decomposed into $\widehat{\mathrm{CL}}$ and the remaining term: $\widehat{\mathrm{DL}}$ (Section B.4), the difference of predictors in CPE performance is clearly presented with 2d plots (Fig.5 and 6), which we call calibration-dispersion maps. For order-2 metrics, both $\widehat{L}_{\phi^{\mathrm{D}}}$ and $\widehat{\mathrm{CE}}_{\phi^{\mathrm{D}}}$ for DPEs were substantially improved by α-calibration, espetially in $\widehat{\mathrm{CE}}_{\phi^{\mathrm{D}}}$, which was not attained with solely applying ensemble-based methods even

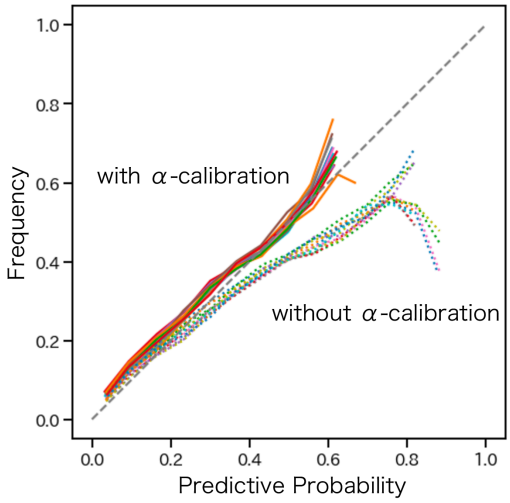
with MDS-full. This improvement of DPE calibration was also visually confirmed with reliability diagrams in Fig. 5 and 6. Though overall characteristics were similar between MDS-1 and MDS-full, a substantial improvement in $\widehat{L}_{\mathrm{sq}}$, $\widehat{\mathrm{EL}}$, and $\widehat{L}_{\phi^{\mathrm{D}}}$ was observed in MDS-full, which seemed the results of enhanced probability predictions with additional training labels.

Table 6: Performance evaluations for MDS data

	MDS-1						MDS-full					
	Order-1 metrics				Order-2 metrics		Order-1 metrics				Order-2 metrics	
Method	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓	$\widehat{L}_{\phi^D}$ ↓	$\widehat{\mathrm{CE}}_{\phi^D}$ ↓	Acc ↑	$\widehat{L}_{\mathrm{sq}}$ ↓	$\widehat{\mathrm{EL}}$ ↓	$\widehat{\mathrm{CE}}$ ↓	$\widehat{L}_{\phi^D}$ ↓	$\widehat{\mathrm{CE}}_{\phi^D}$ ↓
Raw	.9006	.2515	.0435	.0600	.1477	.0628	.8990	.2460	.0380	.0590	.1448	.0539
Raw+α	.9006	.2515	.0435	.0600	**.1454**	**.0406**	.8990	.2460	.0380	.0590	**.1430**	**.0320**
Raw+ts	.9006	**.2509**	**.0430**	**.0575**	.1482	.0663	.8990	**.2459**	**.0379**	**.0587**	.1449	.0545
Raw+ts+α	.9006	**.2509**	**.0430**	**.0575**	**.1445**	**.0261**	.8990	**.2459**	**.0379**	**.0587**	**.1427**	**.0269**
Raw+vs	**.9010**	.2515	.0435	.0579	.1489	.0696	**.8996**	.2461	.0381	.0602	.1452	.0563
Raw+vs+α	**.9010**	.2515	.0435	.0579	**.1449**	**.0318**	**.8996**	.2461	.0381	.0602	**.1431**	**.0310**
Raw+ms	.8992	.2532	.0453	.0656	.1484	.0663	.8978	.2480	.0400	.0710	.1451	.0563
Raw+ms+α	.8992	.2532	.0453	.0656	**.1453**	**.0355**	.8978	.2480	.0400	.0710	**.1428**	**.0274**
MCDO	.8983	.2517	.0437	.0586	.1470	.0562	.8989	.2460	.0380	**.0561**	.1446	.0505
MCDO+α	.8996	.2518	.0438	.0610	**.1450**	**.0346**	.9003	**.2458**	**.0378**	.0568	**.1426**	**.0237**
MCDO+ts	.8986	**.2515**	**.0435**	.0579	.1479	.0635	.8995	**.2458**	**.0378**	.0579	.1447	.0521
MCDO+ts+α	.8991	**.2515**	**.0435**	.0586	**.1442**	**.0186**	.8997	.2460	.0380	.0581	**.1423**	**.0180**
MCDO+vs	.8997	.2521	.0441	.0589	.1487	.0664	**.9004**	.2461	.0381	.0593	.1448	.0531
MCDO+vs+α	**.9009**	.2519	.0439	**.0575**	**.1446**	**.0244**	.9002	.2460	.0380	.0579	**.1426**	**.0228**
MCDO+ms	.8970	.2536	.0456	.0678	.1479	.0612	.8997	.2477	.0397	.0704	.1448	.0529
MCDO+ms+α	.8986	.2534	.0454	.0667	**.1449**	**.0279**	.8986	.2478	.0399	.0695	**.1424**	**.0188**
TTA	.9013	.2458	.0378	.0642	.1441	.0488	.9069	.2402	.0322	.0645	.1425	.0444
TTA+α	.9025	**.2456**	**.0376**	.0684	**.1428**	**.0334**	.9077	.2402	.0322	.0650	**.1413**	**.0277**
TTA+ts	.9012	.2459	.0379	.0646	.1448	.0553	.9074	**.2401**	**.0321**	.0636	.1425	.0456
TTA+ts+α	.9011	.2458	.0378	**.0632**	**.1422**	**.0197**	.9055	.2403	.0323	**.0633**	**.1410**	**.0204**
TTA+vs	**.9031**	.2471	.0392	.0671	.1458	.0598	**.9078**	.2409	.0329	.0666	.1431	.0483
TTA+vs+α	.9025	.2470	.0391	.0657	**.1428**	**.0247**	.9068	.2409	.0329	.0664	**.1416**	**.0252**
TTA+ms	.9001	.2489	.0410	.0756	.1456	.0561	.9034	.2429	.0349	.0778	.1431	.0481
TTA+ms+α	.8991	.2487	.0407	.0750	**.1431**	**.0268**	.9030	.2432	.0352	.0766	**.1414**	**.0213**

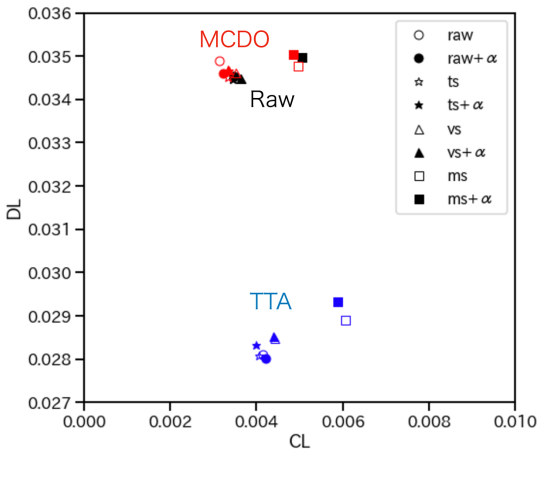


(a) Calibration-dispersion map

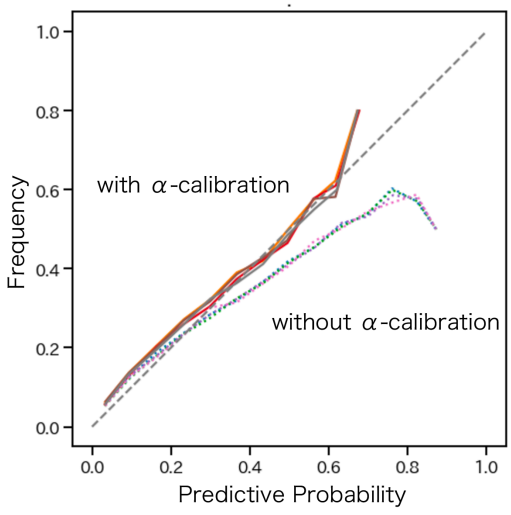


(b) Reliability diagram for disagreement probability

Figure 5: (a) Calibration-dispersion map for the experiments with MDS-1 data. (b) Reliability diagram of disagreement probability estimates (DPEs) for MDS-1 data, where all the predictive methods in Table 6 were compared. The dashed diagonal line corresponds to a calibrated prediction. Calibration of DPEs was significantly enhanced with α-calibration (solid lines) from the original ones (dotted lines).



(a) Calibration-dispersion map



(b) Reliability diagram for disagreement probability

Figure 6: (a) Calibration-dispersion map for the experiments with MDS-full data. (b) Reliability diagram of disagreement probability estimates (DPEs) for MDS-full data, where all the predictive methods in Table 6 were compared. The dashed diagonal line corresponds to a calibrated prediction.