Supplementary Material of Meta-Cal

Xingchen Ma [1] **Matthew B. Blaschko** [1]

A. Proofs of Theoretical Results

In this section, we give proofs of propositions in the main paper.

Proof of Proposition 2. For a given $g \in \mathcal{G}_a$, the first step is to find the supremum of the binned estimator $\widehat{\mathrm{ECE}}_B(g \circ f)$ across all binning schemes $B \in \mathcal{B}$. Let $B = \{I_1, \cdots, I_b\}$, then on the j-th bin I_j, by Minkowski inequality, we have:

$$\begin{aligned}&\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{1}(Y_{j,i} = \widehat{Y}_{j,i}) - \frac{1}{n_j}\sum_{i=1}^{n_j}\widehat{Z}_{j,i}\right| \\ \le&\frac{1}{n_j}\sum_{i=1}^{n_j}\left|\mathbb{1}(Y_{j,i} = \widehat{Y}_{j,i}) - \widehat{Z}_{j,i}\right|,\end{aligned}$$

where n_j is the number of samples falling into I_j, $Y_{j,i}, \widehat{Y}_{j,i}, \widehat{Z}_{j,i}$ are the i-th label, the i-th prediction and the i-th confidence score in the i-th bin respectively. This implies that to maximize the binned estimator, I_j can be further divided into n_j sub-partitions so that each sub-partition contains exactly one sample. Let B^* denotes this implied binning scheme. Then we find the lower bound of $\widehat{\mathrm{ECE}}_{B^*}(g \circ f)$. Since $g \in \mathcal{G}_a$, this imposes a constraint on $\widehat{Z}_{j,i}$:

$$1 > \widehat{Z}_{j,i} > \frac{1}{k}.$$

For the i-th sample in I_j, if $Y_{j,i} = \widehat{Y}_{j,i}$, we have:

$$\left|\mathbb{1}(Y_{j,i} = \widehat{Y}_{j,i}) - \widehat{Z}_{j,i}\right| \in \left(0, 1 - \frac{1}{k}\right).$$

If $Y_{j,i} \neq \widehat{Y}_{j,i}$, we have:

$$\left|\mathbb{1}(Y_{j,i} = \widehat{Y}_{j,i}) - \widehat{Z}_{j,i}\right| > 1/k.$$

Let $n_{j,1}$ be the number of wrongly classified samples in I_j, then:

$$\frac{1}{n_j}\sum_{i=1}^{n_j}\left|\mathbb{1}(Y_{j,i} = \widehat{Y}_{j,i}) - \widehat{Z}_{j,i}\right| > \frac{n_{j,1}}{n_j k}.$$

Since the above inequality holds for every bin, then:

$$\begin{aligned}\sup_{B\in\mathcal{B}}\widehat{\mathrm{ECE}}_B(g \circ f) &= \widehat{\mathrm{ECE}}_{B^*}(g \circ f) \\ &> \sum_{j=1}^{b}\frac{n_j}{n}\frac{n_{j,1}}{n_j k} = \frac{1-\hat{\pi}_0}{k},\end{aligned}$$

where $\hat{\pi}_0$ is the empirical accuracy of f. □

Proof of Proposition 6. Using the construction rules described in Proposition 3, for any $B \in \mathcal{B}$, we have:

$$\begin{aligned}\widehat{\mathrm{ECE}}_B(g \circ f) = &\left[(1-\hat{R}_0(\phi))\cdot\hat{\pi}_0 + \hat{R}_1(\phi)\cdot\hat{\pi}_1\right] \\ &\cdot\left|1 - \frac{(1-\hat{R}_0(\phi))\cdot\hat{\pi}_0}{(1-\hat{R}_0(\phi))\cdot\hat{\pi}_0 + \hat{R}_1(\phi)\cdot\hat{\pi}_1}\right| \\ &+ \left[\hat{R}_0(\phi)\hat{\pi}_0 + (1-\hat{R}_1(\phi))\hat{\pi}_1\right]\cdot\left|\frac{1}{k} - \frac{1}{k}\right| \\ = &\hat{R}_1(\phi)(1-\hat{\pi}_0).\end{aligned}$$

Since the above equality holds for $\forall B \in \mathcal{B}$, it also holds for B^*, which is defined in the proof of Proposition 2. □

Proof of Proposition 7. Under Assumption 1, from the definition of the miscoverage rate in Definition 4, we have:

$$\left(\mathbb{1}(h(X) > r_{(v)}) \mid Y = \widehat{Y}\right) \sim \mathrm{Bern}(F_0(g)),$$

where $F_0(g)$ is the population miscoverage rate of g, $v = \lceil (n_1+1)(1-\alpha)\rceil$, $r_{(v)}$ is the v-th order statistic of $\{r_i\}_{i=1}^{n_1}$. Without loss of generality, we suppose the first m_1 inputs have negative labels and the last m_2 inputs have positive labels, where $m = m_1 + m_2$. Then the empirical miscoverage rate is:

$$\widehat{F}_0(g) = \frac{1}{m_1}\sum_{i=1}^{m_1}\mathbb{1}(h(x_i) > r_{(v)}).$$

Thus $m_1\widehat{F}_0(g)$ follows a Binomial distribution, that is, $m_1\widehat{F}_0(g) \sim \mathrm{Binom}(m_1, F_0(g))$. For large m_1, we can approximate it by $\mathcal{N}(m_1F_0(g), m_1F_0(g)(1-F_0(g)))$ using the central limit theorem. Let R_0 be a random variable following a Normal distribution $\mathcal{N}(F_0(g), \sigma^2)$, where

$\sigma^2 = F_0(g)(1 - F_0(g))/m_1$. Applying the Chernoff bound for a Gaussian variable:

$$\mathbb{P}\left(|R_0 - F_0(g)| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right).$$

Combining the above results finishes the proof of Proposition 7. □

Proof of Proposition 8. This follows directly from Proposition 2. □

Proof of Proposition 12. Under Assumptions 1 and 2, from the definition of coverage accuracy in Definition 5 and Proposition 11, we have:

$$\left(\mathbb{1}(Y = \widehat{Y}) \mid h(X) < l^{-1}(\beta)\right) \sim \mathrm{Bern}(\beta).$$

Without loss of generality, we suppose ranking scores of the first $m_1 \leq m$ inputs are smaller than $l^{-1}(\beta)$. The empirical coverage accuracy is:

$$\widehat{F}_1(g) = \frac{1}{m_1}\sum_{i=1}^{m_1} \mathbb{1}(y_i = \hat{y}_i).$$

Thus $m_1\widehat{F}_1(g)$ follows a Binomial distribution and can be approximated by a Gaussian distribution for large m_1. The remaining proof is similar to the proof of Proposition 7. □

B. A Trivial Construction

In this section, we show bounding $\widehat{\mathrm{ECE}}_B(g \circ f)$ instead of its supremum makes little sense by giving a trivial construction. Suppose we are to evaluate the binned estimator on 4 samples and their predicted confidence scores, predictions, targets are $\{0.6, 0.7, 0.8, 0.81\}$, $\{0, 0, 0, 0\}$, $\{0, 0, 0, 1\}$ respectively. Let the binning scheme be the interval $B = [0, 1]$, we can construct a accuracy-preserving calibration map:

$$g(x) = \begin{cases} 0.9 & , \quad \text{if } x = 0.81 \\ x & , \quad \text{otherwise.} \end{cases}$$

The binned estimator using such a calibration map on this data set is 0. Obviously, this calibration map is not practical.

C. Training Details

To train networks on CIFAR-10 and CIFAR-100, we use stochastic gradient descent with momentum (0.9) using mini-batches of 128 samples for 200 epochs. We also add a L2-weight decay regularization term, which is set to be 0.0001. The start learning rate is set to be 0.1 and is decreased to 0.01 and 0.001 in the beginning of the 80-th epoch and 150-th epoch respectively. Horizontal flipping and cropping are used as data augmentation.