A Affine Calibration with $\alpha = 1$

In this appendix we derive an expression for β for logistic regression calibration when fixing $\alpha = 1$ (see Section 4 in the main paper). That is, we assume the following expression for the calibrated posterior:

$$\log \tilde{P}(y_k|\mathbf{q}, \mathbf{e}) = \gamma + \log P(y_k|\mathbf{q}, \mathbf{e}) + \beta_k \quad (13)$$

where the γ is simply a scaling factor so that the resulting posteriors add to 1 (see Section 4 for the expression for γ).

Given a set $\mathcal{C}^{\text{train}} = \{(\mathbf{q}^{(1)}, y^{(1)}), \ldots, (\mathbf{q}^{(N)}, y^{(N)})\}$ where $\mathbf{q}^{(i)}$ and $y^{(i)}$ are the query and the class of sample i, the logistic regression approach estimates the β_k parameters as the values that minimize

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \tilde{P}(y^{(i)}|\mathbf{q}^{(i)}, \mathbf{e}). \quad (14)$$

To obtain an expression for the β_k we can set to zero the derivative $\frac{\partial \mathcal{L}}{\partial \beta_k}$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k}\left(-\frac{1}{N}\sum_{i=1}^{N} \log \tilde{P}(y^{(i)}|\mathbf{q}^{(i)}, \mathbf{e})\right) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \beta_k} \log \tilde{P}(y^{(i)}|\mathbf{q}^{(i)}, \mathbf{e}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial}{\partial \beta_k}\gamma^{(i)} + \mathbb{1}_{\{y^{(i)}=y_k\}}\right) \quad (15)\end{aligned}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\gamma^{(i)}$ is defined in equation 8. Then,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_k} &= -\frac{1}{N}\sum_{i=1}^{N}\left(-e^{\gamma^{(i)}}P(y_k|\mathbf{q}^{(i)},\mathbf{e})e^{\beta_k} + \mathbb{1}_{\{y^{(i)}=y_k\}}\right) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left(-\tilde{P}(y_k|\mathbf{q}^{(i)}, \mathbf{e}) + \mathbb{1}_{\{y^{(i)}=y_k\}}\right).\end{aligned}$$

Setting this derivative to zero, we get:

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{P}(y_k|\mathbf{q}^{(i)}, \mathbf{e}) = \frac{N_k}{N} \quad (16)$$

where N_k is the number of samples that belongs to class k. Using Equation Equation (13), we obtain the expression for the optimal β_k

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} P(y_k|\mathbf{q}^{(i)}, \mathbf{e})e^{\beta_k}e^{\gamma^{(i)}} &= \frac{N_k}{N} \\ e^{\beta_k}\frac{1}{N}\sum_{i=1}^{N} P(y_k|\mathbf{q}^{(i)}, \mathbf{e})e^{\gamma^{(i)}} &= \frac{N_k}{N} \\ \beta_k = \log \frac{N_k}{N} - \log &\left[\frac{1}{N}\sum_{i=1}^{N} P(y_k|\mathbf{q}^{(i)}, \mathbf{e})e^{\gamma^{(i)}}\right]\end{aligned}$$

B Additional Results

Figure 4 shows the results when the number of training samples is set to 40. Some of the curves like the one corresponding to the cross-entropy of the calibrated model in the DBPedia dataset that could not be fully plotted because cross-entropy could not be computed for a class with zero train probability. This may occur because some classes were never seen in the train set that was used in the calibration process.

The trends are similar to those shown in section 7 with the exception that SUCPA works worse than UCPA in most cases due to a bad estimate of the class priors. In addition, there are some cases such as TREC, in which adaptation improves the performance of the model but it still present a cross-entropy close to 1, which means that performance remains close to random.

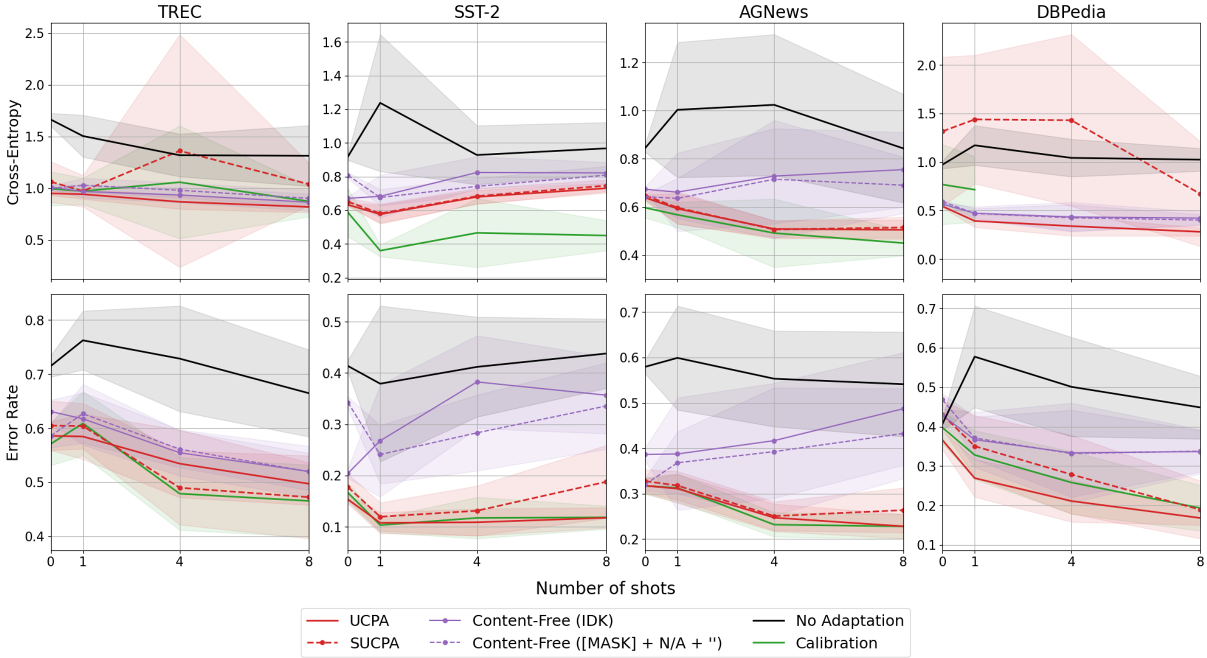

Figure 4: Cross-Entropy and Error Rate (1-Accuracy) vs. the number examples (shots) contained in the prompt for 40 training samples. Red lines show the iterative approach for UCPA and SUCPA. Lines in purple show the results for content-free adaptation and green line is the calibration using parameters α and β. As before, black line shows the case for which no adaptation has been performed.