.

A Penalty functions for ALM: axioms

We provide here the requirements for a penalty function in the Augmented Lagrangian Multiplier (ALM) method.

A function $P : \mathbb{R} \times \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R}$ is a Penalty-Lagrangian function such that $P'(z, \rho, \lambda) \equiv \frac{\partial}{\partial z} P(z, \rho, \lambda)$ exists and is continuous for all $z \in \mathbb{R}$, $\rho \in \mathbb{R}_{++}$ and $\lambda \in \mathbb{R}_{++}$. In addition, a penalty function P should satisfy the following four axioms [3]:

- **Axiom 1:** $P'(z, \rho, \lambda) \geq 0 \quad \forall z \in \mathbb{R}, \rho \in \mathbb{R}_{++}, \lambda \in \mathbb{R}_{++}$
- **Axiom 2:** $P'(0, \rho, \lambda) = \lambda \quad \forall \rho \in \mathbb{R}_{++}, \lambda \in \mathbb{R}_{++}$
- **Axiom 3:** If, for all $j \in \mathbb{N}$, $\lambda^{(j)} \in [\lambda_{\min}, \lambda_{\max}]$, where $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$, then:
  $$\lim_{j\to\infty} \rho^{(j)} = \infty \text{ and } \lim_{j\to\infty} y^{(j)} > 0 \text{ imply that } \lim_{j\to\infty} P'(y^{(j)}, \rho^{(j)}, \lambda^{(j)}) = \infty$$
- **Axiom 4:** If, for all $j \in \mathbb{N}$, $\lambda^{(j)} \in [\lambda_{\min}, \lambda_{\max}]$, where $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$, then:
  $$\lim_{j\to\infty} \rho^{(j)} = \infty \text{ and } \lim_{j\to\infty} y^{(j)} < 0 \text{ imply that } \lim_{j\to\infty} P'(y^{(j)}, \rho^{(j)}, \lambda^{(j)}) = 0.$$

While the first two axioms guarantee that the derivative of the Penalty-Lagrangian function P *w.r.t.* z is positive and equals to λ when $z = 0$, the last two axioms guarantee that the derivative tends to infinity when the constraint is not satisfied, and zero otherwise.

B Additional details on evaluation metrics

- Expectation Calibration Error (ECE). ECE measures the correctness of the predictions by taking the weighted average of the error between accuracy and confidence. Let N be the total number of data samples, B be the total number of bins available for grouping. Then, ECE is given by:

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left|\text{acc}(b) - \text{conf}(b)\right|,$$

Here, n_b, $acc(b)$, and $conf(b)$ denotes the number of samples, accuracy, and the confidence specific to that particular bin (b).

- Threshold Adaptive Calibration Error (TACE). In order to get the best estimate of the overall calibration, it is better to focus on the bins where most of the prediction are made. This prevent the outcome to be skewed, and it is generally decided by an adaptive calibration range, give by r. Then, ACE is given by:

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} \left|\text{acc}(r, k) - \text{conf}(r, k)\right|.$$

To further resolve the problem with tiny confidence scores making the calibration metric infinitesimal, TACE uses a threshold to skip the minimum ones.

- Friedman rank. This metric is employed when there exist multiple metrics and settings to compare several methods. It can be defined as $\text{rank}_F = \frac{1}{m}\sum_{i=1}^{m}\text{rank}_i$, with m being the number of evaluation settings ($m = 16$ in our work, 8 metrics $\times$ 2 datasets), and rank_i the rank of a method in the i-th setting. Thus, the lower the rank obtained by an approach, the better this method is.

C Additional results: logit distributions

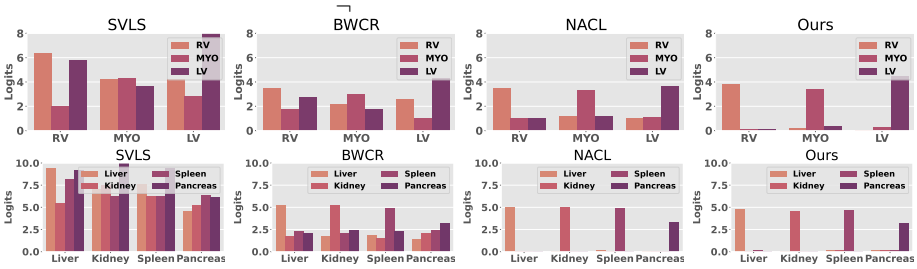

Fig. 2: **Distribution of logit values.** A desirable logit distribution exhibits lower winner logit magnitudes, which facilitate the training of a well-calibrated model, while pushing the remaining logit values to a considerable distance, and thus preserve a high discriminative power. An interesting observation from this figure is that, while NACL seems to generate desirable logit distributions for one dataset (FLARE), it may require fine-tuning of the λ hyperparameter. In contrast, CRaC integrates an explicit mechanism to learn these values automatically, which facilitates a better compromise between segmentation and calibration.