Sparse Activations as Conformal Predictors

Margarida M. Campos[1,2] **João Cálem**[2] **Sophia Sklaviadis**[1,2]

Mário A.T. Figueiredo[1,2,3] **André F.T. Martins**[1,2,3,4]

[1] Instituto de Telecomunicações
[2] Instituto Superior Técnico, Universidade de Lisboa
[3] ELLIS Unit Lisbon
[4] Unbabel
margarida.campos@tecnico.ulisboa.pt

Abstract

Conformal prediction is a distribution-free framework for uncertainty quantification that replaces point predictions with sets, offering marginal coverage guarantees (*i.e.*, ensuring that the prediction sets contain the true label with a specified probability, in expectation). In this paper, we uncover a novel connection between conformal prediction and sparse "softmax-like" transformations, such as sparsemax and γ-entmax (with $\gamma > 1$), which may assign nonzero probability only to a subset of labels. We introduce new non-conformity scores for classification that make the calibration process correspond to the widely used temperature scaling method. At test time, applying these sparse transformations with the calibrated temperature leads to a support set (*i.e.*, the set of labels with nonzero probability) that automatically inherits the coverage guarantees of conformal prediction. Through experiments on computer vision and text classification benchmarks, we demonstrate that the proposed method achieves competitive results in terms of coverage, efficiency, and adaptiveness compared to standard non-conformity scores based on softmax.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

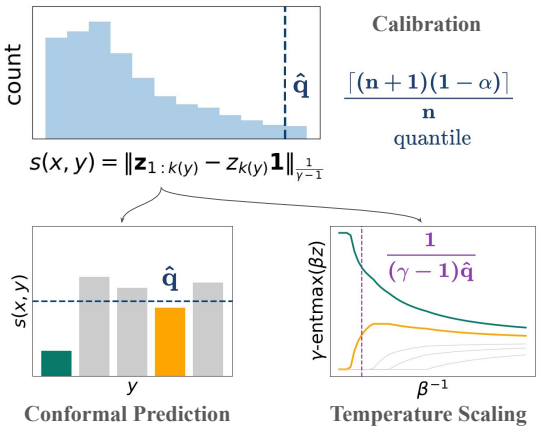

Figure 1: **Conformal prediction meets temperature scaling:** we derive new non-conformity scores $s(x, y)$ that make conformal prediction equivalent to γ-entmax temperature scaling.

1 INTRODUCTION

The use of learned predictive models in many high-stakes applications (*e.g.*, medical, legal, or financial) has stimulated extensive research on uncertainty quantification as a way to enhance predictions with reliable confidence estimates (Silva-Filho et al., 2023; Gawlikowski et al., 2023). In classification tasks, this corresponds to providing an estimate of the posterior probabilities of each of the classes, given the sample to be classified. In regression, this would correspond to providing, not only a point estimate, but also a confidence interval around that estimate.

Unfortunately, modern deep networks are not well calibrated: the class probability estimates may be significantly different from the posterior probability, and thus cannot be trusted as confidence levels (Guo et al.,

2017; Wenger et al., 2020). This observation has motivated considerable work on developing methods to obtain well-calibrated class probability estimates.

To address this problem, **conformal prediction** (§2.1) has emerged as a powerful framework for uncertainty quantification that comes with strong theoretical guarantees under minimal assumptions, as it is model-agnostic and distribution-free (Vovk et al., 2005; Angelopoulos and Bates, 2023). Conformal predictors are *set* predictors, *i.e.*, their outputs are either sets of labels (for classification tasks) or intervals (for regression tasks). Most common variants, namely *split* conformal prediction (Papadopoulos et al., 2002), do not require model retraining: prediction rules are created using a separate calibration set through the design of a non-conformity score, which is a measure of how unlikely an input-output pair is.

A parallel research direction concerns **sparse** alternatives to the softmax activation (§2.2), *i.e.*, capable of producing probability outputs where some labels receive zero probability, yet are differentiable, thus usable in backpropagation (Martins and Astudillo, 2016; Peters et al., 2019). The class of Fenchel-Young losses (Blondel et al., 2020) provides a framework to learn such sparse predictors, generalizing the softmax activation and the cross-entropy loss. An added benefit of sparse probability outputs is that the labels with nonzero probability can be directly interpreted as a *set prediction*, eliminating the need for creating these sets by thresholding probability values. Such set predictions have been used by Martins and Astudillo (2016) for multi-label classification, but not in the scope of uncertainty quantification.

This paper connects the two lines of work above by **conformalizing sparse activations** (Figure 1; §3). We focus on the γ-entmax family[1] (Peters et al., 2019), which includes softmax ($\gamma = 1$) and sparsemax ($\gamma = 2$) as particular cases, and is sparse for $\gamma > 1$. We create set predictions by including all labels receiving nonzero probability under those activations—these sets can be made larger or smaller by controlling a *temperature* parameter. We leverage the calibration techniques of conformal prediction to choose the temperature that leads to the desired coverage level.

Our main contributions are the following:

- A new non-conformity score that makes the conformal predictor equivalent to temperature scaling of the sparsemax activation, establishing a formal link between the two lines of work and ensuring statistical coverage guarantees (§3.1).
- Generalization of the construction above to the γ-entmax family of activations, with the same guarantees (§3.2). This yields a calibration strategy that can also be used with softmax ($\gamma = 1$) and where γ can be tuned as a hyperparameter (*e.g.* to optimize prediction set size).
- Empirically validation of our method on a range of computer vision and text classification tasks across several datasets, comparing the resulting coverage, efficiency, and adaptiveness with commonly used non-conformity scores (§4).
- Expansion of the family of non-conformity scores for conformal prediction tasks with new scores that are competitive with state-of-the-art techniques.

The code to reproduce all the reported experiments is publicly available.[2]

2 BACKGROUND

This section provides background by briefly reviewing basic concepts of conformal prediction and sparse activations.

2.1 Conformal Prediction

Consider a supervised learning task with input set $\mathcal{X}$, output set $\mathcal{Y}$, and a predictor, $f : \mathcal{X} \to \mathcal{Y}$. Given a new observation, $x_{\text{test}} \in \mathcal{X}$, a standard predictor produces a single point estimate: $\hat{y}_{\text{test}} = f(x_{\text{test}})$. A conformal predictor, $\mathcal{C}_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$, outputs instead a prediction set $\mathcal{C}_\alpha(x_{\text{test}}) \subseteq \mathcal{Y}$, which is expected to include the correct target, y_{test}, with a given probability $1 - \alpha$, where α is a user-chosen error rate.

Procedure We focus on split conformal prediction, which does not involve model retraining (Papadopoulos et al., 2002). Given the predictor, f, and a collection of calibration samples, $\mathcal{D}_{\text{cal}} = \big((x_1, y_1), ..., (x_n, y_n)\big)$, a conformal predictor is defined by a non-conformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The non-conformity score measures how unlikely an input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is, compared to the remaining calibration data. Ideally, predictions $\hat{y} \in \mathcal{Y}$ yielding pairs $(x_{\text{test}}, \hat{y})$ that are likely to occur in the data should have a low non-conformity score, and should thus be included in the prediction set $\mathcal{C}_\alpha(x_{\text{test}})$.

The procedure for generating the prediction set for an unobserved sample, x_{test}, is as follows:

1. During calibration, the non-conformity scores are computed for $\mathcal{D}_{\text{cal}}$, $(s_1, ..., s_n)$, where $s_i = s(x_i, y_i)$.

[1] Called α-entmax by Peters et al. (2019). We choose γ in this paper to avoid clash with the confidence level α, commonly used in the conformal prediction literature.

[2] https://github.com/deep-spin/sparse-activations-cp

2. A threshold $\hat{q}$ is set to the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of these scores.
3. At test time, the following prediction set is output:

$$\mathcal{C}_\alpha(x_{\text{test}}) = \{y \in \mathcal{Y} : s(x_{\text{test}}, y) \leq \hat{q}\}. \quad (1)$$

Coverage guarantees If the calibration data and the test datum $((X_1, Y_1), ..., (X_n, Y_n), (X_{\text{test}}, Y_{\text{test}}))$ are *exchangeable*,[3] conformal predictors obtained with this procedure yield prediction sets satisfying

$$\mathbb{P}\big(Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\big) \geq 1 - \alpha, \quad (2)$$

as proved by Vovk et al. (2005), without any other assumptions on the predictive model or the data distribution. Ideally, we would like the predictor with the required coverage to be *efficient*, *i.e.*, to have a small average prediction set size $\mathbb{E}[|C_\alpha(X)|]$, and *adaptive*, *i.e.*, instances that are harder to predict should yield larger prediction sets, representing higher uncertainty.

2.2 Sparse Activations

Consider a classification task with K classes, $|\mathcal{Y}| = K$, and let $\triangle_K = \{\boldsymbol{p} \in \mathbb{R}^K : \boldsymbol{p} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{p} = 1\}$ denote the $(K-1)$–dimensional probability simplex. Typically, predictive models output a vector of label scores, $\boldsymbol{z} \in \mathbb{R}^K$, which is converted to a probability vector in $\triangle_K$ through a suitable transformation, usually softmax:

$$\mathsf{softmax}(\boldsymbol{z})_j := \frac{\exp(z_j)}{\sum_i \exp(z_i)}. \quad (3)$$

Since the exponential function is stricly positive, the softmax transformation never outputs any zero. Martins and Astudillo (2016) introduced an alternative transformation, *sparsemax*, which can produce sparse outputs while being almost-everywhere differentiable:

$$\mathsf{sparsemax}(\boldsymbol{z}) := \arg\min_{\boldsymbol{p} \in \triangle_K} \|\boldsymbol{p} - \boldsymbol{z}\|^2. \quad (4)$$

The solution of (4), which corresponds to the Euclidean projection of $\boldsymbol{z}$ onto $\triangle_K$, can be computed efficiently through Algorithm 1. This algorithm, which sorts the label scores and seeks a threshold which ensures normalization, motivates our proposed nonconformity score to be presented in §3.1.

As shown by Peters et al. (2019) and Blondel et al. (2020), sparsemax and softmax are particular cases of a family of γ-entmax transformations, defined as

$$\gamma\text{-}\mathsf{entmax}(\boldsymbol{z}) := \arg\max_{\boldsymbol{p} \in \triangle_K} \boldsymbol{p}^\top \boldsymbol{z} + H_\gamma(\boldsymbol{p}), \quad (5)$$

[3] A sequence $(Z_1, ..., Z_n)$ is said to be *exchangeable* if $(Z_1, ..., Z_n) \overset{d}{=} (Z_{\pi(1)}, ..., Z_{\pi(n)})$ for any permutations π of $\{1, ..., n\}$, where $\overset{d}{=}$ stands for *identically distributed*.

Algorithm 1 Sparsemax evaluation.

Require: $\boldsymbol{z} \in \mathbb{R}^K$

1: Sort $z_{(1)} \geq z_{(2)} \geq ... \geq z_{(K)}$

2: Find $k(\boldsymbol{z}) = \max\left\{j \in [K] : 1 + j z_{(j)} > \sum_{k=1}^{j} z_{(k)}\right\}$.

3: Compute $\tau = \frac{\left(\sum_{k=1}^{k(\boldsymbol{z})} z_{(k)}\right) - 1}{k(\boldsymbol{z})}$.

4: **return** $\mathsf{sparsemax}(\boldsymbol{z}) = (\boldsymbol{z} - \tau \mathbf{1})_+$.

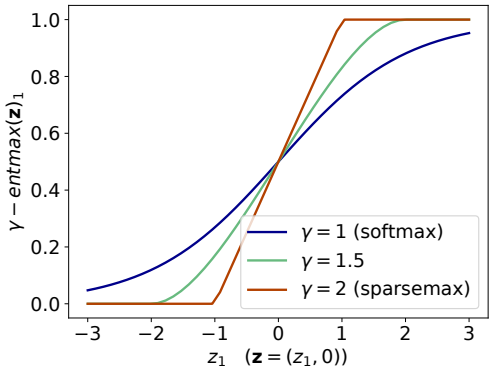

Figure 2: Illustration of entmax in the two-dimensional case γ-entmax$([t, 0])_1$.

where H_γ is a generalized entropy function (Tsallis, 1988), defined, for $\gamma > 0$, as

$$H_\gamma(\boldsymbol{p}) = \begin{cases} \frac{1}{\gamma(\gamma-1)}\left(1 - \sum_{j=1}^K p_j^\gamma\right), & \gamma \neq 1, \\ -\sum_{j=1}^K p_j \log p_j, & \gamma = 1. \end{cases} \quad (6)$$

This family of entropies is continuous with respect to γ and recovers the Shannon entropy for $\gamma = 1$. Softmax and sparsemax are particular cases of the γ-entmax transformation (5) for $\gamma = 1$ and $\gamma = 2$, respectively. Crucially, setting $\gamma > 1$ enables sparsity in the output, *i.e.*, for certain inputs $\boldsymbol{z}$, the solution $\boldsymbol{p}^*$ in (5) will be a sparse probability vector. Figure 2 depicts the behavior of γ-entmax for three different values of γ.

Peters et al. (2019) show that the solution of (5) has the form

$$\gamma\text{-}\mathsf{entmax}(\boldsymbol{z}) = [(\gamma - 1)\boldsymbol{z} - \tau \mathbf{1}]_+^{\frac{1}{\gamma - 1}}, \quad (7)$$

where $\mathbf{1}$ denotes a vector of ones, $[x]_+ = \max\{x, 0\}$, and τ is the (unique) constant ensuring normalization. A bisection algorithm for computing γ-entmax for general γ, as well as a more efficient algorithm for $\gamma = 1.5$, have been proposed by Peters et al. (2019). We make use of these algorithms in our experiments in §4.

Adjusting sparsity Temperature scaling (Guo et al., 2017; Platt et al., 1999) is a simple and popular technique to calibrate the output probabilities from a

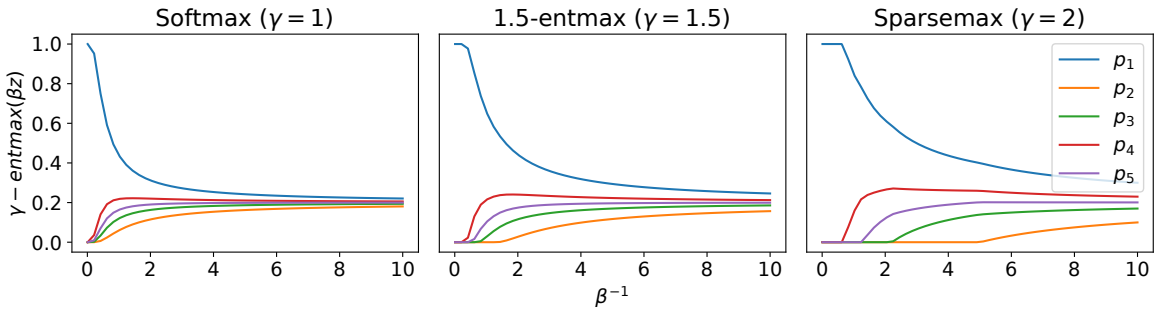

Figure 3: Output of different γ-entmax transformations on label scores $\boldsymbol{z} = [1, -1, -0.2, 0.4, -0.5]$, as a function of temperature parameter β^{-1}, where $[p_1, p_2, p_3, p_4, p_5] = \gamma\text{-entmax}(\beta \boldsymbol{z})$.

predictive model. By multiplying the label scores $\boldsymbol{z}$ by a scaling factor $\beta > 0$, which can be interpreted as an inverse temperature parameter, the probabilities resulting from the subsequent softmax transformation can be made more or less peaked. When $\beta = 0$, $\text{softmax}(\beta \boldsymbol{z})$ becomes a uniform distribution, whereas in the limit $\beta \to +\infty$ (the "zero temperature limit"), $\text{softmax}(\beta \boldsymbol{z})$ tends to a one-hot vector, assigning probability 1 to the label with the largest score.[4]

Analogous limit cases apply also to $\gamma\text{-entmax}(\beta \boldsymbol{z})$, but there is a key qualitative difference (Martins and Astudillo, 2016; Blondel et al., 2020). For any $\gamma > 1$, there is a *finite* threshold β^* (corresponding to a "non-zero temperature") above which $\gamma\text{-entmax}(\beta \boldsymbol{z})$ saturates as the zero temperature limit: $\beta \geq \beta^* \Rightarrow \gamma\text{-entmax}(\beta \boldsymbol{z}) = \lim_{\beta \to +\infty} \gamma\text{-entmax}(\beta \boldsymbol{z})$. Increasing β, which may be seen as a regularization coefficient, from 0 to β^* draws a regularization path where the support of γ-entmax$(\beta \boldsymbol{z})$ varies from the full set of labels to a single label (assuming no ties), as shown in Figure 3. Therefore, using temperature scaling with γ-entmax, for $\gamma > 1$, allows the sparsity of the output to be controlled. We exploit this fact in our proposed method described in §3.

3 CONFORMALIZING SPARSE TRANSFORMATIONS

We now present a connection between sparse transformations and conformal prediction, via the design of a non-conformity score for which the prediction sets given by conformal prediction are the same as the support of the sparsemax output (more generally, γ-entmax with $\gamma > 1$) where the calibration step corresponds to setting β in temperature scaling.

[4]If there are multiple labels tied with the same largest score, $\lim_{\beta \to +\infty} \text{softmax}(\beta \boldsymbol{z})$ becomes a uniform distribution over those labels, assigning zero probability to all other labels.

Sparsity as set prediction Considering that, for $\gamma > 1$, the γ-entmax transformations can produce sparse outputs and that the level of sparsity can be controlled through the coefficient β, the support of the output can be interpreted as a prediction set for a given input $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{z} = f(\boldsymbol{x}) \in \mathbb{R}^K$ is the vector of label scores produced by a trained model. For notational convenience, we assume throughout that the entries of $\boldsymbol{z}$ are sorted in descending order, $z_1 \geq z_2 \geq \ldots \geq z_K$.

3.1 Conformalizing sparsemax

Let us start with sparsemax. Taking a closer look at Algorithm 1, we start by noting that the inequality in line 2 can be equivalently written as[5] $\sum_{k=1}^{j-1}(z_k - z_j) < 1$. Let $S(\boldsymbol{z}) := \{j \in [K] : \text{sparsemax}_j(\boldsymbol{z}) > 0\}$ denote the support of $\text{sparsemax}(\boldsymbol{z})$, *i.e.*, the labels that are assigned strictly positive probability. Scaling the label scores by the coefficient $\beta > 0$, we obtain a necessary and sufficient condition for the j^{th} highest ranked label to be in the support $S(\beta \boldsymbol{z})$:

$$j \in S(\beta \boldsymbol{z}) \iff \sum_{k=1}^{j-1}(z_k - z_j) < \beta^{-1}. \tag{8}$$

This fact leads to the following proposition.

Proposition 1. *Let $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ be a conformal predictor (as described in §2.1). Define the following nonconformity score:*

$$s(x, y) = \sum_{k=1}^{k(y)-1} (z_k - z_{k(y)}), \tag{9}$$

where $k(y)$ is the index of label y in the sorted array $\boldsymbol{z}$, and let $\hat{q}$ be the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical

[5]By noticing that $1 + j z_j > \sum_{k=1}^{j} z_k \Leftrightarrow \left(\sum_{k=1}^{j-1} z_k\right) + z_j - z_j - (j-1) z_j < 1 \Leftrightarrow \sum_{k=1}^{j-1}(z_k - z_j) < 1$.

quantile of the set of calibration scores. Then, setting the sparsemax temperature as $\beta^{-1} := \hat{q}$ at test time leads to prediction sets $C_\alpha(x) = S(\beta \boldsymbol{z})$ achieving the desired $(1-\alpha)$ coverage in expectation.

Proof. This follows directly from the coverage guarantee of conformal prediction (2) and condition (8). □

The non-conformity score (9) has an intuitive interpretation: it sums the score differences between each of the top ranked labels and the label y. The lower the rank of y, the larger the number of terms in this sum will be; conversely, the lower the score of y is compared to the scores of the top labels, the larger the total sum will be. If this sum is larger than a threshold, the label will not be included in the prediction set.

3.2 Conformalizing γ-entmax

We next turn to the more general case of γ-entmax. Let $S(\beta \boldsymbol{z}; \gamma) := \{j \in [K] : \gamma\text{-entmax}_j(\beta \boldsymbol{z}) > 0\}$ denote the support of the distribution induced by the γ-entmax transformation with temperature β^{-1} on the score vector $\boldsymbol{z}$. We start with the following result, which generalizes (8).

Proposition 2. *The following condition characterizes the set of labels in the support $S(\beta \boldsymbol{z}; \gamma)$:*

$$j \in S(\beta \boldsymbol{z}; \gamma) \iff \sum_{k=1}^{j-1} \left[(\gamma - 1)\beta(z_k - z_j)\right]^{\frac{1}{\gamma - 1}} < 1. \tag{10}$$

Proof. The proof can be found in Appendix A. □

Defining $\delta = 1/(\gamma - 1)$ and letting $\|\boldsymbol{v}\|_\delta := (\sum_{k=1}^K |v_k|^\delta)^{\frac{1}{\delta}}$ be the δ-norm in $\mathbb{R}^K$, we can equivalently express (10) as

$$j \in S(\beta \boldsymbol{z}, \gamma) \iff \|\boldsymbol{z}_{1:(j-1)} - z_j \mathbf{1}\|_\delta < \delta \beta^{-1}. \tag{11}$$

The sparsemax case (8) is recovered by setting $\gamma = 2$ (equivalently, $\delta = 1$) in (11). We then have the following result, generalizing Proposition 1:

Proposition 3. *Let $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$, be a conformal predictor (as described in §2.1). Define the following nonconformity score:*

$$s(x, y) = \|\boldsymbol{z}_{1:k(y)} - z_{k(y)} \mathbf{1}\|_\delta, \tag{12}$$

where $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the set of calibration scores. Then, setting the γ-entmax temperature as $\beta^{-1} := \delta^{-1}\hat{q}$ at test time leads to prediction sets $C_\alpha(x) = S(\beta \boldsymbol{z}; \gamma)$ achieving the desired $(1-\alpha)$ coverage in expectation.

Proof. This again follows directly from the conformal prediction coverage guarantee (2) and condition (11). □

Proposition 3 states that, through an appropriate choice of non-conformity score, a conformal predictor generates prediction sets corresponding to the support of the γ-entmax transformation with temperature $\beta^{-1} = \delta^{-1}\hat{q}$, as illustrated in Figure 1. Intuitively, the nonconformity score accumulates the differences between the largest label scores and the score of the correct label, returning the δ-norm of the vector of differences. For example, when $\gamma = 1.5$ (1.5-entmax), we have $\delta = 2$, which corresponds to the Euclidean norm. The non-conformity score (12) is always non-negative and it is zero if $k(y) = 1$.

Log-margin Interestingly, the non-conformity score (12) also works for softmax ($\gamma = 1$, *i.e.*, $\delta = +\infty$):

$$\begin{aligned} s(x, y) &= \|\boldsymbol{z}_{1:k(y)} - z_{k(y)} \mathbf{1}\|_\infty = z_1 - z_{k(y)} \\ &= \log \frac{p_1}{p_{k(y)}}, \end{aligned} \tag{13}$$

which is the log-odds ratio between the most probable class and the true one. Since the log function is monotonic, calibration of this non-conformity score leads to thresholding the odds ratio $p_1/p_{k(y)}$, which has an intuitive interpretation: labels whose probability is above a fraction of that of the most probable label are included in the prediction set.[6] However, the interpretation as temperature calibration no longer applies, since softmax is not sparse. We also experiment with the non-conformity score (13) in §4.

4 EXPERIMENTS

To assess the performance of the non-conformity scores introduced in §3 we report experiments on several classification tasks, comparing the proposed strategies with standard conformal prediction over several dimensions: coverage, efficiency (prediction set size), and adaptiveness, at several different confidence levels.

4.1 Experimental Setup

Datasets We evaluate all approaches on tasks of varying difficulty: image classification on the CIFAR10, CIFAR100, and ImageNet datasets

[6] This non-conformity score, although reminiscent of the previously used *margin score* (Johansson et al., 2017; Linusson et al., 2018), uses the margin (distance to the maximum) on model output, not on the probability estimates.

(Krizhevsky, 2009; Deng et al., 2009) and text classification on the 20 Newsgroups dataset (Mitchell, 1997). For each dataset, the original test data is split into calibration (40%) and test sets (60%) and the results reported are averaged over 5 random splits.

Models For the CIFAR100 and ImageNet datasets, we finetune the *vision transformer* (ViT) model (Kolesnikov et al., 2021), obtaining average accuracies of approximately 0.86 and 0.81 on the test set, respectively. As for the 20 Newsgroups dataset, we finetune a BERT base model (Devlin et al., 2019) for sequence classification, obtaining an average test accuracy of 0.74. Given the simplicity of the task on the CIFAR10 dataset, we train a convolutional neural network from scratch with final average test accuracy of 0.84. More model evaluation details can be found in Appendix C.1.

Conformal procedures We experiment with two commonly used conformal procedures, both involving the softmax transformation:

- InvProb: this is the standard conformal prediction (§2.1) with non-conformity score $s(x, y) = 1-\hat{p}(y|x)$, where $\hat{p}(y|x)$ is the softmax output corresponding to class y;
- RAPS: this is a modification of the standard procedure called *regularized adaptive prediction sets* (Angelopoulos et al., 2021), which makes use of two regularization parameters to improve the efficiency of the *adaptive predictive sets* (APS) predictors (Romano et al., 2020). In prior work, this has shown good efficiency and adaptiveness properties. Implementation of the RAPS procedure is done following the original paper[7] (additional details can be found in Appendix B.1).

We compare these approaches with the following proposed ones:

- γ-entmax: this applies conformal prediction with the non-conformity score (12) as described in Proposition 3, with $1 < \gamma \leq 2$, which has a temperature scaling interpretation. When $\gamma = 2$, we call the procedure sparsemax, corresponding to non-conformity score (9);
- log-margin: this uses the score defined in (13), which corresponds to $\gamma = 1$ (softmax), also a particular case of the non-conformity score (12), but without an interpretation in terms of temperature scaling;
- opt-entmax: a procedure where γ is treated as a hyperparameter ($1 < \gamma < 2$), tuned to minimize the average prediction set size. This is done by splitting the original calibration set in two: one for conformal prediction calibration and the second for the choice of optimal γ. Additional setup details can be found in Appendix B.2.

[7] Since using the same set to both calibrate the conformal predictor and find the optimal regularization hyperparameters would violate the exchangeability assumption, we split the original calibration data into two subsets for this procedure.

4.2 Coverage and Efficiency

Table 1 shows the average empirical coverage obtained by the different methods on the test set for two confidence levels $1 - \alpha$, with $\alpha \in \{0.01, 0.1\}$. We observe that all methods achieve a marginal coverage close to the theoretical bound of $1 - \alpha$, as expected (see Appendix C.2 for coverage analysis for more values of α).

In order to evaluate the efficiency of the predictors, we use two common metrics: the *average set size* and the *singleton ratio* (fraction of prediction sets containing a single element). We vary $\alpha \in [0.01, 0.1]$ to study different confidence levels.

Average set size As can be seen in Figure 4, opt-entmax and InvProb are the most efficient set predictors across almost all tasks and confidence levels (with opt-entmax superior to InvProb in NewsGroups), both dominating RAPS. The 1.5-entmax is in general very competitive, and so is the log-margin predictor (except for ImageNet). The sparsemax predictor generally yields larger sets on average, especially for high confidence levels.

In γ-entmax, γ can be seen as an additional hyperparameter, since for each α, it can be tuned on the calibration set for a given metric, as opt-entmax does for average set size. An example of the behavior of γ-entmax with varying γ is shown in Figure 5.

Singleton ratio Considering the ratio of singleton predictions, shown in Figure 6, we observe significant differences between methods: RAPS and log-margin have the lowest and highest ratio, respectively, across all tasks and confidence levels; InvProb, log-margin and opt-entmax output singletons even for $\alpha = 0.01$ for all tasks except for 20 Newsgroups, where no method produces singletons, which could be related to the lower accuracy of the original predictive model. The coverage of singleton predictions, also shown in Figure 6, is approximately $1 - \alpha$, as desired, for all methods.

4.3 Adaptiveness

Given that conformal predictors are unable to provide *conditional* coverage theoretical guarantees (Vovk, 2012), *i.e.*, there is no lower bound on $\mathbb{P}\big(Y_{\text{test}} \in$

Table 1: Empirical coverage of different conformal procedures on the test set (averaged over 5 splits).

α	Dataset	RAPS	1.5-entmax	log-margin	opt-entmax	InvProb	sparsemax
0.01	CIFAR10	0.990	0.990	0.990	0.989	0.990	0.989
	CIFAR100	0.991	0.991	0.991	0.991	0.990	0.991
	ImageNet	0.990	0.990	0.990	0.990	0.990	0.990
	NewsGroups	0.989	0.988	0.989	0.990	0.989	0.989
0.10	CIFAR10	0.899	0.900	0.901	0.900	0.900	0.900
	CIFAR100	0.897	0.899	0.900	0.899	0.899	0.900
	ImageNet	0.899	0.899	0.899	0.899	0.899	0.900
	NewsGroups	0.895	0.900	0.902	0.900	0.901	0.900

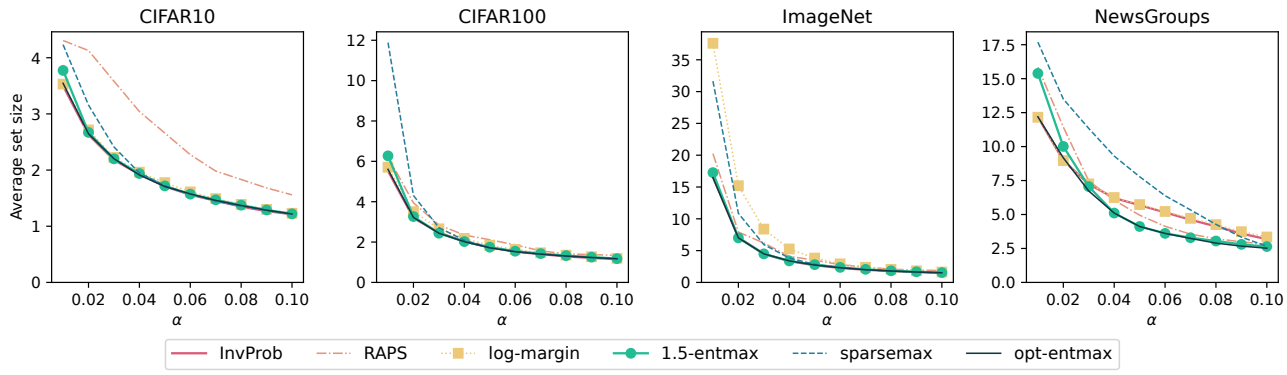

Figure 4: Average prediction set size as a function of significance level α.

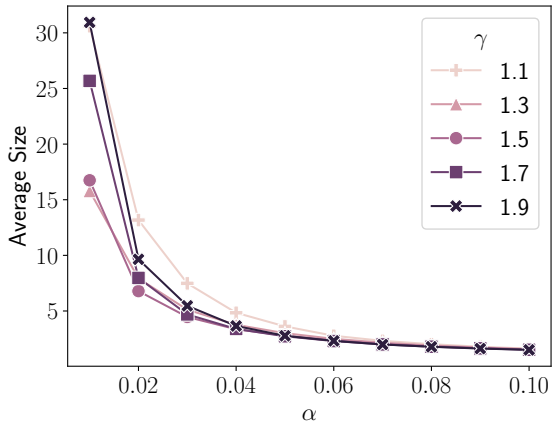

Figure 5: Average set size as a function of α, for the ImageNet dataset with varying γ for entmax.

$\mathcal{C}_\alpha(X_{\text{test}})|X_{\text{test}}\big)$ without extra assumptions, different measures have been proposed to calculate proximity to conditional coverage. Ideally, for any given partition of the data, we would have a coverage close to the $1-\alpha$ bound. Angelopoulos et al. (2021) introduced an adaptiveness criterion based on *size-stratified coverage*, by evaluating coverage in a partition based on the size of the predicted sets. Table 2 shows the size-stratified coverage for the ImageNet dataset, with $\alpha = 0.01$ and $\alpha = 0.1$. For $\alpha = 0.1$, log-margin stands out from other methods, achieving a lower deviation from exact coverage on all set cardinalities. Size-stratified coverage improves for lower values of α, and InvProb, log-margin and opt-entmax exhibit the desired behavior—achieving expected coverage while still predicting small sets for easier examples.

5 DISCUSSION

Choosing a score The choice of non-conformity score plays a crucial role in the efficiency of a conformal predictor (Aleksandrova and Chertov, 2021); consequently, finding useful model-agnostic scores is key for the adoption of conformal prediction as an actionable uncertainty quantification framework. In §3, we introduced a new family of non-conformity measures that we show to be competitive with standard scores on different classification tasks (§4). Expanding the range of possible non-conformity measures is particularly relevant, considering that the optimal score choice is known to be task-dependent (Laxhammar and Falkman, 2010). Additionally, we show that these scores can be used to tune the adaptiveness of a conformal predictor.

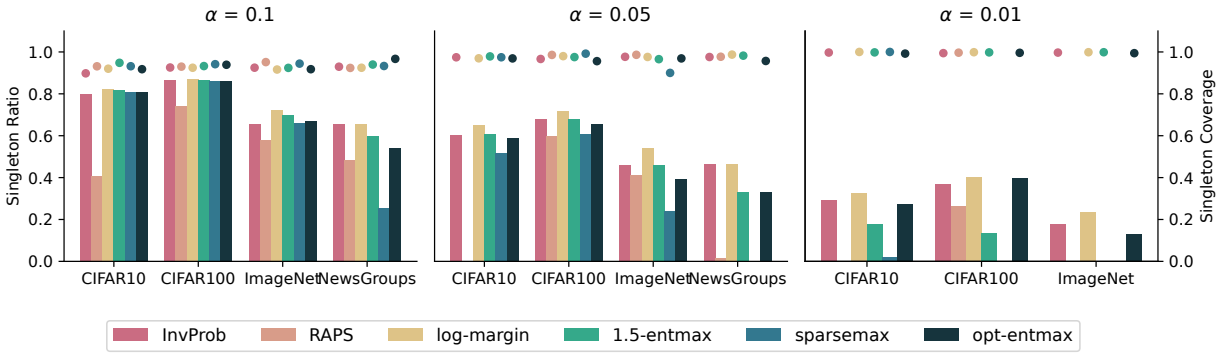

Figure 6: Ratio of singleton prediction sets (bars) and their coverage (points) for different values of confidence level $1 - \alpha$ (NewsGroups is absent for $\alpha = 0.01$ because no singletons were predicted by either method).

Table 2: Size-stratified coverage for the ImageNet dataset with $\alpha = 0.01$ (top) and $\alpha = 0.1$ (bottom).

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	5339	0.994	0	–	7088	0.992	44	1.000	0	–	3899	0.996
2 - 3	7665	0.992	0	–	7970	0.989	1559	1.000	0	–	7383	0.994
4 - 6	5403	0.990	0	–	4821	0.986	5196	0.998	0	–	6159	0.991
7 -10	3221	0.990	0	–	2604	0.982	7585	0.997	177	1.000	3814	0.992
11 - 1000	8372	0.981	30000	0.99	7517	0.987	15616	0.982	29823	0.989	8745	0.978

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	19610	0.932	17408	0.948	21664	0.924	20939	0.933	19820	0.942	20074	0.940
2 - 3	9321	0.851	8862	0.888	6346	0.859	7358	0.849	9098	0.838	8731	0.839
4 - 6	917	0.808	2696	0.774	1404	0.767	1407	0.728	1048	0.618	1140	0.636
7 - 10	4	0.500	753	0.616	373	0.743	262	0.588	34	0.6	50	0.519
11 - 1000	0	–	0	–	213	0.653	34	0.559	0	–	1	0

Temperature scaling connection Conformal prediction and temperature scaling are two popular, yet very distinct, approaches to the task of model calibration and confidence estimation. The fact that there is an equivalence between the two methods for a family of activation functions is not only intrinsically interesting but also promising, as it opens some research threads: designing new non-conformity scores from sparse transformations, or comparing an activation natural calibration capacity through the evaluation of the correspondent conformal predictor.

6 CONCLUSION

In this paper, we showed a novel connection between conformal prediction and temperature scaling in sparse activation functions. Specifically, we derived new non-conformity scores which make conformal set prediction equivalent to temperature scaling of the γ-entmax family of sparse activations. This connection allows γ-entmax with the proposed calibration procedure to inherit the strong coverage guarantees of conformal prediction. Experiments with computer vision and text classification benchmarks show the efficiency and adaptiveness of the proposed non-conformity scores, showing their practical usefulness.

7 ACKNOWLEDGEMENTS

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (NextGenAI - Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

T. Silva-Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112:3211—3260, 2023. URL https://doi.org/10.1007/s10994-023-06336-7.

J. Gawlikowski, C. Tassi, and M. Ali et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023. URL https://doi.org/10.1007/s10462-023-10562-9.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the Internation Conference on Machine Learning (ICML)*, 2017.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *13th European Conference on Machine Learning (ECML)*, pages 345–356, 2002.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In M. Balcan and K. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, 20–22 Jun 2016. URL https://proceedings.mlr.press/v48/martins16.html.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1504–1519, 2019.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Leaning Research*, 21:1–69, 2020.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 07 1988. doi: 10.1007/BF01016429.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Ulf Johansson, Henrik Linusson, Tuve Löfström, and Henrik Boström. Model-agnostic nonconformity functions for conformal classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2072–2079, 2017. doi: 10.1109/IJCNN.2017.7966105.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Classification with reject option using conformal prediction. In Dinh Phung, V. Tseng, G. Webb, B. Ho, M. Ganji, and L. Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 94–105, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, pages 4171–4186, 2019.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive cov-

erage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, 2012. URL https://proceedings.mlr.press/v25/vovk12.html.

Marharyta Aleksandrova and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 151–170, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/aleksandrova21a.html.

Rikard Laxhammar and Göran Falkman. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, page 47–55, 2010.

A Proof of Proposition 2

Let π be the permutation over $[K]$ such that $z_{\pi(1)} \geq z_{\pi(2)} \geq ... \geq z_{\pi(K)}$ are the entries of $\boldsymbol{z}$ sorted in ascending order. We have $\gamma\text{-entmax}_j(\beta\boldsymbol{z}) = [(\gamma-1)\beta z_j - \tau]_+^{\frac{1}{\gamma-1}}$, where τ satisfies

$$\sum_{k=1}^{|S(\beta\boldsymbol{z})|} [(\gamma-1)\beta z_{\pi(k)} - \tau]^{\frac{1}{\gamma-1}} = 1. \tag{14}$$

We prove the desired equivalence by showing implication in both directions, as follows.

- $\boxed{\pi(j) \in S(\beta\boldsymbol{z}) \implies \sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)})\right]^{\frac{1}{\gamma-1}} < 1}$

  To prove this direction, note that, if $\pi(j) \in S(\beta\boldsymbol{z})$, we must have by (7) that $(\gamma-1)\beta z_{\pi(j)} > \tau$. Therefore, we have that, for all $k \in [K]$, $(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)}) < (\gamma-1)\beta z_{\pi(k)} - \tau$. As a consequence,

$$\begin{aligned}\sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)})\right]^{\frac{1}{\gamma-1}} &\leq \sum_{k=1}^{|S(\beta\boldsymbol{z})|} \left[(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)})\right]^{\frac{1}{\gamma-1}} \\ &< \sum_{k=1}^{|S(\beta\boldsymbol{z})|} \left[(\gamma-1)\beta z_{\pi(k)} - \tau\right]^{\frac{1}{\gamma-1}} = 1,\end{aligned} \tag{15}$$

  where the last equality comes from (14).

- $\boxed{\pi(j) \in S(\beta\boldsymbol{z}) \impliedby \sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_{\pi(j)} - z_{\pi(k)})\right]^{\frac{1}{\gamma-1}} < 1}$

  We show the reverse implication by showing that $\pi(j) \notin S(\beta\boldsymbol{z}) \implies \sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_{\pi(j)} - z_{\pi(k)})\right]^{\frac{1}{\gamma-1}} \geq 1$. If $\pi(j) \notin S(\beta\boldsymbol{z})$, we must have by (7) that $(\gamma-1)\beta z_{\pi(j)} \leq \tau$. Therefore, we have that, for all $k \in [K]$, $(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)}) \geq (\gamma-1)\beta z_{\pi(k)} - \tau$. As a consequence,

$$\begin{aligned}\sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)})\right]^{\frac{1}{\gamma-1}} &\geq \sum_{k=1}^{|S(\beta\boldsymbol{z})|} \left[(\gamma-1)\beta(z_{\pi(k)} - z_{\pi(j)})\right]^{\frac{1}{\gamma-1}} \\ &\geq \sum_{k=1}^{|S(\beta\boldsymbol{z})|} \left[(\gamma-1)\beta z_{\pi(k)} - \tau\right]^{\frac{1}{\gamma-1}} = 1.\end{aligned} \tag{16}$$

B Additional Experimental Details

B.1 RAPS

The RAPS procedure was implemented according to the original protocol introduced by Angelopoulos et al. (2021) by adapting the code provided by those authors. The method uses two hyperparameters to regularize the original *adaptive prediction sets* procedure (Romano et al., 2020): λ_{reg} — a regularization penalty added to discourage inclusion in the prediction set, and k_{reg} — the order from which classes receive the penalty term. We split the original calibration data into two sets: calibration data (60%) and hyperparameter tuning data (40%). For each split of the dataset, we find the pair of hyperparameters that minimizes the average prediction set size on hyperparameter tuning set, with $\lambda_{\text{reg}} \in \{0.001, 0.01, 0.1, 1\}$ and $k_{\text{reg}} \in \{1, 5, 10, 50\}$.

B.2 opt-entmax

The opt-entmax procedure splits the calibration data into two sets: calibration data (60%) to perform conformal prediction with γ-entmax for varying values of $\gamma \in \{1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9\}$; and a hyperparameter tuning set (40%) where the average set size is measured and from which the optimal γ is chosen.

Table 3: Accuracy of each trained model: calibration and test set sizes, average and standard deviation of accuracy over the 5 different splits.

	Test			Calibration		
	Size	Accuracy		Size	Accuracy	
		avg	std		avg	std
CIFAR10	6000	0.838	0.005	4000	0.841	0.003
CIFAR100	6000	0.860	0.003	4000	0.858	0.002
ImageNet	30000	0.805	0.001	20000	0.805	0.001
NewsGroups	2261	0.744	0.006	1508	0.742	0.004

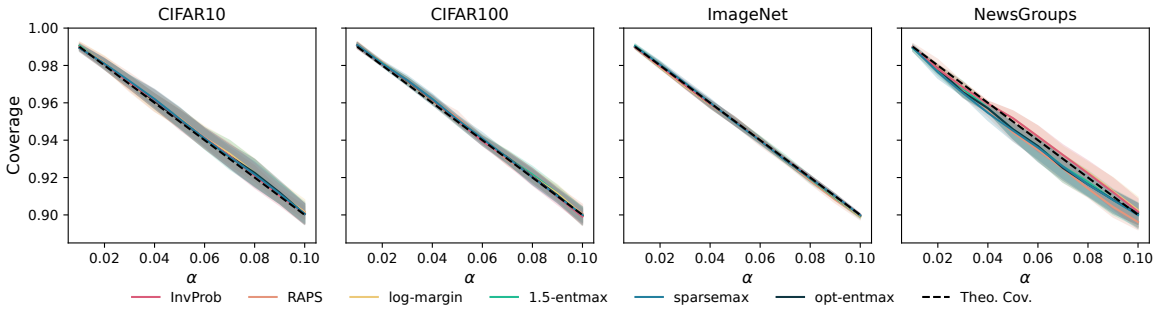

Figure 7: Coverage on the test set as a function of α, over the 5 splits of calibration and test data.

C Experimental Results

C.1 Model Accuracy

All models were evaluated on each of the 5 calibration-test set splits. The average and standard deviation of model accuracy over splits can be found in Table 3, along with the calibration and set sizes for each dataset.

C.2 Coverage

The coverage results of all methods over the 5 splits of calibration is shown in Figure 7.

C.3 Analysis of opt-entmax

Figure 8 shows how varying the value of γ for the non-conformity score affects the average prediction set size for a fixed value of α. It is clear that the optimal value of γ depends on both the task and the confidence level.

C.4 Adaptiveness

In Figure 9 we can see the size-stratified coverage violation (SSCV) — introduced by Angelopoulos et al. (2021), it measures the maximum deviation from the desired coverage $1-\alpha$. Partitioning the possible size cardinalities into G bins, $B_1, ..., B_G$, let $\mathcal{I}_g$ be the set of observations falling in bin g, with $g = 1, ..., G$, the SSCV of a predictor C_α, for that bin partition is given by:

$$\text{SSCV}(C, \{B_1, ..., B_G\}) = \sup_g \left| \frac{\{i : Y_i \in \mathcal{C}_\alpha(X_i), i \in \mathcal{I}_g\}}{|\mathcal{I}_j|} - (1-\alpha) \right| \tag{17}$$

Size-stratified coverage for the ImageNet dataset ($\alpha = 0.05$) can be found in Table 4. An equivalent analysis for datasets CIFAR100 and 20 NewsGroups can be found in Tables 5 and 6, respectively, for different values of α.

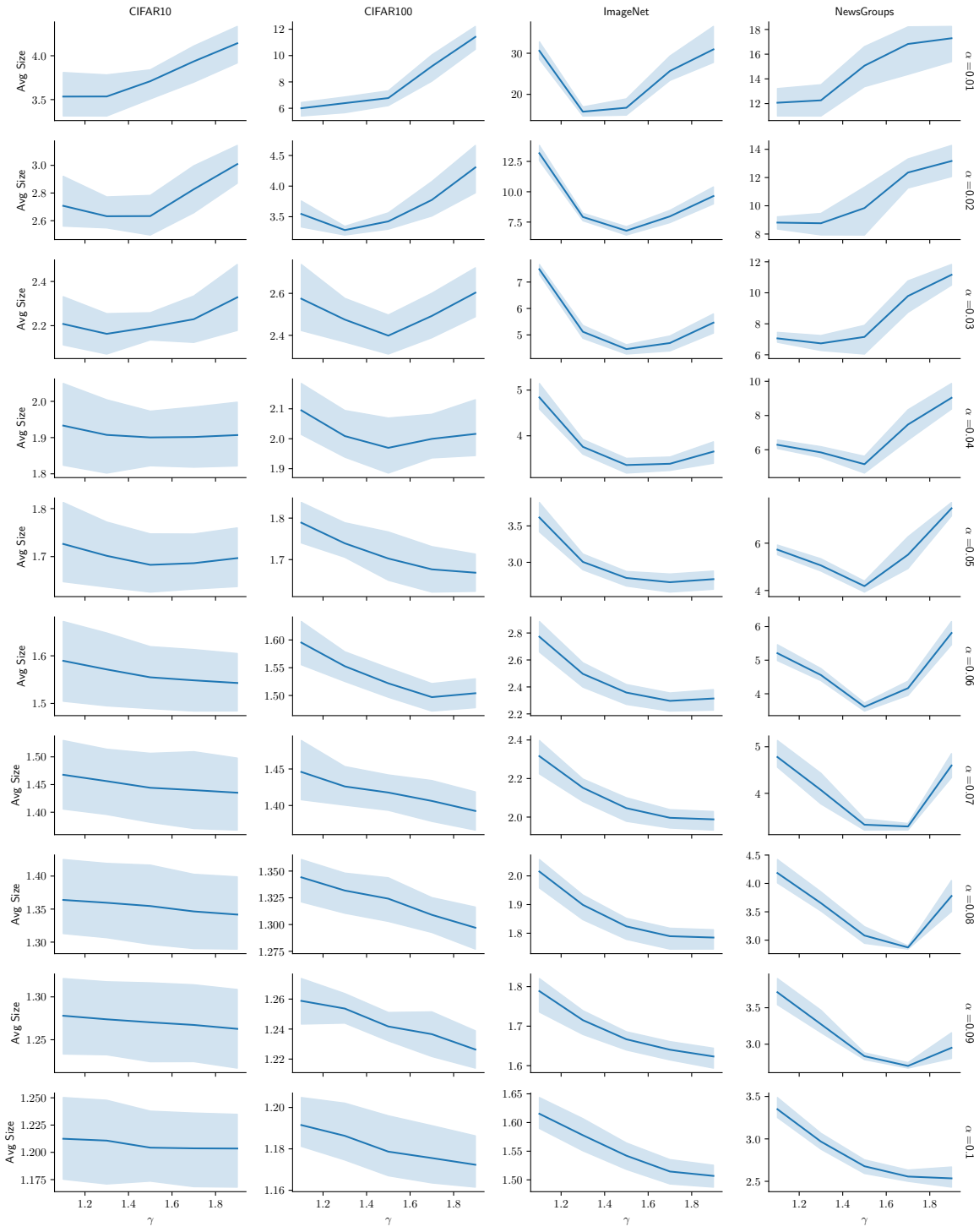

Figure 8: Average set size varying with γ parameter, on separate calibration set for different datasets and values of α.

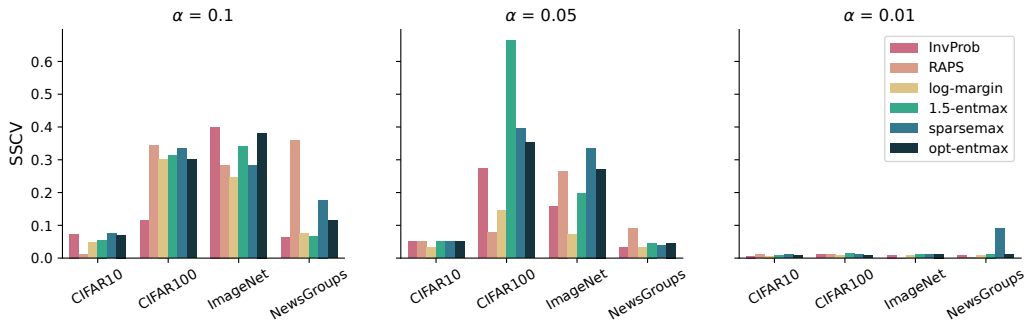

Figure 9: Size-stratified coverage violation (SSCV) for all methods and datasets, for $\alpha \in \{0.01, 0.05, 0.1\}$

Table 4: Size-stratified coverage for the ImageNet dataset with $\alpha = 0.05$.

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	13745	0.975	12321	0.979	16286	0.966	13730	0.975	7260	0.992	11805	0.982
2 - 3	9579	0.958	4774	0.977	8144	0.945	10132	0.956	15693	0.968	11788	0.959
4 - 6	4189	0.903	8654	0.959	2653	0.912	3671	0.905	5681	0.885	4397	0.898
7 - 10	2046	0.835	3494	0.865	1244	0.916	1542	0.858	1226	0.742	1510	0.819
11 - 1000	440	0.793	570	0.684	1673	0.877	925	0.751	140	0.614	500	0.678

Table 5: Size-stratified coverage for the CIFAR100 dataset with $\alpha = 0.01$ (top), $\alpha = 0.05$ (middle) and $\alpha = 0.1$ (bottom).

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	2222	0.999	1577	0.999	2422	0.998	795	0.999	0	NaN	2397	0.998
2 - 3	1476	0.995	1015	0.998	1437	0.993	1464	0.999	12	1.000	1457	0.993
4 - 6	810	0.998	567	1.000	750	0.996	1416	1.000	190	1.000	762	0.997
7 - 10	457	0.991	898	0.998	414	0.990	977	0.993	1410	0.999	424	0.988
11 - 100	1035	0.980	1943	0.984	977	0.986	1348	0.976	4388	0.991	960	0.984

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	4089	0.976	3587	0.986	4312	0.970	4065	0.977	3634	0.987	3934	0.980
2 - 3	1357	0.917	1055	0.967	1139	0.913	1391	0.918	1954	0.916	1569	0.917
4 - 6	453	0.841	1316	0.872	346	0.867	423	0.830	393	0.756	432	0.806
7 - 10	99	0.677	0	–	121	0.826	114	0.728	18	0.556	62	0.597
11 - 100	2	1.000	0	–	82	0.805	7	0.286	1	1.000	3	0.667

size	InvProb		RAPS		log-margin		1.5-entmax		sparsemax		opt-entmax	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	5194	0.916	4456	0.952	5225	0.920	5182	0.924	5170	0.925	5170	0.925
2 - 3	786	0.785	1240	0.819	683	0.779	752	0.769	807	0.742	804	0.749
4 - 6	1	1.000	153	0.556	82	0.598	63	0.587	23	0.565	25	0.600
7 - 10	0	–	0	–	9	0.667	3	0.667	0	–	1	0.000
11 - 100	0	–	0	–	1	1.000	0	–	0	–	0	–

Table 6: Size-stratified coverage for the 20 Newsgroups dataset with $\alpha = 0.01$ (top), $\alpha = 0.05$ (middle) and $\alpha = 0.1$ (bottom).

size	**InvProb**		**RAPS**		**log-margin**		**1.5-entmax**		**sparsemax**		**opt-entmax**	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	0	–	0	–	0	–	0	–	0	–	0	–
2 - 3	77	0.987	0	–	78	0.987	0	–	0	–	13	1.000
4 - 6	338	0.982	0	–	345	0.983	0	–	0	–	279	0.978
7 - 10	718	0.986	0	–	720	0.986	45	0.978	10	0.900	704	0.989
11 - 20	1128	0.994	2261	0.988	1118	0.994	2216	0.990	2251	0.988	1265	0.992

size	**InvProb**		**RAPS**		**log-margin**		**1.5-entmax**		**sparsemax**		**opt-entmax**	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	1045	0.957	40	0.900	1047	0.956	745	0.969	0	NaN	745	0.969
2 - 3	413	0.935	451	0.973	411	0.937	660	0.947	125	0.944	660	0.947
4 - 6	166	0.946	1441	0.948	167	0.946	393	0.954	588	0.912	393	0.954
7 - 10	97	0.948	213	0.859	95	0.947	229	0.904	1126	0.952	229	0.904
11 - 20	540	0.983	116	0.957	541	0.983	234	0.923	422	0.981	234	0.923

size	**InvProb**		**RAPS**		**log-margin**		**1.5-entmax**		**sparsemax**		**opt-entmax**	
	n	cov	n	cov	n	cov	n	cov	n	cov	n	cov
0 - 1	1481	0.918	1095	0.944	1481	0.918	1356	0.933	575	0.967	1226	0.939
2 - 3	320	0.856	311	0.945	312	0.853	471	0.885	1336	0.912	642	0.894
4 - 6	121	0.868	794	0.851	113	0.867	208	0.837	263	0.722	234	0.786
7 - 10	92	0.946	61	0.541	77	0.961	102	0.833	27	0.778	92	0.793
11 - 20	247	0.964	0	–	278	0.975	124	0.944	60	0.883	67	0.985