Sparse Activations as Conformal Predictors

Margarida Campos[1,2], João Calém[1], Sophia Sklaviadis[1,2], Mário A. T. Figueiredo[1,2,3], André F. T. Martins[1,2,3,4]

[1]Instituto Superior Técnico, University of Lisbon, [2]Instituto de Telecomunicações, [3]ELLIS Unit Lisbon, [4] Unbabel **Portugal**

Conformal prediction (CP)

- **Uncertainty quantification** framework: strong coverage **guarantees**, model-agnostic, and distribution-free.
- Given a **non-conformity score**, CP constructs **prediction sets** for test instances using calibration data.

Vovk et al., 2005; Angelopoulos and Bates, 2023

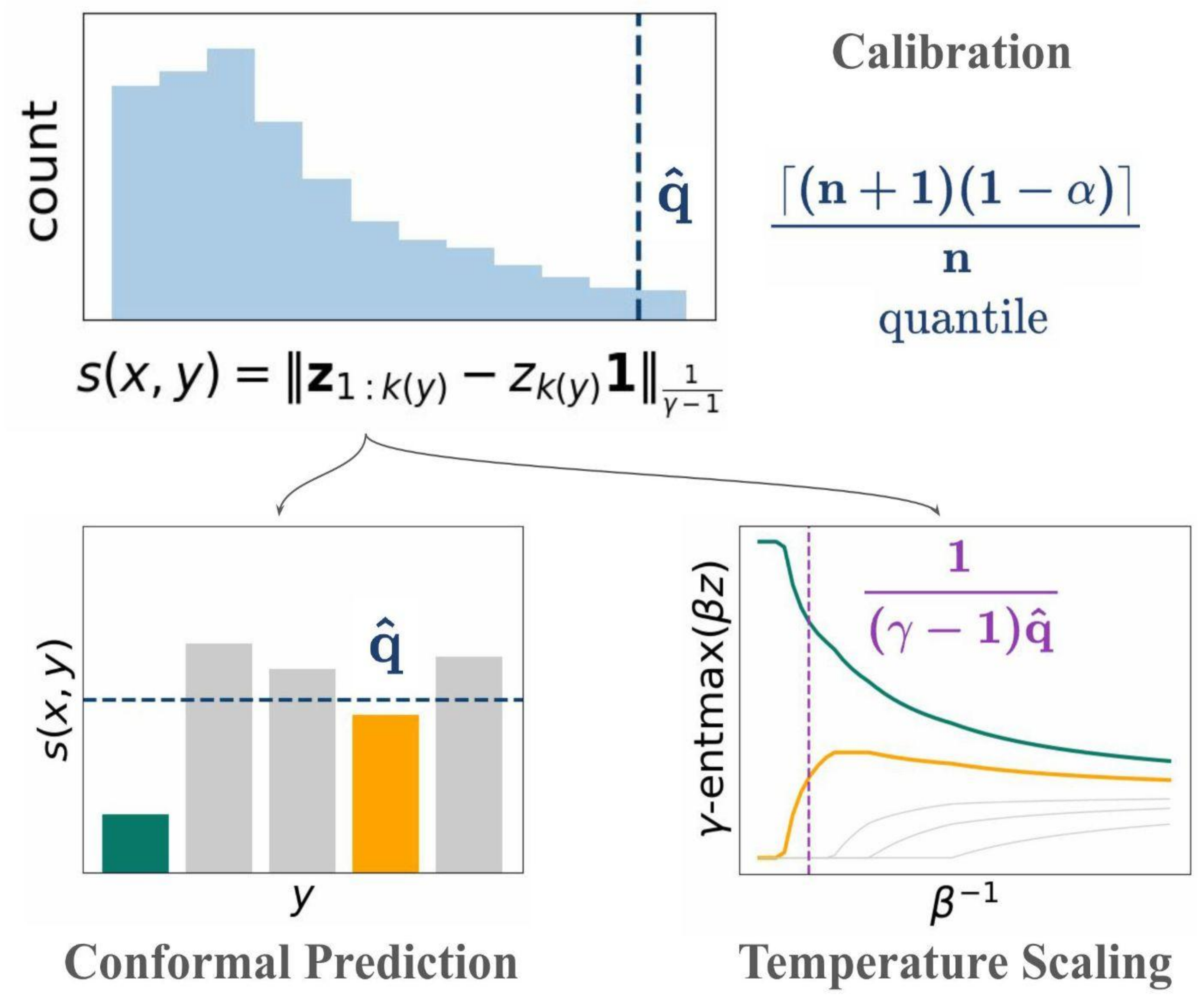
Sparse activation functions

- Fenchel-Young losses: framework for sparse activations interpretable as set prediction, such as:
- **γ-entmax** family: softmax (γ=1) and sparsemax (γ=2); sparse for γ > 1; temperature-controlled sparsity

Martins and Astudillo, 2016; Peters et al., 2019; Blondel et al., 2020

new **conformal predictor** ⇔ **temperature scaling** of sparsemax

generalization to the **γ-entmax** family



expanded family of non-conformity scores

empirical validation on different tasks

💻 **See it live!**

tinyurl.com/sparseconformal

Conformalizing entmax

Let $z_1 > ... > z_K$, support is $j \in S(\beta \mathbf{z}; \gamma) \iff \sum_{k=1}^{j-1} \left[(\gamma-1)\beta(z_k - z_j)\right]^{\frac{1}{\gamma-1}} < 1$

Building a conformal predictor, $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$, with score:

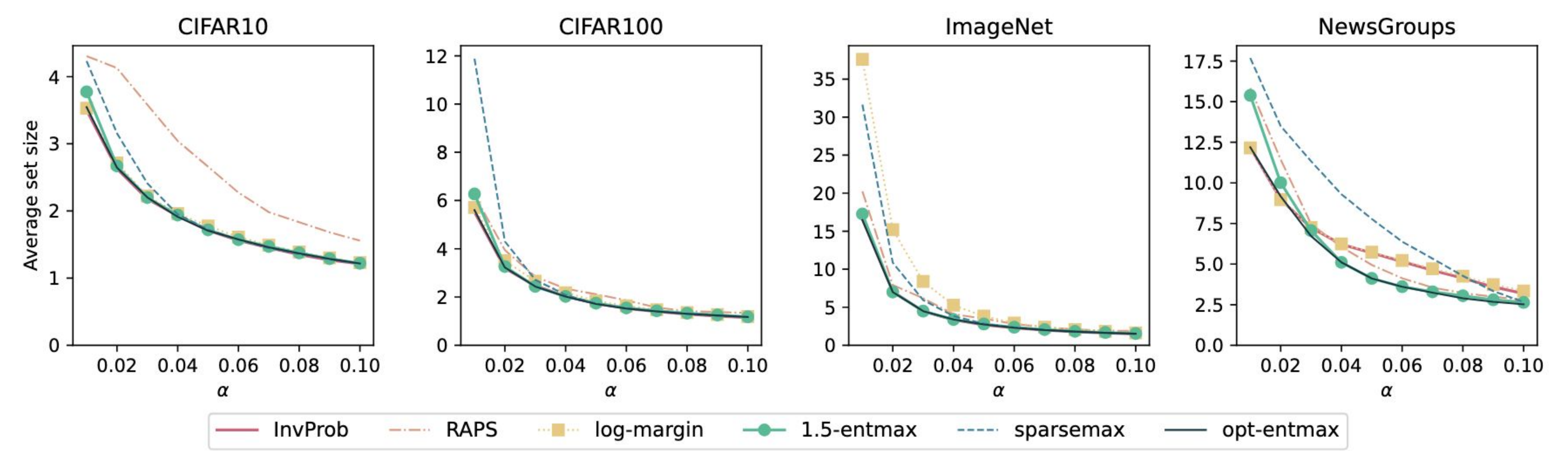
$s(x,y) = \|\mathbf{z}_{1:k(y)} - z_{k(y)}\mathbf{1}\|_\delta$, yields: $C_\alpha(x) = S(\beta \mathbf{z}; \gamma)$

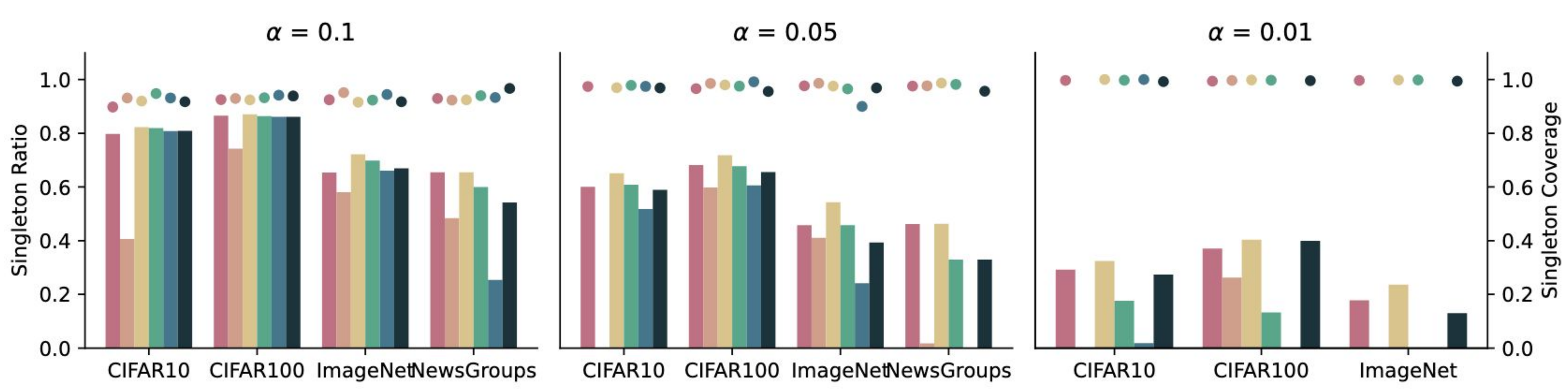
Interesting findings

- γ can be tuned, yielding method `opt-entmax`
- limit case (softmax)→ log-odds ratio (`log-margin`):

$$s(x,y) = \|\mathbf{z}_{1:k(y)} - z_{k(y)}\mathbf{1}\|_\infty = z_1 - z_{k(y)} = \log \frac{p_1}{p_{k(y)}}$$

Results





- choice of score is crucial and **task-dependent**
- new methods (**log-margin** and **opt-entmax**):
  - **efficiency** ✔
  - **adaptiveness** ✔
  - **Interpretability** ✔
  - **natural calibration** assessment

Acknowledgments: EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), DECOLLAGE (ERC-2022-CoG 101088763), Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), and Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

TÉCNICO LISBOA · instituto de telecomunicações · UTTER · Unbabel · ellis European Laboratory for Learning and Intelligent Systems · UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

Full Paper