Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis

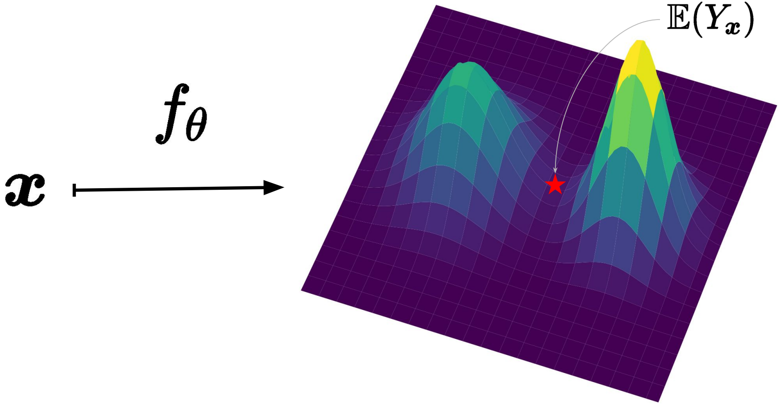
NeurIPS 2023

Victor Letzelter
Co-authors: Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, Gaël Richard

The problem ?

☞ Several tasks are ill-posed and ambiguous by nature.

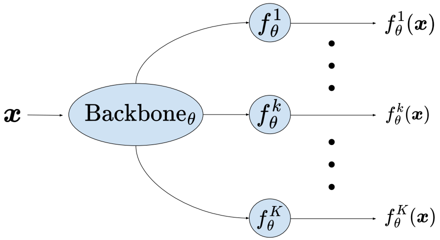
☞ If $p(\boldsymbol{y} \mid \boldsymbol{x})$ is multimodal, the conditional mean $\mathbb{E}(Y_{\boldsymbol{x}})$, where $Y_{\boldsymbol{x}} \sim p(\boldsymbol{y} \mid \boldsymbol{x})$, may be not informative enough.



Multiple choice learning

☞ Consider several *hypotheses* [Guzman-Rivera et al., 2012]

$$f_\theta \triangleq \left(f_\theta^1, \ldots, f_\theta^K\right) \in \mathcal{F}\left(\mathcal{X}, \mathcal{Y}^K\right).$$



☞ Winner-Takes-All (WTA) loss for a set of hypotheses (sMCL, [Lee et al., 2016])

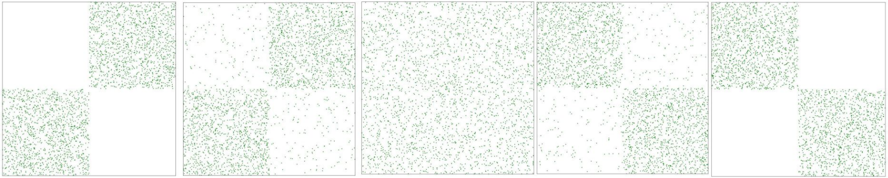
$$\mathcal{L}\left(f_\theta\left(\boldsymbol{x}_s\right), \boldsymbol{y}_s\right) \triangleq \min_{k \in [\![1, K]\!]} \ell\left(f_\theta^k\left(\boldsymbol{x}_s\right), \boldsymbol{y}_s\right).$$

☞ If a *set* of targets $\boldsymbol{Y}_s$ is available for each $\boldsymbol{x}_s$: same for each $\boldsymbol{y} \in \boldsymbol{Y}_s$ [Firman et al., 2018].

How does it work in practice ?

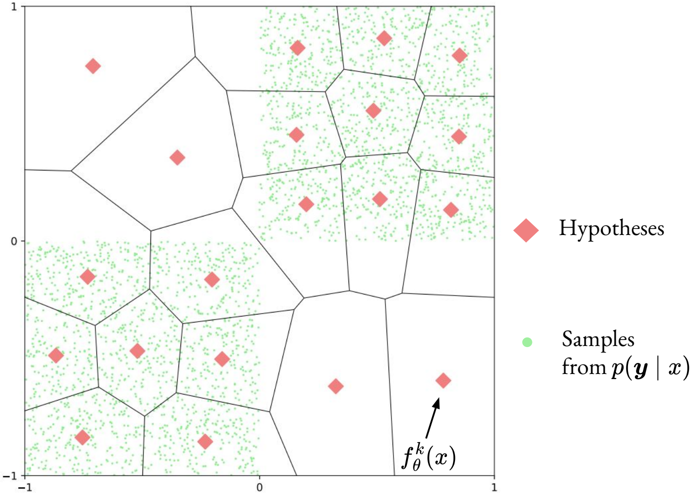
Let $\mathcal{X} = [0, 1]$, and $\mathcal{Y} = [-1, 1]^2$.

- 2D dist. to predict from input scalar $x \in \mathcal{X}$ [Rupprecht et al., 2017].
- Input-output pairs available $\{(x_N, \boldsymbol{y}_N)\}$ where $\boldsymbol{y}_N \sim p(\boldsymbol{y} \mid x_N)$.
- Below: ground-truth dist. (green points) for several inputs.



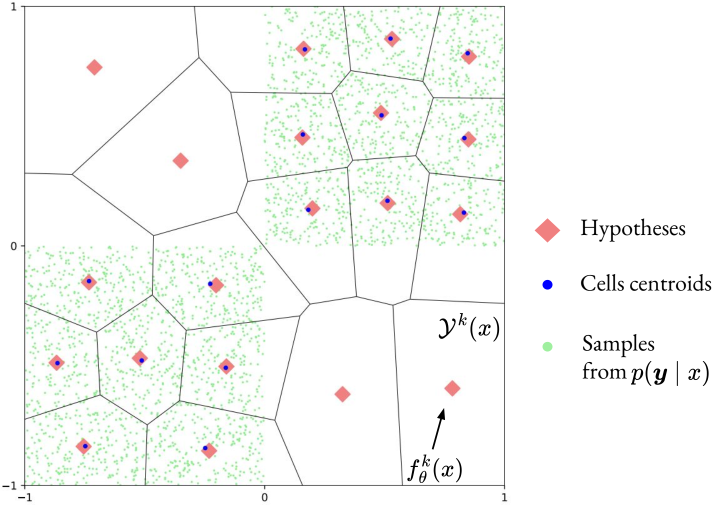
$x = 0$ $x = 0.1$ $x = 0.5$ $x = 0.9$ $x = 1$

Properties



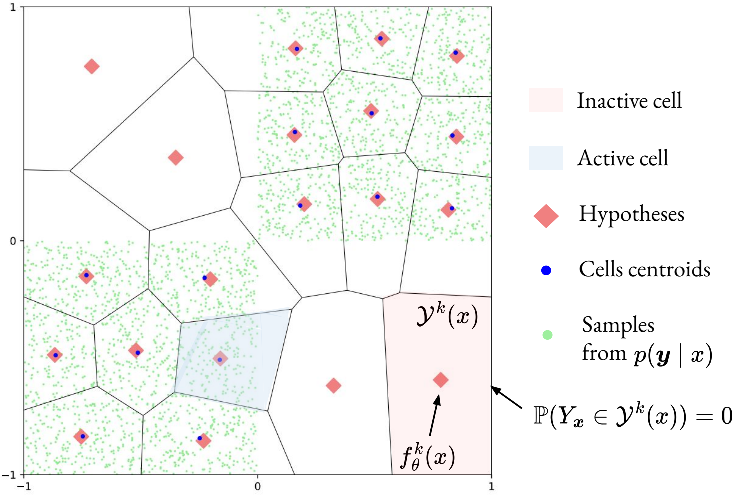
Zoomed prediction of sMCL at $x = 0$

Properties



Centroidal property: $f_\theta^k(\boldsymbol{x}) = \mathbb{E}\left[Y_{\boldsymbol{x}} \mid Y_{\boldsymbol{x}} \in \mathcal{Y}^k(\boldsymbol{x})\right]$

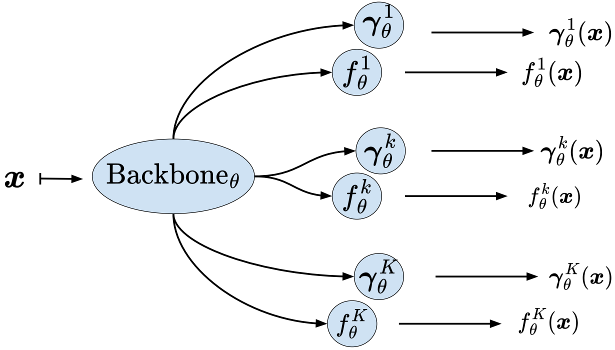
Properties



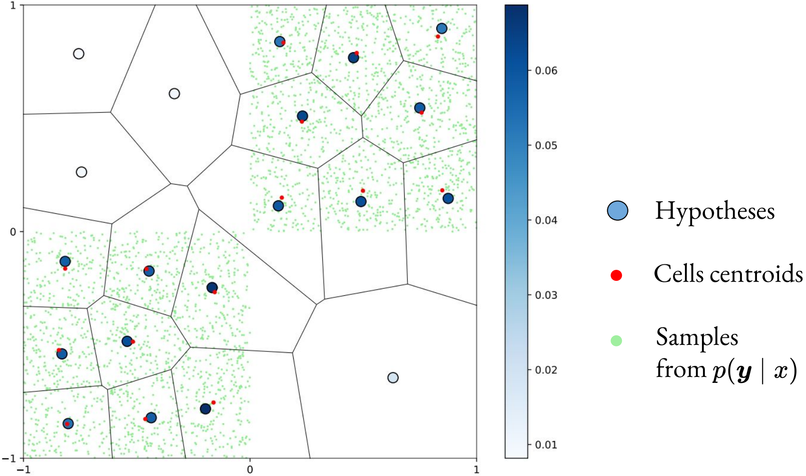
In inactive cells the predictions $f_\theta^k(\boldsymbol{x})$ are meaningless (*overconfidence*).

Proposed solution

- Optimization criterion adapted for overconfidence solving.
- Proposition: *hypothesis scoring* heads $\gamma_\theta^1, \ldots, \gamma_\theta^K \in \mathcal{F}(\mathcal{X}, [0,1])$, to predict $\mathbb{P}(Y_{\boldsymbol{x}} \in \mathcal{Y}^k(\boldsymbol{x}))$ ([Tian et al., 2019] adapted for regression).



rMCL output



Overconfidence solving in rMCL (with scores displayed in the colorbar).

Probabilistic interpretation proposed at inference

If $Y_{\boldsymbol{x}} \sim p(\boldsymbol{y} \mid \boldsymbol{x})$, interpret the heads **inference** time predictions as

$$\boldsymbol{\gamma}_\theta^k(\boldsymbol{x}) = \mathbb{P}\left(Y_{\boldsymbol{x}} \in \mathcal{Y}^k(\boldsymbol{x})\right), \tag{1}$$

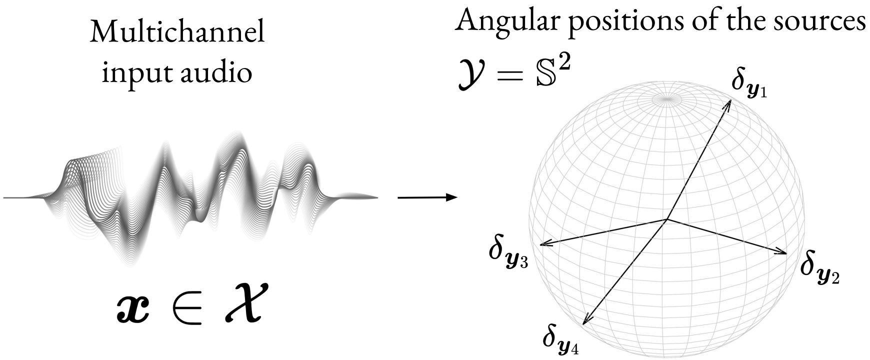
and for $k \in [\![1, K]\!]$ such that $\boldsymbol{\gamma}_\theta^k(\boldsymbol{x}) > 0$

$$f_\theta^k(\boldsymbol{x}) = \mathbb{E}\left[Y_{\boldsymbol{x}} \mid Y_{\boldsymbol{x}} \in \mathcal{Y}^k(\boldsymbol{x})\right]. \tag{2}$$

Example of probabilistic interpretation (justified in the paper)

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{\gamma}_\theta^k(\boldsymbol{x}) \delta_{f_\theta^k(\boldsymbol{x})}(\boldsymbol{y}). \tag{3}$$

Audio application: Sound source localization



Sound source localization (SSL).

Audio application: Sound source localization

- With rMCL: No permutation / imbalance spatial data (smart grid).
- No need to know the number of sources in advance.
- Probabilistic output interpretation.

Target dist. $p(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \sum_{\boldsymbol{y}_t \in \boldsymbol{Y}_t} \delta_{\boldsymbol{y}_t}(\boldsymbol{y})$.

Predicted dist. (rMCL) $\hat{p}(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \sum_{k=1}^{K} \boldsymbol{\gamma}_\theta^k(\boldsymbol{x}) \delta_{f_\theta^k(\boldsymbol{x})}(\boldsymbol{y})$

Experimental setup

Datasets. Several datasets (anechoic, reverberant conditions) [Adavanne et al., 2018].
Metrics. *'Oracle'* (↓): Quality of the *best* hypotheses.
Earth Mover's Distance (↓) between $\hat{p}(\boldsymbol{y} \mid \boldsymbol{x})$ and $p(\boldsymbol{y} \mid \boldsymbol{x})$
Neural network backbone CRNN [Adavanne et al., 2018].
Baselines IE, WTA variants, PIT variant [Lee et al., 2016, Rupprecht et al., 2017, Adavanne et al., 2018, Yu et al., 2017, Schymura et al., 2021, Makansi et al., 2019].

Experiments

- Comparisons in unimodal and multimodal conditions.
- rMCL: solves overconfidence issue of sMCL (vanilla WTA). Competitive, esp. in multimodal setting.
- rMCL: orthogonal to WTA variants (e.g., top-n-WTA, ε-WTA).
- Sensitivity analysis performed: metrics trade-off when K increases.

Thank You!

Poster#1220
Arxiv: `arxiv.org/abs/2311.01052`
Code: `github.com/Victorletzelter/code-rMCL`

Bibliography I

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2018).
Sound event localization and detection of overlapping sources using convolutional recurrent neural networks.
IEEE Journal of Selected Topics in Signal Processing, 13(1):34–48.

Firman, M., Campbell, N. D., Agapito, L., and Brostow, G. J. (2018).
Diversenet: When one right answer is not enough.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5598–5607.

Guzman-Rivera, A., Batra, D., and Kohli, P. (2012).
Multiple choice learning: Learning to produce multiple structured outputs.
Advances in neural information processing systems, 25.

Bibliography II

Lee, S., Purushwalkam Shiva Prakash, S., Cogswell, M., Ranjan, V., Crandall, D., and Batra, D. (2016).
Stochastic multiple choice learning for training diverse deep ensembles.
Advances in Neural Information Processing Systems, 29.

Makansi, O., Ilg, E., Cicek, O., and Brox, T. (2019).
Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153.

Bibliography III

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. (2017).
Learning in an uncertain world: Representing ambiguity through multiple hypotheses.
In *Proceedings of the IEEE international conference on computer vision*, pages 3591–3600.

Schymura, C., Ochiai, T., Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., and Kolossa, D. (2021).
Exploiting attention-based sequence-to-sequence architectures for sound event localization.
In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 231–235. IEEE.

Bibliography IV

Tian, K., Xu, Y., Zhou, S., and Guan, J. (2019).
Versatile multiple choice learning and its application to vision computing.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6349–6357.

Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017).
Permutation invariant training of deep models for speaker-independent multi-talker speech separation.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.