Learning under ambiguity through multiple hypotheses and quantization

Victor Letzelter, PhD Student at **Valeo** and **Telecom Paris.**

Machine Learning and Modeling Seminar, May 2025, Prague

About

Victor Letzelter

PhD Student at **Valeo.ai** and **Institut Polytechnique de Paris**, since March 2023

Working on multi-hypotheses models

Under the supervision of Andrei Bursuc, Slim Essid, Mathieu Fontaine, Gaël Richard, and Patrick Pérez

Presentation

Papers:

V. Letzelter, M. Fontaine, M. Chen, P. Pérez, S. Essid, and G. Richard, Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis. *NeurIPS 2023*

V. Letzelter*, D. Perera*, C. Rommel, M. Fontaine, S. Essid, G. Richard, and P. Pérez. Winner-takes-all learners are geometry-aware conditional density estimators. In *ICML 2024*

C. Rommel, **V. Letzelter**, N. Samet, R. Marlet, M. Cord, P. Pérez, and E. Valle. "ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation". In *NeurIPS* 2024.

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Y. Xu*, **V. Letzelter***, M. Chen, E. Zablocki, and M. Cord. "Annealed Winner-Takes-All for Motion Forecasting." *ICRA 2025.*

A. Cortés*, R. Rehm, **V. Letzelter***. "Winner-takes-all for Multivariate Probabilistic Time Series Forecasting". To appear at *ICML 2025.*

(Equal Contribution)*

Presentation

Papers:

V. Letzelter, M. Fontaine, M. Chen, P. Pérez, S. Essid, and G. Richard, Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis. *NeurIPS 2023*

V. Letzelter*, D. Perera*, C. Rommel, M. Fontaine, S. Essid, G. Richard, and P. Pérez. Winner-takes-all learners are geometry-aware conditional density estimators. In *ICML 2024*

C. Rommel, **V. Letzelter**, N. Samet, R. Marlet, M. Cord, P. Pérez, and E. Valle. “ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation”. In *NeurIPS* 2024.

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. “Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing”. In *NeurIPS. 2024.*

Y. Xu*, **V. Letzelter***, M. Chen, E. Zablocki, and M. Cord. "Annealed Winner-Takes-All for Motion Forecasting." *ICRA 2025.*

A. Cortés*, R. Rehm, V. Letzelter*. “Winner-takes-all for Multivariate Probabilistic Time Series Forecasting”. To appear at ICML 2025.

(Equal Contribution)*

Learning under ambiguity through multiple hypotheses and quantization

What is an ambiguous task ?

Ambiguity ?

Several ML tasks can ambiguous by nature.

For an input x, there may be multiple plausible outputs y.

Ambiguous tasks ?

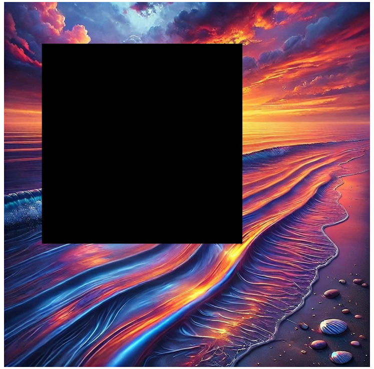
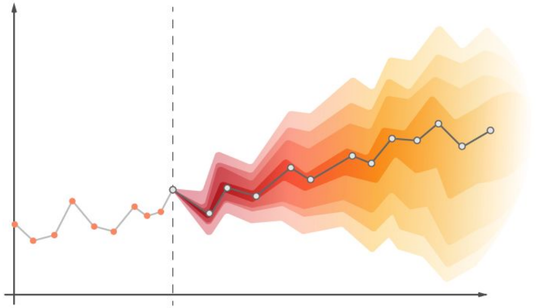
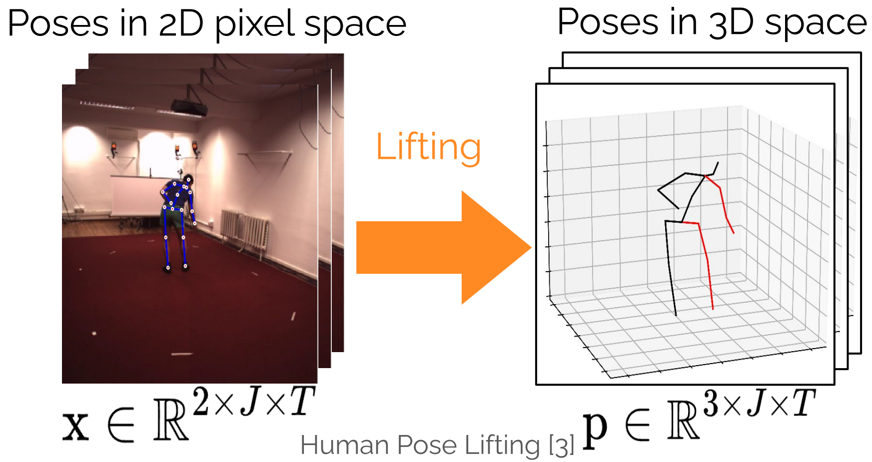


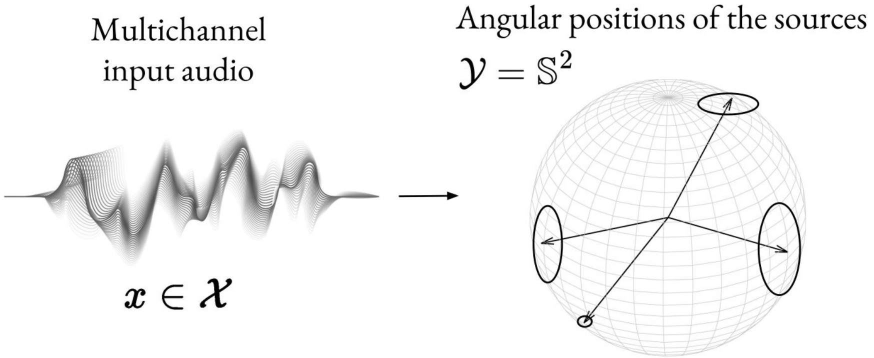
Image inpainting [1]



Time-series forecasting [2]



Human Pose Lifting [3]



Sound source localization [4]

Images sources:

[1] https://neurips.cc/media/PosterPDFs/NeurIPS%202024/94844.png?t=1732973175.403999

[2] https://medium.com/analytics-vidhya/time-series-forecasting-c73dec0b7533

[3] https://neurips.cc/virtual/2024/poster/93050

[4] https://neurips.cc/media/neurips-2023/Slides/70945.pdf

Ambiguity ?

Notations:

- $\mathcal{X}$, $\mathcal{Y}$: Input and Output spaces.
- Joint data distribution: $x_i, y_i \sim \rho(x, y)$ on $\mathcal{X} \times \mathcal{Y}$
- Conditional data distribution for each input: $\rho_x(y) = \rho(y \mid x)$

Idea: modeling ρ_x using a small set of diverse predictions.

Traditional Risk Minimization

Predictor to train: $f_\theta : \mathcal{X} \to \mathcal{Y}$

Loss function: $\ell : (\hat{y}, y) \in \mathcal{Y} \times \mathcal{Y} \to \ell(\hat{y}, y) \in \mathbb{R}_+$

Empirical Risk ~ True Risk:

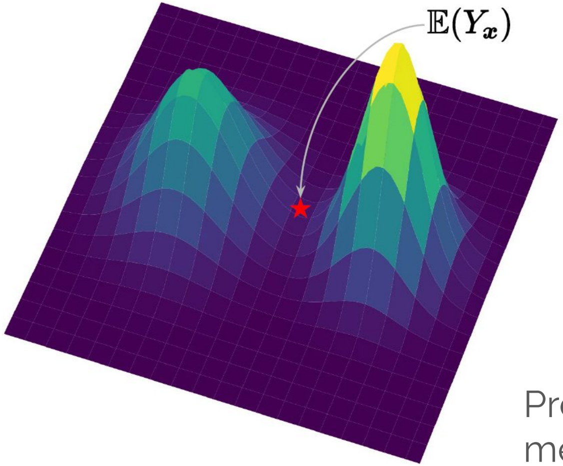
$$\frac{1}{N} \sum_{(x,y) \in \mathcal{D}} \ell(f_\theta(x), y) \underset{N \to \infty}{\simeq} R(f_\theta) \triangleq \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(f_\theta(x), y) \mathrm{d}\rho(x, y).$$

Traditional Risk Minimization

When $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$,

$$f_\theta : x \mapsto \mathbb{E}(Y \mid x) \triangleq \int_{y \in \mathcal{Y}} y \rho(y \mid x) \mathrm{d}y \in \underset{\theta}{\operatorname{argmin}} \, R(f_\theta).$$

$$x \mapsto$$



Predicting a conditional mean is not enough

Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf

Learning under ambiguity through multiple hypotheses and quantization

How are we using multiple hypotheses ? and how does it relates to quantization ?

Winner-takes-all training

- Instead of a training single predict f_θ , we are learning several ones:

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

Winner-takes-all training

- Instead of a training single predictor f_θ, we are learning several ones:

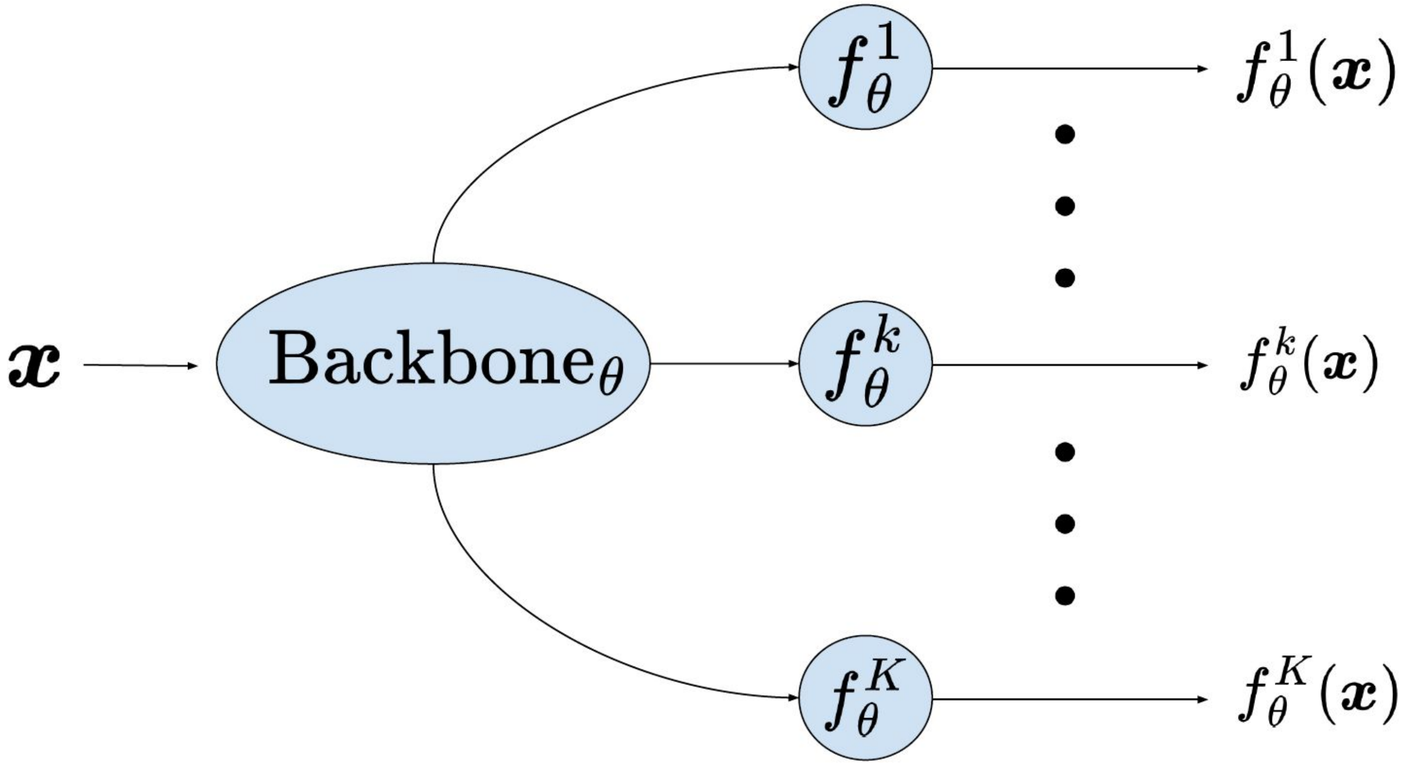
$$(f_\theta^1, \ldots, f_\theta^K) \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^K)$$

Trained as follows for each pair $(x, y) \sim \rho(x, y)$ (computed batch-wise)

1. Compute $f_\theta^1(x), \ldots, f_\theta^K(x)$
2. Select the *Winner* model $k^\star \in \underset{k \in [1,K]}{\operatorname{argmin}} \, \ell\left(f_\theta^k(x), y\right).$
3. Compute the loss of the Winner: $\mathcal{L}^{\mathrm{WTA}}(\theta) = \ell(f_\theta^{k^\star}(x), y)$

Guzman-Rivera, Abner, Dhruv Batra, and Pushmeet Kohli. "Multiple choice learning: Learning to produce multiple structured outputs." In *NeurIPS 2012*
Lee, Stefan, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. "Stochastic multiple choice learning for training diverse deep ensembles." In *NeurIPS 2016*

Practical implementation with multiple heads



Lee, Stefan, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. "Why m heads are better than one: Training a diverse ensemble of deep networks." *arXiv* (2015).

Winner-takes-all training

Training dynamics ?

$$\mathcal{L}^{\mathrm{WTA}}(\theta) = \int_{\mathcal{X}\times\mathcal{Y}} \min_{k=1,\ldots,K} \ell\left(f_\theta^k(x), y\right)\rho(x,y)\mathrm{d}x\,\mathrm{d}y$$

$$= \int_{\mathcal{X}} \underbrace{\left(\int_{\mathcal{Y}} \min_{k=1,\ldots,K} \ell\left(f_\theta^k(x), y\right)\rho(y \mid x)\mathrm{d}x\right)}_{\textbf{Conditional Distortion}} \rho(x)\,\mathrm{d}y$$

-> WTA Optimization performs **quantization** of the conditional distribution $\rho(y \mid x)$ for each *x*.

Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML*, 2024.

Winner-takes-all training

$$\mathcal{L}^{\mathrm{WTA}}(\theta) = \int_{\mathcal{X}} \sum_{k=1}^{K} \int_{\mathcal{Y}_\theta^k(x)} \ell\left(f_\theta^k(x), y\right) \rho(x, y) \mathrm{d}x \; \mathrm{d}y$$

With Voronoi Cell:

$$\mathcal{Y}_\theta^k(x) \triangleq \left\{y \in \mathcal{Y},\; \ell\left(f_\theta^k(x), y\right) < \ell\left(f_\theta^r(x), y\right),\; \forall r \neq k\right\}$$

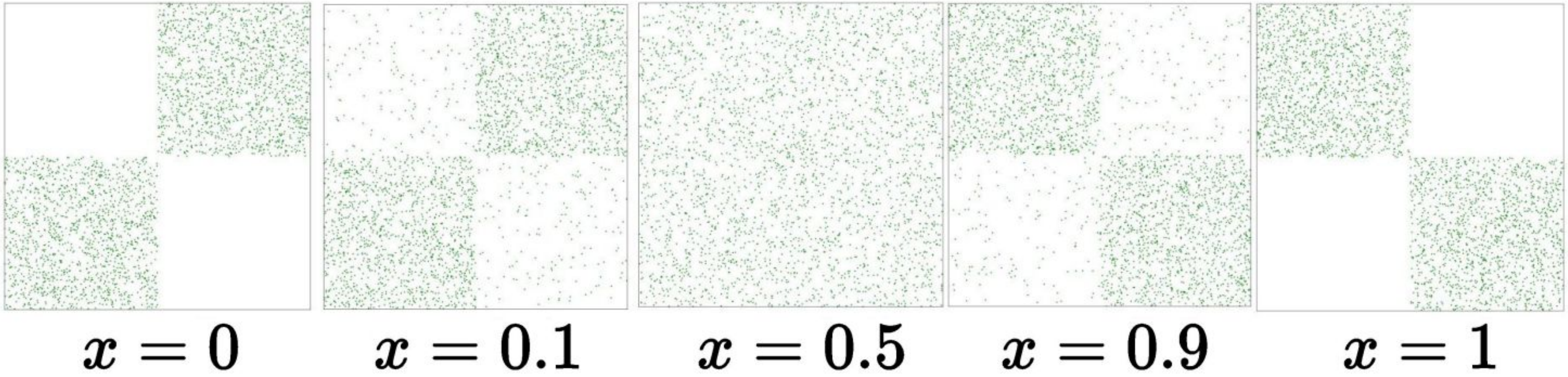
A **necessary condition** to optimize the risk with square quadratic loss is the **centroidal** property

$$f_\theta^k(x) = \mathbb{E}[Y_x \mid Y_x \in \mathcal{Y}_\theta^k(x)]$$

Rupprecht, Christian, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. "Learning in an uncertain world: Representing ambiguity through multiple hypotheses." In *ICCV 2017.*
Du, Qiang, Vance Faber, and Max Gunzburger. "Centroidal Voronoi tessellations: Applications and algorithms." *SIAM review* 41, no. 4 (1999): 637-676.

How does it perform ?

Examples of dataset: Changing Damier.

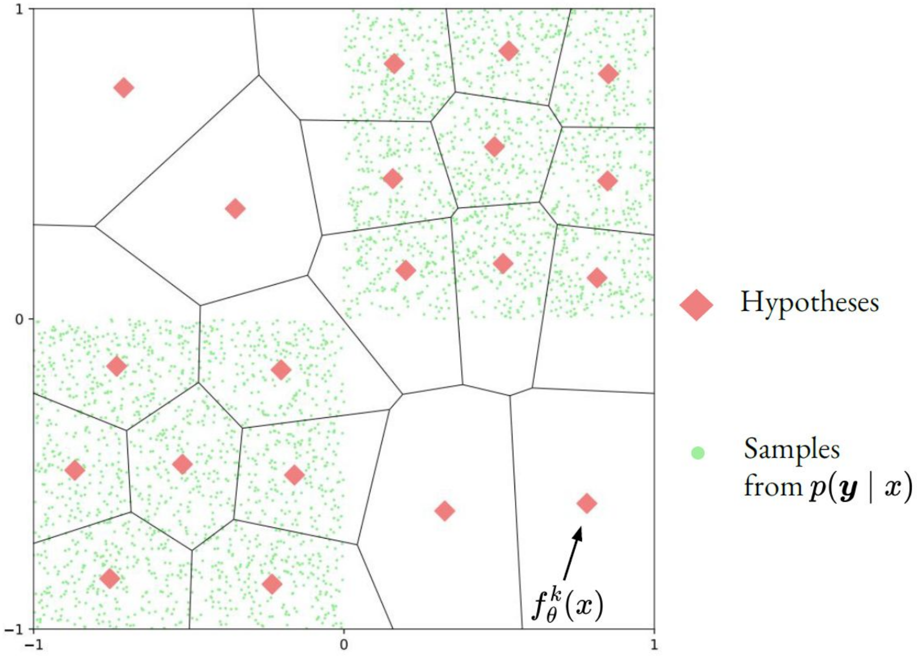


We train a three-layer MLP on such dataset.

Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf
Rupprecht, Christian, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. "Learning in an uncertain world: Representing ambiguity through multiple hypotheses." In *ICCV 2017*
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

How does it perform ?

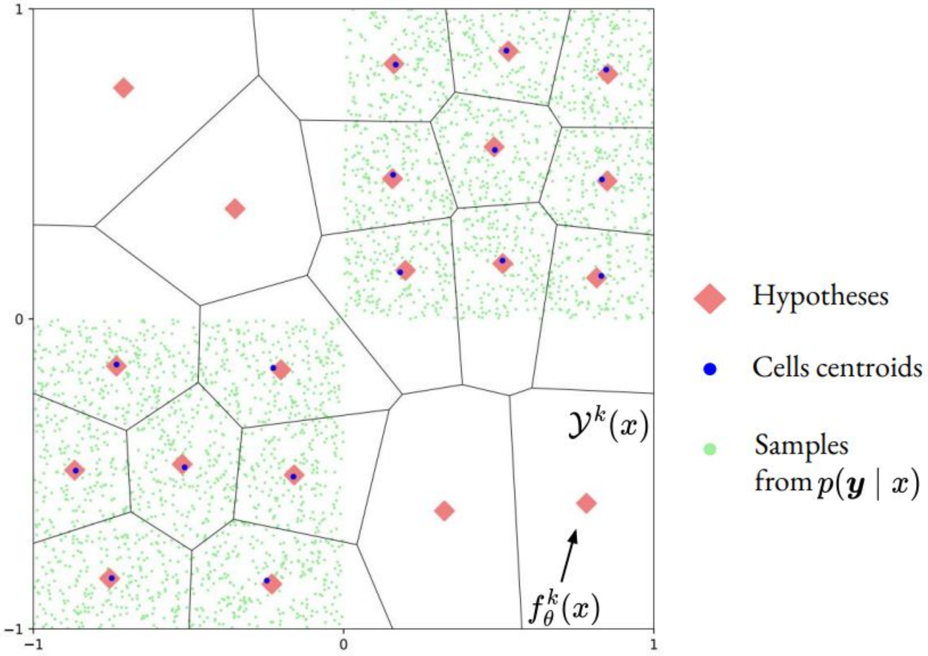
Training dynamics.



Zoomed prediction at x = 0

Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

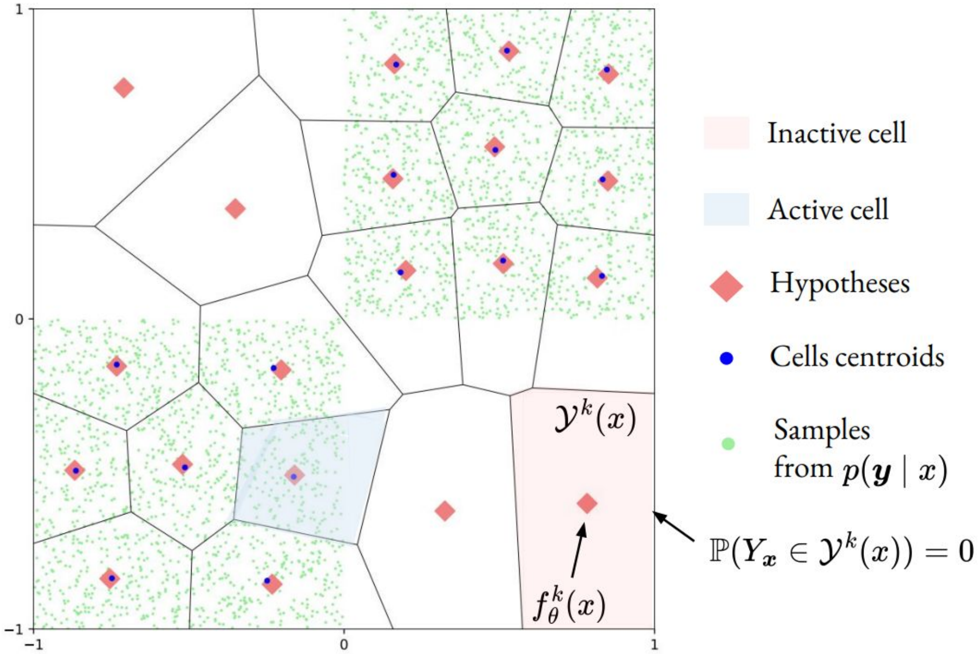
How does it perform ?



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

Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

How does it perform ?

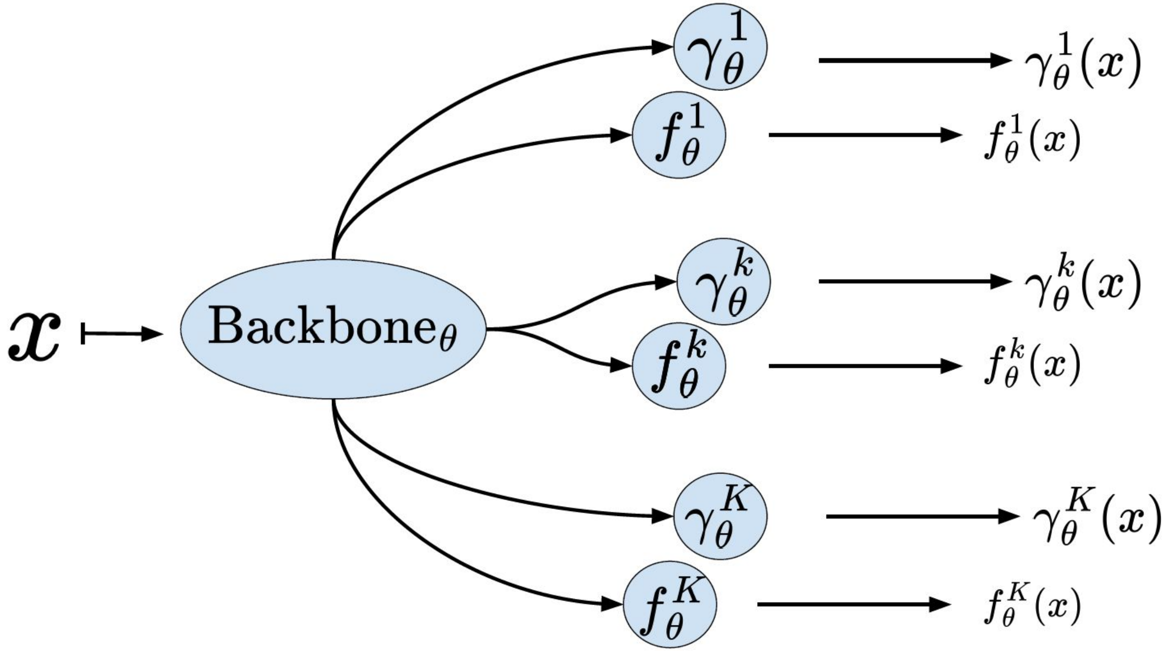


Overconfidence issue

Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023*.

Scoring heads

- Optimization criterion adapted for overconfidence solving.
- Proposition: Scoring Heads: $\gamma_\theta^1, \ldots, \gamma_\theta^K \in \mathcal{F}(\mathcal{X}, [0,1])$ to predict $\mathbb{P}\left(Y_x \in \mathcal{Y}^k(x)\right)$



Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

Scoring heads

Scoring loss:

$$\mathcal{L}^{\text{scoring}}(\theta) \triangleq \sum_{k=1}^{K} \mathrm{BCE}\left(\mathbb{1}\left[y \in \mathcal{Y}_{\theta}^{k}(x)\right], \gamma_{\theta}^{k}(x)\right)$$

Final Loss:

$$\mathcal{L} = \mathcal{L}^{\text{WTA}} + \beta \mathcal{L}^{\text{scoring}}$$

Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

Scoring heads

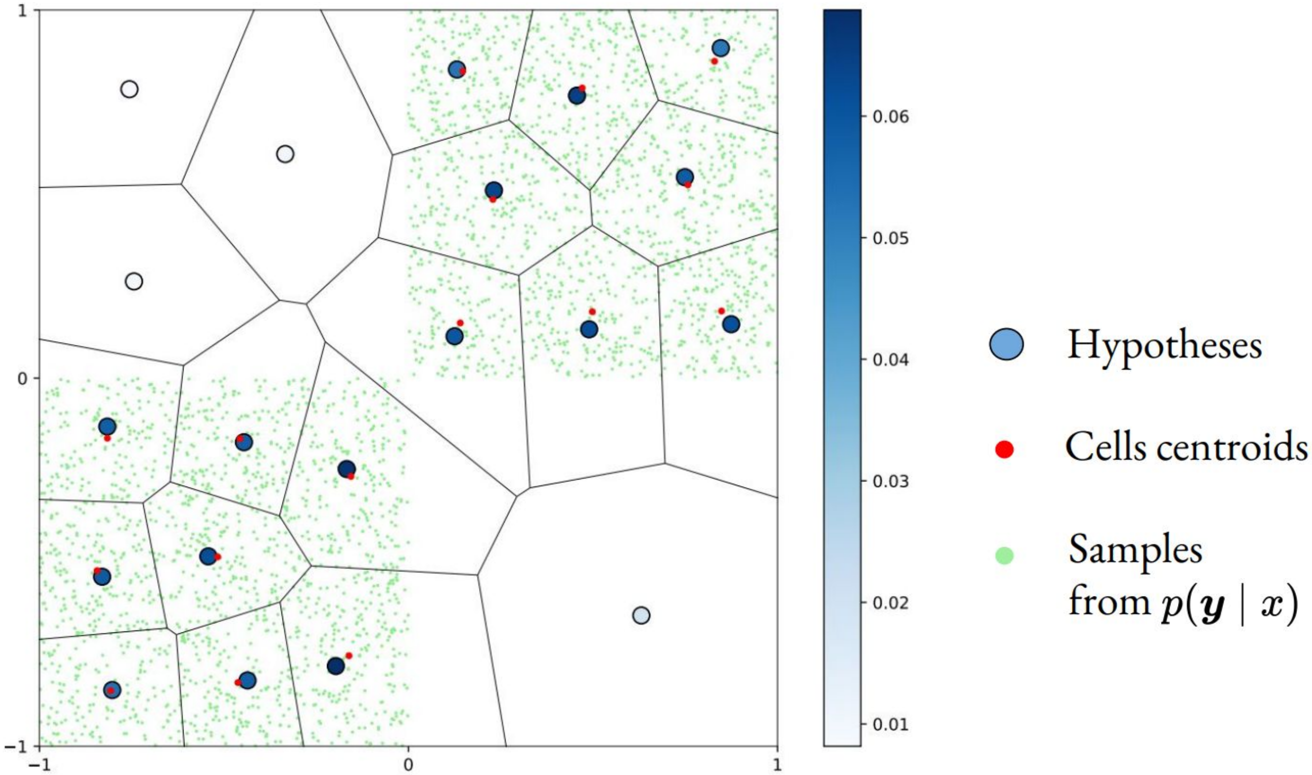
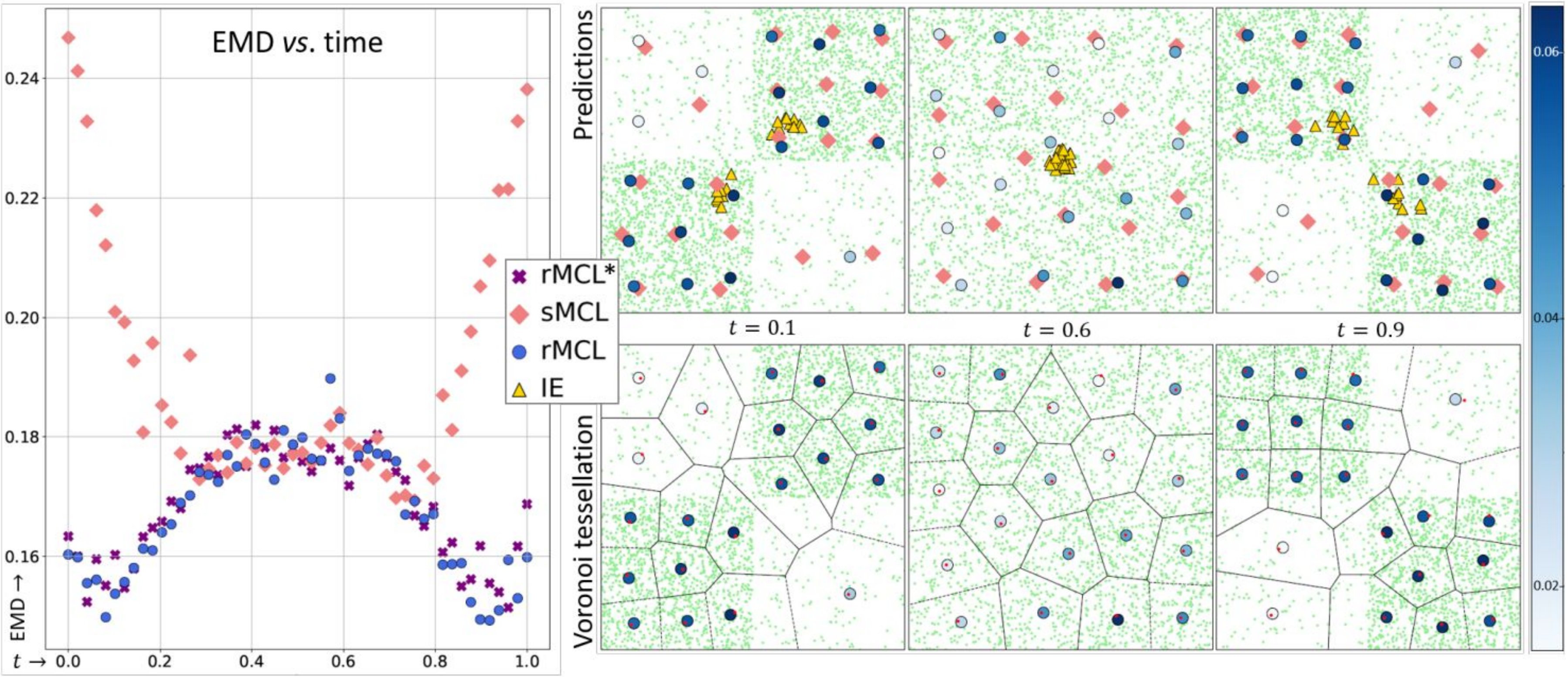


Image from https://neurips.cc/media/neurips-2023/Slides/70945.pdf

Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

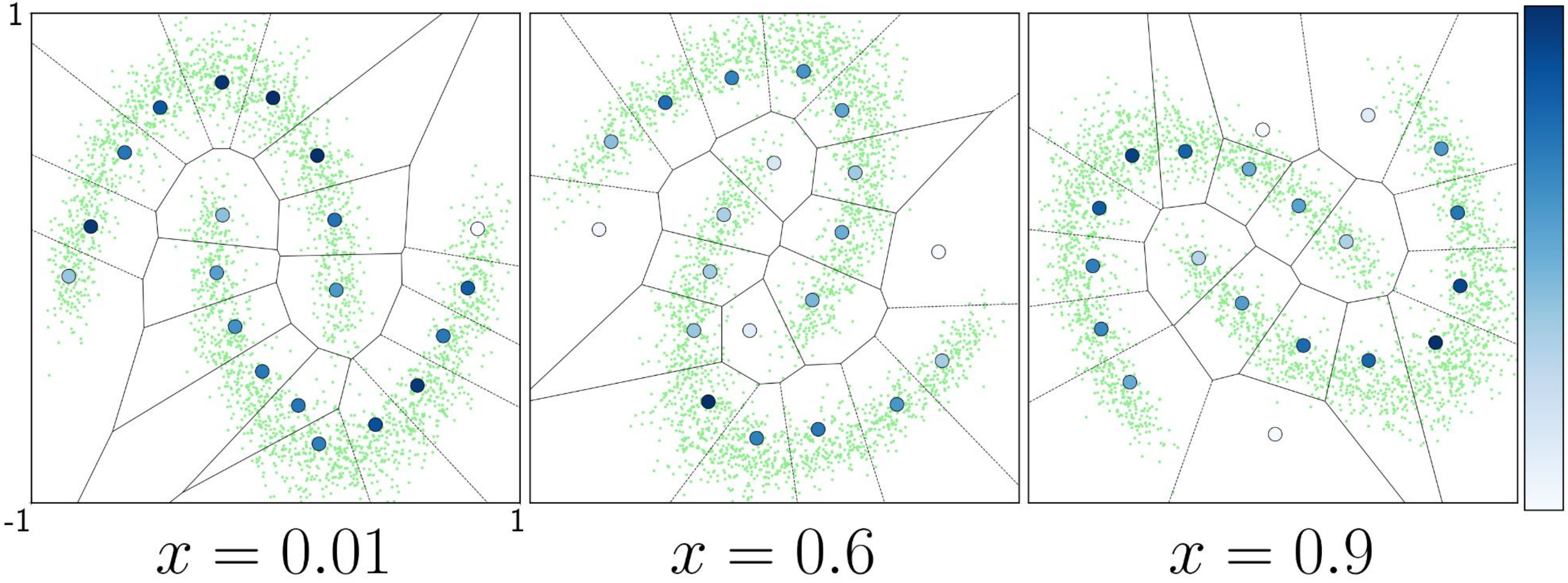
Solving the overconfidence issue



Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

Training dynamics ?

Here: Dataset consisting of a Rotating moon whose angles depends on x.



Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

How to evaluate a multi-hypotheses model ?

Once trained, for a given *x* , we are predicting.

$$(f_\theta^1(x), \ldots, f_\theta^K(x), \gamma_\theta^1(x), \ldots, \gamma_\theta^K(x))$$

We can interpret the model as a conditional distribution estimator, with a mixture of Dirac deltas:

$$\hat{\rho}_x(y) = \sum_{k=1}^{K} \gamma_\theta^k(x) \delta_{f_\theta^k(x)}(y)$$

Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

How to evaluate this estimator ?

If the target distribution ρ_x is known ?

Earth Mover's Distance.

$$\mathrm{EMD}\left(\hat{\rho}_x, \rho_x\right) = \min_{\psi \in \Psi} \sum_{y_s \sim \rho_x} \sum_{\hat{y}_k \sim \hat{\rho}_x(y)} \psi_{s,k} d\left(y_s, \hat{y}_k\right)$$

Where $\psi \in \Psi$ is a (valid) transport plan.

Leonid V Kantorovich. On the translocation of masses. In Dokl. Akad. Nauk. USSR (NS), 1942.

Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

How to evaluate that ?

However, in practice, we just have access to pairs $\{(x_i, y_i)\}$

If ρ_x is not known ?

Distortion Risk. $$\mathcal{R} = \frac{1}{N} \sum_{i=1}^{N} \min_{k \in [1,K]} \left\| y_i - f_\theta^k(x_i) \right\|^2$$

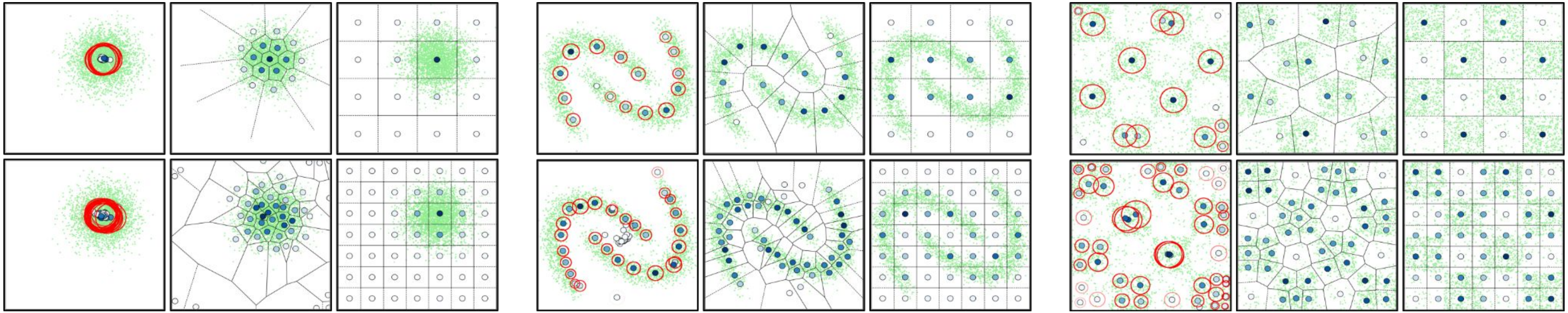
Can be viewed as a (conditional) quantization error

NLL. $$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{\rho}_{x_i}(y_i)$$

N/A if we use Dirac Mixture. We considered **truncated kernels** instead.

Polianskii, Vladislav, Giovanni Luca Marchetti, Alexander Kravberg, Anastasiia Varava, Florian T. Pokorny, and Danica Kragic. "Voronoi density estimator for high-dimensional data: Computation, compactification and convergence." In *UAI*, 2022.
Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

How does it compares with other methods ?

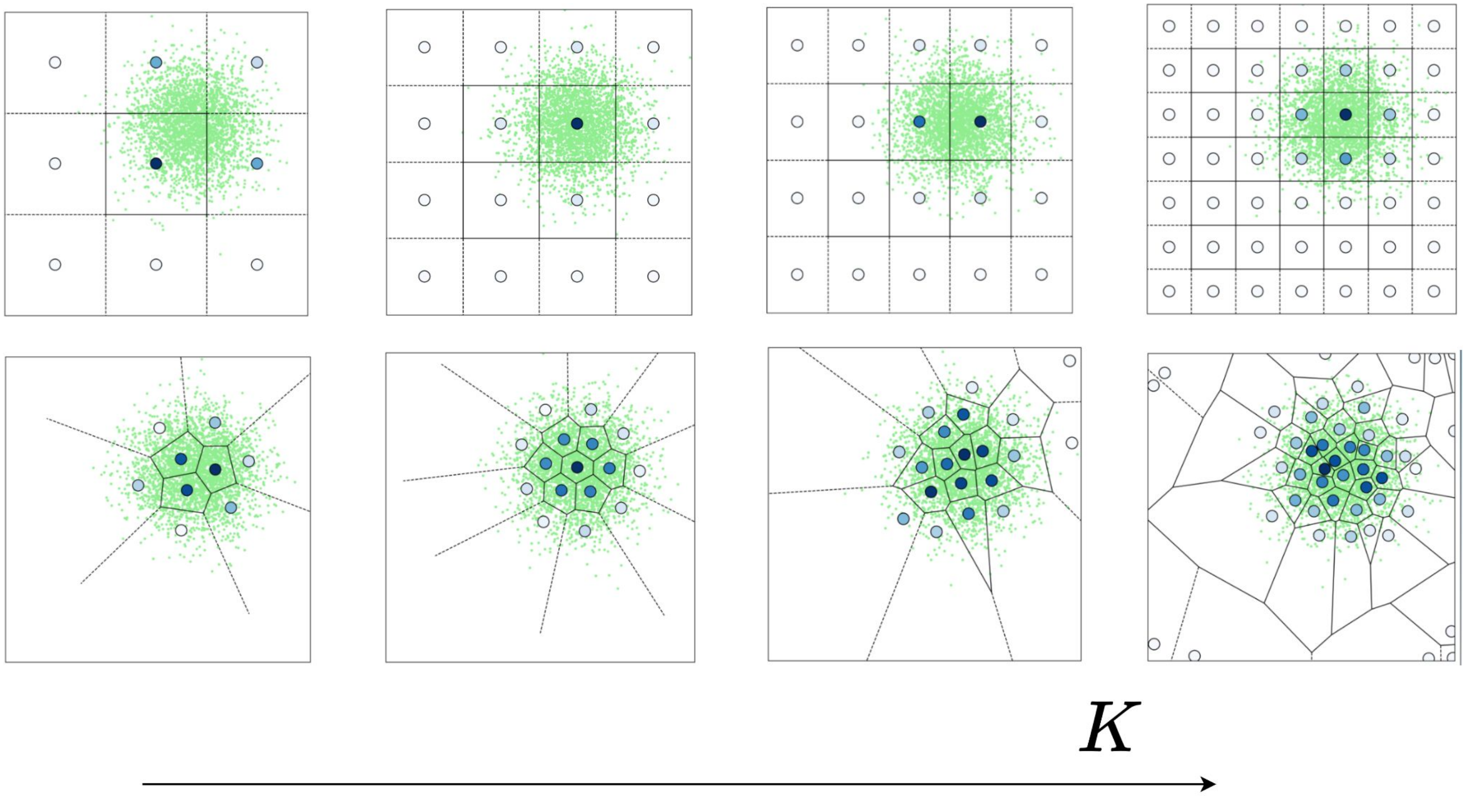


Qualitative results. Datasets: Single Gaussian, Rotated Two Moons and Changing Damier. Within each panel, columns correspond to predictions made by: MDN, Score-based WTA and Histogram (left to right). Dots: predicted hypotheses. *1st row*: 16 hypotheses, *2nd row:* 49 hypotheses.

Bishop, Christopher M. "Mixture density networks." (1994).
Imani, Ehsan, and Martha White. "Improving regression performance with distributional losses." In *ICML*, 2018.
Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

Effect of the number of hypotheses

Using Zador's theorem, we can estimate the asymptotic Risk, and compare it with the one of the histogram.



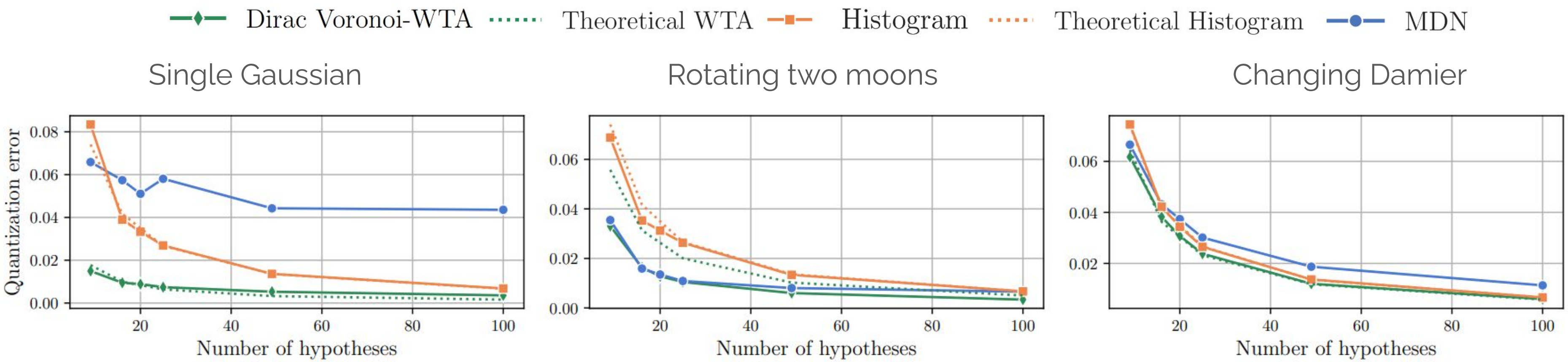
Zador, Paul. "Asymptotic quantization error of continuous signals and the quantization dimension." *IEEE Transactions on Information Theory*, 1982

Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

How does it compares with other methods ?

Metrics.

- Earth Mover's Distance (EMD)
- Negative Log-Likelihood (NLL), which assesses the probabilistic quality of the predictions
- Quantization Error



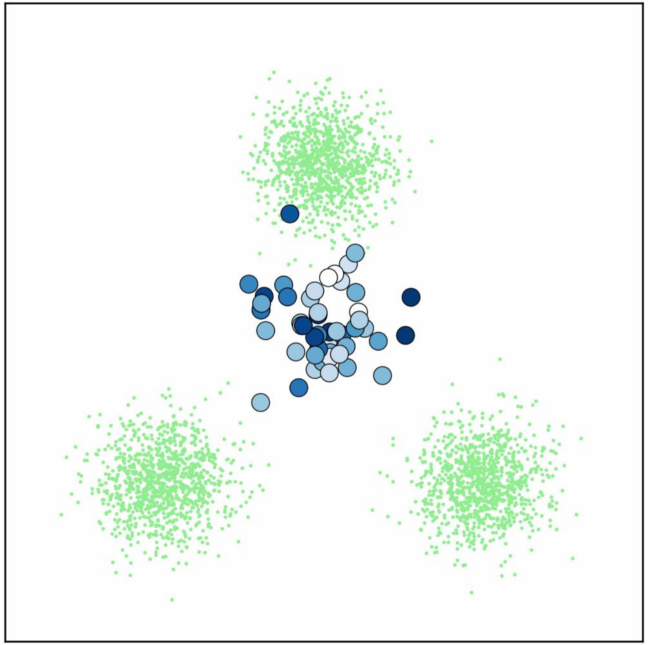
Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

Can we do better in terms of **quantization** ?

TLDR; Yes, using deterministic annealing

Limitations of WTA

$$t = 0$$



Gif from https://github.com/Victorletzelter/annealed_mcl

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Relaxed Winner-takes-all training

Gradient to non-Winner heads ?

$$\mathcal{L}^{\mathrm{R-WTA}}(\theta) = (1-\varepsilon)\ell\left(f_\theta^{k^\star}(x), y\right) + \frac{\varepsilon}{K-1}\sum_{k \neq k^\star} \ell\left(f_\theta^{k}(x), y\right)$$

Rupprecht, Christian, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. "Learning in an uncertain world: Representing ambiguity through multiple hypotheses." In *Proceedings of the IEEE international conference on computer vision*, pp. 3591-3600. 2017.

Annealed Winner-takes-all training

Trained as follows for each pair $(x, y) \sim \rho(x, y)$ at training step t, with e.g., $T(t) = T_0\, \rho^t$

1. Compute $f_\theta^1(x), \ldots, f_\theta^K(x)$

2. Compute a boltzmann distribution on the assignments:

$$q_{T(t)}(k \mid x, y) \triangleq \frac{1}{Z_{x,y}} \exp\left(-\frac{\ell(f_\theta^k(x), y)}{T(t)}\right) \qquad Z_{x,y} \triangleq \sum_{k=1}^{K} \exp\left(-\frac{\ell(f_\theta^k(x), y)}{T(t)}\right)$$

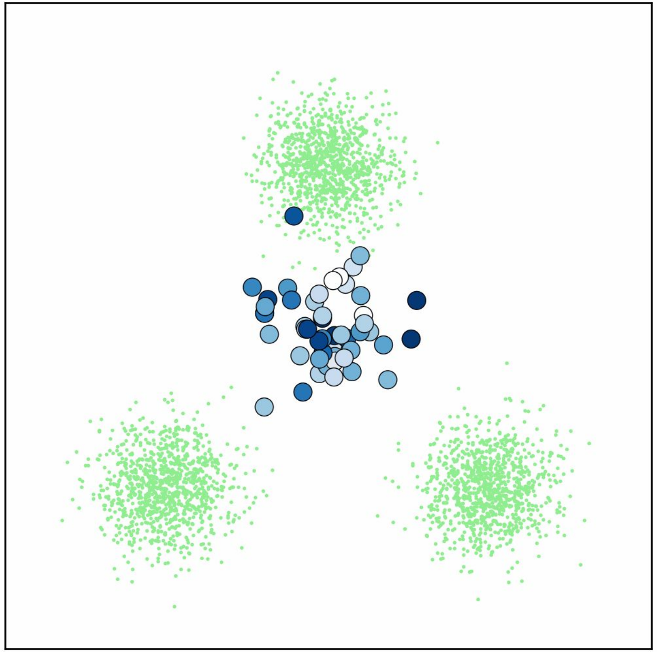
3. Compute the aMCL loss:

$$\mathcal{L}_t^{\mathrm{aMCL}}(\theta) = \sum_{k=1}^{K} q_{T(t)}(k \mid x, y)\, \ell(f_\theta^k(x), y)$$

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Limitations of WTA

$$t = 0, \quad T(t) = 0.6$$



Gif from https://github.com/Victorletzelter/annealed_mcl

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Analysis of aMCL

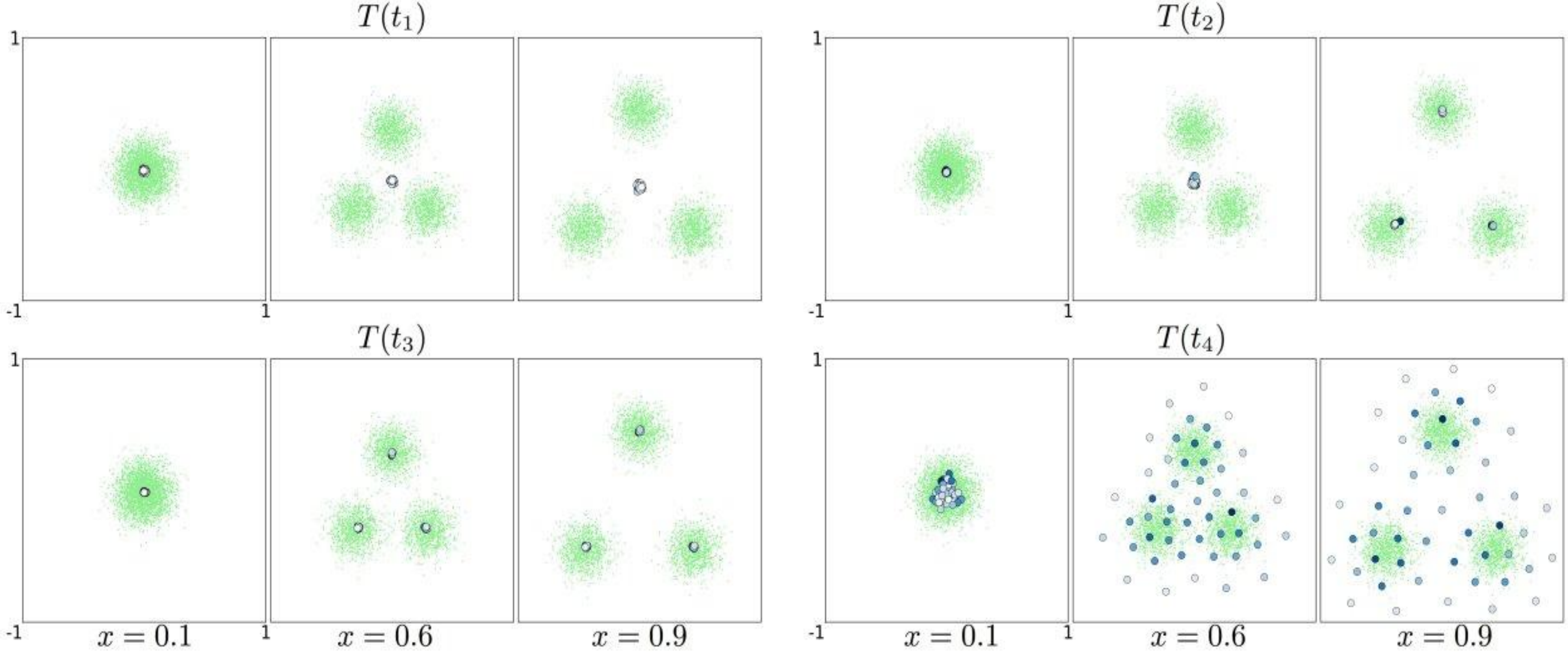
- **Result**: aMCL is an entropy-constrained block coordinate descent on the soft distortion.

- **Result**: Gradient directions and optimality conditions. 1) **gradient towards** soft barycenter and 2) **soft barycenter** property

- **Result**: the hypotheses iteratively **merge** and **splits** into **sub-groups**.

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Training trajectory shows phase transitions

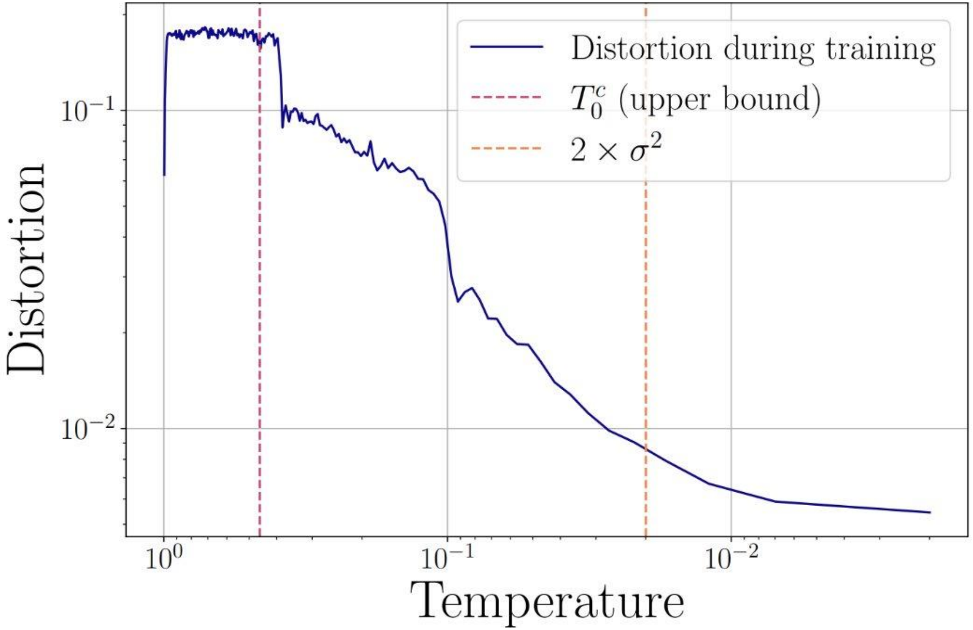


D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Training trajectory shows phase transitions

Idea: Predictions **fuze** at high temperatures, and split at transition temperatures studied here in the **conditional** case.

Rate-Distortion: Count the optimal number of **distinct** hypotheses for each **distortion** level for each input.



D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

Synthetic data

Goal : 1) **qualitative** analysis 2) comparison against **MCL variants** (vanilla-MCL, Relaxed-WTA...)
Result : solve the issues of MCL variant relative to **hypothesis collapse** and **barycenter bias.**

UCI Dataset

Goal : 1) **standard benchmark** 2) comparison against **baselines** (Deep Ensembles, MC Dropout...), MCL-variants.
Result : Very competitive on **distortion** without compromising **RMSE**. Good trade-off between the metrics.

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

UCI Benchmark

Table 1: **Results on UCI regression benchmark datasets comparing Distortion.** Experimental setup is described in Section 5.1.1. Relaxed-WTA results were computed with $\varepsilon = 0.1$ which strikes a good tradeoff between RMSE and Distortion (see Table 5 in Appendix). The rows are ordered by dataset size N. Best results are in **bold**, second bests are underlined.

	Distortion (↓)			
Datasets	Relaxed-WTA ($\varepsilon = 0.1$)	MCL	aMCL	N
Year	9.09 ± NA	4.82 ± NA	**4.46 ± NA**	515345
Protein	1.67 ± 0.16	0.80 ± 0.02	**0.77 ± 0.03**	45730
Naval	**4.21e-7 ± 2.36e-7**	1.84e-6 ± 2.42e-6	5.37e-7 ± 3.83e-7	11934
Power	2.95 ± 0.91	2.31 ± 0.49	**2.18 ± 0.64**	9568
Kin8nm	9.32e-4 ± 7.97e-5	1.00e-3 ± 1.47e-4	**6.81e-4 ± 8.14e-5**	8192
Wine	0.06 ± 0.02	**0.02 ± 0.01**	0.03 ± 0.01	1599
Concrete	6.91 ± 2.81	**5.13 ± 1.23**	5.71 ± 1.72	1030
Energy	0.30 ± 0.12	1.25 ± 1.25	**0.28 ± 0.09**	768
Boston	3.32 ± 2.84	**2.14 ± 0.49**	2.69 ± 1.39	506
Yacht	1.34 ± 0.93	3.09 ± 2.41	**1.15 ± 0.97**	308

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

UCI Benchmark

Table 2: **Results on UCI regression benchmark datasets comparing RMSE.** Best results are in **bold**, second bests are underlined. * corresponds to reported results from [38].

Datasets	PBP*	MC Dropout*	Deep Ensembles*	Relaxed-WTA ($\varepsilon = 0.1$)	MCL	aMCL	N
	RMSE (↓)						
Year	8.88 ± NA	**8.85 ± NA**	8.89 ± NA	8.97 ± NA	9.09 ± NA	9.08 ± NA	515345
Protein	4.73 ± 0.01	4.36 ± 0.04	4.71 ± 0.06	4.38 ± 0.02	4.39 ± 0.10	**4.25 ± 0.02**	45730
Naval	0.01 ± 0.00	0.01 ± 0.00	**0.00 ± 0.00**	1.80e-3 ± 5.66e-4	2.08e-3 ± 1.18e-3	**8.00e-4 ± 4.04e-4**	11934
Power	4.12 ± 0.03	**4.02 ± 0.18**	4.11 ± 0.17	**4.02 ± 0.18**	4.18 ± 0.16	4.08 ± 0.20	9568
Kin8nm	0.10 ± 0.00	0.10 ± 0.00	0.09 ± 0.00	**0.08 ± 0.00**	0.10 ± 0.01	**0.08 ± 0.00**	8192
Wine	0.64 ± 0.01	**0.62 ± 0.04**	0.64 ± 0.04	0.63 ± 0.04	0.63 ± 0.04	0.63 ± 0.04	1599
Concrete	5.67 ± 0.09	**5.23 ± 0.53**	6.03 ± 0.58	5.28 ± 0.58	6.02 ± 0.65	5.47 ± 0.67	1030
Energy	1.80 ± 0.05	1.66 ± 0.19	2.09 ± 0.29	1.64 ± 0.36	2.53 ± 0.99	**1.35 ± 0.97**	768
Boston	3.01 ± 0.18	2.97 ± 0.85	3.28 ± 1.00	**2.85 ± 0.72**	3.54 ± 1.16	3.05 ± 0.91	506
Yacht	**1.02 ± 0.05**	1.11 ± 0.38	1.58 ± 0.48	2.52 ± 1.04	3.28 ± 1.39	1.62 ± 0.53	308

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

UCI Datasets

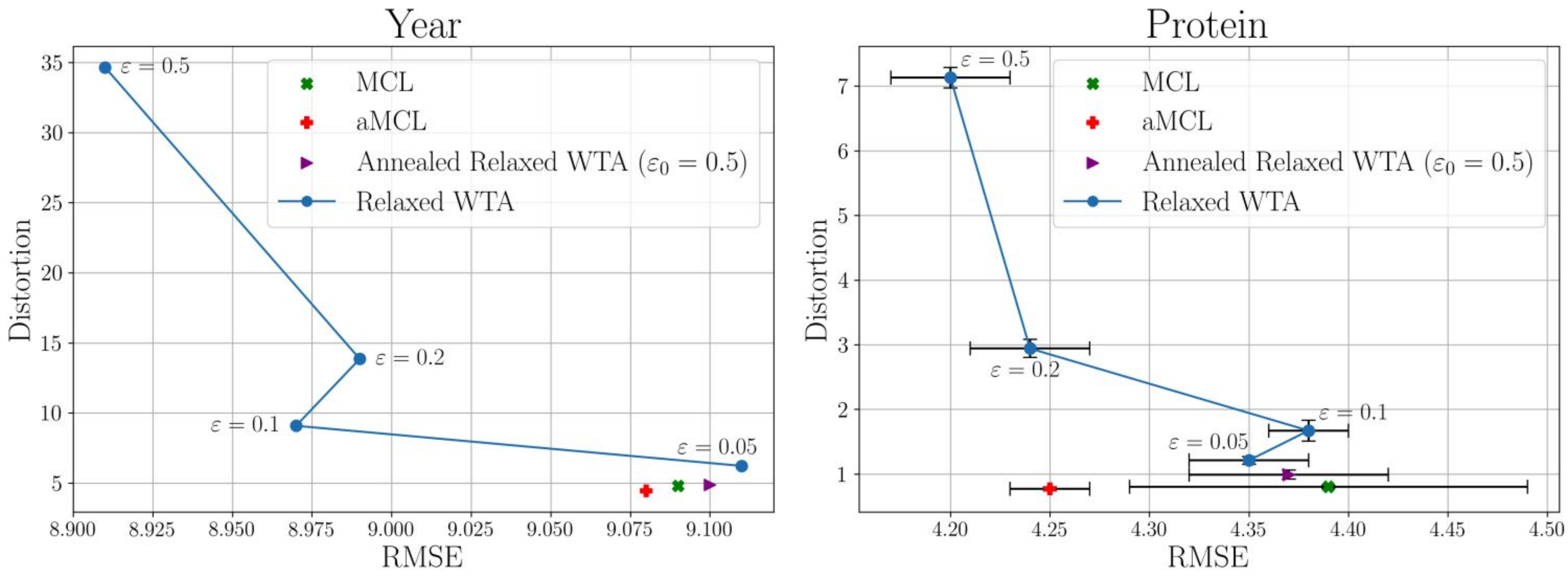


Figure 6: **(RMSE,Distortion) performance on Year and Protein datasets**, for Relaxed-WTA with $\varepsilon \in \{0.05, 0.1, 0.2, 0.5\}$ (blue points, with black error bars for std), Relaxed-WTA with annealed ε (purple) and aMCL (red). Blue lines link the Relaxed-WTA score points. We see that aMCL is below and left of these lines, which indicates a good tradeoff on those two datasets.

D. Perera*, **V. Letzelter***, T. Mariotte, A. Cortés, M. Chen, S. Essid, and G. Richard. "Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing". In *NeurIPS. 2024.*

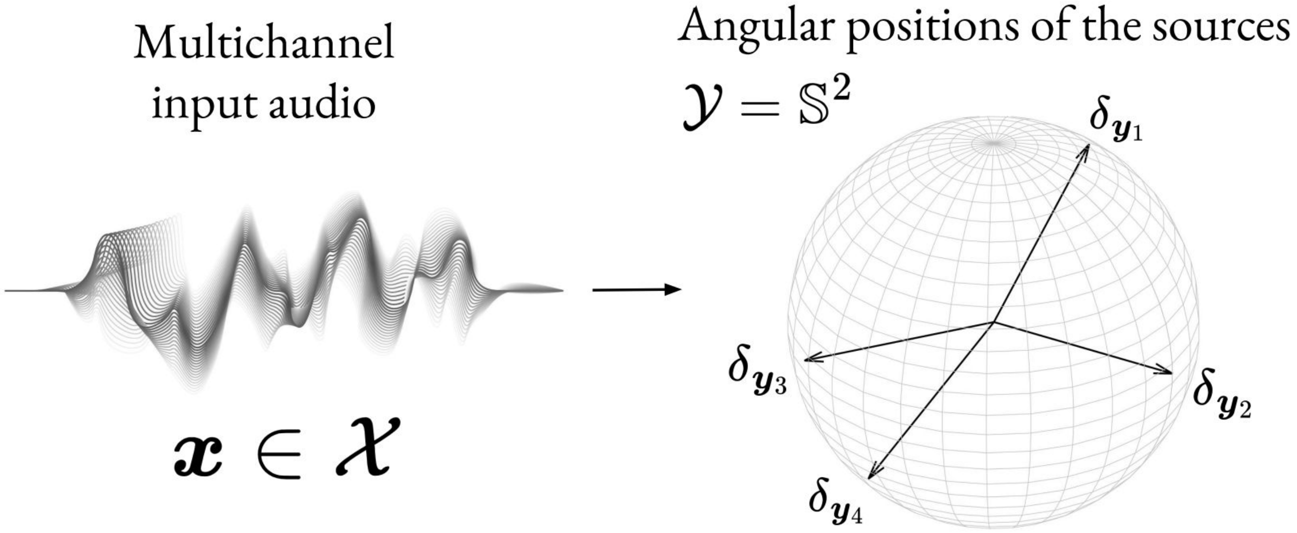
Some applications of this work

TLDR; We applied (Score-based) WTA Based training to:

- Sound source localization
- Human Pose Estimation.
- Motion forecasting
- Multivariate time-series forecasting

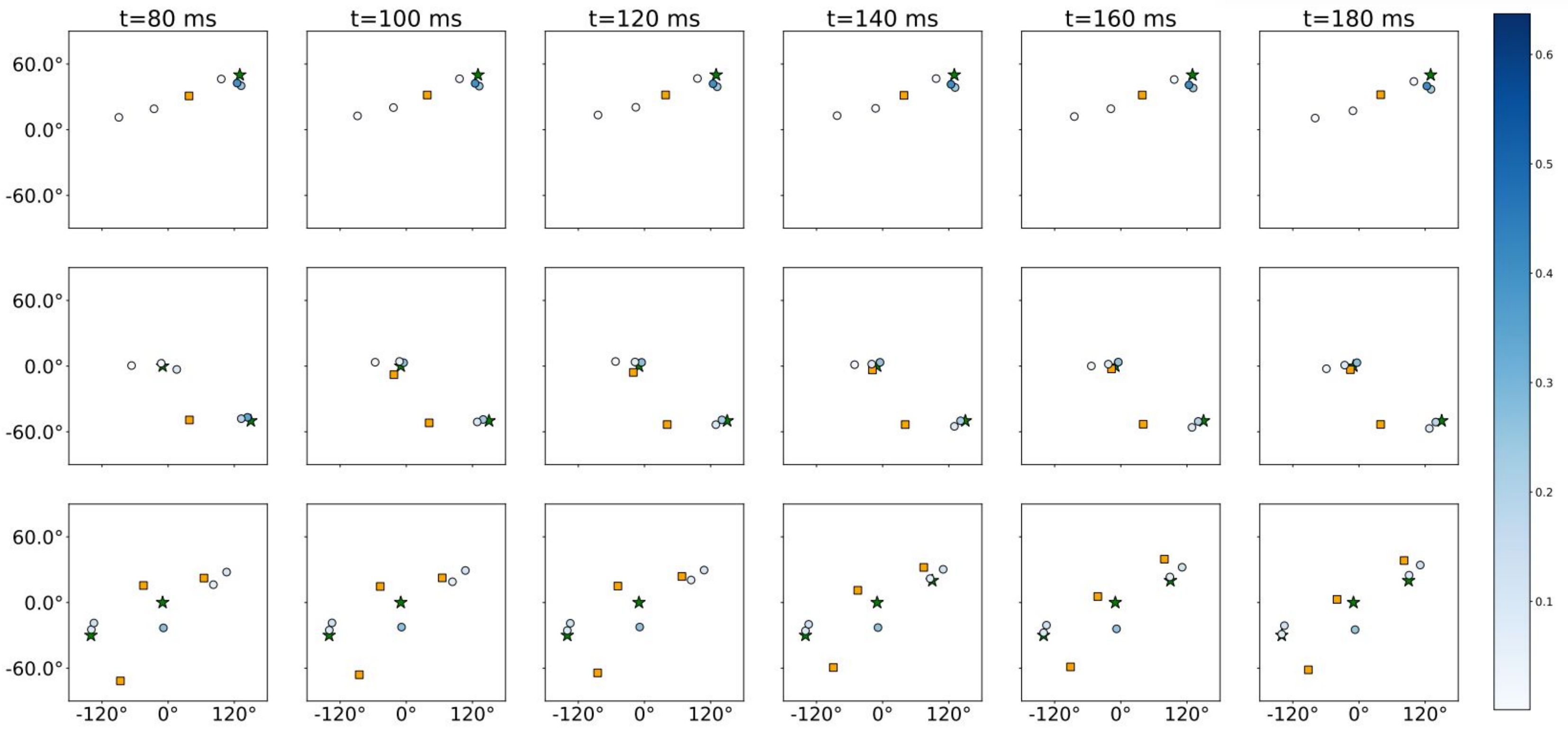
Sound Source Localization

Locating **spatial** sound events -> Ambiguous / ill-posed.



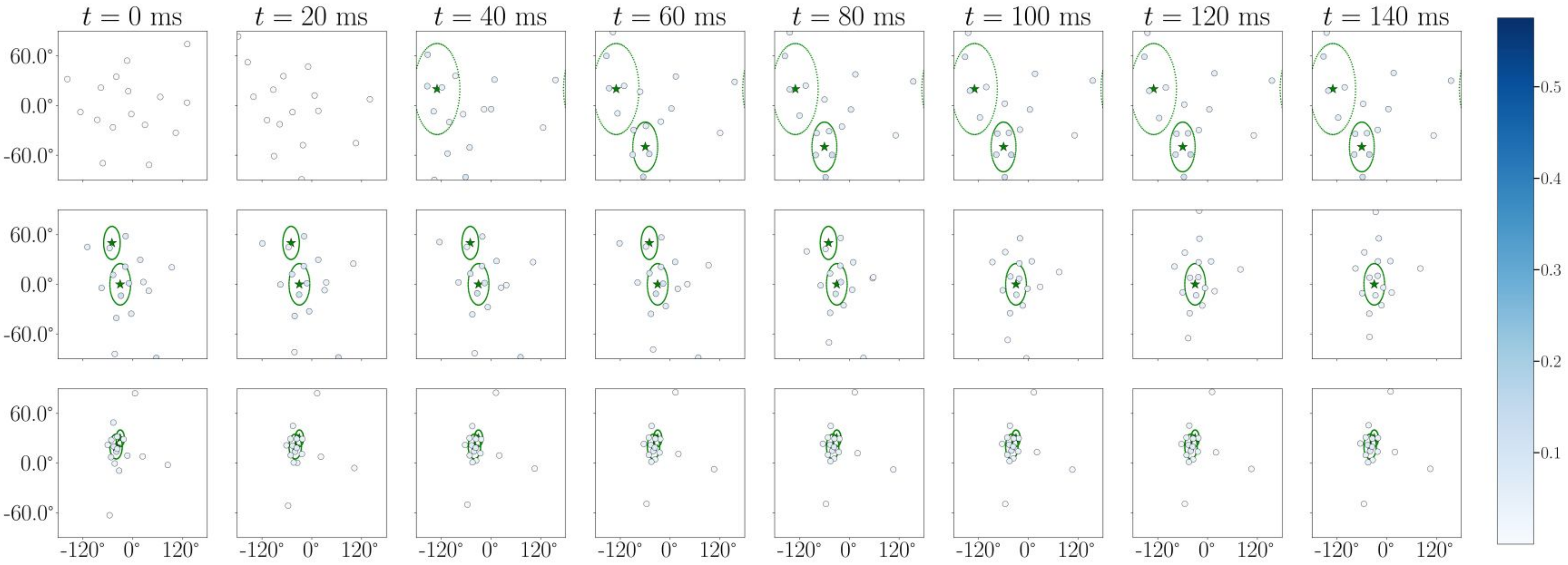
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023*.
Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

Sound Source Localization



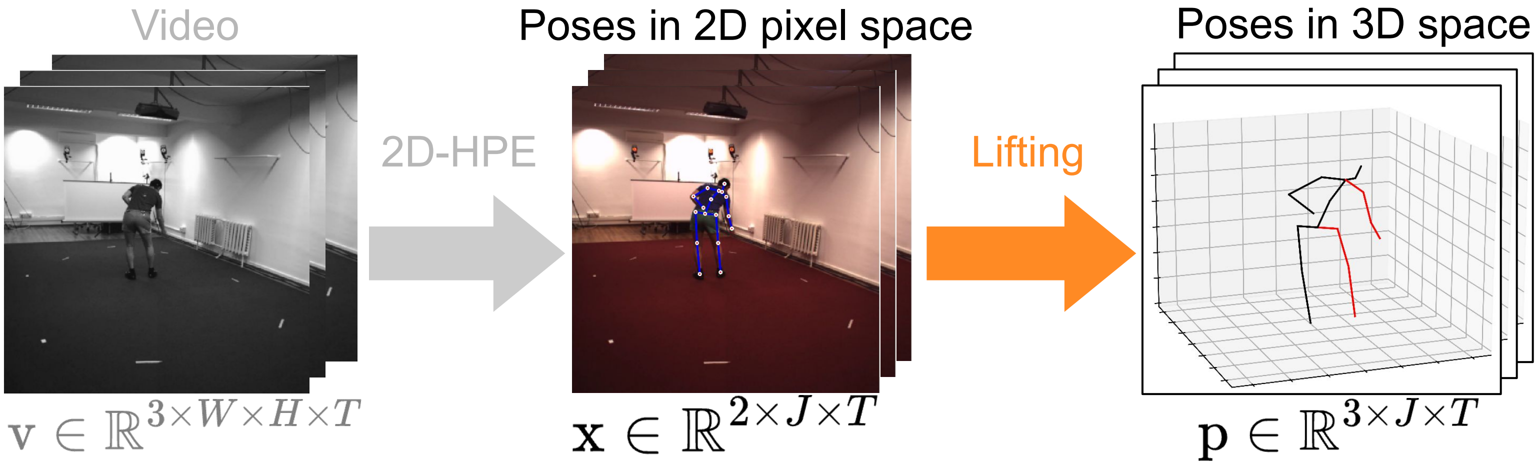
Letzelter, Victor, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gaël Richard. "Resilient Multiple Choice Learning: A learned scoring scheme with application to audio scene analysis." In *NeurIPS 2023.*

Sound Source Localization



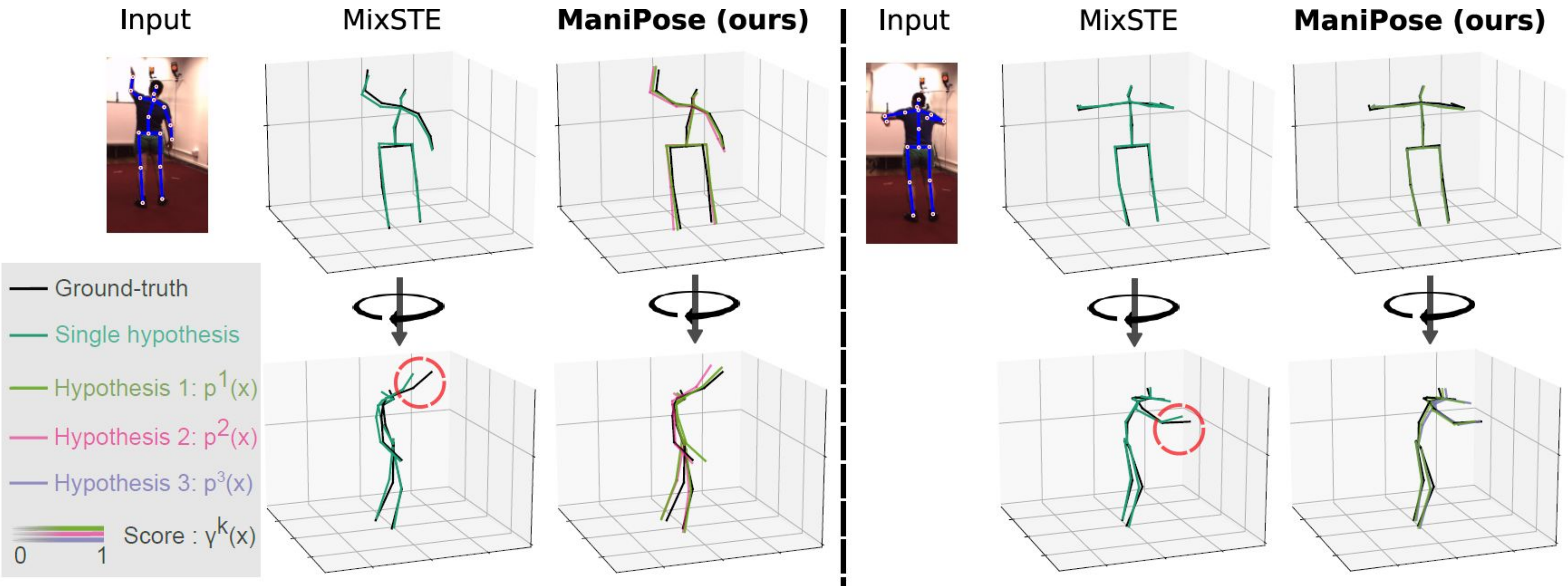
Letzelter, Victor, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. "Winner-takes-all learners are geometry-aware conditional density estimators." In *ICML* 2024.

Monocular 3D Human Pose Estimation



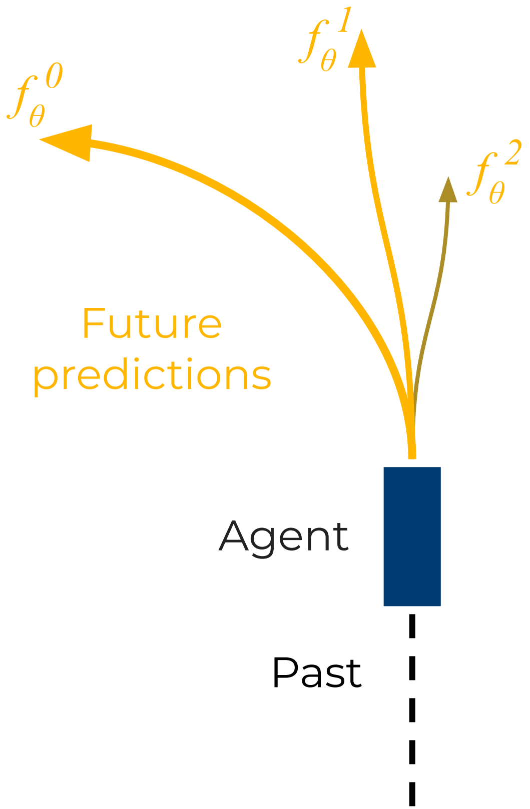
Rommel, Cédric, **Victor Letzelter**, Nermin Samet, Renaud Marlet, Matthieu Cord, Patrick Pérez, and Eduardo Valle. "ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation." In *NeurIPS 2024*

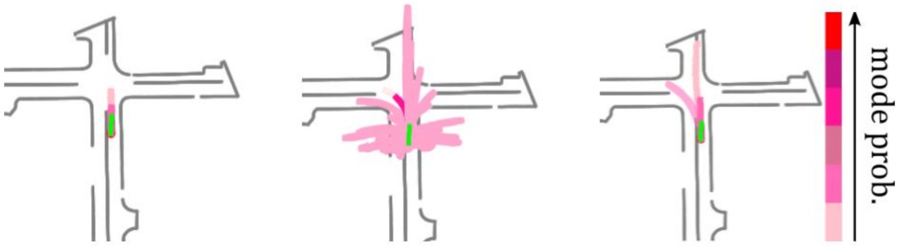
Monocular 3D Human Pose Estimation



Rommel, Cédric, **Victor Letzelter**, Nermin Samet, Renaud Marlet, Matthieu Cord, Patrick Pérez, and Eduardo Valle. "ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation." In *NeurIPS 2024*

Motion forecasting

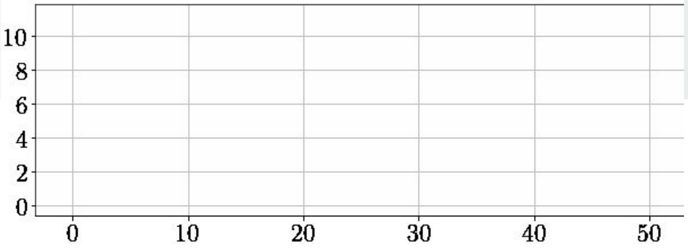




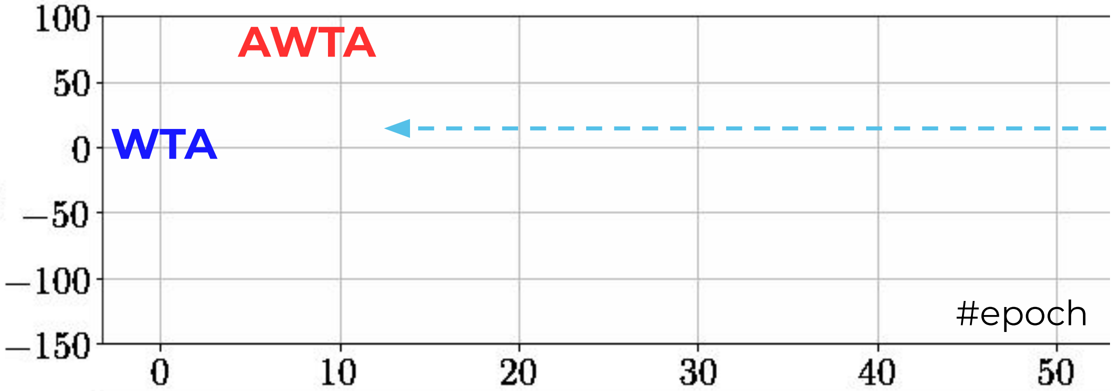
(a) WTA 6 hyp. (b) WTA 64 hyp. (c) **aWTA** 6 hyp.

Y. Xu, **V. Letzelter**, M. Chen, E. Zablocki, and M. Cord. "Annealed Winner-Takes-All for Motion Forecasting." *ICRA 2025.*

Motion forecasting

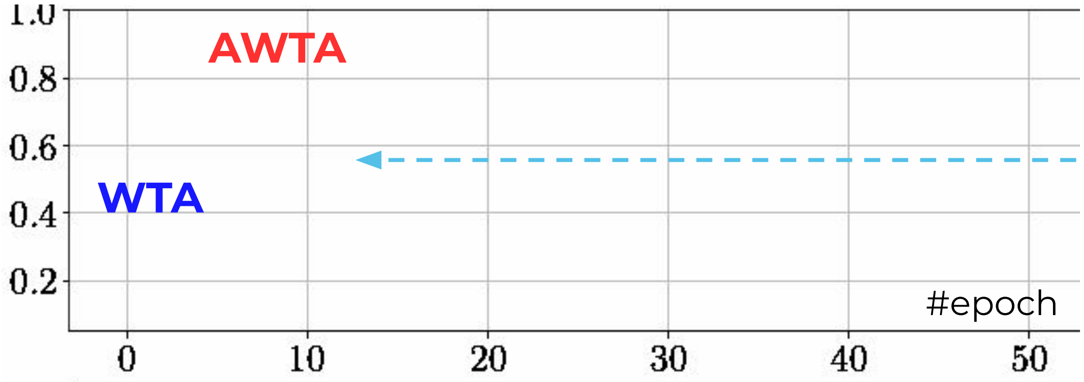


Regression Loss



Sudden loss drop

Miss Rate Metric



Sudden metric improvement

Animation from https://www.youtube.com/watch?v=EOslayPi7Lw

Y. Xu, **V. Letzelter**, M. Chen, E. Zablocki, and M. Cord. "Annealed Winner-Takes-All for Motion Forecasting." *ICRA 2025*.

Thank you for you attention !