Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing

David Perera, Victor Letzelter, Théo Mariotte, Adrien Cortés, Mickael Chen, Slim Essid, Gaël Richard

Annealed Multiple Choice Learning: Overcoming limitations of Winner-takes-all with annealing

David Perera[*1] david.perera@telecom-paris.fr

Victor Letzelter[*1 2] victor.letzelter@telecom-paris.fr

Théo Mariotte[1] **Adrien Cortés**[3] **Mickael Chen**[2] **Slim Essid**[1] **Gaël Richard**[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris
[2] Valeo.ai
[3] Sorbonne Université

Abstract

We introduce Annealed Multiple Choice Learning (aMCL) which combines simulated annealing with MCL. MCL is a learning framework handling ambiguous tasks by predicting a small set of plausible hypotheses. These hypotheses are trained using the Winner-takes-all (WTA) scheme, which promotes the diversity of the predictions. However, this scheme may converge toward an arbitrarily suboptimal local minimum, due to the greedy nature of WTA. We overcome this limitation using annealing, which enhances the exploration of the hypothesis space during training. We leverage insights from statistical physics and information theory to provide a detailed description of the model training trajectory. Additionally, we validate our algorithm by extensive experiments on synthetic datasets, on the standard UCI benchmark, and on speech separation.

1 Introduction

Ambiguous prediction tasks arise in deep learning when the target y is ill-defined from the input x. Predicting y directly from x can be challenging due to the partial predictability of y from the information contained in x. Multiple Choice Learning (MCL) [25, 41] addresses these challenges by providing a small set of plausible *hypotheses*, each representing a different possible outcome given the input. MCL learns these hypotheses using a competitive training scheme that promotes the specialization of the hypotheses in distinct regions of the output space $\mathcal{Y}$. The framework iteratively partitions $\mathcal{Y}$ into a Voronoi tesselation and guides each hypothesis toward the barycenter of its respective Voronoi cell [63]. This mechanism makes MCL akin to a gradient-descent-based and conditional variant of the popular K-means algorithm [45]. Like K-means, MCL is sensitive to initialization, subject to hypothesis collapse [49], and more generally may converge toward arbitrarily suboptimal hypothesis configurations [63]. While there is a substantial body of literature addressing the limitations of K-means [3, 20, 58], relatively little has been done to address these challenges in the context of MCL [63, 49, 54]. The core issue of MCL lies with its greedy gradient-based update of the hypotheses. This greediness precludes the exploration of the hypothesis space, preventing MCL from optimally capturing the ambiguity of y. We propose to incorporate annealing into this gradient descent update in order to improve the robustness of MCL.

Simulated annealing, inspired by the gradual cooling of metals, was originally introduced for statistical mechanics applications [28, 52] and was later applied to combinatorial problems [35]. It is

[*]Equal Contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

a random exploration process concurrent to the popular stochastic gradient descent, with a significant difference: gradient descent always tries to improve performance while annealing also accepts to temporarily degrade it for the sake of exploration. The range of this exploration is controlled by a temperature parameter: with a high temperature, annealing explores wide regions of the search space; when the temperature decreases, the exploration becomes narrow, and the system is able to refine its performance. This strategy has been shown to converge to an optimal state, provided that the cooling is sufficiently slow [26].

Deterministic annealing [62] is a variant of simulated annealing. In simulated annealing, exploration relies on a sequence of random moves across the search space, whereas deterministic annealing seeks greater efficiency by replacing this random process with the exact minimization of a deterministic functional, namely the free energy of the system. It has been shown that deterministic annealing can be efficiently applied to clustering [61, 62]. In this article, we show that it can be further adapted to the conditional and gradient-based setting of MCL. The resulting algorithm, which we name aMCL for *annealed MCL*, addresses the main issues of MCL and achieves strong performance in practical settings while being straightforward to implement and amenable to analysis. Specifically, we make the following contributions.

We introduce Annealed Multiple Choice Learning (aMCL), a novel algorithm that incorporates annealing into the multiple choice learning framework (Section 3).

We propose a theoretical analysis of aMCL, to understand its advantages in comparison to vanilla MCL. We characterize the training trajectory of the model, by establishing an analogy with statistical physics and information theory (Section 4).

We provide extensive experimental validation, by applying this method i) to illustrative synthetic examples; ii) to a standard distribution estimation benchmark (UCI datasets); and also iii) to the challenging audio task of speech separation (Section 5). The accompanying code is made available.[1]

2 Related Work

Multiple choice learning. MCL has been successfully applied to various machine learning tasks, typically using multi-head neural networks, with each head providing a prediction [40, 22, 42]. Several works observed the phenomenon of *hypothesis collapse* [8, 19, 33, 40, 67, 63], where some hypotheses are left unused during training. Various solutions have been proposed to tackle collapse [63, 49, 54]. Notably, [63] introduces Relaxed-WTA, which updates non-winning hypotheses with a gradient scaled by a small constant ε. However, this small gradient biases the hypotheses toward the global barycenter of the target distribution, which can be shown to be suboptimal [14].

Simulated and deterministic annealing. Deterministic annealing is a variant of simulated annealing [28, 52, 35]. Rose *et al.* extensively investigated its properties, particularly in relation to statistical physics and clustering [61, 62, 59, 60]. We are, to the best of our knowledge, the first to combine this technique with Winner-takes-all training in a conditional setting.

Information theory and quantization. Quantization [64] consists of discretizing continuous variables over a finite set of symbols. The rate-distortion theory studies the minimal number of bits necessary to encode information at a given level of quantization error [5, 2, 4]. Recently, a relation has been established between optimal quantization of conditional distributions [14] and multiple choice learning [63, 42, 43]. In this paper, we propose to integrate annealing for conditional quantization.

3 Annealed Multiple Choice Learning

3.1 Winner-takes-all loss and its limitations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote subsets of Euclidean vector spaces. We are interested in so-called *ambiguous tasks*, *i.e.*, for any given input $x \in \mathcal{X}$, there may be several plausible outputs $y \in \mathcal{Y}$. Formally, let $p(x, y)$ denote a joint distribution on $\mathcal{X} \times \mathcal{Y}$. Multiple Choice Learning (MCL) [25, 41] was proposed to train neural networks in this setting, and has proven its effectiveness in a wide range of machine vision [40], natural language [22] and signal processing tasks [42].

[1] https://github.com/Victorletzelter/annealed_mcl

MCL consists in training several predictors $(f_1, \dots, f_n) \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$, typically a multi-head neural network derived from a common backbone, such that for each input $x \in \mathcal{X}$, the predictions $(f_1(x), \dots, f_n(x))$ provide an efficient *quantization* of the conditional distribution $p(y \mid x)$ [63, 42, 43]. This goal is achieved by minimizing the *distortion*

$$D(f) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \min_{k\in[\![1,n]\!]} \ell\left(y, f_k(x)\right) p(x,y)\mathrm{d}x\mathrm{d}y \;, \tag{1}$$

where $\ell : \mathcal{Y}^2 \to \mathbb{R}$ is an underlying loss function, for instance, the squared Euclidean distance $\ell(\hat{y}, y) = \|y - \hat{y}\|^2$. Eq. (1) can be seen as a generalization of the conditional distortion [56].

More specifically, MCL training is an iterative procedure optimizing (1) by alternating the two following steps.

1. Assign each y to the closest hypothesis $f_k(x)$ to build the Voronoi cells:

$$\mathcal{Y}_k(x) \triangleq \{y \in \mathcal{Y} \mid \forall l \in [\![1,n]\!],\ \ell(y, f_k(x)) \leq \ell(y, f_l(x))\} \;. \tag{2}$$

2. Minimize the distortion within each cell by taking a gradient step on the WTA loss:

$$\mathcal{L}^{\mathrm{WTA}}(f) \triangleq \int_{\mathcal{X}} \sum_{k=1}^{n} \left(\int_{\mathcal{Y}_k(x)} \ell(f_k(x), y) p(y \mid x)\mathrm{d}y \right) p(x)\mathrm{d}x \;. \tag{3}$$

The prediction models can be paired with scoring models $(\gamma_1, \dots, \gamma_n) \in \mathcal{F}(\mathcal{X}, [0,1]^n)$, which are trained to estimate the Voronoi regions' probability mass using the scoring loss [42]

$$\mathcal{L}^{\mathrm{scoring}}(\gamma) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} \mathrm{BCE}\left(\mathbb{1}\left[y \in \mathcal{Y}_k(x)\right], \gamma_k(x)\right) p(x,y)\mathrm{d}x\mathrm{d}y \,, \tag{4}$$

where $\mathrm{BCE}(p,q) \triangleq -p\log(q) - (1-p)\log(1-q)$. In practice, the two losses (3) and (4) are optimized in a compound objective $\mathcal{L} = \mathcal{L}^{\mathrm{WTA}} + \mathcal{L}^{\mathrm{scoring}}$. A probabilistic interpretation of such trained predictors has been developed [63, 42]. It shows that the predictions and scores can be interpreted as a mixture model approximating the conditional density $p(y \mid x)$ by $\sum_{k=1}^{n} \gamma_k(x)\delta_{f_k(x)}(y)$.

It has been shown that WTA is sensitive to initialization [54], and often leads to suboptimal hypothesis positions, similarly to K-means. Indeed, WTA is a greedy procedure that updates only the best hypotheses: if one hypothesis falls outside the support of the data density $p(\cdot \mid x)$, it may be isolated from its competitors at initialization, and remain so across training (the *collapse* issue).

Our method improves the WTA training scheme by addressing the inherent greediness of gradient descent and introducing variability in the exploration of the hypothesis space through deterministic annealing. Figure 1 illustrates the limitations of the aforementioned algorithms and the comparative advantage of aMCL.

3.2 Combining deterministic annealing with Multiple Choice Learning

We introduce aMCL, which combines MCL and annealing. Let $t \mapsto T(t)$ denote a temperature schedule decreasing with the training step t, and vanishing at the end of the training. Similarly to MCL, aMCL alternates between an assignation and a minimization step, as follows:

1. Softly assign each y to all $f_k(x)$ using the $\mathrm{softmin}$ operator (or Boltzman distribution $q_{T(t)}$):

$$q_{T(t)}(f_k \mid x, y) \triangleq \frac{1}{Z_{x,y}} \exp\left(-\frac{\ell(f_k(x), y)}{T(t)}\right), \quad Z_{x,y} \triangleq \sum_{s=1}^{n} \exp\left(-\frac{\ell(f_s(x), y)}{T(t)}\right), \tag{5}$$

2. Minimize the distortion within each soft cell by taking a gradient step on the aWTA loss:

$$\mathcal{L}^{\mathrm{aWTA}}_{T(t)}(f) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} \ell(f_k(x), y) q_{T(t)}(f_k \mid x, y) p(x,y)\mathrm{d}x\mathrm{d}y \;, \tag{6}$$

where $q_{T(t)}$ is kept constant (*i.e.,* the $\mathrm{stop\_gradient}$ operator is applied).

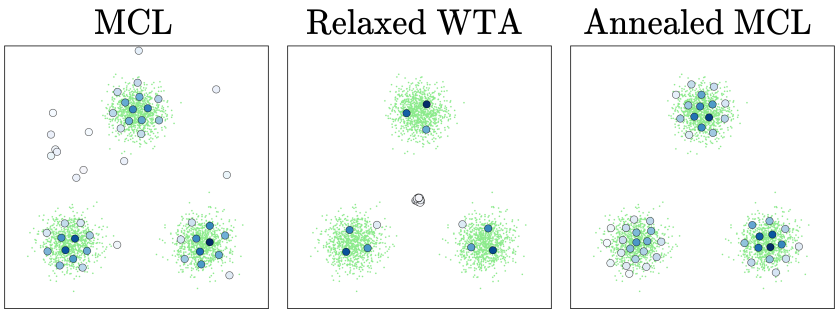


Figure 1: **Overcoming limitations of Winner-Takes-All training with annealing**. Illustrations of the test-time predictions on a Mixture of three Gaussians (green points) with 49 hypotheses. Shaded blue circles represent the hypothesis predictions, with intensity corresponding to the predicted scores. *(Left)* Predictions of MCL as proposed in [41, 42]. *(Middle)* Predictions of Relaxed WTA [63] with $\varepsilon = 0.1$. *(Right)* Annealed MCL with initial temperature $T_0 = 0.6$. Each model was trained with the same backbone (a three-layer MLP). We see that WTA leaves out some hypotheses, achieving a higher quantization error than aMCL. Moreover, we see that Relaxed-WTA is biased toward the barycenter of the distribution, in contrast with aMCL.

Therefore, at the lowest level, aMCL simply consists of replacing the $\min$ operator from (2) by $\mathrm{softmin}$. aMCL introduces the temperature schedule as an additional hyperparameter. As highlighted by the literature on simulated annealing [26], it is crucial to ensure that the temperature decreases slowly enough to benefit from the advantages of annealing. In practice, we experimented with both linear and exponential schedulers (see also Section 5).

On a higher level, we can interpret the objective of aMCL as a smoothed version of the MCL objective. Smoothing with high temperature simplifies the optimization problem (6), making the loss landscape easier to navigate: we can conjecture from this analysis that aMCL will find a global minimum at high temperature, and we can expect it to stay optimal as long as the temperature decreases slowly enough [9]. We can also see aMCL as an input-dependent version of deterministic annealing [61, 62]. In this view, a high temperature encourages the exploration of the hypothesis space and mitigates the greediness of the gradient descent update (3). Moreover, following [26], we can posit that there exists an optimal temperature schedule striking a balance between exploration and optimization. Yet another interpretation is that aMCL constitutes an adaptative extension of Relaxed-MCL [63], as $q_{T(t)}(f_k \mid x, y)$ depends both on the distance between the hypothesis $f_k(x)$ and the target y, and the training step t. These interpretations shed light on the inner workings of aMCL. However, the complete training dynamic of the algorithm appears when we analyze aMCL through the lens of information theory and statistical physics, which is the purpose of the next section.

4 Theoretical analysis

In this Section, we theoretically investigate the properties of our algorithm. Specifically, we detail its training dynamic in Section 4.1. In Section 4.2, we explore how this dynamic relates to the rate-distortion curve. This relationship allows us to study in Section 4.3 the phenomenon of *phase transition*, where hypotheses merge and split into sub-groups depending on the temperature. Throughout this Section, we will focus on the squared Euclidean distance $\ell(\hat{y}, y) = \|y - \hat{y}\|^2$.

4.1 Soft assignation and entropy constraints

Minimizing (1) is NP-hard [1, 10, 48]. Unsurprisingly, MCL can get trapped in local minima during training. In this Section, we discuss why the aMCL training scheme in Section 3.2 is more resilient to this pitfall. The first step toward our analysis is to observe that the $\mathrm{stop\_gradient}$ operator used in (6) of the aMCL update effectively turns the algorithm into an alternating optimization of the soft distortion

$$D(q, f) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} \ell(f_k(x), y)\; q(f_k|x, y)\; p(x, y)\mathrm{d}x\mathrm{d}y \;, \tag{7}$$

where the variables q and f are treated as independent, a procedure similar to Expectation Maximization [12]. This observation is captured by Proposition 1, where $\mathcal{F} = \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$ denotes the set of functions from $\mathcal{X}$ to $\mathcal{Y}^n$, Δ_n the set of all distributions on n items conditioned by points on $\mathcal{X} \times \mathcal{Y}$, $H(\pi) = -\sum_{k=1}^{n} \pi_k \log \pi_k$ the entropy of a discrete distribution π, $H_T = H(q_T)$ the entropy of the Boltzmann distribution at temperature T, and λ_t the learning rate of the gradient descent at step t.

Proposition 1 (Entropy-constrained alternated minimization)**.** *The assignation* (5) *and optimization* (6) *steps of aMCL correspond to an entropy-constrained block coordinate descent on the soft distortion.*

$$q \leftarrow \underset{\substack{q \in \Delta_n \\ H(q) \geq H_{T(t)}}}{\operatorname{argmin}} \; D(q, f) \;, \qquad f_k \leftarrow f_k - \lambda_t \nabla_{f_k} D(q, f) \;, \quad \forall k \in [\![1, n]\!]. \tag{8}$$

This is a corollary of Proposition 2, which provides additional insights on the training dynamics of aMCL.

Proposition 2 (aMCL training dynamic)**.** *See Proposition 5 in Appendix. The following statements are true for all $T > 0$, $f \in \mathcal{F}$, strictly positive $q \in \Delta_n$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

$$(i) \quad \underset{\substack{q \in \Delta_n \\ H(q) \geq H_T}}{\operatorname{argmin}} D(q, f) = q_T \;, \qquad q_T(f_k | x, y) = \frac{\exp\left(-\ell(f_k(x), y)\,/T\right)}{\sum_{s=1}^{n} \exp\left(-\ell(f_s(x), y)\,/T\right)} \;, \quad \forall k \in [\![1, n]\!]$$

$$(ii) \quad \operatorname{argmin}_{f \in \mathcal{F}} D(q, f) = f^\star \;, \qquad f_k^\star(x) = \frac{\int_{\mathcal{Y}} y \; q(f_k^\star | x, y) p(y \mid x) \mathrm{d}y}{\int_{\mathcal{Y}} q(f_k^\star | x, y) p(y \mid x) \mathrm{d}y} \;, \qquad \forall k \in [\![1, n]\!]$$

$$(iii) \quad \nabla_{f_k} D(q, f) = \gamma_k^\star (f_k - f_k^\star) \;, \quad \gamma_k^\star = \int_{\mathcal{Y}} q(f_k^\star | x, y) p(y \mid x) \mathrm{d}y \;, \qquad \forall k \in [\![1, n]\!]$$

Part (i) states that the $\mathrm{softmin}$ operator is the solution of the entropy-constrained minimization of the soft distortion. Part (ii) states that a necessary condition for minimizing the soft distortion is that each f_k is a soft barycenter of the assignation distribution for each temperature T. Part (iii) states that each gradient update moves f_k toward this soft barycenter $f^\star$, and that the update speed depends on the probability mass $\gamma_k^\star$ of the points softly assigned to f_k. Together, they describe the training dynamics of aMCL. Note that as $T \to 0$, q_T converges to a one-hot vector, and the soft barycenter in (ii) becomes a hard barycenter. This is consistent with the necessary optimal condition for MCL, $f_k^\star(x) = \mathbb{E}_{Y \sim p(y|x)}[Y \mid Y \in \mathcal{Y}_k(x)]$, proved by Rupperecht *et al.* [63].

4.2 Rate-distortion curve

We have established in Section 4.1 that the aMCL training scheme is equivalent to an entropy-constrained alternating optimization of the soft-distortion (7), with each hypothesis f_k moving toward a soft barycenter. In this Section, we describe the impact of temperature cooling on this training dynamic.

First, observe that when the temperature is high, the Boltzmann distribution q_T becomes uniform. Therefore, the soft Voronoi cells merge into a single cell $\mathcal{Y}$, the hypotheses f_k converge toward the barycenter of $\mathcal{Y}$, and they all fuse into a single hypothesis:

$$f_k^\star(x) = \mathbb{E}_{(X,Y) \sim p(x,y)}[Y \mid X = x] \;, \quad \forall k \in [\![1, n]\!] \;. \tag{9}$$

Remarkably, this phenomenon occurs even at finite temperatures (see Appendix, Proposition 9). As the temperature decreases, a phenomenon of *bifurcation* occurs [61, 59]. During this process, the hypotheses iteratively split into sub-groups, as shown in Figure 4. The virtual number of hypotheses [60] for each x at a given distortion level is captured by the *conditional rate-distortion function*

$$R_x(D^\star) \triangleq \min_{\substack{q \in \Delta_n, f \in \mathcal{F} \\ D_x(q,f) \leq D^\star}} I_x(\hat{Y}; Y) \;. \tag{10}$$

In Eq. (10), $Y \sim p(y \mid x)$ follows the target distribution, the hypothesis position $\hat{Y} \sim q(f_k \mid x)$ follows a distribution over $\mathcal{Y}$ with $q(f_k \mid x) = \int_{\mathcal{Y}} q(f_k \mid x, y) p(y \mid x) \mathrm{d}x$, $I_x(\hat{Y}; Y)$ is their mutual information, and $D_x(q, f) = \int_{\mathcal{Y}} \sum_k \ell(f_k(x), y) q(f_k \mid x, y) p(y \mid x) \mathrm{d}y$ is the distortion for input x.

The rate-distortion function $R_x(D^\star)$ has the following key properties.

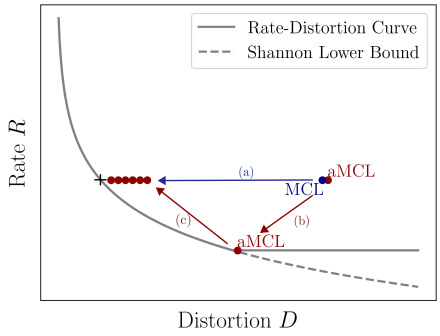


Figure 2: **Illustration of the training trajectory in the Rate-Distortion curve.** Training trajectories of MCL (blue) and aMCL (red) in the case of a single Gaussian. The optimal reachable distortion ('+') is the distortion $D^\star$ satisfying $R(D^\star) = \log_2(n)$ (See Section 4.2).

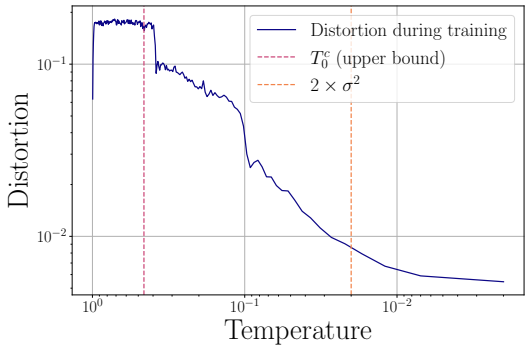


Figure 3: **Regimes in the distortion (1) vs. temperature training curve on the setup of Figure 4.** At first, the hypotheses converge to the conditional mean. It is followed by a plateau phase where performance stagnates. Transition begins at T_0^c: the hypotheses migrate toward the barycenter of each Gaussian. Then, they split and we observe a last phase transition. For reference, $2\sigma^2$ is the critical temperature for a Gaussian with variance σ^2.

Proposition 3 (Rate-distortion properties)**.** *See Proposition 7 in Appendix.*

For each $x \in \mathcal{X}$, let $D_x^{\max}$ (31) denote the optimal conditional distortion when using a single hypothesis, we have the following results.

(i) For each $T > 0$, minimizing the free energy

$$\mathscr{F} = D(q_T, f) - TH(q_T) \,, \tag{11}$$

over all hypotheses positions $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$ comes down to solving the optimization problem that defines (10) for each $x \in \mathcal{X}$.

(ii) R_x is a non-increasing, convex and continuous function of $D^\star$, $R_x(D^\star) = 0$ for $D^\star \geq D_x^{\max}$, and for each x the slope can be interpreted as $R'_x(D^\star) = -\frac{1}{T}$ when it is differentiable.

(iii) For each x, $R_x(D^\star)$ is bounded below by the Shannon Lower Bound (SLB)

$$R_x(D^\star) \geq \mathrm{SLB}(D^\star) \triangleq H(Y) - H(D^\star) \,, \tag{12}$$

where $Y \sim p(y \mid x)$ and $H(D^\star)$ is the entropy of a Gaussian with variance $D^\star$.

Part (i) establishes that the rate-distortion function is tightly linked with our problem. Provided that the hypotheses f_k perfectly optimize the soft distortion (7) at any temperature level, the hypothesis configuration f will follow the optimal parametric curve $(D^\star, R_x(D^\star))$ for each $x \in \mathcal{X}$. Part (ii) describes the shape of the parametric curve $(D^\star, R_x(D^\star))$. The Shannon Lower Bound described in part (iii) is a lower bound on the virtual number of hypotheses reached by aMCL.

The rate-distortion curve effectively describes the training trajectory of our algorithm, with the deterministic annealing procedure consisting of ascending along this curve [61, 59]. Interestingly, there is a set of critical temperatures at which the hypotheses suddenly split, increasing the number of sub-groups they form. By analogy with statistical physics, the behavior at these points has been coined *phase transitions* [23]. An illustration of the trajectory of MCL and aMCL in the rate-distortion space is shown in Figure 2. We see that MCL evolves along a constant rate $R = \log_2(n)$ (in bits) following (a). In contrast, aMCL initially reaches the critical state at $D^\star = D_x^{\max}$ following (b). After the transition, the trajectory of aMCL returns to the maximal rate following (c). We expect the optimization at a lower rate to be simpler and this training trajectory to provide a better initialization for the set of hypotheses compared to the vanilla MCL.

4.3 Phase transitions

During training, as the temperature decreases, the hypotheses f undergo a sequence of phase transitions at specific critical temperatures. The right tool to exhibit these critical temperatures is the Hessian of the free energy $f \mapsto \mathscr{F}(f, q_T)$. Transitions occur when the minimum of $\mathscr{F}$ is no longer stable, and it can be shown [61] that this relates to the eigenvalues of the following block-diagonal covariance matrix:

$$C_{k,k}(f, q|x) = \int_{\mathcal{Y}} (f_k(x) - y)(f_k(x) - y)^t \; q(y \mid f_k, x) p(y \mid x) \mathrm{d}y \; , \tag{13}$$

where $q(y \mid f_k, x)$ denotes the posterior probability of assigning a point y to the hypothesis k, calculated using Bayes's rule [60]. At high temperatures, all hypotheses merge into the conditional barycenter of the distribution (9). In this setting, all the matrices $C_{k,k}$ are equal to the data covariance matrix $C(x) \triangleq \mathbb{C}\mathrm{ov}_{(X,Y)\sim p(x,y)}[Y \mid X = x]$, $\mathscr{F}$ has a unique minimizer, and the stability of this global optimum is conditioned on the strict positivity of the Hessian of $\mathscr{F}$. The first critical temperature T_0^c is defined as the first temperature for which the Hessian of $\mathscr{F}$ is no longer positive definite.

This temperature is connected to $D_x^{\max}$, the vanishing point of the rate-distortion function $R_x(D^\star)$, which also corresponds to the optimal 1-hypothesis distortion [4]. Indeed, $R_x(D^\star)$ measures the virtual number of hypotheses, so that the first splitting of the hypotheses from a single point to several groups will coincide with the moment when $R_x(D^\star) > 0$. We summarize these observations in the following Proposition 4, which is illustrated in Figure 3.

Proposition 4 (First critical temperature)**.** *See Definition 1, Propositions 6 (ii) and 8 in Appendix. Let $\lambda_{\max}(\cdot)$ denote the maximum eigenvalue of a matrix, $D_x^*(T) \triangleq \inf_{f\in\mathcal{F}} D_x(q_T, f)$, $D^*(T) \triangleq \inf_{f\in\mathcal{F}} D(q_T, f)$, and $D_{\max} \triangleq \int_{x\in\mathcal{X}} D_x^{\max} p(x)\mathrm{d}x$. Then the two following properties hold.*

(i) D_x^* and D^* are non-decreasing functions of T and admit generalized inverses (with the convention $g^{-1}(\beta) = \inf\{\alpha \mid g(\alpha) \geq \beta\}$ for the generalized inverse of a function g).

(ii) The conditional (resp. non-conditional) first critical temperature $T_0^c(x)$ (resp. T_0^c) satisfy

$$T_0^c(x) \triangleq (D_x^*)^{-1}(D_x^{\max}) = 2\lambda_{\max}(C(x)) \; , \tag{14}$$

$$T_0^c \triangleq (D^*)^{-1}(D_{\max}) \leq 2 \sup_{x\in\mathcal{X}} \lambda_{\max}(C(x)) \; . \tag{15}$$

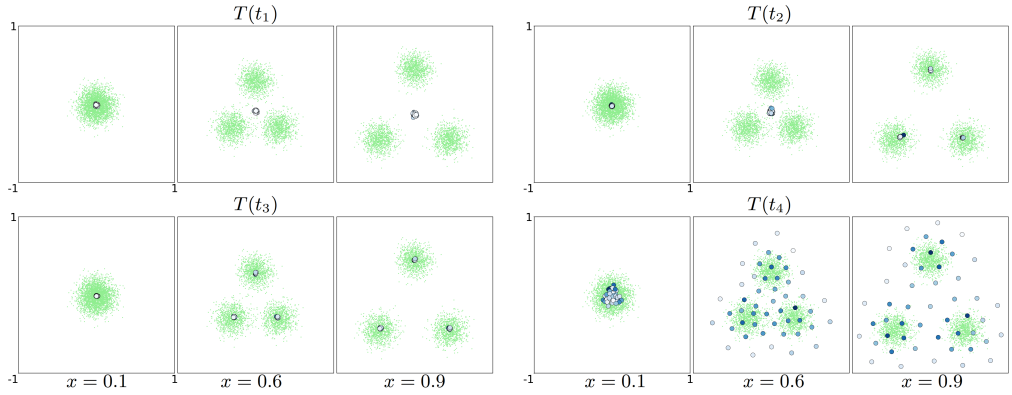


Figure 4: **Conditional phase transitions with $t_1 < t_2 < t_3 < t_4$.** Results for a conditional version of the dataset in Figure 1, where the Gaussian moves linearly and increases with the input $x \in [0, 1]$. aMCL was trained during 1000 epochs, using a linear scheduler with $T_0 = 1.0$, and using 49 hypotheses. Each subplot group corresponds to the predictions of the model at a given temperature, evaluated at different x values. At temperature $T(t_1)$, the hypotheses are at the barycenter. As temperature decreases they undergo a first phase transition (at temperature $T(t_2)$ for $x = 0.9$ and $T(t_3)$ for $x = 0.6$), moving toward each Gaussian's barycenter, followed by a second phase transition at $T(t_4)$. We see that earlier splits in the cooling schedule correspond to conditional distributions with higher variances.

After the first transition, an interesting phenomenon occurs for some distributions $p(x, y)$. Instead of splitting in all directions, the hypotheses f_k continue to form a small number of subgroups. The hypotheses f_k may undergo many phase transitions before they all split apart from each other: this is illustrated in Figure 4. Generally, this recursive splitting reaches an end: under mild conditions [36], there is a critical temperature below which the hypotheses all separate from each other. This defines the last critical temperature T_∞^c. Remarkably, T_∞^c is associated with the Shannon Lower Bound: the temperature at which $R_x(D^\star)$ hits the lower bound $\mathrm{SLB}(D^\star)$ corresponds to the moment when the hypotheses f_k completely separate from each other (see Theorem 1 and 3 in [59]).

5 Experimental validation

In this Section, we experimentally investigate the advantage of our algorithm in practical settings. Specifically, we evaluate it on the standard UCI benchmark in Section 5.1, and we apply it to the challenging task of speech separation in Section 5.2.

5.1 UCI datasets

5.1.1 Setup

General setup. We followed the experimental protocol described by [29] for the UCI benchmark [15]. Specifically, we used the official train-test splits, with 20 folds except for the Protein dataset, which is split into 5 folds, and the Year dataset, which uses a single fold.

Baselines. In our result tables, we also include data from three baseline methods detailed in Table 1 of Lakshminarayanan *et al.*'s paper [38] ('Deep Ensembles'), which serves as a reference. These baselines include Probabilistic Back Propagation [29] (denoted 'PBP'), and Monte Carlo Dropout [21] (denoted 'MC-dropout'). As additional baselines, we include the standard score-based MCL (*e.g.*, [42]). We also include the Relaxed-MCL variant [63] with $\varepsilon = 0.1$. The impact of ε is discussed in Appendix D.2. Our method (aMCL), was trained with an exponential scheduler of the form $T(t) = T_0\rho^t$, with $\rho = 0.95$ and $T_0 = 0.5$. Comparison with a linear scheduler is also provided in Appendix D.1. Both aMCL and Relaxed-MCL were trained for 1,000 epochs. Each MCL system was trained with $n = 5$ hypotheses.

Metrics. We computed the following metrics on the UCI datasets. Let d denote the squared Euclidean distance: $d(\hat{y}_i, y_i) = \|y_i - \hat{y}_i\|^2$, and N the number of samples in each dataset. The RMSE ($\downarrow$) is defined as $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i d(\hat{y}_i, y_i)}$, where $\hat{y}_i$ denotes the estimated conditional mean. The latter was computed with $\sum_{k=1}^n \gamma_k(x_i) f_k(x_i)$ for the MCL variants. The results of this experiment are summed up in Table 1 (Distortion), and also in Table 2 (RMSE). Rows are ordered by dataset size N, with the intensity of the grey color proportional to N (excluding the Year dataset).

5.1.2 Results

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 ($\downarrow$)			
Datasets	Relaxed-WTA ($\varepsilon = 0.1$)	MCL	aMCL	N
Year	7.73 ± NA	4.8 ± NA	**4.39 ± NA**	515345
Protein	1.67 ± 0.16	0.79 ± 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	2.36 ± 0.43	2.26 ± 0.38	**2.06 ± 0.45**	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	6.63 ± 2.51	**5.13 ± 1.2**	5.71 ± 1.72	1030
Energy	0.3 ± 0.12	1.25 ± 1.22	**0.28 ± 0.09**	768
Boston	3.32 ± 2.84	**2.14 ± 0.48**	2.69 ± 1.39	506
Yacht	1.34 ± 0.93	3.09 ± 2.35	**1.15 ± 0.97**	308

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].

	RMSE (↓)						
Datasets	PBP*	MC Dropout*	Deep Ensembles*	Relaxed-WTA ($\varepsilon = 0.1$)	MCL	aMCL	N
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-4	**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

Comparison of aMCL and MCL. We can observe that aMCL performs comparably or outperforms vanilla MCL in most settings, especially for large dataset sizes, both in terms of distortion and RMSE. This outcome supports the claims made in the paper and is especially promising, given that the temperature scheduler was not specifically optimized for each dataset.

Comparison of aMCL and standard UCI baselines. We observe that aMCL performs on par with, and in some cases exceeds, standard baselines on the RMSE metric. This is noteworthy, as aMCL is not explicitly optimized for RMSE during training—its primary focus is on quantization. While perfect quantization would naturally result in optimal RMSE, achieving low RMSE is not the main objective of aMCL. For example, the RMSE performance of MCL is significantly worse across most datasets.

Comparison with Relaxed-MCL. Finally, we compare aMCL and Relaxed-MCL, since aMCL can be interpreted as an adaptative version of Relaxed-MCL. Our results indicate that aMCL generally outperforms Relaxed-MCL in terms of distortion across nearly all datasets. However, we also observe that Relaxed-MCL demonstrates strong performance in terms of RMSE. We attribute these findings to the bias of Relaxed-MCL toward the conditional barycenter of the target distribution, which seems to improve RMSE at the expense of increased distortion. This trade-off arises because RMSE evaluates the accuracy of the barycenter estimation, while distortion measures the quantization performance. The trade-off between distortion and RMSE can be adjusted by tuning the value of ε. Further analysis of this parameter is presented in Appendix D.1, where we show that aMCL strikes a good balance between these two metrics.

These results strongly support the use of aMCL as a quantization algorithm in practical settings. To further evaluate its effectiveness, we also apply aMCL to a more challenging task, namely speech separation.

5.2 Application to speech separation

Speech separation consists of isolating each speaker's signal from a mixture in which they are simultaneously active. This task is of major interest for automatic speech processing applications [50, 44, 71]. In these experiments, we explore the application of MCL and the proposed aMCL to the task of speech separation. An extensive description of the experiments is proposed in Appendix E.2.

5.2.1 Experimental setting

General purpose. Speech separation consists in obtaining the source signals $y_1, \dots, y_m \in \mathbb{R}^l$ from a mixture $x = \sum_{s=1}^{m} y_s$. Hence, the task is to provide estimates $\hat{y}_1, \dots, \hat{y}_m$ of the isolated speech tracks from the mixture x.

Dataset. Source separation experiments are conducted on the Wall Street Journal dataset [30] (WSJ0-mix), a standard benchmark for speech separation. We focus on the 2- and 3- speaker mixture scenarios, with each scenario including 20,000 training, 5,000 validation, and 3,000 testing mixtures.

Model architecture. The source separation task is solved using the Dual-Path Recurrent Neural Network (DPRNN) architecture [46]. DPRNN has been extensively used in speech separation, as it strikes a good balance between performance and number of trainable parameters [73, 65].

Separation metrics. We use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) to measure the separation quality [70, 39]. There is an ambiguity in finding the best assignment between predicted and active sources. The PIT SI-SDR loss [37] initially addresses this issue. We propose to use MCL and our new variant aMCL to perform this matching (See Appendix E).

5.2.2 Results

Comparing PIT, MCL, Relaxed-WTA and aMCL. Table 3 presents the source separation results in the 2- and 3-speaker scenarios. First, aMCL demonstrates performance equivalent to or better than MCL. Both methods can be used for the separation task. However, we observed that MCL is subject to hypothesis collapse for some training seeds, while aMCL is more robust to initialization. This translates into a lower inter-seed standard deviation for aMCL. Second, we observe the advantage of aMCL over Relaxed-WTA, which is consistent with our previous analysis of the barycenter bias of this method. Third, aMCL performs equivalently to PIT in both scenarios. Note that by using MCL or aMCL, the number of predictions n could exceed the number of sources m. This could be leveraged to improve the separation metrics (cf. Appendix E.3.2). Finally, aMCL improves the algorithmic complexity of PIT from $\mathcal{O}(m^3)$ to $\mathcal{O}(mn)$ (cf. Appendix E.3.1). These results make aMCL stand as a good alternative to PIT.

Table 3: **2- and 3- speaker source separation** with PIT (topline), MCL, aMCL and Relaxed-WTA ($\varepsilon = 0.05$). PIT SI-SDR metric ($\uparrow$) on the WSJ0-mix eval set. Results over three training seeds, with mean and standard deviation reported.

Method	2 speakers	3 speakers
PIT	16.88 ± 0.10	10.01 ± 0.04
MCL	16.30 ± 0.59	10.06 ± 0.21
Relaxed-WTA	16.70 ± 0.08	9.43 ± 0.21
aMCL	16.85 ± 0.13	10.00 ± 0.21

Observing phase transition. When the metric is the Euclidean distance, the theoretical analysis of Section 4.3 and the synthetic experiments (see Figure 3) have highlighted a phenomenon of phase transition for aMCL. Here, we analyze the validation loss trajectory as a function of the temperature for different initial temperatures, and using an exponential scheduler. The curves with the two higher initial temperatures in Figure 8 exhibit a plateau until a given temperature. After this critical point, the loss decreases. Although the SI-SDR is non-Euclidean, this behavior resembles that observed for the Euclidean metric. This is detailed in Section E.3.3 of the Appendix.

6 Conclusion

This article introduces aMCL, a novel training method that combines deterministic annealing and the Winner-Takes-all training scheme to address two key issues of MCL: hypothesis collapse and convergence toward a suboptimal local minimum of its quantization objective. We provide a detailed analysis of aMCL's training dynamics. Moreover, drawing on statistical physics and information theory, we provide insights into the trajectory of the aMCL predictions during training. In particular, we exhibit a phase transition phenomenon and establish its connection to the rate-distortion curve. We validate our analysis with experiments on synthetic data, on the UCI datasets, and on a real-world speech separation task. This demonstrates that aMCL is a theoretically grounded alternative to MCL in diverse settings. Future work includes a detailed analysis of the temperature schedule's impact, the derivation of performance bounds, a thorough examination of our algorithm's convergence, particularly at finite temperature, as well as an evaluation of its generalization capabilities on out-of-distribution samples.

Limitations. First, aMCL introduces a temperature schedule: this is a challenging hyperparameter tightly linked to the optimizer and its learning rate. The derivation of optimal schedules is left to future work. Second, annealing requires a slow temperature schedule to maintain model performance. This potentially leads to longer training times.

Broaden impact. This paper introduces research aimed at progressing the field of Machine Learning. While our work has numerous potential societal implications, we believe there are no specific consequences that need to be emphasized in this paper.

7 Acknowledgments

This work was funded by the French Association for Technological Research (ANRT CIFRE contract 2022-1854) and Hi! PARIS through their PhD in AI funding programs, and was performed using HPC resources from GENCI–IDRIS (Grant 2021-AD011013406R1). We are grateful to the reviewers for their insightful comments.

References

[1] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 2009. 4, 15

[2] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 1972. 2

[3] David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2007. 1

[4] Toby Berger. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*, 2003. 2, 7, 20

[5] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 1972. 2

[6] Richard E Blahut. *Principles and practice of information theory*. Addison-Wesley Longman Publishing Co., Inc., 1987. 20

[7] David P Bourne and Steven M Roper. Centroidal power diagrams, lloyd's algorithm, and applications to optimal location problems. *SIAM Journal on Numerical Analysis*, 2015. 15

[8] Mike Brodie, Chris Tensmeyer, Wes Ackerman, and Tony Martinez. Alpha model domination in multiple choice learning. In *IEEE International Conference on Machine Learning and Applications, ICMLA*, 2018. 2

[9] Imre Csiszár. On the computation of rate-distortion functions (corresp.). *IEEE Transactions on Information Theory*, 1974. 4

[10] Sanjoy Dasgupta. The hardness of k-means clustering. 2008. 4, 15

[11] L Davisson. Rate distortion theory: A mathematical basis for data compression. *IEEE Transactions on Communications*, 1972. 20

[12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1977. 5

[13] Qiang Du, Maria Emelianenko, and Lili Ju. Convergence of the lloyd algorithm for computing centroidal voronoi tessellations. *SIAM Journal on Numerical Analysis*, 2006. 15

[14] Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 1999. 2

[15] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017. 8, 24

[16] Jack Edmonds and Richard M Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM, JACM*, 1972. 29

[17] Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 2013. 20

[18] Maria Emelianenko, Lili Ju, and Alexander Rand. Nondegeneracy and weak global convergence of the lloyd algorithm in $\mathbb{R}^d$. *SIAM Journal on Numerical Analysis*, 2008. 15

[19] Michael Firman, Neill DF Campbell, Lourdes Agapito, and Gabriel J Brostow. Diversenet: When one right answer is not enough. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2018. 2

[20] Pasi Fränti and Sami Sieranoja. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 2019. 1

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian international workshop on approximation algorithms for combinatorial optimization, approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning, ICML*, 2016. 8, 24

[22] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, 2021. 2

[23] Thore Graepel, Matthias Burger, and Klaus Obermayer. Phase transitions in stochastic self-organizing maps. *Physical Review E*, 1997. 6

[24] Robert M Gray. *Source coding theory*. Springer Science & Business Media, 1989. 20

[25] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in Neural Information Processing Systems, NeurIPS*, 2012. 1, 2

[26] Bruce Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 1988. 2, 4, 22, 23

[27] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 24

[28] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970. 1, 2, 22, 23

[29] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning, ICML*, 2015. 8, 24

[30] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2016. 9, 27

[31] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2012. 26

[32] Mikaela Iacobelli. Asymptotic quantization for probability measures on riemannian manifolds. *ESAIM: Control, Optimisation and Calculus of Variations*, 2016. 20

[33] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision, ECCV*, 2018. 2

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015. 24

[35] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 1983. 1, 2

[36] Tobias Koch. The shannon lower bound is asymptotically tight. *IEEE Transactions on Information Theory*, 2016. 8

[37] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017. 10

[38] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems, NeurIPS*, 2017. 8, 9, 24

[39] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019. 10, 26

[40] Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. In *International Conference on Machine Learning, ICML*, 2017. 2

[41] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems, NeurIPS*, 2016. 1, 2, 4

[42] Victor Letzelter, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gael Richard. Resilient multiple choice learning: A learned scoring scheme with application to audio scene analysis. *Advances in Neural Information Processing Systems, NeurIPS*, 2023. 2, 3, 4, 8, 15, 16, 23

[43] Victor Letzelter, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Perez. Winner-takes-all learners are geometry-aware conditional density estimators. In *International Conference on Machine Learning, ICML*, 2024. 2, 3, 15, 23

[44] Chenda Li, Jing Shi, Wangyou Zhang, Aswin Shanmugam Subramanian, Xuankai Chang, Naoyuki Kamo, Moto Hira, Tomoki Hayashi, Christoph Boeddeker, Zhuo Chen, et al. Espnet-se: End-to-end speech enhancement and separation toolkit designed for asr integration. In *IEEE Spoken Language Technology Workshop, SLT*, 2021. 9, 26

[45] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 1982. 1

[46] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020. 9, 27, 28

[47] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019. 27, 28

[48] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical Computer Science*, 2012. 4, 15

[49] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 1, 2, 29

[50] Amparo Marti, Maximo Cobos, and Jose J Lopez. Automatic speech recognition in cocktail-party situations: A specific training for separated speech. *The Journal of the Acoustical Society of America, JASA*, 2012. 9, 26

[51] Neri Merhav. Rate–distortion function via minimum mean square error estimation. *IEEE Transactions on Information Theory*, 2011. 20

[52] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 1953. 1, 2, 22, 23

[53] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012. 22

[54] Sriram Narayanan, Ramin Moslemi, Francesco Pittaluga, Buyu Liu, and Manmohan Chandraker. Divide-and-conquer for lane-aware diverse trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021. 1, 2, 3

[55] Elias Nehme, Rotem Mulayoff, and Tomer Michaeli. Hierarchical uncertainty exploration via feedforward posterior trees. *Advances in Neural Information Processing Systems, NeurIPS*, 2024. 22, 23

[56] Gilles Pagès and Jacques Printems. Optimal quadratic quantization for numerics: the gaussian case. *Monte Carlo Methods and Applications*, 2003. 3, 15

[57] Gilles Pages and Jun Yu. Pointwise convergence of the lloyd i algorithm in higher dimension. *SIAM Journal on Control and Optimization*, 2016. 15

[58] José M Pena, Jose Antonio Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 1999. 1

[59] Kenneth Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 1994. 2, 5, 6, 8, 20, 22

[60] Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 1998. 2, 5, 7, 21

[61] Kenneth Rose, Eitan Gurewitz, and Geoffrey C Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 1990. 2, 4, 5, 6, 7, 15, 21

[62] Kenneth Rose, Eitan Gurewitz, and Geoffrey C Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 1992. 2, 4, 15

[63] Christian Rupprecht, 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 *International Conference on Computer Vision, ICCV*, 2017. 1, 2, 3, 4, 5, 8, 15, 23, 29

[64] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *International Convention Record, IRE*, 1959. 2, 20

[65] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021. 9

[66] Hideyuki Tachibana. Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021. 29

[67] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 2

[68] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis. Sudo rm-rf: Efficient networks for universal audio source separation. In *International Workshop on Machine Learning for Signal Processing, MLSP*, 2020. 28

[69] Andrea Vattani. K-means requires exponentially many iterations even in the plane. In *Annual symposium on Computational geometry, SoCG*, 2009. 15

[70] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2006. 10, 26

[71] Thilo Von Neumann, Keisuke Kinoshita, Marc Delcroix, Shoko Araki, Tomohiro Nakatani, and Reinhold Haeb-Umbach. All-neural online source separation, counting, and diarization for meeting analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019. 9, 26

[72] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018. 26

[73] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023. 9

[74] Kenneth Zeger, Jacques Vaisey, Allen Gersho, et al. Globally optimal vector quantizer design by stochastic relaxation. *IEEE Transactions on Signal Processing*, 1992. 23

Organisation of the Appendix

The Appendix is organized as follows. Appendix A presents the theoretical analysis of our algorithm. It begins with the introduction of the notations and the definition of the training scheme in Appendices A.1 and A.2 respectively. This is followed by an interpretation of the algorithm in terms of entropy-constrained alternate optimization in Appendix A.3. Appendix A.4 provides an analysis of the training dynamics of the algorithm in relation to rate-distortion theory, and Appendix A.5 discusses phase transitions. Connection with the literature and additional discussions are provided in Appendix B. Additional details and results from experiments on synthetic data and UCI datasets are provided in Appendices C and D. Finally, Appendix E offers an extensive description of the Source Separation experiment, including the impact of the number of hypotheses in Appendix E.3.2, and the analysis of phase transitions for this task in Appendix E.3.3.

A Theoretical analysis

A.1 Notations and motivations

Following the main paper notations, let $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ denote the input and target spaces respectively. We will assume that $\mathcal{X}$ and $\mathcal{Y}$ are bounded. We note $p(x, y)$ the continuous density of a joint data distribution on $\mathcal{X} \times \mathcal{Y}$. Let $f = (f_1, \dots, f_n) \in \mathcal{F}_\Theta \subset \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$ denote the hypothesis models, which are families of n neural networks with parameters in Θ. Likewise, let $\gamma = (\gamma_1, \dots, \gamma_n) \in \mathcal{F}_{\Theta'} \subset \mathcal{F}(\mathcal{X}, [0, 1]^n)$ denote the score models, with parameters in Θ'. We will sometimes write $\mathcal{F} = \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$.

We denote by Δ_n the set of discrete distributions on n items conditioned by points on $\mathcal{X} \times \mathcal{Y}$ (typically representing a soft assignment of a target to the hypotheses). We denote by $H(q)$ the entropy of a distribution q. If $q \in \Delta_n$, we write $H(q) = -\int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^n q(f_k|x, y) \log q(f_k|x, y) p(x, y) \mathrm{d}x\mathrm{d}y$. If Z is a random variable with density $p(z)$ and Y is another random variable, we define the differential entropy of Z as $H(Z) = -\mathbb{E}_Z[\log(p(Z))]$ and its conditional entropy as $H(Z \mid Y) = -\mathbb{E}_{(Z,Y)}[\log(p(Z \mid Y))]$. Mutual information $I(Z, Y)$ between two random variables Z and Y is defined as $I(Z; Y) = H(Z) - H(Z \mid Y)$.

Let $t \in \mathbb{N}$ denote a training step, and $t_{\text{epoch}} \in \mathbb{N} \cup \{\infty\}$ denote the number of training epochs. Let $T : t \mapsto T(t) \geq 0$ be a temperature schedule that decreases with the training step and verifies $\lim_{t \to t_{\text{epoch}}} T(t) = 0$. Note that we consider the temperature to be constant across an epoch, with linear or exponential decrease between epochs. Let $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$ denote an underlying loss function. We will restrict our analysis to the Euclidean squared distance $\ell(\hat{y}, y) = \|y - \hat{y}\|^2$ unless otherwise stated. For any vector z, we note z^t as its vector transposition. Gradient descent involves a learning rate schedule, that we note λ_t.

Multiple Choice Learning (MCL) aims at training the hypotheses $(f_1, \dots, f_n)$ such that, for each input $x \in \mathcal{X}$, the hypothesis predictions $(f_1(x), \dots, f_n(x))$ provide an efficient *quantization* of the conditional distribution $p(y \mid x)$ [63, 42, 43]. The quality of the quantization can be evaluated using the *distortion*:

$$D(f) = \int_{\mathcal{X}\times\mathcal{Y}} \min_{k \in [\![1,n]\!]} \ell(f_k(x), y) p(x, y) \mathrm{d}x\mathrm{d}y \ , \tag{16}$$

which can be seen as a generalization of the conditional distortion [56]. The two definitions coincide when $p(y \mid x)$ is independent of x. Minimizing (16) is NP-hard [1, 10, 48]. The K-means algorithm, which is the standard approach for this task, can take an exponential number of steps before converging [69], and is only guaranteed to find a local minimum in general settings [13, 7]. The same applies to WTA: due to its greedy nature, it can be trapped, at any point of its training trajectory, into a local minimum that it will never manage to escape. One common issue is the *hypothesis collapse* [57, 18], a situation in which some of the hypotheses are left unused, therefore leading to a suboptimal configuration (see Figure 1).

To mitigate these issues, we introduce an annealed version of MCL. It can be seen as an input-dependent version of deterministic annealing [61, 62]. Our training scheme is described in Appendix A.2.

A.2 Training scheme of aMCL

Let (x, y) be a training example sampled from p at step t. Our proposed annealed MCL training process is defined through the following subsequent steps:

1. Perform a forward pass through $f_1(x), \dots, f_n(x)$.
2. Compute the Boltzmann distribution $q_{T(t)}$ defined for each f_k by

$$q_{T(t)}(f_k \mid x, y) \triangleq \frac{1}{Z_{x,y}} \exp\left(-\frac{\ell(f_k(x), y)}{T(t)}\right) , \quad Z_{x,y} \triangleq \sum_{k=1}^{n} \exp\left(-\frac{\ell(f_k(x), y)}{T(t)}\right) , \tag{17}$$

   and **detach** it from the computational graph.
3. Perform a gradient step on the loss

$$\mathcal{L}^{\text{aWTA}} \triangleq \sum_{k=1}^{n} q_{T(t)}(f_k \mid x, y)\ell(f_k(x), y) , \tag{18}$$

Note that we used the same scoring loss $\mathcal{L}_{\text{scoring}}$ (4) as in rMCL, the score-based method in [42].

In this paper, we extend the distortion $D(f)$ to account for a distribution q on the hypotheses, and define the resulting *soft distortion* as

$$D(q, f) = \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} \ell(f_k(x), y))q(f_k \mid x, y)p(x, y)\mathrm{d}x\mathrm{d}y . \tag{19}$$

When $q = q_{T(t)}$, the soft distortion $D(q_{T(t)}, f)$ corresponds to the expectation of the aWTA loss (18).

We will now introduce a *decoupling* assumption, which states that the family of models $\mathcal{F}_\Theta$ is expressive enough to encompass the global minimizer of the soft distortion. Noting $D_x(q, f)$ the integrand of (19) over the $\mathcal{X}$ integral, we can formalize this assumption as follows.

Assumption 1 (Decoupling)**.** *We assume that the model family is perfectly expressive, i.e.* $\mathcal{F}_\Theta = \mathcal{F}$.

Note that under this assumption, the global minimizer of the soft distortion $D(q, f)$, minimizes the integrand of (19) for all $x \in \mathcal{X}$:

$$\forall q \in \Delta_n, \quad \inf_{f\in\mathcal{F}_\Theta} D(q, f) = \inf_{f\in\mathcal{F}(\mathcal{X},\mathcal{Y}^n)} D(q, f) = \int_{\mathcal{X}} \inf_{f\in\mathcal{F}(\mathcal{X},\mathcal{Y}^n)} D_x(q, f)\ p(x)\mathrm{d}x . \tag{20}$$

In the subsequent Sections, our goal is to analyze the aMCL training scheme and justify its design. Specifically, we show that annealing simplifies the optimization problem defined by (16). In Appendix A.3, we first focus on studying the properties of the soft distortion $D(q, f)$.

In the following, if $x \in \mathcal{X}$ is fixed and when the context is clear, we will omit the dependency on the input and denote the hypotheses position $(f_1, \dots, f_n) \in \mathcal{Y}^n$.

A.3 Soft assignation and entropy constraint

In this Section, we rewrite the assignation and optimization steps of aMCL as an entropy-constrained block optimization of the soft distortion $D(q, f)$. Noting $H_T = H[q_T]$, we formalize this result with the following Theorem.

Theorem 2 (Entropy constrained distortion minimization)**.** *The assignation* (17) *and optimization* (18) *steps of aMCL correspond to a block coordinate descent on the soft distortion. For all* $k \in [\![1, n]\!]$*,*

$$q \leftarrow \operatorname*{argmin}_{\substack{q\in\Delta_n \\ H(q)\geq H_{T(t)}}} D(q, f) , \tag{21}$$

$$f_k \leftarrow f_k - \lambda_t \nabla_{f_k} D(q, f) . \tag{22}$$

Theorem 2 is a corollary of the following Proposition, which provides additional intuition into the dynamic of aMCL.

Proposition 5 (aMCL training dynamic)**.** *For all $T > 0$, $f \in \mathcal{F}$, strictly positive $q \in \Delta_n$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the following statements are true.*

$$(i) \quad \operatorname*{argmin}_{\substack{q\in\Delta_n \\ H(q)\geq H_T}} D(q,f) = q_T \;, \qquad q_T(f_k|x,y) = \frac{\exp\left(-\ell(f_k(x),y)\,/T\right)}{\sum_{s=1}^n \exp\left(-\ell(f_s(x),y)\,/T\right)} \;, \quad \forall k \in [\![1,n]\!]$$

$$(ii) \quad \operatorname{argmin}_{f\in\mathcal{F}} D(q,f) = f^\star \;, \qquad f_k^\star(x) = \frac{\int_{\mathcal{Y}} y\; q(f_k^\star|x,y)p(y\mid x)\mathrm{d}y}{\int_{\mathcal{Y}} q(f_k^\star|x,y)p(y\mid x)\mathrm{d}y} \;, \qquad \forall k \in [\![1,n]\!]$$

$$(iii) \quad \nabla_{f_k} D(q,f) = \gamma_k^\star(f_k - f_k^\star) \;, \quad \gamma_k^\star = \int_{\mathcal{Y}} q(f_k^\star|x,y)p(y\mid x)\mathrm{d}y \;, \qquad \forall k \in [\![1,n]\!]$$

Part (i) states that the $\mathrm{softmin}$ operator is the solution of the entropy-constrained minimization of the soft distortion. Therefore, gradient-based optimization of $D(q,f)$ along axis q is unnecessary, and we can directly use the closed-form expression in (i) for each temperature level $T(t)$, or equivalenty each corresponding entropy level $H_{T(t)}$. Consequently, the optimization of $D(q,f)$ reduces to the gradient-based minimization $D(q_{T(t)},f)$ along axis f for each temperature level $T(t)$.

Part (ii) states that a necessary condition for the minimization of the soft distortion is that each f_k is a soft barycenter of the assignation distribution $q_{T(t)}$ for each temperature T. Part (iii) states that each gradient update moves f_k towards this soft barycenter $f_k^\star$, and that the update speed depends on the probability mass $\gamma_k^\star$ of the points softly assigned to f_k. Together, these results describe the training dynamics of aMCL. Let us prove these results.

Proof. (i) First observe that under the decoupling assumption, minimizing $D(q,f)$ amounts to minimizing the integrand $D_x(q,f) = \int_{\mathcal{Y}} \sum_{k=1}^n \ell(f_k(x),y))q(f_k\mid x,y)p(y\mid x)\mathrm{d}y$ for each $x \in \mathcal{X}$ (see (20)). When x is fixed, the distribution q becomes a tuple of n scalars $(q_1,\ldots,q_n)$, and the Lagragian of (21) writes

$$\mathcal{L} = D(q,f) - T[H(q) - H_T] + \lambda\left(\sum_{k=1}^n q_k - 1\right) \;.$$

Using the first-order necessary optimization conditions, we find that

$$\ell(f_k,y) + T[\log(q_k)+1] + \lambda = 0 \,, \qquad H_T - H[q] = 0 \,,$$
$$\sum_{k=1}^n q_k - 1 = 0 \,.$$

From this it immediately follows that $\lambda = 1 + 1/\sum_{k=1}^n \exp\left(-\frac{\ell(f_k,y)}{T}\right)$ is a normalization factor, $q_k = \exp\left(-\frac{\ell(f_k,y)}{T}\right) / \sum_{s=1}^n \exp\left(-\frac{\ell(f_s,y)}{T}\right)$ is the Boltzmann distribution, and $H_T = H[q_T]$ correspond to its entropy.

(ii) Using the same reasoning than in (i) we set $x \in \mathcal{X}$. Then each $f_k \in \mathcal{Y}$ becomes a simple vector. Observing that $\nabla_{f_k} D_x(q,f)$ vanishes at the minimum of $D_x(q,f)$, a necessary condition to minimize $D(q,f)$ is

$$\nabla_{f_k} D_x(q,f) = \int_{\mathcal{Y}} \frac{\partial \ell}{\partial f_k}(f_k,y)\, q(f_k|x,y)p(y\mid x)\mathrm{d}y = 0 \tag{23}$$

$$= 2\int_{\mathcal{Y}} (f_k - y) q(f_k|x,y)p(y\mid x)\mathrm{d}y = 0 \,. \tag{24}$$

$$\tag{25}$$

Then, as we assumed $q > 0$, we have $\int_{\mathcal{Y}} q(f_k\mid x,y)p(y\mid x)\mathrm{d}y > 0$, and

$$f_k(x) = \frac{\int_{\mathcal{Y}} y\; q(f_k\mid x,y)p(y\mid x)\mathrm{d}y}{\int_{\mathcal{Y}} q(f_k\mid x,y)p(y\mid x)\mathrm{d}y} \;. \tag{26}$$

(iii) It follows immediately from (ii) that

$$\nabla_{f_k} D_x(q,f) = \left(\int_{\mathcal{Y}} q(f_k|x,y)p(y \mid x)\mathrm{d}y\right)\left(f_k - \frac{\int_{\mathcal{Y}} y\, q(f_k|x,y)p(y \mid x)\mathrm{d}y}{\int_{\mathcal{Y}} q(f_k|x,y)p(y \mid x)\mathrm{d}y}\right) = \gamma_k^\star(f_k - f_k^\star) \ .$$

□

Before further analyzing the training dynamics of aMCL, let us pause to examine a few properties of the soft distortion $D(q,f)$. First, observe that the hard distortion $D(f)$ given by (16) is a particular case of soft distortion $D(q,f)$ where $q(f_k|x,y) = \mathbb{1}[k \in \operatorname{argmin}_{s\in[\![1,n]\!]}\ell(f_s(x),y)]$. Second, $\inf_{f\in\mathcal{F}} D_x(q_T,f)$ and $\inf_{f\in\mathcal{F}} D(q_T,f)$ are non-decreasing functions of T. Third, entropy-constrained soft distortion minimization (21) is equivalent to the minimization of the free energy

$$\mathscr{F}(q,f) = D(q,f) - TH(q) \ , \tag{27}$$

a quantity defined in statistical physics, that will be analyzed in more detail in the following Section. Moreover, optimal free energy has a closed-form formula in our case. We group these observations in the following Proposition.

Proposition 6 (Additional properties of the distortion and the free energy)**.** *For all $T>0$, $f\in\mathcal{F}$, $q\in\Delta_n$, $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, the following statements are true.*

(i) $\quad q(f \mid x,y) = (\ \mathbb{1}[k \in \operatorname{argmin}_{s\in[\![1,n]\!]}\ \ell(f_s(x),y)]\ )_{k\in[\![1,n]\!]} \Rightarrow \forall q'\in\Delta_n,\ D(q,f)\le D(q',f)$.

(ii) $\quad D_x^*: T\mapsto \inf_{f\in\mathcal{F}} D_x(q_T,f)$ *and* $D^*: T\mapsto \inf_{f\in\mathcal{F}} D(q_T,f)$ *are non-decreasing on* $]0,\infty[$.

(iii) $\quad \min_{q\in\Delta_n} \mathscr{F}(q,f) = \mathscr{F}(q_T,f) = -T\int_{\mathcal{X}\times\mathcal{Y}} \log\sum_{k=1}^n \exp\left(-\frac{\ell(f_k(x),y)}{T}\right)p(x,y)\mathrm{d}x\mathrm{d}y$.

Proof. (i) Notice that $\operatorname{argmin}_{s\in[\![1,n]\!]}\ell(f_s(x),y)$ selects the hypothesis f_s with lowest loss $\ell(f_s(x),y)$. Therefore, for all $q'\in\Delta_n$, omitting the x dependency in the notations,

$$\sum_{k=1}^n q(f_k|y)\ell(f_k,y) = \min_{k\in[\![1,n]\!]}\ell(f_k,y) = \sum_{k=1}^n q'(f_k|y)\min_{s\in[\![1,n]\!]}\ell(f_s,y) \le \sum_{k=1}^n q'(f_k|y)\ell(f_k,y) \ .$$

We conclude by multiplying by $p(x,y)$ and integrating over $\mathcal{X}\times\mathcal{Y}$.
(ii) Let us first show that for any $f\in\mathcal{F}$, $T\mapsto D_x(q_T,f)$ is non-drecreasing. For that, let us define $\varphi(T) \triangleq \mathbb{E}_{k\sim q_T(f_k|x,y)}[\ell(f_k(x),y)] = \sum_{k=1}^n q_T(f_k \mid x,y)\ell(f_k(x),y)$ for each $T>0$, $x,y\in\mathcal{X}\times\mathcal{Y}$, and $f\in\mathcal{F}$. φ is a differentiable function of T, with derivative

$$\begin{aligned}\varphi'(T) &= \frac{1}{T^2}\left[\sum_{k=1}^n q_T(f_k \mid x,y)\ell(f_k(x),y)^2 - \left(\sum_{s=1}^n q_T(f_k \mid x,y)\ell(f_s(x),y)\right)^2\right] \\ &= \frac{1}{T^2}\mathbb{V}_{k\sim q_T(f_k|x,y)}[\ell(f_k(x),y)] \ge 0 \ .\end{aligned}$$

Therefore, φ is non-decreasing. By the growth of the integral, we deduce that $T\mapsto D_x(q_T,f)$ is non-decreasing.
Let $T'\ge T>0$. Then $D_x(q_{T'},f) \ge D_x(q_T,f) \ge \inf_{f'\in\mathcal{F}} D_x(q_T,f')$ for all $f\in\mathcal{F}$, and by definition of the infimum, we have $\inf_{f\in\mathcal{F}} D_x(q_{T'},f) \ge \inf_{f\in\mathcal{F}} D_x(q_T,f)$. So D_x^* is a non-decreasing function of T, and the same reasoning applies to D^*.
(iii) There are two equalities to prove. First observe that

$$\mathscr{F}(q,f) = \underbrace{D(q,f) - T[H(q) - H_T]}_{\text{Lagragian of (21)}} - \underbrace{T\,H_T}_{\text{independent of } q} \ .$$

We deduce

$$\operatorname{argmin}_{q\in\Delta_n}\ \mathscr{F}(q,f) = \operatorname{argmin}_{q\in\Delta_n}\ (\ D(q,f) - T[H(q) - H_T]\ ) = q_T \ ,$$

which establishes the first equality. Then we conclude by substituting the definition of q_T into $\mathscr{F}(q_T, f)$ and simplifying the expression. Again, we drop the x dependency in the notations for readability.

$$\begin{aligned}
\mathscr{F}(q_T, f) &= \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} q_T(f_k \mid y) \left[\ell(f_k, y) + T \log q_T(f_k \mid y)\right] p(x, y) \mathrm{d}x\mathrm{d}y \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^{n} \underbrace{q_T(f_k|y)}_{\text{sums to 1}} \left[\underbrace{\ell(f_k, y) - T\frac{\ell(f_k, y)}{T}}_{=0} - \underbrace{T \log \sum_{s=1}^{n} \exp\left(-\frac{\ell(f_s, y)}{T}\right)}_{\text{independent of k}} \right] p(x, y) \mathrm{d}x\mathrm{d}y \\
&= -T \int_{\mathcal{X}\times\mathcal{Y}} \log \sum_{k=1}^{n} \exp\left(-\frac{\ell(f_k, y)}{T}\right) p(x, y) \mathrm{d}x\mathrm{d}y \ .
\end{aligned}$$

□

Note that the last equality can be written as:

$$\mathscr{F}(q_T, f) = -T \int_{\mathcal{X}\times\mathcal{Y}} \log(Z_{x,y}) p(x, y) \mathrm{d}x\mathrm{d}y \ = \int_{\mathcal{X}} \mathscr{F}_x(q_T, f) p(x) \mathrm{d}x \ , \tag{28}$$

where we define the conditional free energy $\mathscr{F}_x(q_T, f)$ as

$$\mathscr{F}_x(q_T, f) \triangleq -T \int_{\mathcal{Y}} \log(Z_{x,y}) p(y \mid x) \mathrm{d}y \ . \tag{29}$$

A.4 Rate distortion curve

We have established in the previous Section that aMCL moves the hypotheses f_k towards the soft barycenter of soft Voronoi cells. We now describe the impact of temperature cooling on the position of these soft barycenters.

At high temperature, we observe that the soft barycenters (hence the hypotheses) converge towards the conditional barycenter $\mathbb{E}[Y|X = x]$ and fuze. When temperature decreases, they iteratively split into sub-groups. To capture the virtual number of hypotheses, we introduce the *conditional rate-distortion function*

$$R_x(D^\star) \triangleq \min_{\substack{q\in\Delta_n, f\in\mathcal{F} \\ D_x(q,f)\leq D^\star}} I_x(\hat{Y}; Y) \ , \tag{30}$$

where the target $Y \sim p(y \mid x)$, the hypothesis position $\hat{Y} \sim q(f_k \mid x)$ follows a distribution over $\mathcal{Y}$ with $q(f_k \mid x) = \int_{\mathcal{Y}} q(f_k \mid x, y) p(y \mid x) \mathrm{d}x$, and $I_x(\hat{Y}; Y)$ is their mutual information.

The following Proposition shows that (30) corresponds in fact to the same optimization problem as minimizing the conditional free energy (29), and therefore as minimizing the entropy constrained distortion (21) conditionally on x. It then describes the shape of the parametric curve $(D^\star, R_x(D^\star))$, and provides a lower bound of the rate distortion. For this purpose, we introduce for each $x \in \mathcal{X}$

$$D_x^{\max} \triangleq \inf_{f_1\in\mathcal{Y}} \int_{\mathcal{Y}} \ell(f_1, y) p(y \mid x) \mathrm{d}y = \int_{\mathcal{Y}} \ell(\mathbb{E}[Y|X = x], y) \, p(y \mid x) \, \mathrm{d}y \ , \tag{31}$$

the optimal conditional distortion when using a single hypothesis.

Proposition 7 (Rate-distortion properties)**.** *For each $x \in \mathcal{X}$, we have the following results.*

(i) For each $T > 0$, minimizing the free energy

$$\mathscr{F} = D(q_T, f) - TH(q_T) \ , \tag{32}$$

over all hypotheses positions $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^n)$ comes down to solving the optimization problem that defines (10) for each $x \in \mathcal{X}$ under decoupling Assumption 1.

(ii) *R_x is a non-increasing, convex and continuous function of $D^\star$, $R_x(D^\star) = 0$ for $D^\star \geq D_x^{\max}$, and for each x the slope can be interpreted as $R_x'(D^\star) = -\frac{1}{T}$ when it is differentiable.*

(iii) *For each x, $R_x(D^\star)$ is bounded below by the Shannon Lower Bound (SLB)*

$$R_x(D^\star) \geq \mathrm{SLB}(D^\star) \triangleq H(Y) - H(D^\star) \,, \tag{33}$$

where $Y \sim p(y \mid x)$ and $H(D^\star)$ is the entropy of a Gaussian with variance $D^\star$.

These results are well-known properties of the rate-distortion function [11, 4, 24, 51, 6].

Proof. (i) Under the decoupling assumption 1, optimizing $\mathscr{F}$ comes down to optimizing $\mathscr{F}_x$ for each $x \in \mathcal{X}$. Consequently, we can use the known relationship between the rate-distortion curve and free energy minimization [4, 24, 59].

(ii) See [4, 24].

(iii) See [64]. □

A key characteristic of the SLB is its asymptotic tightness in the low distortion regime for a broad range of probability distributions [32]. For some distributions, such as the Gaussian case, the SLB and the rate-distortion curve align perfectly for $D \leq D_x^{\max}$.

The Rate-Distortion interpretation of aMCL applies for each x, but defining a notion of global rate that is independent of the input is more challenging. This is because the distortion optimal distortion at a given temperature level T, $\inf_{f\in\mathcal{F}} D_x(q_T, f)$, can vary across inputs x. This is left for future work.

A.5 First phase transition

In the previous Section, we mentioned that aMCL predictions fuze at high temperature into the conditional barycenter $\mathbb{E}[Y|X = x]$. Moreover, these fuzed predictions iteratively split as the temperature (or equivalently the distortion) decreases during training. With each split, the number of sub-groups formed by the hypotheses increases. The number of sub-groups, measured in bits, is captured by the rate-distortion $R_x(D^\star)$. In this Section, we focus on the critical temperature corresponding to the first of these splits.

The predictions fuze into a single point at high temperatures so that the number of sub-groups is 1 and $R_x(D^\star) = 0$ in this regime. The first splitting therefore occurs when $R_x(D^\star) > 0$. Correspondingly, we define the first critical temperature in Definition 1.

Definition 1. *The first critical temperature $T_0^c(x)$ for each x and the defined global variant T_0^c are defined as:*

$$T_0^c(x) \triangleq \inf \left\{ T \mid \inf_{f\in\mathcal{F}} D_x(q_T, f) \geq D_x^{\max} \right\} \,, \tag{34}$$

$$T_0^c \triangleq \inf \left\{ T \mid \inf_{f\in\mathcal{F}} D(q_T, f) \geq D_{\max} \right\} \,, \tag{35}$$

in the conditional and the non-conditional setting, where $D_{\max} \triangleq \int_{\mathcal{X}} p(x) D_x^{\max} \mathrm{d}x$ and $D_x^{\max}$ is defined in (7).

Note that we have also equality in the definition $T_0^c(x) = \inf\{T \mid \inf_{f\in\mathcal{F}} D_x(q_T, f) = D_x^{\max}\}$, since $\inf_{f\in\mathcal{F}} D_x(q_T, f) \leq D_x^{\max}$. Moreover, recalling that $D_x^* : T \mapsto \inf_{f\in\mathcal{F}} D_x(q_T, f)$ is a non-decreasing function (Proposition 6 (ii)), we can equivalently define $T_0^c(x)$ as the generalized inverse $T_0^c(x) = (D_x^*)^{-1}(D_x^{\max})$ (*e.g.*, Definition 1 in [17]). The same observations hold for T_0^c.

We also know that aMCL's predictions implicitly minimize the free energy (27) (see Proposition 7, part (i)). Therefore, the conditional barycenter $\mathbb{E}[Y|X = x]$ is stable as long as the Hessian of the free energy is positive definite. This observation can alternatively be used to define the first critical temperature.

Interestingly, all these definitions are equivalent, as shown by the next Proposition. For this purpose, let us introduce the following covariance matrices:

$$C_{k,k}(f,q|x) = \int_{\mathcal{Y}} (f_k(x)-y)(f_k(x)-y)^t \; q(y \mid f_k, x) p(y \mid x) \; \mathrm{d}y \; ,$$

where $q(y \mid f_k, x)$ denotes the posterior probability of assigning the point y to the hypothesis k, calculated using Bayes's rule [60]. At high temperatures, all hypotheses merge into the conditional barycenter of the distribution (9), and the matrices $C_{k,k}$ are equal to the data covariance matrix $C(x) \triangleq \mathbb{C}\mathrm{ov}_{(X,Y)\sim p(x,y)}[Y \mid X = x]$.

Proposition 8 (First critical temperature)**.** *We have the two following results.*

(i) We know from [61] that for all $x \in \mathcal{X}$, $T_0^c(x) = 2\lambda_{\max}(C(x))$.

(ii) Under decoupling assumption 1, $T_0^c \leq 2 \sup_{x\in\mathcal{X}} \lambda_{\max}(C(x))$.

Proof. (i) See [61].

(ii) Let $T = \sup_{x\in\mathcal{X}} T_0^c(x)$. By definition of $T_0(x)$ and due to Proposition 6 (ii), $\inf_{f\in\mathcal{F}} D_x(q_T, f) \geq \inf_{f\in\mathcal{F}} D_x(q_{T_0^c(x)}, f) \geq D_x^{\max}$ for all $x \in \mathcal{X}$. Then, we deduce from (20):

$$\inf_{f\in\mathcal{F}} D(q_T, f) = \inf_{f\in\mathcal{F}} \int_{\mathcal{X}} D_x(q_T, f) p(x) \mathrm{d}x = \int_{\mathcal{X}} \inf_{f\in\mathcal{F}} D_x(q_T, f) p(x) \mathrm{d}x \geq \int_{\mathcal{X}} D_x^{\max} p(x) \mathrm{d}x \; ,$$
$$\inf_{f\in\mathcal{F}} D(q_T, f) \geq D_{\max} \; .$$

Using (i), we can write:

$$T_0^c = \inf \left\{ T \mid \inf_{f\in\mathcal{F}} D(q_T, f) \geq D_{\max} \right\} \leq T = \sup_{x\in\mathcal{X}} T_0^c(x) \; .$$

□

We now analyze why the predictions converge towards the conditional barycenter $\mathbb{E}[Y|X = x]$. In the next Proposition, we first focus on the asymptotic regime, when the temperature T increases to infinity, and then extend this analysis at finite temperature.

Proposition 9 (Training Dynamics with Temperature)**.** *For all $x \in \mathcal{X}$ and all $k \in [\![1, n]\!]$, we have the following properties.*

(i) If $T = \infty$ and $f \in \mathrm{argmin}_{f\in\mathcal{F}} D(q_T, f)$, then $f_k(x) = \mathbb{E}[Y \mid X = x]$.

(ii) The gradient of the soft distortion can be re-written for each $T > 0$ as

$$\nabla_{f_k} D_x = 2 \int_{\mathcal{Y}} \sum_{r=0}^{\infty} \frac{1}{r!} \left(-\frac{\|f_k(x) - y\|^2}{T} \right)^r \frac{(f_k(x) - y)}{Z_{x,y}(T)} p(y \mid x) \mathrm{d}y \; . \tag{36}$$

Therefore, the first order approximation of $\nabla_{f_k} D_x$ in T writes:

$$\nabla_{f_k} D_x \underset{T\to\infty}{=} \frac{2}{n} \int_{\mathcal{Y}} (f_k(x) - y) p(y \mid x) \mathrm{d}y + o\left(1\right) \; . \tag{37}$$

When $T \to \infty$, aMCL reduces to standard risk minimization, so that a necessary condition of optimization is that all the hypotheses $f(x)$ are located at the conditional barycenter $\mathbb{E}[Y|X = x]$ (Proposition 9, part (i)). This conditional barycenter is stable and corresponds to the global minimizer of the aMCL training objective when $T \to \infty$. We further analyze the *force* that pushes all the hypotheses toward this conditional barycenter, by looking at an expansion (36) of $\nabla_{f_k} D$, which provides the direction of the hypotheses updates for each $T > 0$ (Proposition 9, part (ii)). As $T \to \infty$, we see that a global driving force, to which all the data points contribute, is pushing the hypotheses toward the barycenter in (37). As T decreases, local interactions corresponding to the higher-order terms appear and increase in amplitude. They are responsible for the phase transitions.

Proof. (i) By definition of the Boltzmann distribution

$$q_T(f_k \mid x, y) = \frac{\exp\left(-\ell(f_k(x), y)/T\right)}{\sum_{s=1}^{n} \exp\left(-\ell(f_s(x), y)/T\right)} \xrightarrow[T\to\infty]{} \frac{1}{n} \ .$$

Therefore $D(q_T, f) \xrightarrow[T\to\infty]{} \frac{1}{n}\sum_{k=1}^{n}\int_{\mathcal{Y}} \ell(f_k(x), y)p(y \mid x)\mathrm{d}y$, and all the hypotheses solve the same quantization problem. A necessary condition to minimize $D(q_T, f)$ is then

$$\forall k \in [\![1, n]\!], \ f_k(x) = \mathbb{E}[Y|X = x] \ .$$

(ii) We note $Z_{x,y}(T) = \sum_{s=1}^{n} \exp\left(-\|f_s(x) - y\|^2/T\right)$. Then, accounting for the stop_gradient operator in (6), the gradient of the soft distortion writes

$$\nabla_{f_k} D_x(q_T, f) = 2\int_{\mathcal{Y}} \frac{\exp\left(-\|f_k(x) - y\|^2/T\right)}{Z_{x,y}(T)}(f_k(x) - y)p(y \mid x)\mathrm{d}y \ .$$

Using the series expansion of the exponential, we have for each $T > 0$,

$$\exp\left(-\frac{\|f_k(x) - y\|^2}{T}\right) = \sum_{r=0}^{\infty} \frac{1}{r!}\left(-\frac{\|f_k(x) - y\|^2}{T}\right)^r \ .$$

We can rewrite the gradient of the soft distortion:

$$\nabla_{f_k} D_x = 2\int_{\mathcal{Y}} \sum_{r=0}^{\infty} \frac{1}{r!}\left(-\frac{\|f_k(x) - y\|^2}{T}\right)^r \frac{(f_k(x) - y)}{Z_{x,y}(T)} p(y \mid x)\mathrm{d}y \ .$$

Keeping only the first term, and observing $Z_{x,y}(T) \xrightarrow[T\to\infty]{} n$ we get:

$$\nabla_{f_k} D_x \underset{T\to\infty}{=} \frac{2}{n}\int_{\mathcal{Y}} (f_k(x) - y)p(y \mid x)\mathrm{d}y + o\left(1\right) \ .$$

□

This Section focuses on the first phase transition. However, multiple phase transitions may occur during training, as shown in Figure 4. It is interesting to note that, for deterministic annealing, the final phase transition occurs when the rate-distortion curve hits the Shannon Lower Bound defined in (33) (see [59]).

B Connection with the literature and discussion

B.1 Annealing at inference time

A promising direction for future research consists in using annealing at inference time. With this scheme, aMCL could be used to perform an input-dependent hierarchical clustering [53] at test time, similarly to the idea proposed in [55]. More precisely, we can store the model's parameters at different times of the training schedule (*i.e.*, at several temperature levels). During inference, this allows replaying the temperature cooling for new test samples, by performing forward passes through each of the trained models.

Replaying this trajectory may have several advantages. Indeed, the hypotheses trajectory follows the rate-distortion curve and consequently explores recursively the modes of the distribution as the temperature decays. Crucially, at each critical temperature, when the hypotheses are about to split, they are exactly located at the barycenter of these modes. If we can track these splitting moments, for instance by counting the number of distinct virtual hypotheses at each step of the cooling schedule, we can perform a hierarchical clustering that iteratively uncovers the modes of the distribution.

B.2 Discussion on Stochastic simulated annealing

Non-deterministic simulated annealing [28, 52] is a promising research direction due to its strong convergence properties (see Hajek theorem [26]). It requires to define the state f of the system and

the optimization objective $D(f)$. At each step, the state f is updated to a neighbor state $\tilde{f}$ based on a stochastic exploration criterion. The probability of accepting a neighbor state $\tilde{f}$ depends on the objective variation $D(\tilde{f}) - D(f)$ and the temperature T.

Our objective $D(f)$ corresponds to the Distortion (1). However, the state of the system can be defined in various ways. In the non-conditional setting, [74] defines the state as the hypothesis positions $(f_k)_{k\in[\![1,n]\!]}$ (similarly to the present work), while [26, 28, 52] defines it as the cluster assignation of each dataset sample.

In both cases, we expect that storing and updating this state using neural networks would be costly. Moreover, evaluating $D(\tilde{f})$ requires going through a validation set, which is time-consuming. Further investigation in this direction is left for future research.

B.3 Comparison with an additional baseline: Relaxed WTA with annealed ε

Relaxed WTA [63] suffers from a bias toward the barycenter. As suggested in [55], a natural idea to tackle this issue would be to anneal ε during training. In Figure 1, we have shown the results of this additional baseline on a Gaussian synthetic dataset, using a linearly decreasing epsilon during training $\varepsilon(t) = \varepsilon_0 \left(1 - \frac{t}{1000}\right)$ with $\varepsilon_0 = 0.98$.

In Figure 5, we show its training trajectory when using the same setup as in Figure 1. All hypotheses initially converge to the barycenter, then the winners gradually move towards the modes as ε decreases. As ε approaches 0, only a few additional hypotheses escape from the barycenter to reach the modes, indicating that annealing does not solve the collapse issue of Relaxed-WTA.

Results on the UCI datasets confirm this qualitative analysis (Table 5 and Figure 6). In Figure 6, aMCL outperforms the best Relaxed-WTA variants on Distortion.

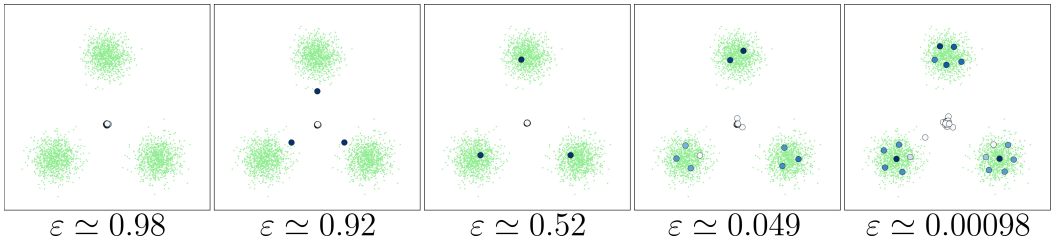


Figure 5: **Training dynamics of Relaxed-WTA with ε annealed.** Predictions during training of Relaxed-WTA with ε annealed linearly over 1000 epochs. We use the same setup as in Figure 1 of the main paper, with 49 hypotheses as shaded blue points (shade proportional to the predicted scores), and the targets as green points. For large ε, the hypotheses converge at the distribution barycenter. As ε decreases, three *Winners* gradually move towards the barycenter of the Gaussians. When ε approaches 0, the Winner-Take-All dynamic prevails and some hypotheses escape the barycenter to reach the Gaussian modes. However, the modes are quickly saturated, the unused hypotheses stop receiving gradient, and they remain stuck at the barycenter.

C Experimental details in the synthetic data experiments

This Section provides a detailed description of the synthetic data experiments presented in the paper, particularly the results and visualizations in Figures 1, 3 and 4. Note that for each of those experiments, a two-hidden-layer Multi-Layer Perceptron (MLP) with 256 neurons per layer and ReLU activation functions was used, except for the last layer. The final layer of the MLP was duplicated to account for multiple hypotheses $\{f_k(x)\}_{k=1}^n$ (followed by a $\tanh$ activation to restrict the output in the square $[-1,1]^2$) and scores $\{\gamma_k(x)\}_{k=1}^n$ (followed by $\mathrm{sigmoid}$ activation), similar to [42, 43], with $n = 49$ hypotheses in this study. At each epoch, 100,000 points were sampled from the corresponding synthetic datasets, and the models were trained for $t_{\text{epoch}} = 1000$ epochs. Note also that the visualization of the output space in Figure 1 and Figure 4 is restricted to the square $[-1,1]^2$.

Settings of Figure 1. In this Figure, the dataset consists of a mixture of three two-dimensional Gaussians located at $\mu_0 = (-0.5, -0.5)$, $\mu_1 = (0, 0.5)$ and $\mu_2 = (0.5, -0.5)$, each with a standard

deviation of 0.1. In this Figure, each model was trained with a Stochastic Gradient Descent optimizer with a constant learning rate of 0.01. For the Relaxed WTA run, we used $\varepsilon = 0.1$, and for the Annealed MCL run, we used an exponential scheduler in the form $T(t) = T_0\rho^t$, with $T_0 = 0.6$ and $\rho = 0.99$.

Settings of Figure 3 and Figure 4. In these experiments, the dataset is a conditional variant of that used in Figure 1, with a ground-truth distribution of the form $p(y \mid x) = \frac{1}{3}\sum_{i=0}^{2}\mathcal{N}(\tilde{\mu}_i(x), \sigma^2)$, with $\sigma = 0.1$, $\tilde{\mu}_i(x) = x\mu_i$ for $x \in [0, 1]$, and μ_i is defined as above for $i \in [\![0, 2]\!]$. Here, aMCL was trained with a linear scheduler defined by $T(t) = T_0(1 - \frac{t}{t_{\text{epoch}}})$, with $T_0 = 1.0$. Note that in the results of Figure 4, we have $(t_1, t_2, t_3, t_4) = (50, 745, 870, 1000)$ with associated temperature values $(T(t_1), T(t_2), T(t_3), T(t_4)) \simeq (0.950, 0.256, 0.131, 0.001)$. Here, aMCL was trained with Adam optimizer [34], with a constant learning rate of 0.01. In Figure 3, T_0^c (upper bound) correspond to the right-hand-side of (15) and was computed as $2\lambda_{\max}(\int_{\mathcal{Y}} yy^t p(y \mid x = 1)\mathrm{d}y) \simeq 0.46$, approximating the covariance matrix over 1,000 samples. Note that the distortion plotted in the curve of Figure 3 corresponds to the hard distortion (1) averaged over a validation set of 25,000 samples every 5 epoch.

D Additional results from UCI Datasets experiments

The UCI Regression Datasets [15] serve as a standard benchmark for conditional distribution estimation. This Section provides additional details and results from the experiments presented in Section 5.1.

Experimental setup. The experiments were conducted using the protocol from [29]. Each dataset is divided into 20 train-test folds, except the protein dataset, which had 5 folds, and the Year Prediction MSD dataset which used a single test-set split. We used the same neural network backbone for each baseline: a one-hidden layer MLP with ReLU activation function, containing 50 hidden units except for the Protein and Year datasets, for which 100 hidden units were used. In the WTA models, the final layer of the MLP was duplicated to account for multiple hypotheses $\{f_k(x)\}_{k=1}^n$ (followed by no activation) and scores $\{\gamma_k(x)\}_{k=1}^n$ (followed by sigmoid activation). The WTA-based methods (Relaxed-MCL, MCL, and aMCL) were trained with $n = 5$. All models were trained with the Adam optimizer over 1,000 epochs with a constant learning rate of 0.01. The data loading pipeline was adapted from the implementation of [27]. The best models for each dataset are highlighted in bold in Table 2 and Table 1, with the second-best models underlined.

Baselines. Table 2 of the main paper includes results from three baselines reported from Table 1 of [38], which we use as references for those benchmarks. 'PBP' stands for Probabilistic Back Propagation [29], and 'MC-dropout' corresponds to Monte Carlo Dropout [21]. Relaxed-MCL was trained with $\varepsilon = 0.1$ in Tables 1 and 2. Other values of ε are experimented in Appendix D.2.

Metrics. We used RMSE and distortion as metrics. RMSE is defined as $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\hat{y}_i - y_i\|^2}$, where $\hat{y}_i$ denotes the estimated conditional mean, defined as $\sum_{k=1}^{n}\gamma_k(x)f_k(x)$ for the WTA variants, and N is the number of samples in each test set. The distortion of the multi-hypotheses models was computed as $\mathrm{Distortion} = \frac{1}{N}\sum_{i=1}^{N}\min_{k\in[\![1,n]\!]}\|\hat{y}_k - y_i\|^2$.

Temperature scheduler. For training aMCL in those experiments, an exponential temperature scheduler was used, defined as $T(t) = T_0\rho^t$, with $\rho = 0.95$ and $T_0 = 0.5$, where t denotes the current epoch and $t_{\text{epoch}} = 1,000$ is the maximum number of epochs. When the temperature of the exponential scheduler was lower than $5e-4$, we set the training back to the vanilla Winner-takes-all mode.

Evaluation details. During the evaluation, following [29], both input and output variables were normalized using the training data's means and standard deviations. For evaluation, the original scale of the output predictions was restored using the transformation $f_\theta^k(x) \mapsto \mu_{\text{train}} + \sigma_{\text{train}} f_\theta^k(x)$ where μ_{train} and σ_{train} are the empirical mean and standard deviation of the output variable, computed across the training set.

The results concerning the distortion metric on the UCI datasets are presented in Table 1, and their analysis is provided in Section 5.1.

D.1 Impact of the scheduler type in aMCL

The impact of the scheduler type on the performance on the UCI dataset is provided in Table 4. We see that for large dataset sizes, the exponential scheduler outperforms the linear one. Further study on the impact of the speed of the scheduler on the performance is left for further work.

Table 4: **Impact of the scheduler type in aMCL.** We display the results on the UCI datasets comparing two types of temperature schedules in aMCL. Here, both use an initial temperature $T_0 = 0.5$. aMCL (exp) uses an exponential scheduler of the form $T(t) = T_0\rho^t$, with $\rho = 0.95$, and aMCL (lin) uses a linear scheduler $T(t) = T_0(1 - \frac{t}{t_{\text{epoch}}})$. The rows are ordered by dataset size N.

	RMSE (↓)		Distortion (↓)		
Datasets	aMCL (exp)	aMCL (lin)	aMCL (exp)	aMCL (lin)	N
Year	**9.08 ± NA**	9.20 ± NA	**4.39 ± NA**	4.46 ± NA	515345
Protein	**4.25 ± 0.02**	4.26 ± 0.04	**0.77 ± 0.03**	0.92 ± 0.05	45730
Naval	**8.00e-4 ± 4.04e-4**	1.72e-3 ± 1.71e-3	**5.37e-7 ± 3.83e-7**	3.97e-6 ± 1.36e-5	11934
Power	4.08 ± 0.2	**4.07 ± 0.16**	**2.06 ± 0.45**	11.93 ± 2.36	9568
Kin8nm	**0.08 ± 0.00**	**0.08 ± 0.00**	**6.81e-4 ± 8.14e-5**	2.94e-3 ± 1.80e-3	8192
Wine	**0.63 ± 0.04**	0.67 ± 0.05	**0.03 ± 0.01**	0.09 ± 0.02	1599
Concrete	5.47 ± 0.67	**4.99 ± 0.63**	**5.71 ± 1.72**	7.97 ± 3.84	1030
Energy	1.35 ± 0.97	**0.80 ± 0.33**	**0.28 ± 0.09**	0.55 ± 0.51	768
Boston	**3.05 ± 0.91**	3.12 ± 0.68	**2.69 ± 1.39**	6.18 ± 2.99	506
Yacht	1.62 ± 0.53	**0.85 ± 0.25**	1.15 ± 0.97	**0.51 ± 0.37**	308

D.2 Impact of ε in Relaxed WTA

RMSE compares the barycenter of the predicted distribution with the target positions. For this metric, Relaxed-MCL outperforms other approaches on the UCI datasets when ε is high (see Table 5 and Figure 6). This outcome is expected, as Relaxed-WTA is biased towards the distribution barycenter under this regime (see Figure 1 of the paper).

However, using RMSE for comparison discards valuable spatial distribution information of the hypotheses. Distortion, defined in (1), addresses this limitation by measuring quantization performance, which our theoretical analysis is based on. When focusing on the Distortion metric, Table 5 shows that lower values of epsilon (*e.g.*, $\varepsilon = 0.1$ or annealed epsilon) consistently perform better than for $\varepsilon = 0.5$ across nearly all settings. This improvement is attributed to the reduced bias toward the barycenter in this configuration. Additionally, for Distortion, aMCL (trained with an exponential scheduler) generally outperforms Relaxed-WTA on the UCI datasets for the tested Relaxed WTA variants (see Tables 1, 5 and Figure 6). This is especially true on large datasets (Year, Protein).

Overall, aMCL demonstrates a strong balance between Distortion and RMSE compared to the baselines, especially on the largest datasets. Figure 6 further supports this trend, providing additional analysis on Year and Protein, where aMCL, MCL, and Relaxed-MCL are compared across different ε values in the (RMSE, Distortion) space.

Table 5: **Impact of ε in Relaxed-WTA on UCI regression benchmark datasets comparing RMSE and Distortion.** Best results are in **bold**, second bests are underlined. The annealed version of Relaxed-WTA (R-WTA) uses a linear scheduler starting at $\varepsilon_0 = 0.5$ and decreasing to 0.

	RMSE (↓)			Distortion (↓)			
	R-WTA ($\varepsilon = 0.5$)	R-WTA ($\varepsilon = 0.1$)	R-WTA (ε annealed)	R-WTA ($\varepsilon = 0.5$)	R-WTA ($\varepsilon = 0.1$)	R-WTA (ε annealed)	N
Year	**8.91 ± NA**	8.97 ± NA	9.1 ± NA	26.17 ± NA	7.73 ± NA	**4.56 ± NA**	515345
Protein	**4.20 ± 0.02**	4.38 ± 0.02	4.37 ± 0.05	7.13 ± 0.14	1.67 ± 0.16	**0.99 ± 0.07**	45730
Naval	**1.47e-03 ± 7.47e-04**	1.80e-03 ± 5.66e-04	2.08e-03 ± 8.49e-04	1.13e-06 ± 1.45e-06	**4.21e-07 ± 2.36e-07**	7.70e-07 ± 7.85e-07	11934
Power	**3.99 ± 0.16**	4.02 ± 0.18	4.1 ± 0.14	6.59 ± 0.62	**2.36 ± 0.43**	3.77 ± 1.94	9568
Kin8nm	**0.08 ± 0.00**	**0.08 ± 0.00**	0.09 ± 0.00	2.82e-03 ± 2.09e-04	9.32e-04 ± 7.97e-05	**9.06e-04 ± 3.89e-04**	8192
Wine	**0.63 ± 0.04**	**0.63 ± 0.04**	0.7 ± 0.06	0.19 ± 0.03	**0.06 ± 0.02**	0.14 ± 0.05	1599
Concrete	**4.91 ± 0.65**	5.28 ± 0.58	5.14 ± 0.56	13.69 ± 3.51	**6.63 ± 2.51**	6.82 ± 3.05	1030
Energy	**1.03 ± 0.21**	1.64 ± 0.36	1.48 ± 0.59	0.40 ± 0.13	**0.30 ± 0.12**	0.33 ± 0.1	768
Boston	2.85 ± 0.57	2.85 ± 0.72	**2.8 ± 0.69**	5.95 ± 2.94	**3.32 ± 2.84**	3.73 ± 2.38	506
Yacht	**1.17 ± 0.32**	2.52 ± 1.04	1.59 ± 0.34	**0.63 ± 0.28**	1.34 ± 0.93	0.73 ± 0.49	308

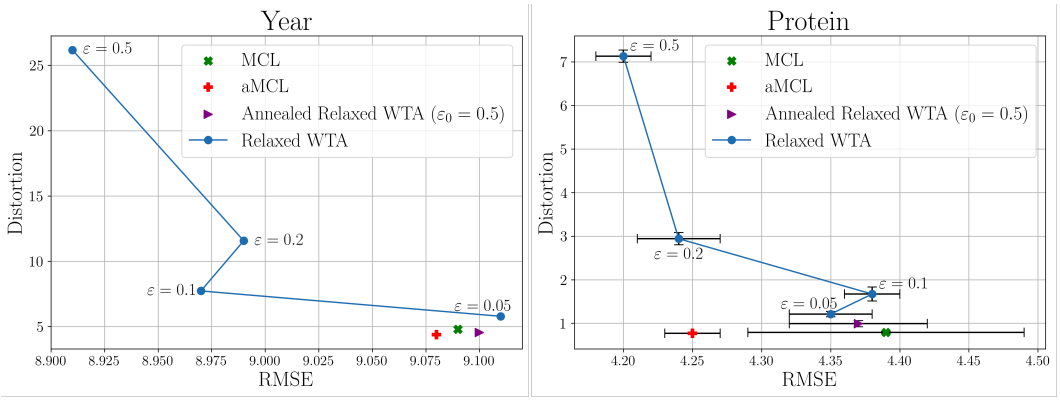


Figure 6: **(RMSE,Distortion) performance on Year and Protein datasets**, for Relaxed-WTA with $\varepsilon \in \{0.05, 0.1, 0.2, 0.5\}$ (blue points, with black error bars for std), Relaxed-WTA with annealed ε (purple) and aMCL (red). Blue lines link the Relaxed-WTA score points. We see that aMCL is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

E Application to speech separation

This Appendix provides a detailed description of the speech source separation experiments. Speech separation consists of isolating the audio signals of each individual speaker from a mixture in which they are simultaneously active. This task is of major interest for speech processing applications such as Automatic Speech Recognition [50, 44], Speaker Diarization [71], or singing voice extraction from music [31]. Most of the recent speech separation systems are deep neural networks trained in a supervised setting, where the ground truth of the separated speaker tracks is known [72].

E.1 Task description

Formally, let $y_1, \dots, y_m \in \mathbb{R}^l$ denote the raw speech signals from m individual speakers, with l denoting the number of time frames in the corresponding audio recordings. Under anechoic (no reverberation) and clean (no background noise) conditions, the mixture can be expressed as $x = \sum_{s=1}^{m} y_s$. Hence, the task is to provide estimates $\hat{y}_1, \dots, \hat{y}_m$ of the isolated speech tracks from the mixture signal x using a neural network f_θ. The quality of the estimation $\ell(\hat{y}_k, y_s)$ is commonly measured using the audio-domain distance Signal-to-Distortion Ratio (SDR) and its scale-invariant (SI) variant SI-SDR [70, 39]. Since there is a fundamental ambiguity concerning the best association $\hat{y}_k \mapsto y_s$, separation systems are trained using the Permutation Invariant Training (PIT) loss:

$$\mathcal{L}_{\text{sep}}^{\text{PIT}}(\hat{y}, y) = \min_{\sigma \in \Sigma} \left(\frac{1}{m} \sum_{s=1}^{m} \ell(\hat{y}_{\sigma(s)}, y_s) \right) , \tag{38}$$

where Σ is the set of permutations of $[\![1, m]\!]$.

Instead, we propose to use the MCL framework to solve this assignment problem. Naming $\mathcal{L}_{\text{sep}}^{\text{MCL}}$ the resulting loss, $\mathcal{L}_{\text{sep}}^{\text{aMCL}}$ its annealed variant, and $\mathcal{L}_{\text{sep}}^{\text{Relaxed-WTA}}$ its relaxed variant, we define

$$\mathcal{L}_{\text{sep}}^{\text{MCL}}(\hat{y}, y) = \frac{1}{m} \sum_{s=1}^{m} \min_{k \in [\![1,n]\!]} \ell(\hat{y}_k, y_s) , \tag{39}$$

$$\mathcal{L}_{\text{sep}}^{\text{aMCL}}(\hat{y}, y) = \frac{1}{m} \sum_{s=1}^{m} \sum_{k=1}^{n} q_{T(t)}(\hat{y}_k \mid y_s) \ell(\hat{y}_k, y_s) , \tag{40}$$

$$\mathcal{L}_{\text{sep}}^{\text{Relaxed-WTA}}(\hat{y}, y) = \frac{1}{m} \sum_{s=1}^{m} \sum_{k=1}^{n} q_{\varepsilon}(\hat{y}_k \mid y_s) \ell(\hat{y}_k, y_s) , \tag{41}$$

where $q_\varepsilon(\hat{y}_k \mid y_s) = \begin{cases} 1 - \varepsilon, & \text{if } k = \operatorname{argmin}_{k' \in [\![1,n]\!]} \ell(\hat{y}_{k'}, y_s) \\ \varepsilon/(n-1), & \text{otherwise} \end{cases}$.

This gives the separation task a new interpretation: the targets y_s are samples drawn from a common distribution $p(y \mid x)$ conditional on the mix x, the estimated speech signals $\hat{y}_k = f_k(x)$ are the output of neural networks f_k, and the mix x is their common input.

Notice that (39) is subject to collapse. Furthermore, when $n = m$, (38) always finds the best assignation, and acts as a topline, while MCL acts as a baseline. The goal of these experiments is therefore threefold:

- applying our proposed aMCL algorithm to a more realistic and data-intensive setting;
- extending the analysis of aMCL to a more general setting where the underlying metric (SI-SDR) is non-Euclidean;
- ensuring that MCL is a valid substitute for PIT in the context of source separation, thus opening up a new avenue for MCL research in this domain.

E.2 Experimental settings

E.2.1 Separation model architecture

We use the Dual-Path Recurrent Neural Network (DPRNN) [46] separation architecture for our experiments. This model follows the encoder-masker-decoder structure, originally proposed in ConvTasNet [47]. Each component of the network is described below:

Encoder. The encoder transforms the raw audio signal $x \in \mathbb{R}^l$, with l the number of time frames, into a sequence of features. It is implemented as a 1-D convolutional layer with F channels, kernel size K, and stride S. It outputs the sequence $\mathbf{X} \in \mathbb{R}^{F \times T}$ where T is the number of encoded frames.

Masker. The masker of the DPRNN model first applies a sliding window of width W to the sequence of encoded features $\mathbf{X}$. The input sequence is transformed from $\mathbb{R}^{F \times T}$ to $\mathbf{X}' \in \mathbb{R}^{F \times P \times W}$, where P is the number of windows. DPRNN is composed of two recurrent layers, processing the sequence $\mathbf{X}'$ over two different directions (dual-path):

- *intra*-chunk: processes the dimension W,
- *inter*-chunk: processes the dimension P.

Thus, the model learns local and global dependencies. The RNN is implemented as a bidirectional Long Short-Term Memory layer (LSTM) with hidden dimension H. The sequence $\mathbf{X}'$ is processed by B successive DPRNN blocks. The output of the last block is denoted $\mathbf{X}'' \in \mathbb{R}^{H \times P \times W}$. Before decoding, the sequence is mapped back to the original dimensions using overlap-and-add. A final convolutional layer maps the sequence $\mathbf{X}''$ to the masks $\mathbf{M} \in \mathbb{R}^{n \times F \times T}$, where n is the number of predictions. The k-th source latent sequence $\hat{\mathbf{Y}}_k = \mathbf{M}_k \odot \mathbf{X}$ is inferred by the element-wise product $\odot$ between the k-th mask $\mathbf{M}_k$ and the encoded sequence $\mathbf{X}$.

Decoder. The decoder is implemented as a 1-D transposed convolutional layer with n groups, F channels, and a kernel of size K and stride S. It maps the latent sequences $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times F \times T}$ to the audio domain, and we note $\hat{y} \in \mathbb{R}^{n \times l}$ the corresponding vector. Therefore, using the MCL framework notations, the output of the prediction model k is $f_k(x) = \hat{y}_k$.

Scoring. We do not use scoring models in these experiments. Indeed, we will focus on the case where $n = m$ and all predictions should be matched to a corresponding source. This ensures a fair comparison with the PIT baseline, which uses the same setting.

The model configuration used for our experiments is presented in Table 6. It corresponds to the model configuration from the original paper. The variables l, T, and P, which depend on the unknown input length l, are not specified.

E.2.2 Training configuration

Dataset. The source separation experiments are conducted on the Wall Street Journal mix dataset (WSJ0-mix) [30] which is the current benchmark dataset for speech source separation. It consists of clean utterances of read speech linearly combined to build synthetic speech mixtures. Several versions of the dataset are available to separate from 2-speaker to 5-speaker mixtures. In the present

work, we focus on the 2- and 3-speaker scenarios. Each version features 20000, 5000, and 3000 mixtures for training, validation, and testing respectively. The audio signals are samples at 8kHz.

Training objectives. Four types of systems are compared in the following experiments:

- PIT: training with $\mathcal{L}_{\text{sep}}^{\text{PIT}}$ defined in (38),
- MCL: training with $\mathcal{L}_{\text{sep}}^{\text{MCL}}$ defined in (39),
- aMCL: training with $\mathcal{L}_{\text{sep}}^{\text{aMCL}}$ defined in (40),
- Relaxed-WTA: training with $\mathcal{L}_{\text{sep}}^{\text{Relaxed}-\text{WTA}}$ defined in (41).

We use the SI-SDR score as the separation quality measure

$$\ell(y, \hat{y}) = 10 \log_{10} \frac{\langle y, \hat{y} \rangle^2}{\|y\|^2 \|\hat{y}\|^2 - \langle y, \hat{y} \rangle^2} \ , \tag{42}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product on $\mathbb{R}^l$.

Training parameters. The separation models are trained on 5-second audio segments. Since the utterances from the training set might have a longer duration, 5-second segments are randomly cropped inside the training utterances. The utterances shorter than 5s are removed. The batch size is set to 22 to reduce training time compared to the standard setup [47, 46, 68], in which a batch size of 4 is used. We verified that our models match the originally reported scores [46] when using this original setting. To analyze robustness to initialization, we train each method with 3 different seeds, and we report the inter-seed average scores and standard deviations. More precisely, the prediction scores are first averaged across the evaluation dataset for each single seed. Then these dataset-wise scores are used to compute the final average and standard deviations. Each model is trained on $t_{\text{epoch}} = 200$ epochs, without early stopping. All the training set is processed during each epoch. Unless otherwise stated, the temperature scheduler is chunk linear:

$$T(t) = T_0 \left(1 - \frac{t}{t_{\max}} \right) \mathbb{1}[t < t_{\max}] \ , \tag{43}$$

with initial temperature $T_0 \approx 0.1$ and $t_{\max} = 100$. Note that at the end of the training, when $t \in [t_{\max}, t_{\text{epoch}}]$, annealing is disabled by setting $T(t) = 0$ in order to fine-tune the hypotheses on the target MCL objective.

The neural network weights are updated using the Adam optimizer, with a learning rate set to 10^{-3}. The learning rate is halved after every 5 epochs with no improvement in the validation metric. Separation models are trained on Nvidia A40 GPU cards. The DPRNN has 3.7 million parameters.

Evaluation. The separation scores are computed on the evaluation subset of the WSJ0-mix datasets. Two types of scores are presented:

- PIT SI-SDR: the assignment is performed with PIT (38) as in standard source separation approaches;
- MCL SI-SDR: the assignment is performed with MCL (39), and this score can be seen as a form of quantization error.

Table 6: **Configuration of the DPRNN model used for source separation experiments.**

Parameter	Symbol	Value
Feature dimension	F	64
Encoder/decoder kernel size	K	16
Encoder/decoder stride	S	8
DPRNN Chunk size	W	100
Hidden dimension	H	128
Number of DPRNN blocks	B	6

E.3 Experimental results

This Section presents the results of the experiments conducted on speech separation with the aMCL framework, and is organized as follows:

- Appendix E.3.1: comparison of the separation performance of each approach;
- Appendix E.3.2: impact of the number of hypotheses on the separation performance with MCL and aMCL;
- Appendix E.3.3: analysis of the phase transitions of aMCL in the context of speech separation.

E.3.1 Separation performance on 2- and 3-speaker mixtures

Table 7 shows the average score of each separation model in the 2- and 3-speaker scenarios.

First, we observe that PIT and aMCL achieve similar performances, both in terms of the average PIT SI-SDR and MCL SI-SDR scores, and both in the 2- and the 3- speaker settings (see lines 1 and 3 of Table 7). Looking at the inter-seed variance, we further see that the differences in performances between these approaches are not significant. Note that MCL and aMCL have a $\mathcal{O}(mn)$ complexity, while PIT is $\mathcal{O}(m^3)$ with the Hungarian algorithm [16]. Therefore aMCL reaches the same performance as PIT with a gain in terms of complexity. Further work includes an analysis of this complexity gap when the number of speakers is high, similarly to [66].

Second, we see that aMCL achieves better performance than MCL in the 2-speaker scenario, both in terms of PIT SI-SDR and MCL SI-SDR metrics (see lines 2 and 3 of Table 7). More precisely, we see that this performance discrepancy between MCL and both PIT and aMCL is due to a higher variance of the MCL training method. In fact, MCL has a higher inter-seed variance than aMCL and PIT in all settings. Indeed this method is known to be sensitive to initialization [49, 63]. In contrast, we see that aMCL is more robust to initialization, having a lower inter-seed variance and higher average score. This phenomenon was highlighted in Section A, which provided a theoretical analysis of the Euclidean case. This experiment suggests that our conclusions hold even for the source separation task, where the underlying loss function ℓ is non-Euclidean.

In the subsequent Sections, we further study these two claims. In Section E.3.2, we demonstrate the advantages of aMCL over PIT by letting the number of predictions n exceed the number of sources m. In Section E.3.3, we show that other conclusions of our theoretical analysis concerning phase transitions also seem to hold in the non-Euclidean setting.

Table 7: **Comparison of the training methods PIT, MCL, and aMCL**, for the task of source separation of 2 and 3 speakers. The performance is measured using the PIT SI-SDR metric *(left)* and the MCL SI-SDR metric *(right)* computed on the WSJ0-mix evaluation subset. Each method is trained using three seeds and we report inter-seed average score and standard deviation.

(a) PIT SI-SDR (↑)

Method	2 speakers	3 speakers
PIT	16.88 ± 0.10	10.01 ± 0.04
MCL	16.30 ± 0.59	10.06 ± 0.21
aMCL	16.85 ± 0.13	10.00 ± 0.21

(b) MCL SI-SDR (↑)

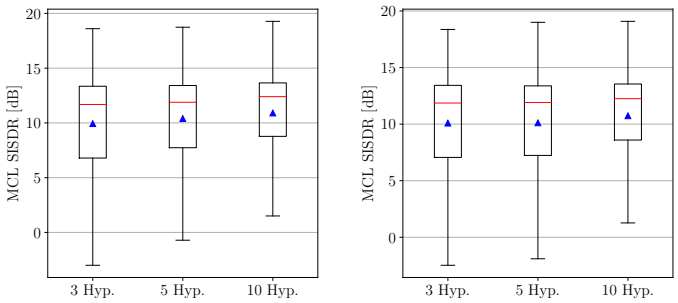
Method	2 speakers	3 speakers
PIT	16.88 ± 0.10	10.04 ± 0.04
MCL	16.30 ± 0.59	10.09 ± 0.21
aMCL	16.86 ± 0.13	10.04 ± 0.20

E.3.2 Impact of the number of hypotheses in MCL and aMCL

In this Section, we study the impact of the number of hypotheses on the source separation performances. We focus on the 3-speaker scenario, and we increase the number of predictions. More precisely, we consider the cases $n \in \{3, 5, 10\}$. If $n = 3$, the number of hypotheses is the same as the number m of sources to separate. If $n \in \{5, 10\}$ case the number of hypotheses is higher. Since the number of predictions and sources differ, it is not possible to compute the PIT SI-SDR metric, which is ill-defined in this setting. Therefore, we only present results for the MCL SI-SDR metric. In this experiment, we use a single seed across the considered methods to ensure a fair comparison. We report dataset-wise score statistics in Figure 7.

First, we observe that the separation performance increases as the number of predictions n increases, both for MCL and aMCL. This shows that the additional hypotheses are effectively used by both methods, despite the risk of collapse or convergence towards a suboptimal configuration. Recalling that MCL, aMCL, and PIT performances are similar in the case $n = m = 3$, this demonstrates a strong advantage of the MCL family compared to PIT. Indeed, we can improve MCL SI-SDR score beyond what is achievable by PIT training, by increasing the number of predictions n. Moreover, this improvement comes at minimal computational complexity cost, as discussed in Section E.3.1. This property can be exploited to handle settings where the number of speakers is unknown. This is left to further work.

Second, we observe that there are no significant discrepancies between MCL and aMCL performances in this setting. This suggests that the main difference between these two approaches is their sensitivity to initialization.



(a) MCL (b) aMCL

Figure 7: **Impact of the number of hypotheses.** Comparison of MCL (left) and aMCL (right) on the WSJ0-3mix dataset, using the MCL SI-SDR metric. Dataset-wise scores average (blue triangle), median (red line), and quartiles (main box) are reported. The whiskers extend the box by 1.5 times the interquartile gap. The number of hypotheses n is selected among $\{3, 5, 10\}$. A higher score indicates better separation performance.

E.3.3 Phase transition in aMCL training

In this Section, we analyze the training trajectory of aMCL in speech separation. We focus on exponential schedulers defined by

$$T(t) = T_0 \, \rho^t \, \mathbb{1}[t < t_{\max}] \, , \tag{44}$$

where $\rho = 0.9$ is the decay factor, and $T_0 \in \{0.1, 5, 23\}$ is the initial temperature. Figure 8 indicates the negative MCL SI-SDR score computed during training on a validation set. In order to compare meaningfully the different temperature schedules, we plot the separation score against the temperature, regardless of the training step when this temperature is reached.

We see that that aMCL exhibits phase transitions in this setting. Indeed, when we use a scheduler with high initial temperature $T_0 \in \{5, 23\}$, the negative MCL SI-SDR metric stays on a plateau and then decreases after some critical temperature T_0^c has been reached. As expected, this critical temperature seems to be independent of the initial temperature $T_0 > T_0^c$. Moreover, when using a scheduler with initial temperature $T_0 \approx 0.1 < T_0^c$, aMCL does not undergo a phase transition. This suggests that the phase transition phenomenon exhibited in Section A also exists in the non-Euclidean case, even though the properties of the barycenter and the shape of the Voronoi cells have to be redefined. Further work will analyze more exhaustively the impact of the scheduler type on this phase transition phenomenon. We hypothesize that the temperature decay speed at the critical temperature T_0^c will play a key role in this analysis.

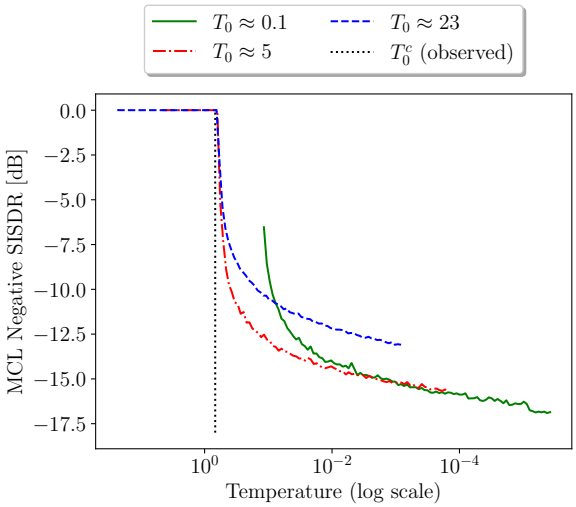


Figure 8: **Phase transition in speech separation training.** Impact of the initial temperature T_0 of the temperature scheduler on the source separation performance during training. The y-axis corresponds to the negative MCL SI-SDR metric. The x-axis corresponds to the temperature $T(t)$ at each training step t, and is displayed in logarithmic scale. Comparison of several initial temperatures $T_0 \approx 0.1$ (green solid line), $T_0 \approx 5$ (red dashed and dotted line), and $T_0 \approx 23$ (blue dashed line). A lower score indicates better separation performance.