Beyond Mixtures and Products for Ensemble Aggregation: A Likelihood Perspective on Generalized Means

Raphaël Razafindralambo[1] **Rémy Sun**[1] **Damien Garreau**[2] **Frédéric Precioso**[1] **Pierre-Alexandre Mattei**[1]

[1]Université Côte d'Azur, Inria, CNRS, I3S/LJAD, Maasai, Nice, France
[2]Julius-Maximilians-Universität Würzburg, Institute for Computer Science / CAIDAS, Würzburg, Germany

Abstract

Density aggregation is a central problem in machine learning, for instance when combining predictions from a Deep Ensemble. The choice of aggregation remains an open question with two commonly proposed approaches being linear pooling (probability averaging) and geometric pooling (logit averaging). In this work, we address this question by studying the normalized generalized mean of order $r \in \mathbb{R} \cup \{-\infty, +\infty\}$ through the lens of log-likelihood, the standard evaluation criterion in machine learning. This provides a unifying aggregation formalism and shows different optimal configurations for different situations. We show that the regime $r \in [0, 1]$ is the only range ensuring systematic improvements relative to individual distributions, thereby providing a principled justification for the reliability and widespread practical use of linear ($r = 1$) and geometric ($r = 0$) pooling. In contrast, we show that aggregation rules with $r \notin [0, 1]$ may fail to provide consistent gains with explicit counterexamples. Finally, we corroborate our theoretical findings with empirical evaluations using Deep Ensembles on image and text classification benchmarks. Our Python implementation is publicly available at `https://github.com/rarazafin/deep_ensemble_generalized`.

1 INTRODUCTION

Modern machine learning (ML) increasingly relies on collections of probabilistic models rather than a single monolithic predictor (Lakshminarayanan et al., 2017; Liu et al., 2024). This naturally raises the question of how to integrate several probability distributions into a single coherent distribution (French, 1983; Genest and Zidek, 1986). Distribution integration arises in many settings in ML, including ensembling classifier outputs (Kuncheva, 2014, Section 5), Mixup training techniques (El-Laham et al., 2025), Bayesian inference (Carvalho et al., 2023), the aggregation of Gaussian process experts (Liu et al., 2018; Cohen et al., 2020), mixture-of-experts architectures in large language models (Liu et al., 2024; Cai et al., 2025), and generative modeling (Liu et al., 2022; Razafindralambo et al., 2026).

Two canonical approaches are particularly widespread. The first consists in forming a mixture of distributions (Jacobs et al., 1991; Jordan and Jacobs, 1994), corresponding to an arithmetic average of densities. The second relies on multiplicative aggregation, where distributions are combined through a normalized product (Hinton, 2002). These methods offer contrasting properties (Winkler, 1968): mixtures act as a logical 'OR', capturing heterogeneity by pooling the supports, whereas multiplicative models act as a logical 'AND', sharpening the density where components overlap and thus favoring regions of consensus. This viewpoint is recurrent in compositional generative modeling (Liu et al., 2022; Du et al., 2023). More generally in ML, the choice between both has been a fundamental and practical question; for instance, Buchanan et al. (2023) empirically compare logit and probability averaging in multi-class classification.

A key theoretical motivation for mixtures lies in their variance-reduction effect, allowing the collective evaluation to outperform the average individual ones (Hastie et al., 2009, Chapter 15; Mattei and Garreau, 2025), a phenomenon related to the broader "wisdom of crowds" principle (Surowiecki, 2004). When predictors are diverse, individual errors tend to compensate for one another, improving collective performance.

Beyond the classical arithmetic and geometric means, several works have sought to go beyond and generalize the notion of averaging probability distributions (Genest and Zidek, 1986; Amari, 2007) to reveal new aggregation properties and fundamental trade-offs underlying different pooling

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5677–5699.

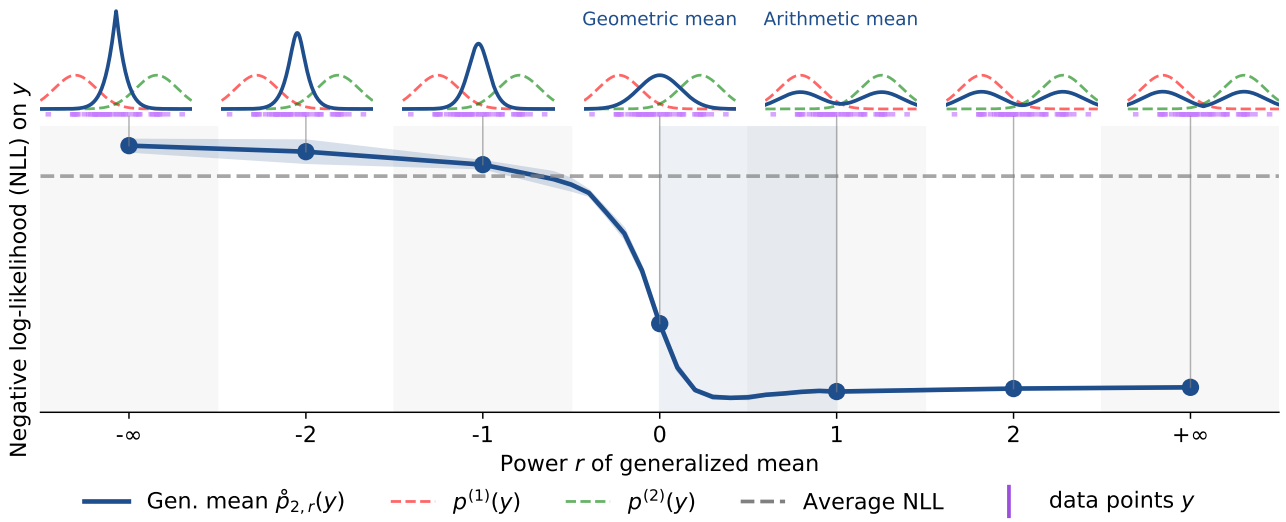


Figure 1: **Only intermediate powers r, particularly $r \in [0, 1]$, yield consistent NLL improvements, whereas extreme values can degrade performance. Top:** aggregated densities $\mathring{p}_2$ obtained from two Gaussian experts $p^{(1)}$ and $p^{(2)}$ for various r. Small r concentrates mass in a single region, while large positive r preserves the bimodal structure of the experts. **Bottom:** NLL evaluated on samples $y \sim \mathcal{N}(0, 4^2)$ (purple ticks indicate test samples). In this setting, negative r (min-type behavior) underperforms the average individual NLL (dashed line), whereas positive r achieves the lowest values. Our theory (Theorem 3.1) identifies $r \in [0, 1]$ as the provably reliable interval. See Appendix B for different behaviors.

rules (Elkin and Pettigrew, 2025). These efforts have led to continuous families of pooling operators that extend standard averages and include a broad spectrum of aggregation rules, ranging from harmonic and arithmetic means to more extreme operators such as max- or min-based pooling. In particular, Amari (2007) introduces an α-family of operators parameterized by a free parameter α, and establishes optimality properties with respect to α-divergences for each member of this family.

We examine whether density aggregation through generalized means of order $r \in \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ (Hardy et al., 1952) provides improved aggregation rules for model ensembling. A natural criterion in ML to assess an aggregated distributions is the data log-likelihood, as it directly measures how well an aggregated model explains observed data. This objective is central to modern learning procedures, as it underlies cross-entropy loss minimization in supervised classification and KL-based objectives in broader settings such as generative modeling and variational inference (Bishop, 2006, Chapter 10; Theis et al., 2015; Goodfellow et al., 2016, Chapter 5). This perspective therefore motivates both a theoretical and empirical investigation into whether variance reduction extends to generalized pooling operators and improve log-likelihood over individual models.

In this work, we show that our defined normalized generalized mean of order r yields a reliable aggregation rule in terms of log-likelihood, with theoretical guarantees and empirical evidence that gains over individual distributions are consistently observed for $r \in [0, 1]$. The idea is summarized in Figure 1, which illustrates how the shape of normalized generalized mean density varies with r along with the unreliable behavior of negative log-likelihood induced by negative r values.

Our main contribution lies in analyzing, through the lens of log-likelihood, the generalized mean of order $r \in \overline{\mathbb{R}}$ which defines valid densities (Section 3.1) with properties determined by r (Sections 2 and 3). More precisely, this translates into the following results:

- We identify in Section 3.2 a theoretical regime $r \in [0, 1]$ ensuring consistent log-likelihood benefits for any set of densities, uniformly across all data points. This regime includes the arithmetic and geometric means as boundary cases, explaining their practical successes.
- We characterize in Section 3.3 failure points outside this interval based on agreement and disagreement of density values, highlighting distinct behaviors for $r < 0$ and $r > 1$. This explains the lack of reliability outside $[0, 1]$ compared to the guarantees within the interval, as observed in Figure 1.
- We empirically corroborate the theoretical findings through Deep Ensemble experiments on classification tasks on vision and sentiment analysis (Section 4). We also observe that the optimal r is non trivial but close to the $[0, 1]$ range.

2 BACKGROUND AND NOTATIONS ON GENERALIZED MEANS

We consider a data space $\mathcal{X}$ equipped with a reference measure $\mathrm{d}\mathbf{x}$. We focus on probability distributions that admit densities with respect to $\mathrm{d}\mathbf{x}$. We denote by $\mathcal{P}$ the set of probability densities $p : \mathcal{X} \to \mathbb{R}_{\geq 0}$ satisfying $\int_{\mathcal{X}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$.

Given $k \geq 1$ densities $p^{(1)}, \ldots, p^{(k)} \in \mathcal{P}$, this work studies the construction of an aggregated density $\mathring{p}$ obtained by combining these distributions through generalized averaging operators, followed in most of the cases by a normalization step to ensure that $\mathring{p} \in \mathcal{P}$.

We first introduce formal notation for the mixture and multiplicative aggregation, as they constitute the most popular reference aggregations for our results. We then turn to the generalized mean of order r in $\overline{\mathbb{R}}$, which forms the central object of this paper and subsumes these operators.

2.1 CANONICAL EXAMPLES: LINEAR AND LOGARITHMIC POOLING

We introduce below the two standard combinations for probability density functions $p^{(1)}$ and $p^{(2)}$ in $\mathcal{P}$, and recall a few of their basic properties.

The *mixture* aggregation, also known as *linear pooling / opinion pool* (Stone, 1961; Genest et al., 1986; Clemen and Winkler, 1999; Elkin and Pettigrew, 2025), is defined for two probability as

$$\mathring{p}_{\text{lin}}(\mathbf{x}) \coloneqq \frac{1}{2}\Big(p^{(1)}(\mathbf{x}) + p^{(2)}(\mathbf{x})\Big). \tag{1}$$

This rule is highly democratic, as each density retains its influence in the aggregation, often resulting in a multimodal distribution. The mixture admits an appealing variational characterization (Amari, 2007): it minimizes the average (forward) Kullback–Leibler divergence,

$$\mathring{p}_{\text{lin}} = \arg\min_{q \in \mathcal{P}} \frac{1}{2} \sum_{i=1}^{2} \mathrm{KL}\Big(p^{(i)} \,\|\, q\Big). \tag{2}$$

The *multiplicative* aggregation is defined as the normalized product

$$\mathring{p}_{\text{mult}}(\mathbf{x}) \coloneqq \sqrt{p^{(1)}(\mathbf{x})\, p^{(2)}(\mathbf{x})}/Z, \tag{3}$$

where $Z = \int_{\mathcal{X}} \sqrt{p^{(1)}(\mathbf{x})\, p^{(2)}(\mathbf{x})}\,\mathrm{d}\mathbf{x}$ is the normalization constant. This operation is also commonly referred to as a *Product-of-Experts* (PoE) (Hinton, 2002; Cohen et al., 2020; Razafindralambo et al., 2026), but is also known in the literature as *logarithmic pooling / opinion pool* (Carvalho et al., 2023; Genest et al., 1986; Clemen and Winkler, 1999; Elkin and Pettigrew, 2025) or *exponential mixture* (Amari, 2007). The PoE is well-defined since the geometric mean of densities in $\mathcal{P}$ remains normalizable (Genest et al., 1986; Carvalho et al., 2023). Geometric averaging reduces to pre-softmax logit averaging in classification; see Appendix A.3.

The PoE enjoys several theoretical properties that sharply contrast with those of the mixture model. The PoE minimizes the average *reverse* KL divergence to the individual models (Amari, 2007),

$$\mathring{p}_{\text{mult}} = \arg\min_{q \in \mathcal{P}} \frac{1}{2} \sum_{i=1}^{2} \mathrm{KL}\Big(q \,\|\, p^{(i)}\Big). \tag{4}$$

Moreover, it typically exhibits a more *pessimistic* (and even dictatorial) behavior: regions assigned low density by any individual model are strongly penalized, and if $p^{(k)}(x) = 0$ for some k, then $\mathring{p}_{\text{mult}}(x) = 0$ as well. As a result, PoE aggregations tend to produce distributions that are more concentrated, and often closer to unimodal compared to mixtures (Winkler, 1968; Genest and Zidek, 1986). Figure 1 provides an illustration (see $r = 0$).

Other notable properties of the PoE include *independence preservation* (Laddaga, 1977; Cooke, 1991, Chapter 11) and the *externally* Bayesian property (Genest et al., 1986; Cooke, 1991, Chapter 11). A more comprehensive overview of the properties for both linear and logarithmic pooling is also provided in Elkin and Pettigrew (2025) (Section 4).

Both aggregation schemes also have information-geometric interpretations: they correspond to the midpoints of two natural geodesics connecting $p^{(1)}$ and $p^{(2)}$ (see, e.g. Amari, 2016, Section 2.4).

2.2 GENERALIZED POWER MEAN

We now discuss a generalized aggregation framework that extends the operators defined above. In this work, we adopt the generalized mean following the definition of Hardy et al. (1952), referred to as the *ordinary mean*, and extend it to the setting of probability densities. Given k positive values $a_1, \ldots, a_k$, we define the *generalized (or power) mean of order* r by

$$M_r(a_1, \ldots, a_k) \coloneqq \left(\frac{1}{k} \sum_{i=1}^{k} a_i^r\right)^{1/r}, \quad r \in \mathbb{R} \setminus \{0\}. \tag{5}$$

The continuous extension at $r = 0$ corresponds to the geometric mean, which naturally arises as the limit of M_r when r tends to zero. We define

$$M_0(a_1, \ldots, a_k) \coloneqq \left(\prod_{i=1}^{k} a_i\right)^{1/k}. \tag{6}$$

The extreme cases recover the minimum and maximum values,

$$M_{-\infty}(a_1, \ldots, a_k) \coloneqq \lim_{r \to -\infty} M_r(a_1, \ldots, a_k) = \min_i a_i,$$

$$M_{+\infty}(a_1, \ldots, a_k) \coloneqq \lim_{r \to +\infty} M_r(a_1, \ldots, a_k) = \max_i a_i.$$

Classical choices recover well-known means: $r = 1$ corresponds to the arithmetic mean (AM), $r = 0$ to the geometric mean (GM), $r = -1$ to the harmonic mean. In fact the parameter r controls the behavior of the generalized mean and can be interpreted as an optimism parameter. Large values of r emphasize the largest inputs and lead to an optimistic aggregation, while small (negative) values emphasize the smallest inputs and result in a more pessimistic behavior. More precisely, larger values of r make the mean a smooth approximation of the maximum, while smaller values make it a smooth approximation of the minimum.

A key property of the generalized mean is its monotonicity with respect to the order r, which will play a central role in our analysis.

Lemma 2.1 (Power mean inequality, Hardy et al., 1952; Chapter 2). *Let $a_1, \ldots, a_k$ be positive values. If $r < s$,*

$$M_r(a_1, \ldots, a_k) \leq M_s(a_1, \ldots, a_k). \tag{7}$$

where equality holds if and only if $a_1 = a_2 = \cdots = a_k$. This result generalizes the inequality chain $\frac{a_1+a_2}{2} \geq \sqrt{a_1 a_2} \geq \frac{2}{\frac{1}{a_1}+\frac{1}{a_2}}$ where the left inequality is known as AM-GM.

Hardy et al. (1952) provide an extensive list of elementary properties of M_r, including fundamental inequalities. Power means can be further generalized to f-means, where aggregation is defined through a transformation function f; see De Carvalho (2016) for a characterization of generalized means as f-means and their statistical properties, and Barczy and Páles (2026) for a loss-function-based analysis.

Generalized means of probability densities. We now extend the concept to the setting of probability densities. Given k probability densities $p^{(1)}, \ldots, p^{(k)} \in \mathcal{P}$, we define

$$M_{k,r}(\mathbf{x}) \coloneqq M_r\big(p^{(1)}(\mathbf{x}), \ldots, p^{(k)}(\mathbf{x})\big). \tag{8}$$

This yields a nonnegative function on $\mathcal{X}$, which does not necessarily integrate to one. For example, taking the geometric mean of two Gaussian densities does not in general produce a normalized density, and a normalization step is therefore required.

Definition 2.1 (Generalized power mean of densities). *The mean of order r associated with $p^{(1)}, \ldots, p^{(k)} \in \mathcal{P}$ is defined as*

$$\overset{\circ}{p}_{k,r}(\mathbf{x}) \coloneqq \frac{1}{Z_{k,r}} M_{k,r}(\mathbf{x}), \tag{9}$$

where the normalization constant is given by

$$Z_{k,r} \coloneqq \int_{\mathcal{X}} M_{k,r}(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{10}$$

By construction, $\overset{\circ}{p}_{k,r} \in \mathcal{P}$ whenever $Z_{k,r}$ is finite, which is true for all r (see Proposition 3.1).

2.3 RELATED WORK

Closely related to our approach, Cooke (1991) study a family of density aggregation operators of the form $\overset{\circ}{p}_{k,r} \propto M_{k,r}$, which coincides with our r-family of aggregations. They derive several immediate properties of this family, including properties analogous to those mentioned in Section 2.1 for the canonical cases, and Equation (14).

Differently, Amari (2007) introduce an *α-family of integrations*, where $\alpha \in \mathbb{Z} \cup \{-\infty, +\infty\}$, based on the α-representation of densities in information geometry (Amari and Nagaoka, 2000), which also includes the classical means as special cases. They relate their mean to the α-divergence $D_\alpha(p \,\|\, q)$ for any $p, q \in \mathcal{P}$ (Chernoff, 1952; Amari and Nagaoka, 2000), that include as special cases of α the KL, reverse KL, or the square of Hellinger. Amari (2007, Theorem 2) show that the α-integration is optimal under the α-divergence criterion, in the sense that it minimizes the weighted average α-divergence to the individual distributions. This result generalizes Equations (2) and (4), and connects to earlier Bayesian formulations of generalized predictive densities (Corcuera and Giummolè, 1999).

Finally, a closely related line of work studies the general theory of distribution aggregation under the name of *opinion pooling* (Elkin and Pettigrew, 2025). In this framework, aggregation is defined abstractly as a parameter-free operator $F : (P_1, \ldots, P_n) \mapsto P^\star$, mapping a profile of probability distributions to a collective one. This literature characterizes desirable axiomatic properties of such operators (e.g., unanimity preservation, independence, compatibility with Bayesian updating, ...).

3 LIKELIHOOD GUARANTEES FOR GENERALIZED MEAN AGGREGATION

We now present the core theoretical contribution of this paper by analyzing the generalized power mean aggregation and by identifying when it guarantees a systematic improvement in log-likelihood. After establishing that the generalized mean defines a normalizable density in Section 3.1, we prove in Section 3.2 that such an improvement holds for orders $r \in [0, 1]$, precisely between the geometric and arithmetic means, explaining their robustness. We finally show in Section 3.3 via counterexamples that this guarantee fails outside this interval, with qualitatively different failure modes for $r < 0$ and $r > 1$.

3.1 DOES EVERY ORDER r DEFINE A DENSITY?

An important point to examine is whether the normalization constant is finite for all values of r, and the following proposition shows that it is indeed the case.

Proposition 3.1 (Generalized means of order r are well defined). *Assume $p^{(1)}, \dots, p^{(k)}$ are k positive probability density functions over $\mathcal{X}$. Then for all $r \in \overline{\mathbb{R}}$*

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty. \tag{11}$$

Proof. We apply Lemma 2.1 followed by the definition of $M_{k,+\infty}$ from Section 2.2. For all $r \in \overline{\mathbb{R}}$,

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\,\mathrm{d}\mathbf{x} \leq \int_{\mathbb{R}^d} M_{k,+\infty}(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{12}$$

$$\leq \int_{\mathbb{R}^d} \sum_{i=1}^{k} p^{(i)}(\mathbf{x})\,\mathrm{d}\mathbf{x} = k < +\infty. \tag{13}$$

Additionally, it is worth noting that when $r \leq 1$ (for example for the geometric mean),

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\mathrm{d}\mathbf{x} \leq \int_{\mathbb{R}^d} M_{k,1}(\mathbf{x})\mathrm{d}\mathbf{x} = 1. \tag{14}$$

□

To the best of our knowledge, no previous work has established a bound on the normalizing constant for all real values of r. Proofs for the geometric pooling ($r = 0$) appear in (Genest et al., 1986, p. 489, Carvalho et al., 2023, Theorem 2.1). The case $r \leq 1$, also mentioned by Cooke (1991, Chapter 11), is of particular interest: the bound on the normalizing constant in Equation (14) implies $-\log Z_{k,r} \geq 0$. The sign of $-\log Z_{k,r}$ is crucial for analyzing $\overset{\circ}{p}_{k,r}$, since the log-likelihood can be written as $\log \overset{\circ}{p}_{k,r} = \log M_{k,r} - \log Z_{k,r}$. It therefore directly determines whether normalization increases or decreases the overall likelihood, which is central to our arguments (see Theorem 3.1).

3.2 WHEN IS THE GENERALIZED MEAN RELIABLY BENEFICIAL?

We examine the important question of when aggregation guarantees an improvement in comparison to the average individual log-likelihood, which would correspond to a "wisdom of crowds" effect, and show that this occurs exactly for $r \in [0, 1]$.

Theorem 3.1 (Wisdom of Crowds on Log-likelihood). *Let $\mathbf{x} \in \mathcal{X}$ and $k \in \mathbb{N}^*$. Assume all density functions $p^{(1)}, \dots, p^{(k)}$ are positive. If $r \in [0, 1]$,*

$$\log \overset{\circ}{p}_{k,r}(\mathbf{x}) \geq \frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(\mathbf{x}). \tag{15}$$

Proof. Let $\mathbf{x} \in \mathcal{X}$, and $0 < r \leq 1$. By Jensen's inequality (concavity of log) and since $r > 0$,

$$\log M_{k,r}(\mathbf{x})/Z_{k,r} = \frac{1}{r} \log \left(\frac{1}{k} \sum_{i=1}^{k} [p^{(i)}(\mathbf{x})]^r \right) - \log Z_{k,r} \tag{16}$$

$$\geq \frac{1}{rk} \sum_{i=1}^{k} \log([p^{(i)}(\mathbf{x})]^r) - \log Z_{k,r}. \tag{17}$$

Hence,

$$\log \overset{\circ}{p}_{k,r}(\mathbf{x}) \geq \frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(\mathbf{x}) - \log Z_{k,r}. \tag{18}$$

Note that for $r = 0$, Equation (18) still holds by using $\log \prod = \sum \log$.

Since $r \leq 1$, it is straightforward from Equation (14) that

$$\log M_{k,r}(\mathbf{x})/Z_{k,r} \geq \frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(\mathbf{x}) \tag{19}$$

since $\log Z_{k,r} \leq 0$. □

Log-likelihood quantifies how well a model p explains data, and larger values indicating better fit. Theorem 3.1 therefore establishes aggregation reliability for $r \in [0, 1]$, by guaranteeing that, for any data point $\mathbf{x}$, the aggregated model achieves a log-likelihood at least as large as the average of individual ones. Thus $[0, 1]$ can be seen as a "safe" interval.

The result is particularly revealing: the geometric mean ($r = 0$) is the most pessimistic and least democratic aggregation that remains reliably beneficial, while the arithmetic mean ($r = 1$) is the least pessimistic and most democratic one to achieve this. Together, they span the entire regime of safe aggregation, explaining why these two classical rules are privileged in the literature.

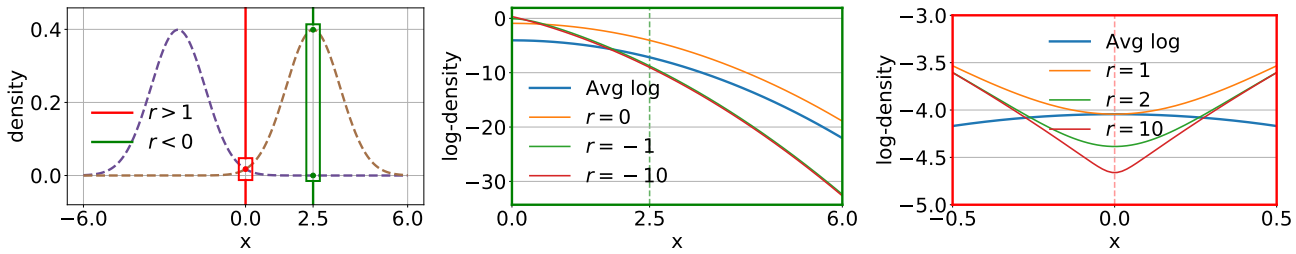
3.3 FAILURE CASES OUTSIDE THE RELIABILITY RANGE

The "wisdom of crowds" effect no longer holds when $r < 0$ or $r > 1$, as Equation (15) fails to hold for all $\mathbf{x}$. This is formalized in the following theorem.

Theorem 3.2 (Failure of the Wisdom of Crowds). *Let $r \notin [0, 1]$. Then there exist two positive densities $p^{(1)}, p^{(2)} \in \mathcal{P}$ with $p^{(1)} \neq p^{(2)}$ and a point $\mathbf{x} \in \mathcal{X}$ such that*

$$\log \overset{\circ}{p}_{2,r}(\mathbf{x}) < \frac{1}{2} \sum_{i=1}^{2} \log p^{(i)}(\mathbf{x}). \tag{20}$$

We establish Theorem 3.2 through **two different types of counter-examples** on $\mathbf{x}$ depending on the **position of** r



(a) The two experts (b) $r < 0$ (c) $r > 1$

Figure 2: **Visual illustration of Theorem 3.2. Aggregation fails at points where densities strongly differ when $r < 0$, and at consensus points when $r > 1$.** We look at the location of the gap in Equation (20) across r regimes. "Avg log" denotes the average of individual log-likelihoods (right-hand side of Equation (20)). Left (a): We consider two gaussians $\mathcal{N}(\pm 2.5, 1)$. Center (b): For $r < 0$, the inequality of Theorem 3.1 is satisfied at $x = 2.5$. We observe that the gap is not confined and amplifies when x increases. Right (c): For $r \geq 1$, the AM ($r = 1$) globally dominates the average log-density, while all $r > 1$ exhibit a very localized (but non-punctual) negative effect around $x = 0$.

relative to the interval $[0, 1]$. More precisely, when $r < 0$, Equation (15) breaks down at disagreement points, while for $r > 1$ it fails at consensus points. As an illustrative example supporting this point and Theorem 3.2, we consider two symmetric Gaussian densities defined as

$$p^{(1)}(x) = \mathcal{N}(x; -m, 1) \text{ and } p^{(2)}(x) = \mathcal{N}(x; m, 1). \quad (21)$$

The aggregation breaks down at disagreement points. When $r < 0$, $x = m$ satisfies Equation (20). This point corresponds to a pronounced imbalance, since for large m

$$p^{(2)}(m) \gg p^{(1)}(m) \approx 0, \quad (22)$$

that is a location where the densities strongly disagree. The gap is in fact exponential, as

$$\frac{p^{(2)}(m)}{p^{(1)}(m)} = e^{2m^2}, \quad (23)$$

and this situation is illustrated in Figure 2a in which $x = m$ corresponds to the mode of $p^{(2)}$ and the tail of $p^{(1)}$. Hence, it results in the average log-likelihood exceeding the log power mean on this point when $r < 0$, as shown in Figure 2b. This behavior arises because pessimistic (min-like) power means heavily penalize regions where at least one density assigns nearly zero probability mass. A detailed proof is provided in Appendix C.1, where we show that this penalization effect for large m reverses Jensen's inequality from Equation (17).

The aggregation breaks down at consensus points. When $r > 1$, we show in contrast that $x = 0$ satisfies Equation (20). This corresponds to a positive consensus point, namely a location where both densities agree and assign nonzero probability mass. Indeed, at this point,

$$p^{(1)}(0) = p^{(2)}(0) > 0, \quad (24)$$

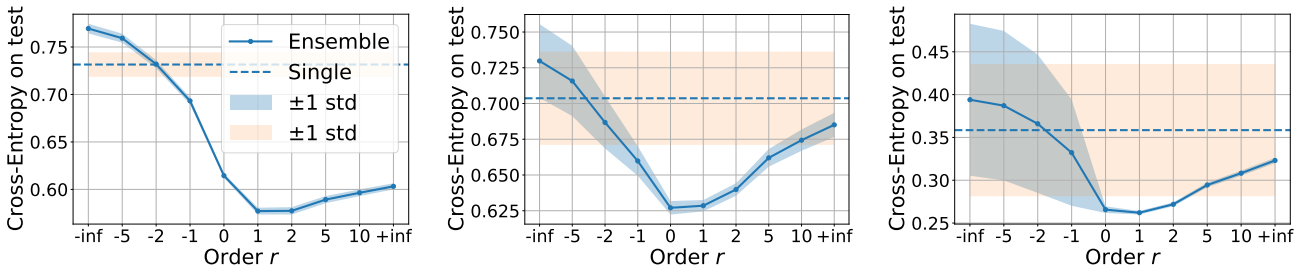
while the two densities differ elsewhere. Such a situation is illustrated in Figure 2a. As shown in Figure 2c, the average log-likelihood exceeds $\mathring{p}_{2,r}$ at this point when $r > 1$. Unlike the previous case, this agreement is localized, resulting in a highly local breakdown of the inequality. The failure originates from the normalization: although max-like aggregation favors large values, normalization redistributes probability mass toward regions where one density dominates, thereby weakening the contribution of the consensus point. A detailed proof is provided in Appendix C.1, where we show explicitly that $Z_{2,r}$ enforces Equation (20) even though Jensen's inequality holds with equality.

Non-continuous counter-examples. The phenomenon is even more pronounced in the extreme regimes and can be illustrated in the classification setting. For $r = -\infty$ (minimum), the binary classifiers $p^{(1)} = (0.9, 0.1)$ and $p^{(2)} = (0.01, 0.99)$, with true label $y = 0$, form a counter-example. Likewise, for $r = +\infty$ (maximum), the ternary classifiers $p^{(1)} = (0.9, 0.03, 0.07)$ and $p^{(2)} = (0.9, 0.07, 0.03)$, both agreeing with the true label $y = 0$, also violate the inequality. See Appendix C.2 for details.

4 EXPERIMENTS

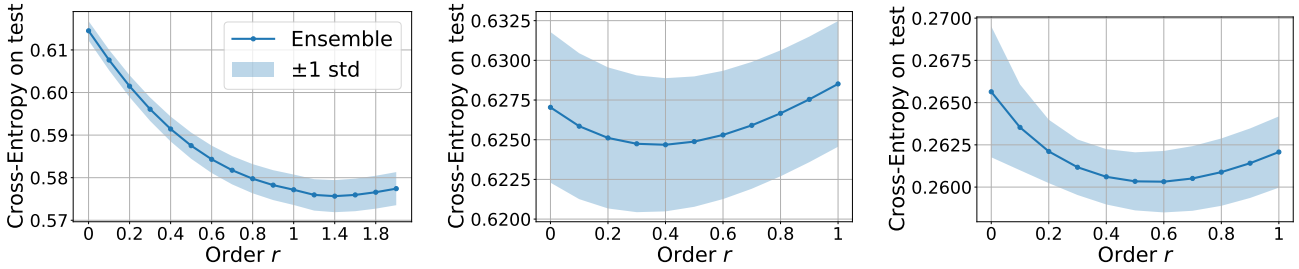
In this section, we empirically test whether generalized mean aggregation within the theoretical safe regime $r \in [0, 1]$ from Theorem 3.1 yields actual improvements in negative log-likelihood across all models and tasks considered, and analyze how behavior degrades outside this range.

How to build models to ensemble. To study aggregation behavior in practice, we train multiple neural networks with same architectures independently on the same dataset and loss, using different random initializations and data shuffles. This setup corresponds to the *deep ensemble* (DE) framework introduced by Lakshminarayanan et al. (2017) and Fort et al. (2019). In this case, each predictive distribution $p^{(i)}$ is represented by the softmax output of an indepen-



(a) CIFAR-100 (b) MedMNIST (c) IMDb

Figure 3: **Illustration of the cross-entropy behavior of the normalized power mean likelihood $\mathring{p}_{k,r}$ (via cross-entropy) in three classification settings.** All plots exhibit a U-shaped performance curve: extreme aggregation amplifies model disagreement, whereas optimal values lie in $[0, 1]$. Consistent with Theorem 3.1, the regime $r \in [0, 1]$ remains reliably below the individual model uncertainty band. By contrast, negative orders ($r < 0$) perform poorly, likely due to disagreements on dominant classes, as discussed in Section 3.3.



(a) CIFAR-100 (b) MedMNIST (c) Imdb

Figure 4: **Illustration of the cross-entropy behavior of the normalized power mean likelihood $\mathring{p}_{k,r}$, focusing on values of r around $[0, 1]$ (zoomed view of Figure 3).** On MedMNIST (b) and IMDb (c), the optimal likelihood lies within the reliability interval $[0, 1]$ (Theorem 3.1), whereas on CIFAR-100 (a) it lies slightly beyond it (≈ 1.4). This shows that while the interval $[0, 1]$ provides a stable and reliable regime, mild optimism outside it can still be beneficial in practice.

dently trained classifier, and aggregation is performed over multiple such models. For each experiment, we build ensembles of size 10 and repeat the process across 5 independent ensemble draws to reduce variance. Inference-time aggregation uses the generalized mean of order r (Definition 2.1).

Datasets and architectures. We evaluate aggregation across vision, medical imaging, and NLP to test the generality of the phenomenon. For multi-class classification, we use CIFAR-100 (Krizhevsky, 2009), a benchmark of 60k low-resolution color images spanning 100 classes with only 600 images per class. The large label space and limited per-class data increase classification difficulty and the likelihood of inter-model disagreement. In the medical domain, we consider DermaMNIST (Yang et al., 2023), which contains dermatoscopic skin lesion images across diagnostic categories. The dataset is strongly imbalanced, with benign melanocytic nevi representing about two thirds of samples while several classes account for only 1–5%. We train standard convolutional residual networks (He et al., 2016; Zagoruyko and Komodakis, 2016) for both vision tasks. Finally, to assess aggregation in NLP, we train transformers (Vaswani et al., 2017) on IMDb (Maas et al., 2011), a balanced binary sentiment classification dataset of 50k movie reviews.

Evaluation. We evaluate performance using test cross-entropy, which corresponds to the negative log-likelihood of the true labels under the predictive distribution. We measure how it varies as a function of r, with particular attention to the predicted safe regime $r \in [0, 1]$.

4.1 INTERMEDIATE ORDERS ARE MORE RELIABLE THAN EXTREMES

By analyzing aggregation across the full range $r \in \overline{\mathbb{R}}$, we show in Figure 3 that performance follows a characteristic U-shaped curve, with reliable improvements confined to the regime $r \in [0, 1]$.

Across all three datasets extreme orders degrade performance, whereas intermediate values, in particular $[0, 1]$ (cf. Theorem 3.1), consistently outperform individual models. This pattern holds even under strong inter-model uncertainty (Figure 3c), where all $r \in [0, 1]$ remain below the

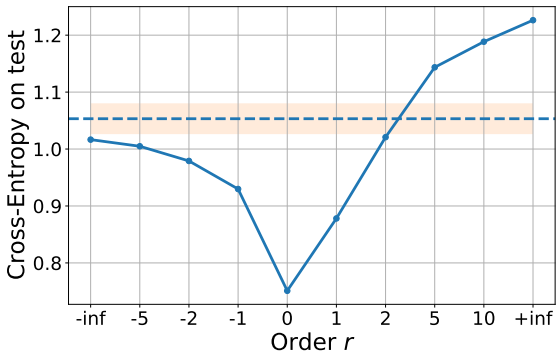


Figure 5: **Illustration of the cross-entropy behavior of the normalized power mean likelihood $\mathring{p}_{k,r}$ on CIFAR-100 in a controlled near-consensus regime.** The trend contrasts with Figure 3: extreme optimistic aggregation becomes harmful and the maximum operator performs worst.

uncertainty band, supporting the reliability of this interval.

At the extremes, $r = -\infty$ yields worse cross-entropy than individual models, likely due to amplified disagreement (Section 3.3), especially on CIFAR-100 where the large label space increases conflicting predictions. Conversely, on DermaMNIST, $r = +\infty$ reaches the individual uncertainty region, suggesting that max-like aggregation can amplify shared confident errors likely induced by class imbalance.

Finally, ensemble variance is lower than that of individual models for all r and decreases as r increases, with intermediate values achieving both the strongest variance reduction and the best log-likelihoods, supporting the "wisdom of crowds" effect.

We report additional evaluation metrics (Accuracy, Brier score, and ECE) in Appendix D. They exhibit a similar U-shaped trend as cross-entropy, with intermediate values $r \in [0, 1]$ being reliable and yielding better performance.

4.2 WHAT IS THE OPTIMAL ORDER r?

We reveal through a finer sweep on each plot (Figure 4) that the optimal r is typically intermediate rather than located at extreme values, but that despite $I = [0, 1]$ being the theoretically reliable regime, the empirically optimal r value can sometimes lie outside of it.

In Figures 4b and 4c, the empirical optimum lies in $[0.3, 0.5] \subset I$ and $[0.5, 0.7] \subset I$, respectively. By contrast, Figure 4a suggests that mild optimism ($r \in [1.2, 1.6] \not\subseteq I$) can still improve model likelihood in practice.

This shows that the interval $[0, 1]$ does not guarantee optimality, as better empirical performance may occur for values of r exceeding 1. Related work by Safont et al. (2019) learns the optimal aggregation parameter directly from data for classification, within an alternative family of aggregation rules derived from α-divergences.

We note that no universal ordering exists between the canonical values $r = 0$ and $r = 1$, as neither consistently dominates across datasets as observed in Figure 4. Interestingly, the arithmetic mean ($r = 1$) appears less favorable on the imbalanced dataset, in line with observations reported by Buchanan et al. (2023) in predictive settings.

4.3 NEAR-CONSENSUS MAKES OPTIMISM HARMFUL

We show in Figure 5 that aggregation behavior can reverse in comparison to Figures 3a to 3c when models assign nearly equal probability to one class while differing elsewhere: extreme optimistic rules become harmful, and the maximum-type aggregation performs worst. This supports the counter-example in Section 3.3.

Such structured near-consensus is difficult to obtain with independently trained predictors, as agreement typically fluctuates across classes and inputs. To isolate this effect (Figure 5), we duplicate a trained CIFAR-100 predictor and add small Gaussian perturbations to all logits except the true-class one, while keeping probabilities normalized

5 CONCLUSION

We studied the impact of aggregating probability distributions through the normalized generalized mean of order r on densities, a continuous family encompassing classical mixture and geometric aggregation rules as special cases. We theoretically show that $r \in [0, 1]$ is the regime ensuring systematic log-likelihood improvements over individual models, while outside orders may fail with distinct mechanisms. Simple Gaussian settings and empirical validation demonstrate that our results hold in the specific case of supervised task; but one can expect them to extend to broader contexts. Notably, we confirm the distinguished roles of the geometric and arithmetic means, whose widespread use aligns with their delimitation of the theoretical safe regime. Our work evaluates density operations through log-likelihood, while prior studies focus on accuracy. For instance, Buchanan et al. (2023) compare averaging normalized logits and softmax outputs (logarithmic vs. linear pooling) and observe dataset-dependent behavior, with logarithmic pooling ($r = 0$) favored on long-tailed datasets. This confirms the role of data in aggregation performance.
While our results identify the interval $r \in [0, 1]$ as a provably reliable regime, experiments suggest that the empirically optimal value may depend on properties of the data or model. Consequently, a natural direction for future work is to better characterize the optimal aggregation order r.

Acknowledgements

This work was supported by the French government through the France 2030 programme managed by the ANR, with the reference numbers "ANR-23-IACL-0001", "ANR-15-IDEX-01", and "ANR-21-CE23-0005-01".

References

Shun-ichi Amari. Integration of stochastic models by minimizing α-divergence. *Neural Computation*, 19(10):2780–2796, 2007.

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Matyas Barczy and Zsolt Páles. Expected loss of quasi-arithmetic means of exchangeable random variables. *arXiv preprint arXiv:2606.22183*, 2026.

Christopher M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

E. Kelly Buchanan, Geoff Pleiss, and John P. Cunningham. The effects of ensembling on long-tailed data. In *NeurIPS Workshop Heavy Tails in Machine Learning*, 2023.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Yanshuai Cao and David J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. In *NeurIPS Workshop Modern Nonparametrics 3: Automating the Learning Pipeline*, 2014.

Luiz M. Carvalho, Daniel A. M. Villela, Flavio C. Coelho, and Leonardo S. Bastos. Bayesian inference for the weights in logarithmic pooling. *Bayesian Analysis*, 18(1): 223–251, 2023. doi: 10.1214/22-BA1311.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Robert T. Clemen and Robert L. Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

Samuel Cohen, Rendani Mbuvha, Tshilidzi Marwala, and Marc P. Deisenroth. Healing products of Gaussian processes. *International Conference on Machine Learning*, 2020.

Roger Cooke. *Experts in uncertainty: opinion and subjective probability in science*. Oxford university press, 1991.

José Manuel Corcuera and Federica Giummolè. A generalized Bayes rule for prediction. *Scandinavian Journal of Statistics*, 26(2):265–279, 1999.

Miguel De Carvalho. Mean, what do you mean? *The American Statistician*, 70(3):270–274, 2016.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: compositional generation with energy-based diffusion models and MCMC. In *International conference on machine learning*, volume 202, pages 8489–8510. PMLR, 2023.

Yousef El-Laham, Niccolò Dalmasso, Svitlana Vyetrenko, Vamsi K Potluru, and Manuela Veloso. Mixup regularization: A probabilistic perspective. *arXiv preprint arXiv:2502.13825*, 2025.

Lee Elkin and Richard Pettigrew. *Opinion Pooling*. Cambridge University Press, 2025.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Simon French. *Group consensus probability distributions: A critical survey*. University of Manchester. Department of Decision Theory, 1983.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Christian Genest, Kevin J. McConway, and Mark J. Schervish. Characterization of externally Bayesian pooling operators. *The Annals of Statistics*, pages 487–501, 1986.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press Cambridge, 2016.

Godfrey H. Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer series in statistics New-York, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14 (8):1771–1800, 2002.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Ludmila I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

Robert Laddaga. Lehrer and the consensus proposal. *Synthese*, 36(4):473–477, 1977.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30:6402–6413, 2017.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Haitao Liu, Jianfei Cai, Yi Wang, and Yew Soon Ong. Generalized robust Bayesian committee machine for large-scale gaussian process regression. In *International Conference on Machine Learning*, pages 3131–3140. PMLR, 2018.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

Andrew Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

Pierre-Alexandre Mattei and Damien Garreau. Are ensembles getting better all the time? *Journal of Machine Learning Research*, 26(201):1–46, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Raphaël Razafindralambo, Rémy Sun, Frédéric Precioso, Damien Garreau, and Pierre-Alexandre Mattei. When are two scores better than one? Investigating ensembles of diffusion models. *Transactions on Machine Learning Research*, 2026.

Gonzalo Safont, Addisson Salazar, and Luis Vergara. Multiclass alpha integration of scores from multiple classifiers. *Neural Computation*, 31(4):806–825, 2019.

Mervyn Stone. The opinion pool. *The Annals of Mathematical Statistics*, pages 1339–1342, 1961.

James Surowiecki. *The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday, 2004.

Cedrique Rovile Njieutcheu Tassi, Jakob Gawlikowski, Auliya Unnisa Fitri, and Rudolph Triebel. The impact of averaging logits over probabilities on ensembles of neural networks. In *AISafety@ IJCAI*, pages 132–141, 2022.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Robert L. Winkler. The consensus of subjective probability distributions. *Management science*, 15(2):B–61, 1968.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, number 87, pages 1–46, 2016.

Beyond Mixtures and Products for Ensemble Aggregation: A Likelihood Perspective on Generalized Mean (Supplementary Material)

Raphaël Razafindralambo[1] **Rémy Sun**[1] **Damien Garreau**[2] **Frédéric Precioso**[1] **Pierre-Alexandre Mattei**[1]

[1]Université Côte d'Azur, Inria, CNRS, I3S/LJAD, Maasai, Nice, France
[2]Julius-Maximilians-Universität Würzburg, Institute for Computer Science / CAIDAS, Würzburg, Germany

This appendix complements the main paper with methodological details, supplementary illustrations, and additional theoretical derivations:

- Appendix A presents detailed experimental settings corresponding to the experiments of Section 4.
- Appendix B explores a broader range of behaviors of generalized means in the Gaussian setting, extending the illustrative example of Figure 1.
- Appendix C provides theoretical derivations of the "wisdom of crowds" counterexamples introduced in Section 3.3.
- Appendix E derives analytical expressions for the normalization constant in the $\mathbb{R}^d$ Gaussian case with $k \geq 1$ densities.

A EXPERIMENTAL DETAILS

We describe the experiments of Section 4 in greater detail, covering both the experimental setup and the numerically stable computation procedures.

A.1 DATASETS AND MODELS

We adopt three datasets, each paired with an architecture suited for the associated classification task.

We use CIFAR-100 (Krizhevsky, 2009), a benchmark dataset of 60,000 color images of size 32×32 spanning 100 object classes, with only 600 images per class. The classes cover a wide variety of natural objects, including animals, vehicles, household items, and everyday scenes. Our training set contains 50k images. We train WideResNet-28-10 models (Zagoruyko and Komodakis, 2016), where 28 denotes the network depth and 10 the widening factor controlling the number of channels. It is a popular architecture for small-resolution image classification tasks.

We also evaluate DermaMNIST (Yang et al., 2023), derived from the HAM10000 collection (Tschandl et al., 2018), a medical imaging dataset of dermatoscopic skin lesion images categorized into seven diagnostic classes. The dataset is relatively small and strongly class-imbalanced: benign melanocytic nevi represents about two thirds of samples while several classes account for only 1–5%. We adopt the official dataset split ($\approx$ 7k train / 1k val / 2k test). We use the standard 28×28 resolution version and adopt a ResNet-18 architecture, following the experimental setup recommended by the MedMNIST benchmark (Yang et al., 2023).

We finally evaluate on the IMDb movie reviews dataset (Maas et al., 2011), a large-scale natural language corpus of 50,000 user-written reviews from the Internet Movie Database, where the task consists in classifying reviews as positive or negative. We train lightweight Transformer-based text classifiers composed of token embeddings, sinusoidal positional encoding, and a small Transformer encoder followed by mean pooling and a linear classification head. We use a 35k/7.5k/7.5k train/validation/test split.

A.2 TRAINING THE ENSEMBLE

We train five independent Deep Ensembles (Lakshminarayanan et al., 2017) per experiment, each composed of ten models trained with cross-entropy loss. Models within each ensemble are trained independently with different random initializations and data shuffling. The best checkpoints are selected based on validation performance. Reported means and standard deviations are computed across the five ensembles to assess statistical stability.

A.3 EVALUATING THE GENERALIZED MEAN FOR CLASSIFICATION

In this section, we explain how generalized mean aggregation is implemented on softmax outputs for likelihood evaluation in classification.

In this setting, each model produces a predictive distribution over L classes. Given k predictors, we denote by

$$p^{(i)}(y = \ell \mid x), \qquad i = 1, \dots, k, \tag{25}$$

the probability assigned to class ℓ for an input x. In this case, averaging densities amounts to aggregating the predicted class probabilities themselves. We therefore apply the generalized (power) mean of order r across model predictions. For each class y, we define

$$M_{k,r}(y \mid x) = \begin{cases} \left(\frac{1}{k} \sum_{i=1}^{k} \left(p^{(i)}(y \mid x) \right)^r \right)^{\frac{1}{r}}, & r \neq 0, \\ \exp\left(\frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(y \mid x) \right), & r = 0. \end{cases} \tag{26}$$

Because this operation is applied independently to each class, the resulting scores do not necessarily sum to one. We therefore follow the necessary normalization step across classes to obtain a valid predictive distribution,

$$\mathring{p}_{k,r}(y \mid x) = \frac{M_{k,r}(y \mid x)}{\sum_{\ell=1}^{L} M_{k,r}(\ell \mid x)}. \tag{27}$$

Stable log-domain implementation. Direct computation in probability space may lead to numerical instability due to extremely small softmax probabilities and the normalization step involving sums of exponentials. We therefore perform aggregation in the log-domain using numerically stable $\mathrm{logsumexp}$ operations (e.g., `torch.logsumexp`[1]).

For finite $r \neq 0$, the generalized mean of predictions writes

$$\log M_{k,r}(y \mid x) = \frac{1}{r} \left(\log \sum_{i=1}^{k} \exp\big(r \log p^{(i)}(y \mid x) \big) - \log k \right), \tag{28}$$

which can be implemented efficiently using a $\mathrm{logsumexp}$ operation over models.

The predictive distribution is finally obtained by exponentiating and normalizing across classes,

$$\mathring{p}_{k,r}(y \mid x) = \exp\big(\log M_{k,r}(y \mid x) - \log Z(x) \big), \tag{29}$$

where the normalization constant is

$$\log Z(x) = \log \sum_{y} \exp\big(\log M_{k,r}(y \mid x) \big). \tag{30}$$

For $r = 0$, corresponding to the geometric mean, it can simply be obtained by applying a softmax to the arithmetic mean of the logits if we consider softmax probabilities. This is a key implementation advantage, which has made logit averaging a common and well-studied practice (Tassi et al., 2022; Buchanan et al., 2023; Wood et al., 2023).

Let $f^{(i)}(x) \in \mathbb{R}^L$ denote the vector of logits produced by the i-th model, so that

$$p^{(i)}(y \mid x) = \left[\mathrm{softmax}\left(f^{(i)}(x) \right) \right]_y. \tag{31}$$

[1] Implemented using the PyTorch library (Paszke et al., 2019).

Let us show that for the geometric mean, corresponding to the case $r = 0$, we have

$$\overset{\circ}{p}_{k,0}(y \mid x) = \left[\operatorname{softmax}\left(\frac{1}{k}\sum_{i=1}^{k} f^{(i)}(x)\right)\right]_y. \tag{32}$$

We express the operator in terms of the softmax as follows:

$$M_{k,0}(y \mid x) = \exp\left(\frac{1}{k}\sum_{i=1}^{k}\log p^{(i)}(y \mid x)\right) \tag{33}$$

$$= \exp\left(\frac{1}{k}\sum_{i=1}^{k}\log\left[\operatorname{softmax}\left(f^{(i)}(x)\right)\right]_y\right). \tag{34}$$

Now, for any vector $\mathbf{z} \in \mathbb{R}^L$, its log-softmax satisfies

$$\log\operatorname{softmax}(\mathbf{z}) = \mathbf{z} - \log\left(\sum_{\ell=1}^{L} e^{\mathbf{z}_\ell}\right)\mathbf{1}, \tag{35}$$

where $\mathbf{1} \in \mathbb{R}^L$ is the vector of ones. Therefore, taking the component associated with class y gives

$$\log\left[\operatorname{softmax}(\mathbf{z})\right]_y = \mathbf{z}_y - \log\left(\sum_{\ell=1}^{L} e^{\mathbf{z}_\ell}\right). \tag{36}$$

Applying this identity with $\mathbf{z} = f^{(i)}(x)$, we obtain

$$M_{k,0}(y \mid x) = C(x)\exp\left(\frac{1}{k}\sum_{i=1}^{k}[f^{(i)}(x)]_y\right), \tag{37}$$

where $C(x)$ does not depend on the class y and thus does not affect the final normalized distribution. Indeed, we finally obtain

$$\overset{\circ}{p}_{k,0}(y \mid x) = \frac{M_{k,0}(y \mid x)}{\sum_{\ell=1}^{L} M_{k,0}(\ell \mid x)} \tag{38}$$

$$= \frac{\cancel{C(x)}\exp\left(\frac{1}{k}\sum_{i=1}^{k}[f^{(i)}(x)]_y\right)}{\sum_{\ell=1}^{L}\cancel{C(x)}\exp\left(\frac{1}{k}\sum_{i=1}^{k}[f^{(i)}(x)]_\ell\right)} \tag{39}$$

$$= \left[\operatorname{softmax}\left(\frac{1}{k}\sum_{i=1}^{k} f^{(i)}(x)\right)\right]_y. \tag{40}$$

Predictive performance: Accuracy, Brier Score, and ECE. We complement the likelihood-based evaluation by assessing the predictive performance of the generalized mean aggregation. Specifically, we study how the order r affects classification accuracy and calibration. Given a test set $\{(x_i, y_i)\}_{i=1}^{N}$ with $y_i \in \{1, \ldots, L\}$, let

$$\hat{y}_{k,r}(x) = \arg\max_{\ell} \overset{\circ}{p}_{k,r}(\ell \mid x)$$

denote the predicted class. We report the classification accuracy,

$$\mathrm{Acc}(r) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\{\hat{y}_{k,r}(x_i)=y_i\}}, \tag{41}$$

the multiclass Brier score,

$$\mathrm{Brier}(r) = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\left(\overset{\circ}{p}_{k,r}(\ell \mid x_i) - \mathbf{1}_{\{y_i=\ell\}}\right)^2, \tag{42}$$

which measures the squared distance between the predicted probabilities and the one-hot target vector, and the Expected Calibration Error (ECE). Let

$$\mathrm{conf}_{k,r}(x) = \max_c \overset{\circ}{p}_{k,r}(\ell \mid x)$$

denote the prediction confidence. Partitioning the test samples into M equal-width confidence bins $B_1, \ldots, B_M$, the ECE is defined as

$$\mathrm{ECE}(r) = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{43}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denote the empirical accuracy and average confidence within bin B_m, respectively. In our implementation, we adopt the `MulticlassCalibrationError` metric from the `torchmetrics.classification` with 15 equally spaced confidence bins and the ℓ_1 norm.

B VISUAL EXPERIMENTS ON THE GAUSSIAN SETTING

In this section, we analyze the continuous setting (that would correspond to regression) to illustrate how both expert configuration and data dispersion influence the effect of the power parameter r of $\overset{\circ}{p}_{k,r}$ on the negative log-likelihood.

We consider symmetric Gaussian experts $p^{(1)}$ and $p^{(2)}$ of means μ_1 and $\mu_2 = -\mu_1$ respectively with fixed standard deviation 1.8, and evaluate aggregation performance on samples drawn from a centered Gaussian distribution with standard deviation σ (with $\sigma = 4$ corresponding to Figure 1). For each setting, the NLL is estimated over several tens of thousands of samples. By varying σ and the separation between experts, we clarify the mechanisms governing the different aggregation regimes.

B.1 HOW DATA CONCENTRATION SHAPES AGGREGATION PERFORMANCE

In this section, we show that concentration of data around agreement points can cause the maximum aggregation to underperform, whereas dispersion of data across the experts' supports can exacerbate the shortcomings of minimum-type aggregation.

In Figure 6, decreasing σ (from $\sigma = 4$ configuration in Figure 1) first accentuates the U-shaped behavior. For intermediate values of σ ($\{3, 2, 1\}$), all orders r outperform the individual baseline (dashed line). When σ becomes very small ($\sigma = 0.2$), the trend reverses: the curve retains a similar structure but with inverted ordering, and minimum-type aggregation underperforms the individual likelihoods. This effect arises because σ governs whether samples predominantly fall in disagreement regions (penalizing min-type aggregation) or agreement regions (penalizing max-type aggregation when $p^{(1)} \neq p^{(2)}$), with intermediate values providing a balanced trade-off.

B.2 HOW EXPERT SEPARATION SHAPES AGGREGATION PERFORMANCE

In this section, we show that getting the experts mean closer to each other in comparison to Figure 1 (where $\mu_1 = 3.5$) reduce the effect of power means since $p^{(1)} \approx p^{(2)}$, and putting the experts more separated correct the problem from Figure 1 encountered on negative values of r.

In Figure 7, when μ_1 and μ_2 are close, the effect of the generalized mean is limited and slightly negative for small r, as test samples predominantly lie in disagreement regions. Increasing the separation between μ_1 and μ_2 reduces these disagreement regions and concentrates samples around the mode of the normalized generalized mean.

C FAILURE OF THE "WISDOM OF CROWDS": COUNTER-EXAMPLES

We provide counter-example in both continuous settings with Gaussian densities, and discrete settings (see Section 3.3). While the former examines general regimes outside the reliability interval ($r < 0$ and $r > 1$), the latter focuses on the extreme aggregation cases corresponding to the minimum and maximum operators.

C.1 SYMMETRIC GAUSSIAN DENSITIES

In this section, we prove Theorem 3.2. We show that we can find two positive densities $p^{(1)}, p^{(2)} \in \mathcal{P}$ with $p^{(1)} \neq p^{(2)}$ and a point $\mathbf{x}$, such that

$$\log \mathring{p}_{2,r}(\mathbf{x}) < \frac{1}{2} \sum_{i=1}^{2} \log p^{(i)}(\mathbf{x}) \tag{44}$$

when $r \notin [0, 1]$.

Proof. We consider two symmetric Gaussian distributions. Let $\varphi(x) : x \mapsto (2\pi)^{-1/2} \exp(-x^2/2)$ be the standard normal density. We let $m > 0$ and we consider

$$p^{(1)}(x) = \varphi(x + m), \qquad p^{(2)}(x) = \varphi(x - m). \tag{45}$$

which correspond to $\mathcal{N}(-m, 1)$ and $\mathcal{N}(m, 1)$ respectively.

To construct counter-examples for $r > 1$ and $r < 0$, we consider two points of interest: respectively $x = 0$, the unique point where the two densities coincide, and $x = m$, which corresponds to the mode of $p^{(2)}$ and lies in the tail of $p^{(1)}$.

The aggregation breaks down at disagreement points when $r < 0$. We show that if $r < 0$ and $x = m$, Jensen's inequality in Equation (17) (and thus Equation (15)) can be reversed as soon as m is large enough. Figure 2b illustrates how negative r reverses Equation (15) for $|x|$ and m large enough. We have

$$p^{(2)}(m) = \varphi(0), \qquad p^{(1)}(m) = \varphi(2m) = \varphi(0)e^{-2m^2}. \tag{46}$$

For any $r < 0$, the power mean at $x = m$ satisfies

$$M_{2,r}(m) = \left(\tfrac{1}{2}\left(p^{(1)}(m)^r + p^{(2)}(m)^r\right)\right)^{1/r} \tag{47}$$

$$= \varphi(0) \left(\tfrac{1}{2}(1 + e^{-2rm^2})\right)^{1/r}. \tag{48}$$

Taking logarithms and comparing with the average log-density yields

$$\begin{aligned} &\log M_{2,r}(m) - \tfrac{1}{2}\Big(\log p^{(1)}(m) + \log p^{(2)}(m)\Big) \\ &\quad = \frac{1}{r} \log\Big(\tfrac{1}{2}(1 + e^{-2rm^2})\Big) + m^2. \end{aligned} \tag{49}$$

Since $r < 0$, the term e^{-2rm^2} grows exponentially in m, thus $\frac{1}{r} \log\Big(\frac{1}{2}(1 + e^{-2rm^2})\Big) + m^2 \sim \frac{1}{r} \log(e^{-2rm^2}) + m^2 = -2m^2 + m^2 = -m^2$ when $m \to +\infty$, so the right-hand side tends to $-\infty$ as $m \to +\infty$. Therefore, for m large enough,

$$\log M_{2,r}(m) < \tfrac{1}{2}\Big(\log p^{(1)}(m) + \log p^{(2)}(m)\Big), \tag{50}$$

which contradicts Equation (17) and Equation (15).

The generalized mean fails to improve agreement points when $r > 1$. This time we show that the failure of Equation (15) on $x = 0$ does not come from reversing Equation (17), which will in fact be an equality here, but from normalization redistributing probability mass away from this point. Figure 2c illustrates this phenomenon, with a concentrated degradation around this point. By using Lemma 2.1, we have for all $r > 1$

$$Z_{2,r} = \int_{\mathbb{R}} M_{2,r}(x)\,\mathrm{d}x > \int_{\mathbb{R}} M_{2,1}(x)\,\mathrm{d}x \tag{51}$$

$$= \int_{\mathbb{R}} \frac{p^{(1)}(x) + p^{(2)}(x)}{2}\,\mathrm{d}x = 1. \tag{52}$$

where Equation (51) holds since the set where $p^{(1)}$ and $p^{(2)}$ differ has strictly positive measure. Now we consider $x = 0$, the symmetric point with respect to the two means $\pm m$. We have

$$p^{(1)}(0) = p^{(2)}(0) = \varphi(m) > 0. \tag{53}$$

Hence, from Equations (51) and (53) the composed density satisfies

$$\overset{\circ}{p}_{2,r}(0) = \frac{M_{2,r}(0)}{Z_{2,r}} = \frac{\varphi(m)}{Z_{2,r}} \tag{54}$$

$$\implies \log \overset{\circ}{p}_{2,r}(0) = \log \varphi(m) - \log Z_{2,r} < \log \varphi(m). \tag{55}$$

Since

$$\frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(0) = \frac{1}{2}\sum_{i=1}^{2} \log \varphi(m) = \log \varphi(m), \tag{56}$$

it contradicts Equation (15) and verifies Equation (44). □

C.2 COUNTER-EXAMPLES IN EXTREME AGGREGATION REGIMES

We provide explicit counter-examples in discrete settings showing that the likelihood inequality established for $r \in [0, 1]$ does not extend to the extreme aggregation regimes $r = -\infty$ (minimum aggregation) and $r = +\infty$ (maximum aggregation). We consider classification, where each predictor outputs a probability vector over classes and aggregation is performed componentwise followed by normalization.

Failure for $r = -\infty$ at disagreement points. Consider binary classification with labels $y \in \{0, 1\}$ and true label $y = 0$. Let two classifiers produce

$$p^{(1)} = (0.9,\ 0.1), \qquad p^{(2)} = (0.01,\ 0.99). \tag{57}$$

Minimum aggregation selects the smallest probability assigned to each class:

$$M_{2,-\infty} = \big(\min(0.9, 0.01), \min(0.1, 0.99)\big) = (0.01,\ 0.1). \tag{58}$$

After normalization,

$$\overset{\circ}{p}_{2,-\infty} = \frac{(0.01,\ 0.1)}{0.01 + 0.1} = (0.0909,\ 0.9091). \tag{59}$$

The aggregated log-likelihood of the true label is therefore

$$\log \overset{\circ}{p}_{2,-\infty}(y = 0) = \log(0.0909) \approx -2.40. \tag{60}$$

The average individual log-likelihood equals

$$\frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(y = 0) = \frac{1}{2}\big(\log 0.9 + \log 0.01\big) \approx -2.36. \tag{61}$$

Hence,

$$\log \overset{\circ}{p}_{2,-\infty}(0) < \frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(0), \tag{62}$$

which contradicts the "wisdom of crowds" principle from Equation (15).

Failure for $r = +\infty$ at agreement points. Consider three-class classification with labels $y \in \{0, 1, 2\}$ and true label $y = 0$. Let two classifiers produce

$$p^{(1)} = (0.9,\ 0.03,\ 0.07), \qquad p^{(2)} = (0.9,\ 0.07,\ 0.03). \tag{63}$$

Maximum aggregation selects the largest probability assigned to each class:

$$M_{2,+\infty} = \big(\max(0.9, 0.9), \max(0.03, 0.07), \max(0.07, 0.03)\big) = (0.9,\ 0.07,\ 0.07). \tag{64}$$

After normalization,

$$\overset{\circ}{p}_{2,+\infty} = \frac{(0.9,\ 0.07,\ 0.07)}{0.9 + 0.07 + 0.07} = (0.865,\ 0.067,\ 0.067). \tag{65}$$

The aggregated log-likelihood of the true label is therefore

$$\log \overset{\circ}{p}_{2,+\infty}(y = 0) = \log(0.865) \approx -0.145. \tag{66}$$

The average individual log-likelihood equals

$$\frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(y = 0) = \frac{1}{2}\big(\log 0.9 + \log 0.9\big) = \log 0.9 \approx -0.105. \tag{67}$$

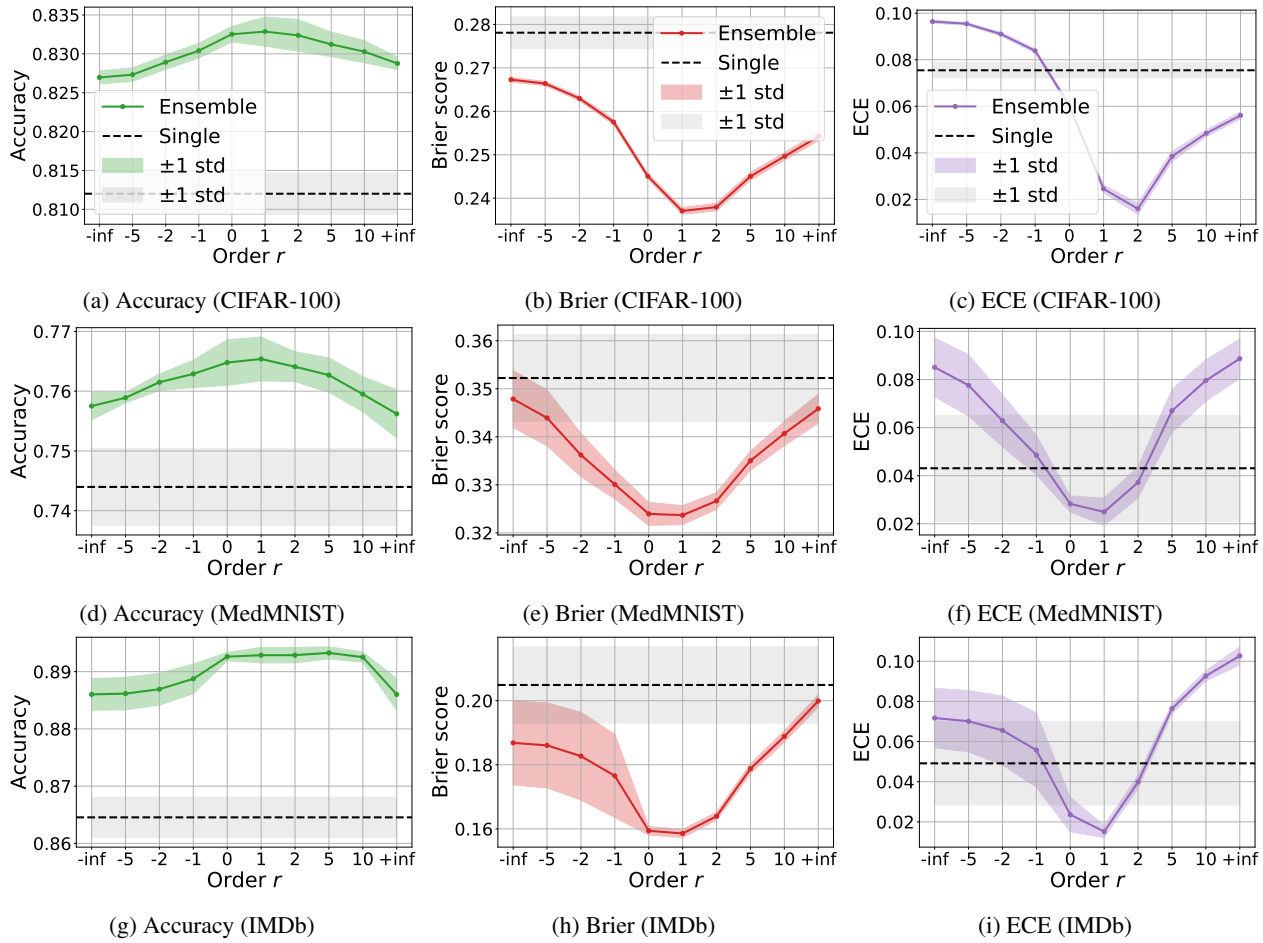
Hence,

$$\log \overset{\circ}{p}_{2,+\infty}(0) < \frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(0), \tag{68}$$

which again contradicts the log-likelihood inequality from Equation (15).

D ADDITIONAL PREDICTIVE METRICS

We report in Figure 8 additional evaluation metrics to complement Figure 3. Accuracy, Brier score, and ECE, which are classical metrics used in predictive settings, exhibit trends consistent with cross-entropy on the test set. This suggests that the wisdom of crowds effects observed along r are not limited to likelihood-based metrics, but also translate into improved predictive performance and calibration. See implementation details in Appendix A.3.



(a) Accuracy (CIFAR-100) (b) Brier (CIFAR-100) (c) ECE (CIFAR-100)

(d) Accuracy (MedMNIST) (e) Brier (MedMNIST) (f) ECE (MedMNIST)

(g) Accuracy (IMDb) (h) Brier (IMDb) (i) ECE (IMDb)

Figure 8: **Accuracy, Brier score, and Expected Calibration Error obtained from the normalized power mean likelihood $\mathring{p}_{k,r}$ across three classification datasets.** All metrics exhibit a U-shaped trend as a function of r, similar to the behavior observed for cross-entropy. Interestingly, most values of r benefit accuracy and Brier score, whereas ECE behaves differently.

E ANALYTICAL EXPRESSIONS OF INTEGRALS: GENERALIZED MEANS OF GAUSSIANS

In this section, we derive analytical expressions for the normalizing constant of the generalized mean for $r \in \mathbb{R}$. We show that closed-form formulas can be obtained within the reliability interval $[0, 1]$ using multinomial expansion (Section 3.2), namely for $r = 0$ and for values of r which are reciprocals of positive integers. Note that the case $r = 1$ is trivial, as the normalizing constant equals one; and for the case $r = 0$ the resulting density has been extensively studied (Cao and Fleet, 2014; Cohen et al., 2020), notably showing that the weighted geometric mean remains Gaussian. On the other hand, we show that for other values of r, including those outside $[0, 1]$, multinomial expansion does not yield to finite closed-form expressions although we know that it is finite (Proposition 3.1). This further emphasizes that the interval $[0, 1]$ constitutes not only a theoretical reliability regime (Theorem 3.1) but also the unique domain where aggregation retains analytical tractability. Thus, moving beyond it provides no analytical or practical advantage.

Motivation in regression. In probabilistic regression, Gaussian densities are typically employed as likelihood functions. Studying the generalized mean of Gaussian densities is therefore of practical interest: if the corresponding normalizing constant is analytically tractable, aggregation directly induces a valid Gaussian-based regression model and tractable model likelihood without requiring numerical integration or approximation schemes, except potentially at prediction time where computing the mean still involves an integral.

Notations and goal. Let $p^{(1)}, \ldots, p^{(k)}$ be Gaussian densities in $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$. Formally, we consider for all $\mathbf{x} \in \mathcal{X}$

$$p^{(i)}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma_i|^{1/2}} \exp\!\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right), \qquad i = 1, \ldots, k, \tag{69}$$

where $\mu_i \in \mathbb{R}^d$ denotes the mean vector and $\Sigma_i \in \mathbb{R}^{d\times d}$ is a symmetric positive definite covariance matrix (we write $\Sigma_i \succ 0$).

The goal is to determine

$$Z_{k,r} = \int \left(\frac{1}{k}\sum_{i=1}^{k} p^{(i)}(\mathbf{x})^r\right)^{\frac{1}{r}} \mathrm{d}\mathbf{x}. \tag{70}$$

We distinguish two cases. First, we consider tractable integrals, which correspond to $r = 0$ (geometric mean) and $r = 1/n$ with $n \in \mathbb{N} \setminus \{0\}$, covering a countable subset of the safe interval $[0, 1]$. Then, we look at the values $r \in \mathbb{R}$ which do not satisfy the former conditions and do not correspond to tractable integrals.

E.1 TRACTABLE INTEGRALS

Assume $n \in \mathbb{N} \setminus \{0\}$. The integral of interest is written as

$$Z_{k,r} = \int \left(\frac{1}{k}\sum_{i=1}^{k} p^{(i)}(\mathbf{x})^{\frac{1}{n}}\right)^{n} \mathrm{d}\mathbf{x}. \tag{71}$$

This formulation allows us to directly apply the following multinomial theorem, defined as follows.

Lemma E.1 (Multinomial theorem). *Let k be a positive integer and n be a non-negative one. Let $a_1, \ldots, a_k \in F$ where F is a field (e.g., $F = \mathbb{R}$). Then,*

$$(a_1 + \cdots + a_k)^n = \sum_{\substack{n_1+\cdots+n_k=n \\ n_i \geq 0}} \binom{n}{n_1, \ldots, n_k} \prod_{i=1}^{k} a_i^{n_i}, \tag{72}$$

where the multinomial coefficient is defined as

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!}. \tag{73}$$

Mainly from Equation (72), we can derive the proposition below.

Proposition E.1. *Let $p^{(1)}, \ldots, p^{(k)}$ be $k \geq 1$ Gaussian densities of parameters (μ_i, Σ_i). If $r = \frac{1}{n}$ with $n \in \mathbb{N} \setminus \{0\}$, we have*

$$\begin{aligned} Z_{k,r} = \frac{1}{k^n} \sum_{\substack{n_1+\cdots+n_k=n \\ n_i \geq 0}} \frac{n!}{n_1! \cdots n_k!} (2\pi)^{\frac{d}{2}} \Big|\sum_{i=1}^{k} \frac{n_i}{n}\Sigma_i^{-1}\Big|^{-\frac{1}{2}} \prod_{i=1}^{k} (2\pi)^{-\frac{dn_i}{2n}} |\Sigma_i|^{-\frac{n_i}{2n}} \\ \cdot \exp\left(\frac{1}{2}\Big(\sum_{i=1}^{k} \frac{n_i}{n}\mu_i^\top \Sigma_i^{-1}\Big)\Big(\sum_{i=1}^{k} \frac{n_i}{n}\Sigma_i^{-1}\Big)^{-1}\Big(\sum_{i=1}^{k} \frac{n_i}{n}\Sigma_i^{-1}\mu_i\Big) - \frac{1}{2}\sum_{i=1}^{k} \frac{n_i}{n}\mu_i^\top \Sigma_i^{-1}\mu_i\right). \end{aligned} \tag{74}$$

If $r = 0$,

$$\begin{aligned} Z_{k,r} = (2\pi)^{\frac{d}{2}} \Big|\sum_{i=1}^{k} \frac{1}{k}\Sigma_i^{-1}\Big|^{-\frac{1}{2}} \prod_{i=1}^{k} (2\pi)^{-\frac{d}{2k}} |\Sigma_i|^{-\frac{1}{2k}} \\ \cdot \exp\left(\frac{1}{2}\Big(\sum_{i=1}^{k} \frac{1}{k}\mu_i^\top \Sigma_i^{-1}\Big)\Big(\sum_{i=1}^{k} \frac{1}{k}\Sigma_i^{-1}\Big)^{-1}\Big(\sum_{i=1}^{k} \frac{1}{k}\Sigma_i^{-1}\mu_i\Big) - \frac{1}{2}\sum_{i=1}^{k} \frac{1}{k}\mu_i^\top \Sigma_i^{-1}\mu_i\right). \end{aligned} \tag{75}$$

Proof. From Lemma E.1, we have

$$Z_{k,r} = \frac{1}{k^n} \sum_{\substack{n_1+\cdots+n_k=n \\ n_i \geq 0}} \binom{n}{n_1, \ldots, n_k} \underbrace{\int \prod_{i=1}^{k} p_i(\mathbf{x})^{\frac{n_i}{n}}}_{I_k}. \tag{76}$$

We remark that I_k is the integral of a weighted geometric mean since $\sum_{i=1}^{k} \frac{n_i}{n} = 1$. Let us denote the weights as $w_i = \frac{n_i}{n}$ for all i. For all $\mathbf{x} \in \mathcal{X}$, we have

$$\prod_{i=1}^{k} p_i(\mathbf{x})^{w_i} = \prod_{i=1}^{k} (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{k} w_i (\mathbf{x}-\mu_i)^\top \Sigma_i^{-1} (\mathbf{x}-\mu_i)\right) \tag{77}$$

$$= C_k \exp\left(-\frac{1}{2}\left(\mathbf{x}^\top Q_k \mathbf{x} - 2 b_k^\top \mathbf{x} + c_k\right)\right), \tag{78}$$

where

$$C_k = \prod_{i=1}^{k} (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}}, \quad Q_k = \sum_{i=1}^{k} w_i \Sigma_i^{-1} \succ 0, \quad b_k = \sum_{i=1}^{k} w_i \Sigma_i^{-1} \mu_i, \quad c_k = \sum_{i=1}^{k} w_i \mu_i^\top \Sigma_i^{-1} \mu_i. \tag{79}$$

Completing the square allows us to factor out a term with an explicit integral. Indeed,

$$\prod_{i=1}^{k} p_i(\mathbf{x})^{w_i} = C_k \exp\left(-\frac{1}{2}\left(\mathbf{x}^\top Q_k \mathbf{x} - 2 b_k^\top \mathbf{x} + c_k\right)\right) \tag{80}$$

$$= C_k \exp\left(-\frac{1}{2}\left(\mathbf{x}^\top Q_k \mathbf{x} - 2 (Q_k^{-1} b_k)^\top Q_k \mathbf{x} + c_k\right)\right) \tag{81}$$

$$= C_k \exp\left(-\frac{1}{2}\left(\mathbf{x}^\top Q_k \mathbf{x} - 2 (Q_k^{-1} b_k)^\top Q_k \mathbf{x} + (Q_k^{-1} b_k)^\top Q_k (Q_k^{-1} b_k)\right)\right) \exp\left(\frac{1}{2} b_k^\top Q_k^{-1} b_k - \frac{1}{2} c_k\right) \tag{82}$$

$$= C_k \exp\left(-\frac{1}{2} (\mathbf{x} - Q_k^{-1} b_k)^\top Q_k (\mathbf{x} - Q_k^{-1} b_k)\right) \exp\left(\frac{1}{2} b_k^\top Q_k^{-1} b_k - \frac{1}{2} c_k\right) \tag{83}$$

$$= C_k (2\pi)^{\frac{d}{2}} |Q_k|^{-\frac{1}{2}} \mathcal{N}(\mathbf{x};\, Q_k^{-1} b_k, Q_k^{-1}) \exp\left(\frac{1}{2} b_k^\top Q_k^{-1} b_k - \frac{1}{2} c_k\right). \tag{84}$$

where Equation (81) follows from the fact that Q_k is invertible and symmetric. Using $\int \mathcal{N}(\mathbf{x};\, Q_k^{-1} b_k, Q_k^{-1}) \mathrm{d}\mathbf{x} = 1$, we have

$$\int \prod_{i=1}^{k} p_i(\mathbf{x})^{w_i} \mathrm{d}\mathbf{x} = (2\pi)^{\frac{d}{2}} |Q_k|^{-\frac{1}{2}} \prod_{i=1}^{k} (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}} \exp\left(\frac{1}{2} b_k^\top Q_k^{-1} b_k - \frac{1}{2} c_k\right). \tag{85}$$

where Q_k, b_k, c_k are defined in Equation (79). We finally use Equation (76) to prove the first result of Proposition E.1.

For the case $r = 0$, we just apply the same reasoning from Equation (77) leading to Equation (85) with $w_i = \frac{1}{k}$. □

E.2 INTEGRALS FOR WHICH WE HAVE NO CLOSED-FORM

As soon as r does not belong to the cases studied in Appendix E.1, and in particular when $r \notin [0, 1]$, we cannot derive explicit integrals $Z_{k,r}$ with a finite sum like before.

To see this, we consider the easy case $k = 2$, and assume α is any real (or complex) number. It can be negative. Without restriction, Lemma E.1 is generalized through Binomial series, as stated in the following lemma

Lemma E.2 (Binomial Series). *Let $a, b \in \mathbb{R}$. Then, if $|a| < |b|$,*

$$(a+b)^\alpha = \sum_{i=0}^{\infty} \binom{\alpha}{i} a^i b^{\alpha-i}. \tag{86}$$

where the generalized binomial coefficient is defined by

$$\binom{\alpha}{i} = \frac{\alpha(\alpha-1)(\alpha-2)\ldots(\alpha-i-1)}{i!}. \tag{87}$$

From this lemma, $Z_{k,r}$ can be expressed for general $r \in \mathbb{R} \setminus \{0\}$ as an integral involving a binomial series, following the same kind of approach as in Appendix E.1. To ensure convergence of the series expansion, the integral must be decomposed into two regions corresponding to the cases $p^{(1)} < p^{(2)}$ and $p^{(2)} < p^{(1)}$. However, unless r^{-1} is a positive integer (as in Section E.1), the resulting binomial expansions are no longer finite sums. Consequently, evaluating $Z_{k,r}$ generally requires approximation or numerical procedures, either for computing the integral itself or for truncating the associated series representation.

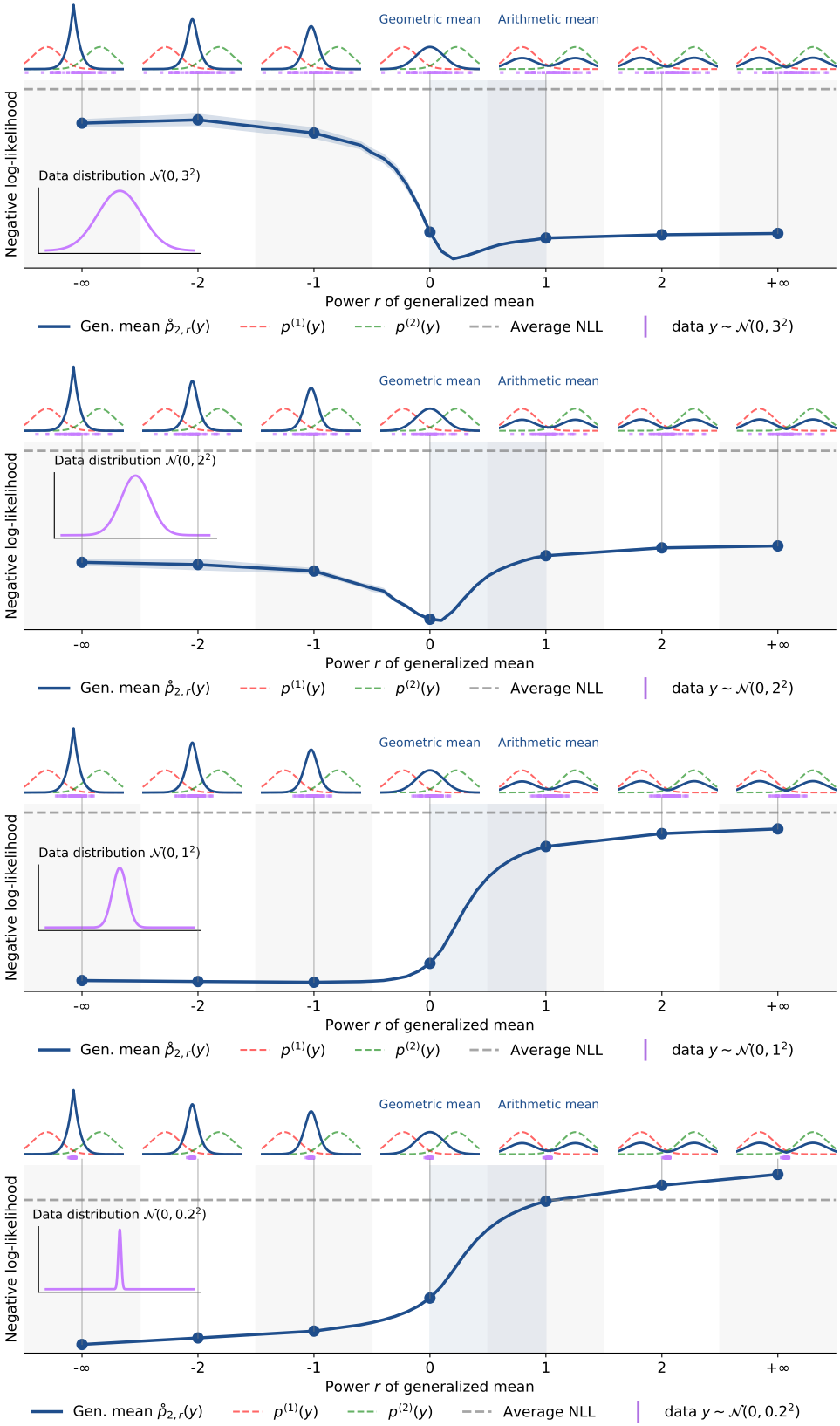


Figure 6: **Effect of decreasing σ on aggregation behavior.** From top to bottom: large ($\sigma = 3$), moderate ($\sigma = 2$), small ($\sigma = 1$), and very small σ ($\sigma = 0.2$). The U-shaped behavior first sharpens, then all r outperform the individual baseline for intermediate σ, before the trend reverses for very small σ.

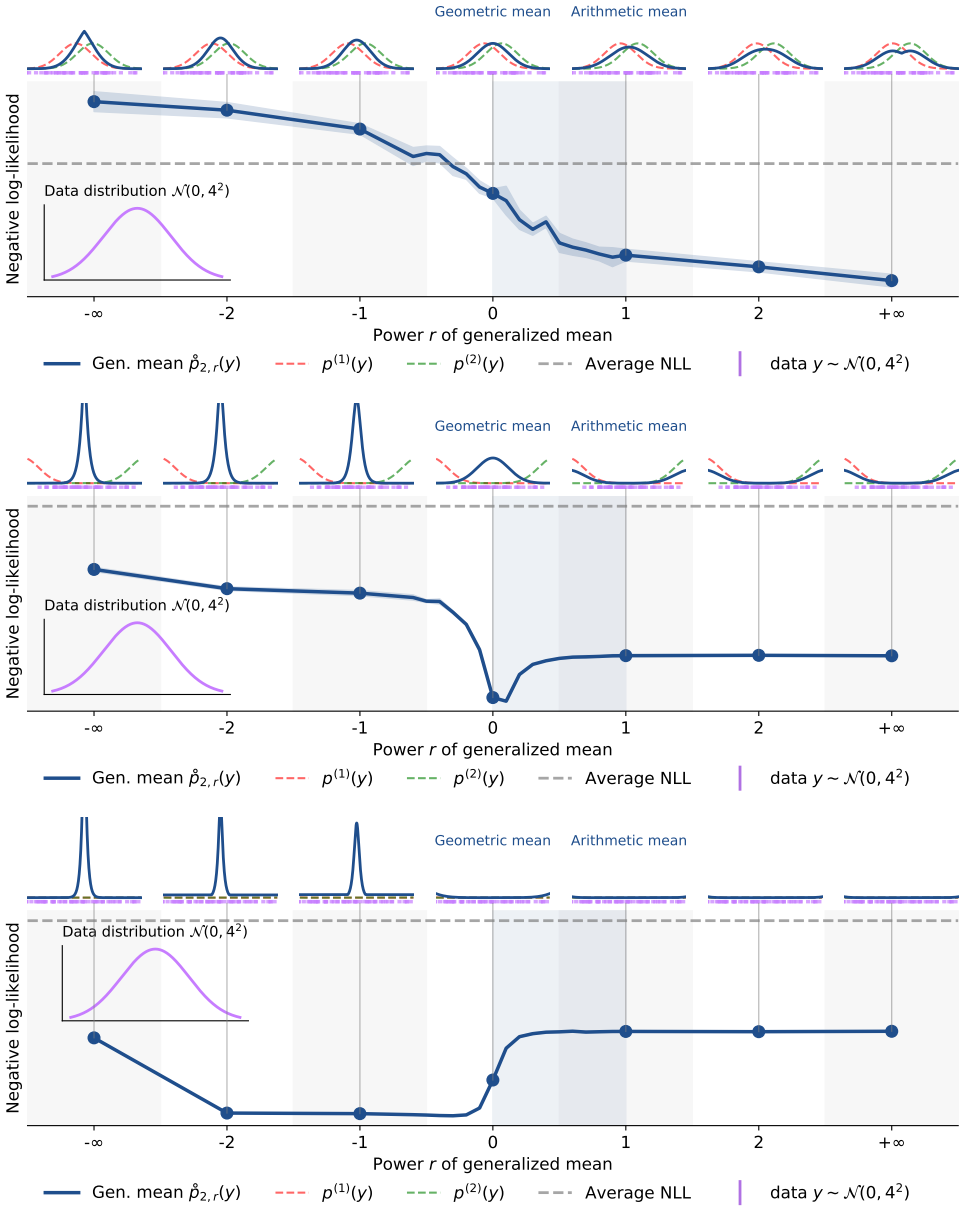


Figure 7: **Effect of μ_1 and μ_2 on aggregation performance.** From top to bottom: merged experts, moderately separated experts, and highly separated experts. When μ_1 and μ_2 are close, disagreement regions are frequent, which explains the poor performance of min-like aggregations. As the means separate, this effect diminishes as test samples concentrate on the modes of the generalized mean.