Variance-Gated Ensembles: An Epistemic-Aware Framework for Uncertainty Estimation

Anonymous authors
Paper under double-blind review

Abstract

Machine learning applications require fast and reliable per-sample uncertainty estimation. A common approach is to use predictive distributions from Bayesian or approximation methods and additively decompose uncertainty into aleatoric (data-related) and epistemic (model-related) components. However, additive decomposition has recently been questioned, with evidence that it breaks down when using finite-ensemble sampling and/or mismatched predictive distributions. This paper introduces Variance-Gated Ensembles (VGE), an intuitive, differentiable framework that injects epistemic sensitivity *via* a signal-to-noise gate computed from ensemble statistics. VGE provides: (i) a Variance-Gated Margin Uncertainty (VGMU) score that couples decision margins with ensemble predictive variance; and (ii) a Variance-Gated Normalization (VGN) layer that generalizes the variance-gated uncertainty mechanism to training *via* per-class, learnable normalization of ensemble member probabilities. We derive closed-form vector-Jacobian products enabling end-to-end training through ensemble sample mean and variance. VGE matches or exceeds state-of-the-art information-theoretic baselines while remaining computationally efficient. As a result, VGE provides a practical and scalable approach to epistemic-aware uncertainty estimation in ensemble models.

1 Introduction

The average accuracy of predictions does not imply reliable decisions. Per-sample uncertainty estimation is therefore essential for identifying when model predictions should not be trusted. A common approach to this problem estimates predictive uncertainty, typically *via* Bayesian model averaging (BMA) in Bayesian neural networks (BNNs) or through approximations such as Monte Carlo Dropout (MCD) (Gal & Ghahramani, 2016), Deep Ensembles (DE) (Lakshminarayanan et al., 2017), and more recently, Last-Layer Ensembles (LLE) (Harrison et al., 2024; Sensoy et al., 2018).

These methods approximate the posterior predictive distribution

$$p(y \mid \mathbf{x}, \mathcal{D}) = \int_{\mathbf{w}} p(y \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{dw} = \mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}[p(y \mid \mathbf{x}, \mathbf{w})] \approx \frac{1}{M}\sum_{m=1}^{M} p(y \mid \mathbf{x}, \mathbf{w}_m), \tag{1}$$

where the intractable integral is approximated by averaging a finite sample of predictions from an ensemble of models with parameters $\{\mathbf{w}_m\}_{m=1}^{M}$ of the posterior or its approximation. In practical terms, the final prediction for y does not come from a single set of model parameters, but rather from many plausible sets of model parameters, where each $\mathbf{w}_m$ represents a version. The term $p(y \mid \mathbf{x}, \mathbf{w}_m)$ is the prediction made by one particular version, and $p(\mathbf{w} \mid \mathcal{D})$ tells us how believable each version is based on the evidence. This way, by averaging predictions, the overall ensemble model (*i.e.*, all sets of model parameters) incorporates both what the data suggests and the remaining uncertainty.

The additive decomposition of the total (TU, predictive), as measured by entropy into Aleatoric (AU, data-related) and Epistemic (EU, model-related) Uncertainty components is considered the standard uncertainty

analysis (Houlsby et al., 2011). The additive decomposition is defined as,

$$\underbrace{H[p(y \mid \mathbf{x}, \mathcal{D})]}_{\text{TU}} = \underbrace{\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}\Big[H[p(y \mid \mathbf{x}, \mathbf{w}])\Big]}_{\text{AU}} + \underbrace{I[p(y \mid \mathbf{x}, \mathcal{D});\ p(\mathbf{w} \mid \mathbf{x}, \mathcal{D})]}_{\text{EU}}, \tag{2}$$

and assumes that predictive entropy (*i.e.*, uncertainty) can be additively decomposed into the expected conditional entropy and conditional mutual information I. TU measures the overall uncertainty of the prediction, AU quantifies the uncertainty that persists even with a perfect model (arising from intrinsic variability or data noise), and EU captures the uncertainty arising from limited knowledge of the model parameters (reflecting what could be reduced if more data or better modeling were available).

The identification of these terms follows from their information-theoretic interpretation. The expected conditional entropy $\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}[H[p(y|\mathbf{x}, \mathbf{w})]]$ averages the predictive entropy of individual models. Even with perfect knowledge of the parameters, each model retains uncertainty due to intrinsic data variability, making this term irreducible and therefore aleatoric. Conversely, the mutual information $I[p(y \mid \mathbf{x}, \mathcal{D});\ p(\mathbf{w} \mid \mathbf{x}, \mathcal{D})]$ measures the expected reduction in predictive uncertainty that would result from observing the model parameters. It is zero when the posterior $p(\mathbf{w} \,|\, \mathcal{D})$ collapses on a single model and increases with posterior spread, reflecting uncertainty that can in principle be reduced with additional data or knowledge (*i.e.*, epistemic).

The additive decomposition in Equation 2 arises from fundamental identities: (i) Shannon entropy H of a random variable Y with probability $p(y)$; (ii) conditional entropy of Y that measures uncertainty after observing another random variable Z; and (iii) mutual information I that quantifies the expected reduction in uncertainty about Y after observing Z:

$$\text{(i) } H(Y) = -\sum_y p(y)\log p(y) \quad \text{(ii) } H(Y \mid Z) = \mathbb{E}_z\left[H(Y \mid Z = z)\right] \quad \text{(iii) } I(Y; Z) = H(Y) - H(Y \mid Z). \tag{3}$$

For the posterior predictive distribution $p(y \mid \mathbf{x}, \mathcal{D})$, obtained by marginalizing over model parameters $\mathbf{w}$ (Equation 1), conditional mutual information between label y and parameters $\mathbf{w}$, given input $\mathbf{x}$ and data $\mathcal{D}$, satisfies

$$I[p(y \mid \mathbf{x}, \mathcal{D});\ p(\mathbf{w} \mid \mathbf{x}, \mathcal{D})] = H[p(y \mid \mathbf{x}, \mathcal{D})] - H[p(y \mid \mathbf{x}, \mathbf{w}, \mathcal{D})]. \tag{4}$$

Under the standard assumption that y is conditionally independent of $\mathcal{D}$ given $(\mathbf{x},\ \mathbf{w})$ (*i.e.*, $p(y \mid \mathbf{x}, \mathbf{w}, \mathcal{D}) = p(y \mid \mathbf{x}, \mathbf{w})$

$$H[p(y \mid \mathbf{w}, \mathbf{x}, \mathcal{D})] = \mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}\Big[H[p(y \mid \mathbf{x}, \mathbf{w})]\Big] \tag{5}$$

which, after substituting into Equation 4 and rearrangement, we arrive at the well-known additive decomposition Equation 2.

While supported by information-theory, additive decomposition has recently come under criticism and can fail under finite-ensemble sampling or when the model posteriors mismatch the true predictive distribution (Wimmer et al., 2023). Recent work by Schweighofer et al. (2023) investigated the standard additive decomposition approach and showed that it mistakenly assumed the BMA distribution is equivalent to the true posterior predictive distribution. The authors subsequently introduced Expected Pairwise Cross-Entropy (EPCE) and Kullback–Leibler (EPKL) divergences of ensemble model predictions as measures for predictive and epistemic uncertainty, which is arguably the current state-of-the-art method for uncertainty estimation. However, these pairwise measures incur $O(M^2C)$ computational complexity for M models and C classes, which becomes prohibitive for large ensembles and many-class problems.

In this work, we introduce VGE, a framework that addresses these limitations through an exponential gating mechanism that modulates probabilities based on variance-scaled (*i.e.*, epistemic) signal-to-noise derived from ensemble moments. We propose a VGMU score using the top-2 ranked classes modulated by epistemic uncertainty. This approach is extended to an end-to-end VGN layer, where per-class parameters $\mathbf{k}$ are adaptively learned during training. The VGN layer suppresses high-variance, low-consensus predictions and re-shapes member distributions, resulting in epistemic-aware mixtures. Importantly, while pairwise divergence measures incur $O(M^2C)$ complexity, VGE achieves $O(MC)$ for uncertainty decomposition and $O(C)$

for VGMU evaluation, enabling real-time deployment without sacrificing epistemic sensitivity. Collectively, VGE unifies an epistemic-aware VGMU estimation with VGN in a single, differentiable training and inference pipeline.

In summary, this research provides the following contributions:

1. **Variance-Gated Ensembles.** A framework that introduces epistemic sensitivity *via* a signal-to-noise gate computed from ensemble statistics.
2. **Variance-Gated Margin Uncertainty.** A margin-based score, using uncertainty that quantifies class separability under predictive spread.
3. **Variance-Gated Normalization.** A differentiable, epistemic-aware calibration layer that modulates member probabilities through a per-class learnable parameter $\mathbf{k}$, enabling efficient end-to-end optimization with $O(MC)$ complexity.

Apart from this section, the remainder of the paper is organized as follows: Section 2 positions this work within the existing literature and contrasts it with related approaches. Section 3 introduces the variance-gated ensemble framework, including the VGMU scoring metric and the formulation of VGN. This section also presents a geometric-based interpretation of ensemble predictions. Section 4 introduces the analytical gradients required for end-to-end training using VGN. Section 5 presents experimental set-up and results, comparing the proposed approach with common uncertainty estimation methods in the machine learning literature. Section 6 provides an axiomatic comparison to prior uncertainty decomposition frameworks, and clarifies the intended scope and limitations of VGN and VGMU. Finally, Section 7 concludes the paper with a summary of findings and recommendations. We now position variance-gated ensembles within prior work on ensemble disagreement, uncertainty decomposition, calibration, and margin-based decision requirements before introducing the proposed framework.

2 Related Work

Uncertainty estimation (Gawlikowski et al., 2023; Hüllermeier & Waegeman, 2021) in classification tasks is often framed through Bayesian predictive distributions; however, an exact Bayesian solution is intractable for modern neural networks. Practical approximations include variational BNN (Radford, 1995), MCD (Gal & Ghahramani, 2016; Gal et al., 2017), DE (Lakshminarayanan et al., 2017), and LLE approaches such as evidential deep learning (Sensoy et al., 2018), variational inference (Harrison et al., 2024; Steger et al., 2024), multihead (Lee et al., 2015), and other ensemble strategies (Huang et al., 2017; Kushibar et al., 2022; Wen et al., 2020). Among these approaches, DE have demonstrated strong performance in calibration and out-of-distribution (OOD) detection, often outperforming approximate Bayesian methods under dataset shift (Ovadia et al., 2019). As a result, many recent uncertainty estimation methods operate on ensemble predictive distributions. Decomposition of predictive entropy into an expected conditional entropy term and a mutual-information term (Houlsby et al., 2011; Depeweg et al., 2018) is a well-established framework. While well-supported by information-theory, recent work has shown that entropy-based epistemic measures can behave unreliably for finite ensembles and posterior mismatch (Wimmer et al., 2023; Schweighofer et al., 2023). In such settings, mutual information and conditional entropy may fail to represent predictive disagreement, motivating alternative formulations that do not rely on entropy identities.

To address limitations of entropy-based epistemic measures, a recent approach to quantify uncertainty *via* disagreement between ensemble members is EPKL divergence (Schweighofer et al., 2023). These measures reflect finite-sample variability across ensemble predictions and have been shown to improve epistemic uncertainty estimation and out-of-domain (OOD) detection. While effective, the disagreement measure incur quadratic costs in ensemble size, which has motivated the development of more computationally efficient alternatives. An alternative research area focuses on uncertainty signals derived from ensemble statistics, such as the mean and variance of predicted class probabilities (Malinin & Gales, 2018). Moment-based approaches are attractive in large-scale or real-time settings. VGE follows this approach by relating epistemic-sensitive uncertainty measures from ensemble variance, enabling linear-time computation while retaining sensitivity to predictive disagreement.

Calibration methods, such as temperature scaling, operate post hoc and adjust predictive confidence without modifying uncertainty structure during training (Guo et al., 2017; Kumar et al., 2022). In contrast, recent work explores integrating uncertainty-aware mechanisms into the training process itself, including diversity-promoting losses and uncertainty-regularized objectives (Fort et al., 2020; Ashukha et al., 2021). VGN aligns within this area by introducing a differentiable normalization layer that modulates ensemble predictions based on epistemic variance, allowing uncertainty re-shaping to be learned end-to-end.

In many applications, risk-based decisions depend on ambiguity between top-ranked classes rather than uncertainty of the full-simplex distribution. Margin-based criteria, such as Best-versus-Second Best (BvSB) scores, are used in active learning and selective prediction (Joshi et al., 2009; Geifman & El-Yaniv, 2017). VGMU extends this method by incorporating ensemble variance into the decision margin, offering a decision-focused epistemic uncertainty measure that emphasizes disagreement among top-ranked classes while remaining computationally efficient.

3 Framework Definition and Setup

This section describes the variance-gated ensemble framework for uncertainty-aware ensemble modeling, defining normalization and analytical properties. We first formalize variance-gated normalization, where member probabilities are modulated by a signal-to-noise gate, and analyze its sensitivity to sample mean confidence, predictive spread, and per-class learnable parameter $\mathbf{k}$. We then provide a geometric-based interpretations, followed by a variance-gated uncertainty decomposition for total uncertainty into aleatoric and epistemic components, and introduce the variance-gated margin uncertainty score to capture class separability. See Table S1 in the supporting information for a listing of symbols and abbreviations used throughout the variance-gated ensemble framework.

3.1 Ensemble Statistics and Variance Gate Definition

We first define the variance gate and analyze its local sensitivities. Let each ensemble member produce a predictive categorical distribution

$$\mathbf{p}_m = p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}_m) \in \Delta^{C-1}, \qquad m \in \{1, \ldots, M\}, \tag{6}$$

where $\mathbf{p}_m = [p_m(1), \ldots, p_m(C)]^\top$ lies in a compact convex region known as the $(C{-}1)$-simplex. The per-class ensemble sample mean and standard deviation are defined as

$$\bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}_m, \qquad \mathbf{s} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{p}_m - \bar{\mathbf{p}})^2 + \varepsilon}, \quad M > 1. \tag{7}$$

Let $\bar{\mathbf{p}} \geq 0$, $\mathbf{s} \geq 0$, $\mathbf{k} > 0$, and $\varepsilon > 0$ (*e.g.*, 1.0×10^{-8}). The variance gate is defined as

$$\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}, \qquad 0 \leq \mathbf{\Gamma} < 1 \in \mathbb{R}^C \tag{8}$$

and the normalized variance-gated member distribution is defined as

$$\mathbf{q}_m = \frac{\mathbf{p}_m \odot \mathbf{\Gamma}}{Z_m}, \quad Z_m = \mathbf{p}_m^\top \mathbf{\Gamma} \qquad 0 \leq \mathbf{q}_m \leq 1 \in \mathbb{R}^C, \tag{9}$$

where the scalar normalization Z_m ensures $\mathbf{1}^\top \mathbf{q}_m = 1$ and $\mathbf{1} \in \mathbb{R}^C$ denotes the all-ones vector. The per-class $\mathbf{k} > 0$ controls the sensitivity of the gate.

The variance gate modulates ensemble predictions based on the scaled signal-to-noise ratio $\mathrm{SNR} = \bar{\mathbf{p}} \,/\, \mathbf{ks}$ and acts as a smooth reliability correction. This means that classes with high mean confidence and low predictive spread receive larger gate values, while classes that are uncertain or highly variable are suppressed before normalization. We now examine the sensitivity of the variance gate and gated member distributions to sample mean confidence $\bar{\mathbf{p}}$, predictive spread $\boldsymbol{s}$, scaling factor $\mathbf{k}$, and its effects on gated distributions $\mathbf{q}_m$.

Proposition 3.1 (Sensitivity to sample mean confidence $\bar{\mathbf{p}}$)**.** *For the exponential gate* $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$*, the per-class derivative with respect to mean confidence* $\partial\mathbf{\Gamma}/\partial\bar{\mathbf{p}} > 0$ *and scales inversely with predictive spread* $\mathbf{s}$*.*

Proof. Consider a single class $c \in \{1, \ldots, C\}$, for which the gate is

$$\Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c}. \tag{10}$$

Differentiating with respect to the mean confidence $\bar{p}_c$ yields

$$\frac{\partial \Gamma_c}{\partial \bar{p}_c} = \frac{1}{k_c s_c}\, e^{-\bar{p}_c/k_c s_c} = \frac{1 - \Gamma_c}{k_c s_c} > 0. \tag{11}$$

The derivative is strictly positive, showing that increasing the mean confidence for class c always increases its gate value. Importantly, the factor $1/k_c s_c$ shows how this sensitivity depends on predictive spread. As the ensemble disagreement s_c increases, the magnitude of the derivative decreases, meaning that the gate becomes less responsive to changes in mean confidence. Thus, for classes with high predictive variance, increases in $\bar{p}_c$ are suppressed. This establishes that the sensitivity of the gate to confidence scales inversely with the predictive spread. □

Remark 3.1.1. *The factor* $(1 - \Gamma_c)$ *acts as a saturation term. As* $\bar{p}_c/(k_c s_c) \to \infty$*, we have* $\Gamma_c \to 1$ *and therefore* $\partial\Gamma_c/\partial\bar{p}_c \to 0$*. Consequently, once a class is deemed sufficiently reliable, further increases in mean confidence produce diminishing effects. This ensures that the variance-gate is most sensitive in intermediate regions and becomes increasingly insensitive as confidence saturates.*

Proposition 3.2 (Sensitivity to predictive spread $\mathbf{s}$)**.** *For the exponential gate* $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$*, the per-class derivative with respect to predictive spread is* $\partial\mathbf{\Gamma}/\partial\mathbf{s} < 0$ *and scales inversely to the squared predictive spread* $\mathbf{s}$ *and* $\mathbf{k}$*.*

Proof. Consider a single class $c \in \{1, \ldots, C\}$, for which

$$\Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c}. \tag{12}$$

Differentiating with respect to the predictive spread s_c yields

$$\frac{\partial \Gamma_c}{\partial s_c} = -\frac{\bar{p}_c}{k_c s_c^2}\, e^{-\bar{p}_c/k_c s_c} = -\frac{(1 - \Gamma_c)\, \bar{p}_c}{k_c s_c^2} < 0. \tag{13}$$

The derivative is strictly negative, indicating that increasing predictive spread decreases the gate value. The factor $1/k_c s_c^2$ shows that the magnitude of this effect decays rapidly as s_c grows. Thus, classes with high ensemble disagreement experience strong suppression, while further increases in large spreads have diminishing influence. □

Remark 3.2.1. *Increasing the predictive spread* s_c *while holding* $\bar{p}_c$ *and* k_c *fixed decreases the gate through two mechanisms: i) A linear dependence on* $\bar{p}_c$ *and a quadratic decay in* s_c*, and ii) the saturation factor* $(1 - \Gamma_c)$ *that ensures once the gate is already small, additional increases in spread have limited effect. This behavior enforces strong suppression for uncertain classes, while the quadratic decay in* s_c *ensures diminishing sensitivity as predictive spread grows.*

Proposition 3.3 (Sensitivity to scalar $\mathbf{k}$)**.** *For the exponential gate* $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$*, where* k_c *may be user-defined or learned, the derivative with respect to* $\mathbf{k}$ *is* $\partial\mathbf{\Gamma}/\partial\mathbf{k} < 0$ *and scales inversely to squared* $\mathbf{k}$ *and* $\mathbf{s}$*.*

Proof. Consider a single class $c \in \{1, \ldots, C\}$, for which

$$\Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c}. \tag{14}$$

Differentiating with respect to the scaling parameter k_c yields

$$\frac{\partial \Gamma_c}{\partial k_c} = -\frac{\bar{p}_c}{k_c^2 s_c}\, e^{-\bar{p}_c/k_c s_c} = -\frac{(1 - \Gamma_c)\, \bar{p}_c}{k_c^2 s_c} < 0. \tag{15}$$

Thus, increasing k_c decreases the gate value. The inverse quadratic dependence on k_c shows that the influence of the scaling parameter rapidly diminishes as k_c grows, resulting in controlled and saturating sensitivity. □

Remark 3.3.1. *Increasing k_c while holding $\bar{p}_c$ and s_c fixed decreases the gate by increasing the effective predictive spread. The quadratic decay in k_c ensures that sensitivity to further increases rapidly diminishes, while the saturation factor $(1 - \Gamma_c)$ prevents excessive attenuation once the gate is already small. Under mild distributional assumptions on ensemble dispersion, the product $k_c s_c$ may be interpreted as a classwise risk-tolerance threshold reflecting typical deviations from the ensemble mean.*

3.1.1 Geometric Interpretation

We now reinterpret variance-gating geometrically in the probability simplex. Each pair $(\bar{\mathbf{p}}, \{\mathbf{p}_m\})$ defines a configuration in the simplex Δ^{C-1}. The geometry of this configuration is defined by: (i) Confidence (or ambiguity): How close (or far) the ensemble sample mean $\bar{\mathbf{p}}$ is positioned near a simplex vertex; and (ii) Certainty (or uncertainty): How close (or far) the set of members cluster around the ensemble sample mean $\bar{\mathbf{p}}$ (*i.e.*, ensemble member agreement/disagreement). These geometric effects are modulated by the local sensitivities of the variance gate, which increases with mean confidence and is progressively attenuated by predictive spread and the classwise risk-tolerance scale $k_c s_c$. The qualitative combinations of these two dimensions correspond to four simplex spaces, which define regions in the probability simplex for all ensemble behaviors:

- **Confident–Certain:** Characterized by a near-deterministic ensemble mean and low variance. The ensemble members form a cluster around a single vertex of the simplex, indicating high confidence and high agreement.

$$\bar{p}_c \uparrow,\ s_c \downarrow \Longrightarrow\ \Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c} \approx 1. \tag{16}$$

- **Ambiguous–Certain:** Occurs when the ensemble mean is near-uniform but the variance is low. Members concentrate near the simplex barycenter, showing ambiguity but mutual certainty.

$$\bar{p}_c \downarrow,\ s_c \downarrow \Longrightarrow\ \Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c}, \quad \text{depends on } \bar{p}_c/k_c s_c. \tag{17}$$

- **Confident–Uncertain:** Defined by a near-deterministic mean but high variance. Members radiate outward from a vertex, reflecting confidence but disagreement among ensemble members.

$$\bar{p}_c \uparrow,\ s_c \uparrow \Longrightarrow\ \Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c} \ll 1, \quad \text{with sensitivity suppressed by large } s_c. \tag{18}$$

- **Ambiguous–Uncertain:** Defined by both a near-uniform mean and high variance. Members are diffused around the barycenter, indicating ambiguity and high disagreement.

$$\bar{p}_c \downarrow,\ s_c \uparrow \Longrightarrow\ \Gamma_c = 1 - e^{-\bar{p}_c/k_c s_c} \approx 0. \tag{19}$$

The variance gate functions as a continuous geometric adjustment within the probability simplex. It maintains the geometry in confident–certain regions, selectively attenuates confident–uncertain areas, and contracts diffuse ensembles' predictions in ambiguous–uncertain cases. In contrast, ambiguous–certain regions remain unaffected, as their geometry already reflects agreement in uncertainty. Collectively, these four simplex spaces define the set of ensemble behaviors through which variance-gating is applied.

Collectively, the sensitivity properties of the variance gate define a controlled signal-to-noise mechanism. The gate increases with mean confidence $\bar{p}_c$, but this effect is progressively attenuated by predictive spread s_c and the class-wise parameter k_c, with sensitivities that decay quadratically in both quantities. As a result, confident and consistent ensemble predictions are preserved, while high-confidence but high-variance predictions are suppressed in a stable and saturating manner (for a risk-based interpretation, see SI S5).

3.2 Variance-Gated Margin Uncertainty

Entropy is known to overestimate uncertainty when probability values are spread across many classes and can underestimate it when a model is highly confident about a few options. Therefore, entropy alone is not sufficient to provide adequate information for decision-making. We posit the following question: *Can we identify a measure that is sensitive to class separation, while incorporating uncertainty awareness?* We want to identify a scoring metric that (i) maintains epistemic awareness and (ii) ideally, provides a user-defined threshold that reflects the degree of acceptable risk.

We propose using the margins of the top-2 mean predictions and their corresponding standard deviations. This approach extends the BvSB score used by Joshi et al. (2009), incorporating a sensitivity scalar k (applied to all classes). Let $\bar{p}_1$ and $\bar{p}_2$ denote the top-1 and top-2 ranked classes of $\bar{\mathbf{p}}$, we define a prediction rule $\hat{y}$ and derive a signal-to-noise ratio (SNR) as:

$$\hat{y} = \begin{cases} i & \text{if } \bar{p}_1 - k\mathrm{S}_1 > \bar{p}_2 + k\mathrm{S}_2 \\ \text{uncertain} & \text{otherwise} \end{cases}; \qquad \mathrm{SNR} = \frac{\bar{p}_1 - \bar{p}_2}{\mathrm{S}_1 + \mathrm{S}_2 + \varepsilon} > k \tag{20}$$

where $\bar{p}_1 - \bar{p}_2$ is the probability margin between the two most likely classes and $\mathrm{S}_1 + \mathrm{S}_2 + \varepsilon$ is the combined standard deviations. In principle, the SNR can be interpreted as a binary decision boundary between classes $\bar{p}_1$ and $\bar{p}_2$, restricted by k. Under mild distributional assumptions on ensemble dispersion, threshold k can be restricted to reflect the fraction of samples that require abstentions or human intervention (*i.e.*, by sorting SNR in increasing values). For example, when $k = 1$, only samples with SNR > 1 will be considered; all others are uncertain. However, this criterion fails to capture cases where a model outputs ambiguous and uncertain predictions. Such outputs artificially inflate the SNR values, leading to misleading classifications. To address this limitation, we introduce VGMU, using a gating function that rescales model predictions by incorporating both confidence and variance (*i.e.*, epistemic) information

$$\mathrm{VGMU} = 1 - \gamma\bar{p}_1, \qquad \gamma = 1 - e^{-\mathrm{SNR}} \tag{21}$$

The VGMU functions as an uncertainty metric, where small values correspond to confident, well-separated predictions (low uncertainty), while large values capture ambiguous or uncertain cases (low separation or high variance).

3.3 Variance-Gated Uncertainty Decomposition

Using the variance-gated distribution $\mathbf{q}$, the total (predictive) uncertainty is measured by

$$\mathrm{TU} \Rightarrow H(\bar{\mathbf{q}}) = -\bar{\mathbf{q}}^\top \log \bar{\mathbf{q}}. \tag{22}$$

Following standard ensemble decomposition (Houlsby et al., 2011), we define the gated aleatoric and epistemic components as

$$\mathrm{AU} \Rightarrow \frac{1}{M}\sum_{m=1}^{M} H(\mathbf{q}_m) = -\frac{1}{M}\sum_{m=1}^{M} \mathbf{q}_m^\top \log \mathbf{q}_m, \qquad \mathrm{EU} \Rightarrow \mathrm{TU} - \mathrm{AU} = \frac{1}{M}\sum_{m=1}^{M} D_{\mathrm{KL}}(\mathbf{q}_m \,\|\, \bar{\mathbf{q}}). \tag{23}$$

To compare our variance-gated decomposition, we computed the corresponding standard decomposition without variance-gating (*i.e.* distributions $\mathbf{p}_m$) and inter-member disagreements as unbounded EPKL (Schweighofer et al., 2023), and bounded Expected Pairwise Jensen-Shannon (EPJS) divergences using each ensemble member pair (i, j), the midpoint $\mathbf{m}_{ij} = \frac{1}{2}(\mathbf{p}_i + \mathbf{p}_j)$, where pairwise measures are calculated as

$$\mathrm{EPKL} = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_i \,\|\, \mathbf{p}_j), \qquad \mathrm{D}_{\mathrm{KL}}(\mathbf{q}_i \,\|\, \mathbf{p}_j) = \mathbf{p}_i^\top(\log \mathbf{p}_i - \log \mathbf{p}_j), \tag{24}$$

$$\mathrm{EPJS} = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} \mathrm{D}_{\mathrm{JS}}(\mathbf{p}_i \,\|\, \mathbf{p}_j), \qquad \mathrm{D}_{\mathrm{JS}}(\mathbf{p}_i \,\|\, \mathbf{p}_j) = \tfrac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathbf{p}_i \,\|\, \mathbf{m}_{ij}) + \tfrac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathbf{p}_j \,\|\, \mathbf{m}_{ij}). \tag{25}$$

4 Analytical Gradients for Variance-Gated Normalization

In this section, we introduce reverse-mode differentiation in the ensemble setting, using VGN to capture epistemic signals during training. We discuss vector-Jacobian products through the variance-gated normalized layer, gradients of the gate $\mathbf{\Gamma}$ with respect to $\bar{\mathbf{p}}, \mathbf{s}$ and a learnable per-class parameter $\mathbf{k}$. By optimizing a negative log-likelihood objective (*e.g.*, cross-entropy), the model implicitly minimizes total (predictive) uncertainty, with a particular emphasis on reducing the epistemic component (Equation 2). The propagation of these gradients back to ensemble members provides a practical approach for variance-aware training within modern automatic differentiation frameworks.

Recall that each ensemble member produces $\mathbf{p}_m \in \Delta^{C-1}$ and all members share the same gating function $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$, with ensemble statistics $\bar{\mathbf{p}}, \mathbf{s}$ and $\mathbf{k}$ defined in Section 3. The normalized gated member distribution and its normalization constant are

$$\mathbf{q}_m = \frac{\mathbf{p}_m \odot \mathbf{\Gamma}}{Z_m}, \qquad Z_m = \mathbf{p}_m^\top \mathbf{\Gamma}. \tag{26}$$

When the training objective is applied to the predicted ensemble mixture distribution (averaged across models), the loss depends only on $\bar{\mathbf{q}}$, where $\mathcal{L} = \mathcal{L}(\bar{\mathbf{q}})$ such that

$$\bar{\mathbf{q}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{q}_m \implies \frac{\partial \bar{\mathbf{q}}}{\partial \mathbf{q}_m} = \frac{1}{M}\mathbf{I}. \tag{27}$$

Then by the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{q}_m} = \frac{\partial \bar{\mathbf{q}}}{\partial \mathbf{q}_m}\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{q}}} \implies \frac{1}{M}\mathbf{u}, \quad \mathbf{u} = \frac{\partial \mathcal{L}}{\partial \bar{\mathbf{q}}} \quad \text{(upstream gradient)}. \tag{28}$$

This quantity is obtained by differentiating the loss objective with respect to the variance-gated mixture distribution $\bar{\mathbf{q}}$ and represents the gradient signal backpropagated through the variance-gated normalization layer.

In the following analysis we decompose the total gradient of the mixture objective into its independent paths and describe gradients that flow through the ensemble statistics (mean and variance) that parameterize the gate. These local gradients are combined to provide a general reverse-mode differentiation rule for shared ensemble gates. This defines a complete analytical foundation for backpropagating uncertainty-aware gradients, in which the supervised cross-entropy loss depends only on $\bar{\mathbf{q}}$. This can be interpreted as optimizing a model by minimizing predictive uncertainty, subject to epistemic constraints (complete analytical derivations and gradient expressions available in SI S2).

4.1 Full Gradient Decomposition

Proposition 4.1. *When the objective depends on the ensemble mixture* $\bar{\mathbf{q}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{q}_m$*, the total gradient received by each ensemble member is*

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m} = \frac{1}{M}\left(\left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\mathbf{\Gamma}} + \left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\bar{\mathbf{p}}} + \left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\mathbf{s}} \right). \tag{29}$$

Proof. Each path represents an independent dependency of $\mathcal{L}$ on $\mathbf{p}_m$:

- **Direct normalization** (with $\mathbf{\Gamma}$ and $\mathbf{k}$ fixed): how $\mathbf{q}_m$ changes when $\mathbf{p}_m$ changes. This captures the effect of normalization on the simplex space;
- **Indirect gating *via* the mean** (with $\mathbf{s}$ and $\mathbf{k}$ fixed): how $\mathbf{q}_m$ changes when the gate $\mathbf{\Gamma}$ changes through the mean $\bar{\mathbf{p}}$; and
- **Indirect gating *via* the variance** (with $\bar{\mathbf{p}}$ and $\mathbf{k}$ fixed): how $\mathbf{q}_m$ changes when the gate $\mathbf{\Gamma}$ changes through the predictive spread $\mathbf{s}$.

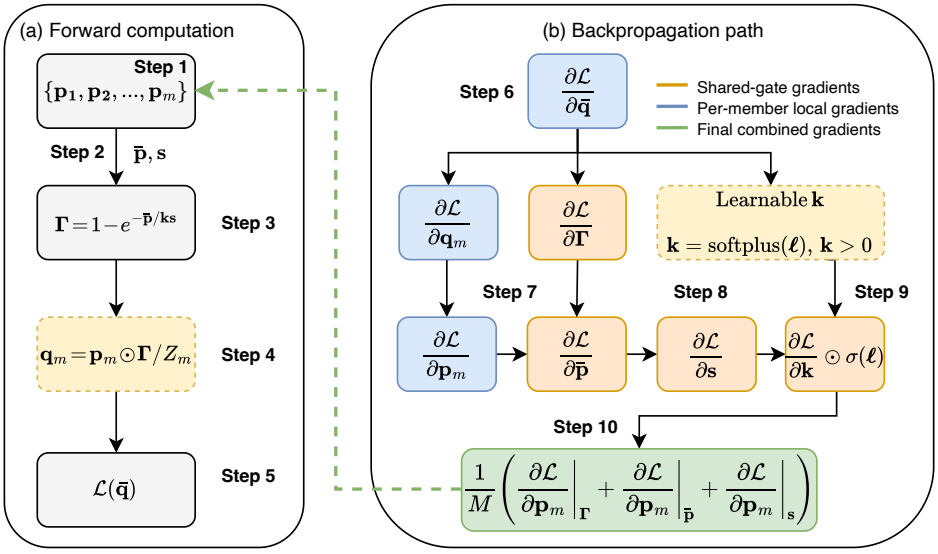


Figure 1: Forward and backward passes of VGN. Panel (a) displays the forward computation, in which ensemble predictions are modulated by a shared variance gate and combined into a mixture distribution. Panel (b) displays the backpropagation path, showing how gradients propagate through the normalization layer and shared gate *via* ensemble mean and predictive spread. See below for further step-by-step discussion.

By the multivariate chain rule, these contributions combine additively to provide the total gradient. □

Remark 4.1.1. *The term "fixed" in the parameters definition denotes that the corresponding variable is held constant during partial differentiation.*

We instantiate these formulations for our variance-gate introduced in Section 3, providing analytic gradients with respect to gating statistics, $(\bar{\mathbf{p}}, \mathbf{s})$ and per-class learnable parameter $\mathbf{k}$. For the exponential variance-gate $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$, gradients with respect to the ensemble mean $\bar{\mathbf{p}}$, predictive spread $\mathbf{s}$, and sensitivity parameter $\mathbf{k}$ follow directly by the chain rule and are summarized in SI S2. Collectively, these derivations establish a complete analytic framework for backpropagating uncertainty-aware gradients through ensemble networks, enabling end-to-end optimization under epistemic-sensitive normalization (Figure 1). The variance-gated normalization layer propagates gradients across ensemble members $\mathbf{p}_m$ through a shared gating mechanism parameterized by ensemble statistics $(\bar{\mathbf{p}}, \mathbf{s})$ and a learnable sensitivity parameter $\mathbf{k}$. Each member's forward activation contributes to the ensemble mean and variance, which adjusts the shared gate $\mathbf{\Gamma}$ during backpropagation. By including a learnable parameter $\mathbf{k} = \text{softplus}(\boldsymbol{\ell})$, models can adaptively modulate the strength of epistemic signals for optimizing predictive uncertainty. See Table S2 for a summary of analytical gradients used for the variance-gated normalization framework.

We summarize the forward and backward passes of VGN in Figure 1 and describe them step-by-step below:

(a) Forward computation

- **Step 1**: Each ensemble member $m \in \{1, \ldots, M\}$ produces a categorical predictive distribution $\mathbf{p}_m \in \Delta^{C-1}$.
- **Step 2**: The ensemble sample mean $\bar{\mathbf{p}}$ and predictive spread $\mathbf{s}$ are computed across members. These statistics summarize ensemble consensus and disagreement and are shared across all ensemble members.
- **Step 3**: Using the ensemble statistics, a shared variance gate is constructed as $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$ where $\mathbf{k} > 0$ controls the sensitivity of the gate.

- **Step 4**: Each member distribution is modulated by the shared gate and re-normalized as: $\mathbf{q}_m = \mathbf{p}_m \odot \mathbf{\Gamma}/Z_m$, $Z_m = \mathbf{p}_m^\top \mathbf{\Gamma}$.
- **Step 5**: The ensemble mixture $\bar{\mathbf{q}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{q}_m$ is formed, and the loss $\mathcal{L}$ is applied to this mixture distribution.

(b) Backpropagation path

- **Step 6**: Gradients with respect to the mixture are distributed equally to ensemble members as: $\frac{\partial \mathcal{L}}{\partial \mathbf{q}_m} = \frac{1}{M}\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{q}}}$.
- **Step 7**: The total gradient with respect to each member prediction $\mathbf{p}_m$ decomposes into three additive contributions (Proposition 4.1): (i) a direct path through the normalization with the gate held fixed, $\partial \mathcal{L}/\partial \mathbf{p}_m|_{\mathbf{\Gamma}}$; (ii) an indirect path through the ensemble mean $\bar{\mathbf{p}}$; and (iii) an indirect path through the predictive spread $\mathbf{s}$.
- **Step 8**: Both indirect paths propagate through the shared variance gate *via* $\partial \mathbf{\Gamma}/\partial \bar{\mathbf{p}}$ and $\partial \mathbf{\Gamma}/\partial \mathbf{s}$, coupling gradients across ensemble members.
- **Step 9**: The sensitivity parameter $\mathbf{k}$, reparameterized as $\mathbf{k} = \mathrm{softplus}(\boldsymbol{\ell})$, receives gradients through its influence on the gate $\mathbf{\Gamma}$.
- **Step 10**: All three gradient contributions are summed and scaled by $1/M$, resulting with the final per-member gradient update.

For completeness, all Jacobians, vector–Jacobian products, and gradients with respect to ensemble statistics and learnable gating parameters are derived in full in Supporting Information Section S1.

5 Experiments

We evaluate the proposed variance-gated ensemble framework on MNIST, SVHN, CIFAR-10, and CIFAR-100 using convolutional backbones and ensemble configurations commonly employed in uncertainty estimation. For MNIST we use a LeNet-5 style network, while for SVHN and CIFAR-10/100 we use WideResNet-28-10. Experiments consider MCD, LLE, and their hybrid variants MCD-LLE, with and without VGN. Models are trained using the Adam optimizer with early stopping, and all results are averaged over three trials with different random seeds. We compare VGMU against entropy-based EU (mutual information) and recent information-theoretic baselines, including EPKL and EPJS pairwise divergence measures. For variance-gated variants, the classwise gating parameter $\mathbf{k}$ is learned end-to-end. Uncertainty scores were assessed *via* rank-based agreement with baseline methods (Spearman's ρ and Kendall's τ), Cumulative Area Under the Curve (AUC_c) for uncertainty mass concentration, and margin–variance geometry visualizations. Predictive performance and calibration are reported using accuracy, F1-score, Expected Calibration Error (ECE), and ensemble diversity. Complete network specifications, training protocols, evaluation definitions, and implementation details are provided in SI S3, and additional results for MNIST and SVHN appear in SI S4. MNIST and SVHN are included as complementary benchmarks to provide uncertainty behavior in low-noise and image settings.

5.1 Rank Consistency and Alignment

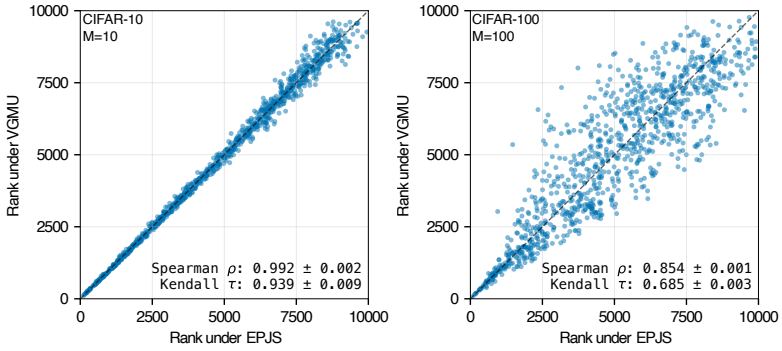
We compare the uncertainty rankings produced by VGMU with those from commonly used information-theoretic uncertainty measures. Table 1 reports Spearman correlations for CIFAR-10 and CIFAR-100 (additional results in SI S4.1) computed for each testing dataset and ensemble configuration.

On CIFAR-10, VGMU align with all baselines, indicating that the margin-based score captures similar uncertainty structure as pairwise divergence measures and mutual information. For the CIFAR-100 dataset, correlations remain strong for MCD variants but diverged for LLE models (*e.g.*, EPKL, $\rho = 0.566$). This

Table 1: Spearman rank correlation (ρ) between VGMU and epistemic uncertainty baselines.[1]

Dataset	Method	VGMU *vs.* EPJS	VGMU *vs.* EPKL	VGMU *vs.* EU
CIFAR-10	MCD (M=10)	0.989 ± 0.002	0.985 ± 0.002	0.986 ± 0.002
	MCD-LLE (M=10)	0.987 ± 0.002	0.986 ± 0.002	0.986 ± 0.002
	LLE (M=10)	0.992 ± 0.002	0.985 ± 0.004	0.988 ± 0.004
	LLE-VGN (M=10)	0.993 ± 0.002	0.987 ± 0.003	0.990 ± 0.002
CIFAR-100	MCD (M=10)	0.874 ± 0.012	0.856 ± 0.011	0.862 ± 0.011
	MCD-LLE (M=10)	0.874 ± 0.015	0.860 ± 0.016	0.864 ± 0.016
	LLE (M=10)	0.854 ± 0.001	0.566 ± 0.035	0.814 ± 0.004
	LLE-VGN (M=10)	0.875 ± 0.022	0.697 ± 0.017	0.843 ± 0.023

[1] Higher values indicates better alignment with information-theoretic measures.



(a) (b)

Figure 2: Rank consistency between VGMU and EPJS on CIFAR-10 (a) and CIFAR-100 (b) for the LLE models. The diagonal denotes perfect agreement between uncertainty rankings.

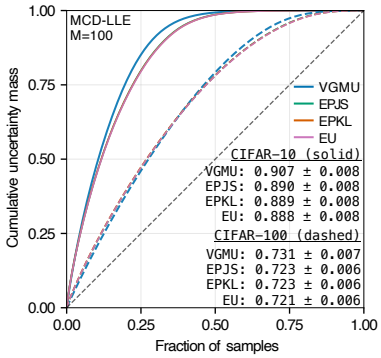


Figure 3: AUCc curves for CIFAR-10/100. The diagonal (AUC_c = 0.5), corresponds to no concentration on difficult samples.

reflects a deliberate design choice rather than a limitation. Pairwise measures capture distributional disagreement across all 100 classes, while VGMU focuses exclusively on the decision-relevant margin between the top-2 predictions. When ensemble members agree on the most likely classes but disagree about the tail distribution, pairwise measures increase substantially while VGMU remains low. For practical decision-making (*e.g.*, whether to trust a prediction or defer to a human for review), disagreement about distant classes is irrelevant; VGMU's insensitivity to such disagreement is a feature, moving away from information-theoretic approaches (Wimmer et al., 2023).

The addition of VGN improves LLE's correlation with EPKL from 0.566 to 0.697, indicating that normalization can partially recover sensitivity to distributional disagreement. Figure 2 illustrates this behavior through rank-rank scatter plots. Points cluster along the diagonal for CIFAR-10, while CIFAR-100 displays off-diagonal deviations for LLE models. These deviations correspond to samples where ensemble members agree on the top predictions but disagree about lower-ranked classes, the cases where VGMU's decision-focused design diverges from entropy-based scores.

5.2 Uncertainty Mass Concentration

For practical deployment in selective prediction or human-in-the-loop systems, uncertainty estimates should concentrate on difficult samples; otherwise they become non-informative for decision-making under uncertainty.

We quantify this through AUC_c, calculated by sorting samples by descending uncertainty and measuring the area under the cumulative distribution. Higher AUC_c values indicate more efficient concentration of

Table 2: Cumulative area of the curve (AUC_c) scores for uncertainty mass concentration. Higher values indicate sharper concentration on difficult samples.

Dataset	Method	VGMU	EPJS	EPKL	EU
CIFAR-10	MCD (M=10)	0.895 ± 0.007	0.883 ± 0.007	0.883 ± 0.007	0.881 ± 0.007
	LLE (M=10)	0.885 ± 0.015	0.873 ± 0.016	0.876 ± 0.015	0.868 ± 0.016
	MCD-LLE (M=100)	0.907 ± 0.008	0.890 ± 0.008	0.889 ± 0.008	0.888 ± 0.008
CIFAR-100	MCD (M=10)	0.719 ± 0.006	0.726 ± 0.005	0.728 ± 0.005	0.725 ± 0.005
	LLE (M=10)	0.628 ± 0.006	0.643 ± 0.008	0.668 ± 0.013	0.651 ± 0.007
	MCD-LLE (M=100)	0.731 ± 0.007	0.723 ± 0.006	0.723 ± 0.006	0.721 ± 0.006

uncertainty on difficult samples, while $\text{AUC}_c = 0.5$ no concentration. Table 2 summarizes AUC_c across datasets (additional information provided in SI S4.2).

On CIFAR-10, VGMU is comparable across all evaluated ensemble configurations. In several cases, the separation in mean values relative to information-theoretic baselines is larger than the observed run-to-run variability, indicating a tendency toward stronger concentration of uncertainty on difficult samples. From a practical perspective, this behavior implies that VGMU can prioritize challenging samples more efficiently, requiring fewer predictions to capture a fixed proportion of the total uncertainty mass. This trend is less pronounced for CIFAR-100. Information-theoretic measures show better performing AUC_c values than VMGU for all settings, reflecting VGMU's insensitivity to disagreement across the full class simplex. This observation is consistent with the rank-correlation analysis, where information-theoretic measures capture forms of uncertainty that VGMU intentionally de-emphasizes. However, AUC_c can be restored using the hybrid MCD-LLE configuration (although difference within observed variability). This pattern suggests that increasing ensemble diversity through the combination of stochastic sampling and multiple classifier heads may help recover decision-relevant uncertainty structure, allowing margin-based estimation to remain competitive on more challenging tasks. Figure 3 shows cumulative uncertainty mass concentration curves for CIFAR-10 and CIFAR-100 using MCD-LLE models ($M = 100$). For CIFAR-10, uncertainty mass is concentrated among a relatively small fraction of sample (*ca.* 25%), VGMU shows a small but consistent upward shift relative to the information-theoretic baselines. In this case of CIFAR-100, the signals are noticeably closer to the diagonal ($\text{AUC}_c \approx 0.5$), indicating substantially weaker concentration. This collapse in concentration is consistent with the increased difficulty of CIFAR-100, where uncertainty is spread across a broader set of samples and full-simplex disagreement becomes more prominent. SI S4.2 results indicate a general trend where VGMU adopts a conservative uncertainty assessment. In these cases, uncertainty is distributed more broadly across samples (*i.e.*, $\text{AUC}_c \approx 0.5$) rather than being concentrated, consistent with the margin-based design of VGMU.

5.3 Margin-Variance Geometry

The VGMU score explicitly couples the predictive margin ($\bar{p}_1 - \bar{p}_2$) with ensemble variance ($S_1 + S_2$), providing a two-dimensional uncertainty landscape. This geometric perspective provides insight into how different ensemble methods populate the margin-variance space and how VGMU responds to these configurations. Figure 4 illustrates the margin-variance landscape for CIFAR-100 (additional visualizations provided in SI S4.3). Several consistent patterns emerge:

- **Ensemble method signatures.** Different ensemble strategies show characteristic distributions in margin-variance space. DE and DE-VGN scatter broadly, showing substantial variance even at high margins. LLE and LLE-VGN occupy regions with moderate variance and moderate-to-high margins, while MCD and MCD-LLE produce more compact clusters with lower overall variance.

- **VGMU response.** The behavior of VGMU across this space confirms its sensitivity to both margin and variance. Low VGMU values (blue) occur predominantly when the predictive margin is large and variance is low, whereas high values (yellow) arise from either small margins or elevated variance.

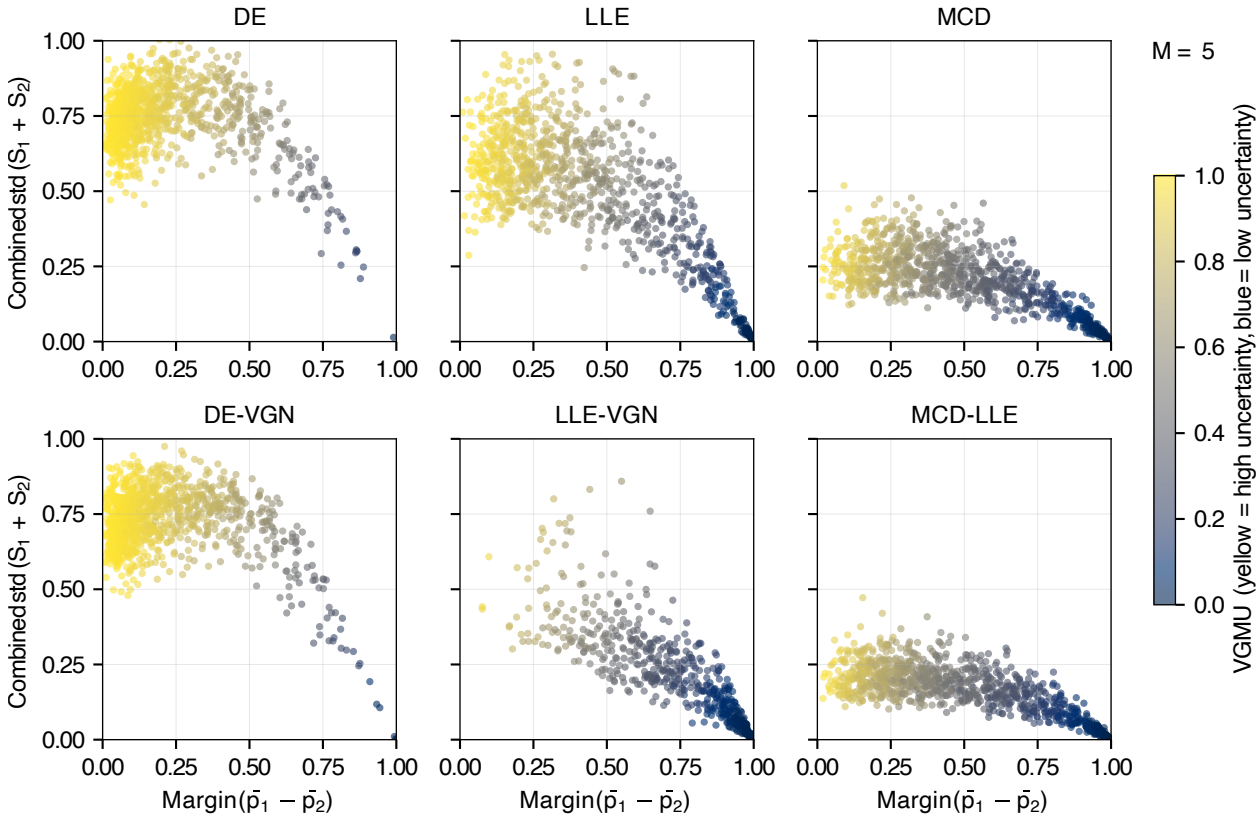


Figure 4: Margin-variance geometry for CIFAR-100 with $M = 5$. Each point represents a test sample; color indicates VGMU value (yellow = high uncertainty, blue = low). DE and DE-VGN show high variance even at large margins, LLE and LLE-VGN show moderate variance, while MCD and MCD-LLE concentrates samples in the high-margin (confident), low-variance (certain) region.

The variance-gated formulation induces smooth transitions rather than hard decision boundaries, enabling graded confidence assessments.

- **Effect of VGN.** Comparing LLE and LLE-VGN reveals a shift toward lower VGMU values, with samples concentrating in the high-margin, low-variance region while high-uncertainty regions remain populated. This reduction in variance is achieved through VGN, which suppresses high-variance class predictions during training. Similar trends are observed across other datasets and configurations in SI S4.3, although in some cases the effects are subtle.

The margin-variance geometry provides an intuitive interpretation of how, predictive margin and epistemic disagreement interact. By coupling margin and variance, VGMU assigns low uncertainty only when predictions are well-separated and consistent across ensemble members, while remaining conservative in regions characterized by ambiguity or disagreement. The effect of VGN is to reshape this geometry during training by suppressing high-variance class predictions, leading to a greater concentration of samples in the high-margin, low-variance region (Figure 4: LLE *vs.* LLE-VGN). This geometric behavior is consistent with the rank-based and concentration analyses, and shows how VGN complements VGMU by promoting decision-relevant uncertainty with conservative assessments in challenging settings.

5.4 Calibration and Performance

Table 3 reports calibration metrics for CIFAR-10 and CIFAR-100 (additional results in SI S4.5). For DE on CIFAR-10, VGN reduces mean ECE by a margin that exceeds the observed run-to-run variability, while also improving classification accuracy, indicating improved alignment with predictive confidence. In

Table 3: Performance and calibration metrics.

Dataset	Method	Accuracy	F1-Score	ECE	Diversity[1]
CIFAR-10	DE (M=5)	0.836 ± 0.012	0.836 ± 0.012	0.049 ± 0.014	$1.9 \times 10^{-2} \pm 0.2$
	DE-VGN (M=5)	0.875 ± 0.005	0.875 ± 0.006	0.023 ± 0.003	$9.8 \times 10^{-3} \pm 0.5$
	LLE (M=10)	0.848 ± 0.003	0.848 ± 0.002	0.058 ± 0.010	$3.6 \times 10^{-3} \pm 0.6$
	MCD-LLE (M=100)	0.853 ± 0.008	0.852 ± 0.008	0.067 ± 0.004	$3.6 \times 10^{-4} \pm 0.2$
CIFAR-100	DE (M=5)	0.487 ± 0.002	0.481 ± 0.003	0.095 ± 0.006	$3.5 \times 10^{-3} \pm 0.1$
	DE-VGN(M=5)	0.507 ± 0.007	0.504 ± 0.007	0.086 ± 0.006	$3.1 \times 10^{-3} \pm 0.2$
	LLE (M=10)	0.543 ± 0.005	0.538 ± 0.005	0.062 ± 0.004	$2.4 \times 10^{-3} \pm 0.1$
	MCD-LLE (M=100)	0.565 ± 0.013	0.563 ± 0.013	0.102 ± 0.001	$2.3 \times 10^{-4} \pm 0.1$

[1] Defined as the ensemble variance averaged across samples and classes, $\mathbb{E}_{i,c}[\mathrm{Var}_M]$.

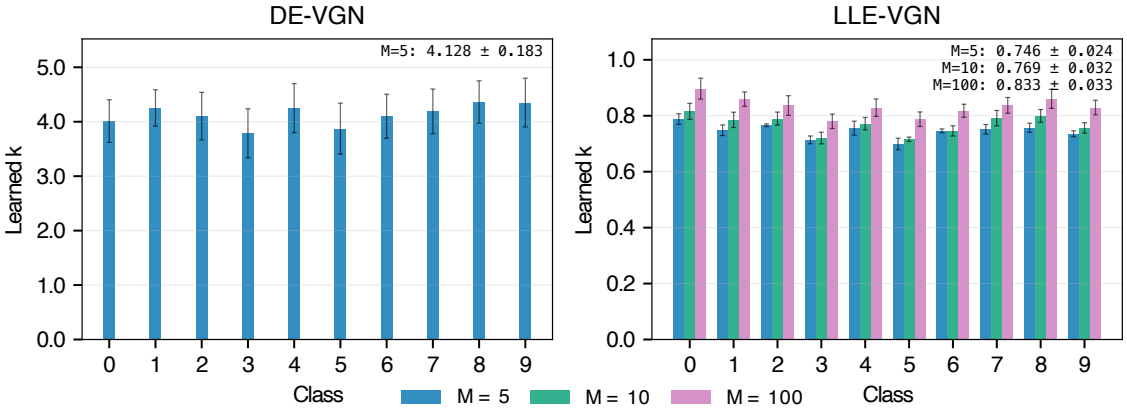


Figure 5: Learned per-class $\mathbf{k}$ values for VGN models on CIFAR-10. DE-VGN learns higher values ($\bar{k} \approx 4.1$) than LLE-VGN ($\bar{k} \approx 0.8$), reflecting adaptation to ensemble diversity.

contrast, for CIFAR-100 and for other ensemble settings, calibration effects are smaller and less consistent, with differences falling within the variability across runs.

Overall, the results indicate that incorporating VGN does not adversely affect calibration and, in several configurations, provides small but consistent improvements. Similar trends are observed for the MCD-LLE setting, where VGN tends to improve or preserve calibration rather than degrade it. Figure 5 displays the per-class $\mathbf{k}$ parameters learned by VGN models (see Figure S5 for CIFAR-100). DE-VGN learns higher values ($\bar{k} \approx 2.5$–4.0) than LLE-VGN ($\bar{k} \approx 0.75$–1.1), reflecting the diversity in deep ensembles. Higher $\mathbf{k}$ reduces gate sensitivity (see Proposition 3.3 for additional details), allowing DE-VGN to tolerate the disagreement among independently trained members. In these examples, the relative consistency of $\mathbf{k}$ across classes suggests that it adapts to ensemble-level diversity rather than class-specific difficulty. However, results on SVHN (see Figure S4) indicate that $\mathbf{k}$ captures both effects, varying across classes while also reflecting overall ensemble diversity.

A key advantage of the VGE framework is computational efficiency. Divergence measures scale as $O(M^2C)$, whereas our epistemic-aware VGE's uncertainty decomposition is $O(MC)$ and $O(C)$ for VGMU evaluations once moments are computed. The quadratic scaling of pairwise methods becomes prohibitive for large ensembles. With $M = 100$ members and $C = 100$ classes, pairwise measures requires 10^6 operations per sample compared to 10^4 for VGN decomposition and 10^2 for VGMU. On CIFAR-100, wall-clock per-sample inference evaluation on a MacBook Pro (Apple M3 Pro; 36 GB) with LLE-VGN ($M = 10$) completes in 0.1 μs, compared to 36.5 μs for EPKL. This corresponds to a 365× speedup, offering real-time uncertainty estimation in deployment.

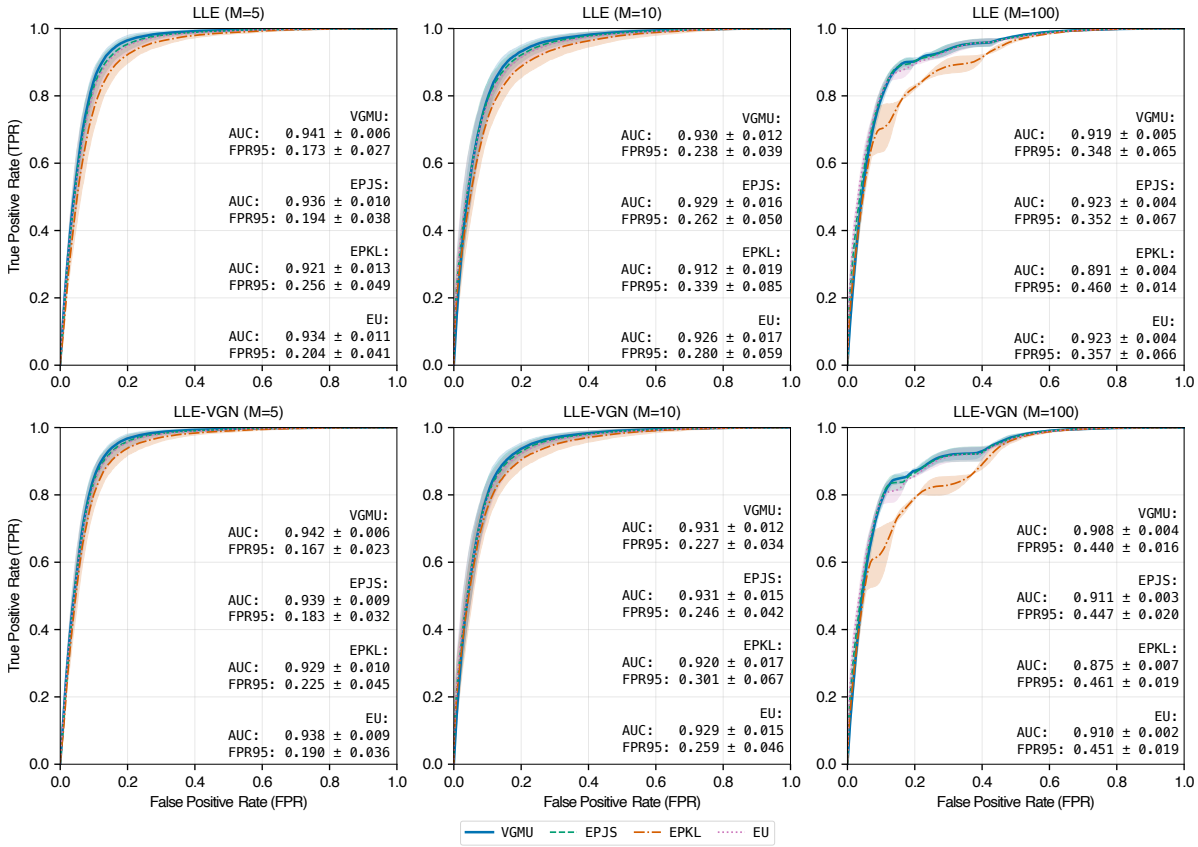


Figure 6: ROC curves for OOD detection (SVHN, ID→ CIFAR-10, OOD) across LLE (top) and LLE-VGN (bottom) with $M \in \{5, 10, 100\}$. At small ensemble sizes ($M = 5$), VGMU shows a visible separation from EPKL in both AUC and FPR@95, while differences with EPJS and EU fall within observed variability. As M increases, FPR@95 degrades for all methods. VGN provides consistent but modest FPR@95 improvements at $M = 5$ and $M = 10$, with degradation at $M = 100$.

5.5 Out-of-Distribution Detection

To evaluate whether VGN and/or the VGMU score provides competitive OOD detection, we train models on SVHN (ID) and evaluate against CIFAR-10 (OOD). Detection performance is measured using the area under the ROC curve (AUC) and the false positive rate at 95% true positive rate (FPR@95). All LLE and LLE-VGN models for SVHN attained *ca.* 95% accuracy with strong calibration (ECE ≈ 0.014). As in prior sections, we interpret differences conservatively, distinguishing trends whose magnitude exceeds run-to-run variability.

LLE and LLE-VGN across ensemble sizes. Figure 6 reports OOD detection results for LLE and LLE-VGN across $M \in \{5, 10, 100\}$. At $M = 5$, VGMU achieves the highest mean AUC (0.941 ± 0.006 for LLE; 0.942 ± 0.006 for LLE-VGN) and lowest mean FPR@95 (0.173 ± 0.027; 0.167 ± 0.023). The AUC differences between VGMU and EPJS or EU fall within overlapping variability, indicating comparable detection performance among these measures. However, the separation from EPKL is more pronounced. For LLE ($M = 5$), the FPR@95 gap between VGMU (0.173 ± 0.027) and EPKL (0.256 ± 0.049) exceeds the observed run-to-run variability. As ensemble size increases to $M = 10$ and $M = 100$, FPR@95 degrades for all methods, reflecting reduced per-member diversity in LLE configurations where heads share a single backbone. With $M = 100$, VGMU and EPJS perform comparably in AUC (≈ 0.919–0.923), while EPKL degrades more substantially (AUC $= 0.891 \pm 0.004$ for LLE; 0.875 ± 0.007 for LLE-VGN), consistent with the known sensitivity of pairwise KL divergence to distributional tail disagreement. Incorporating VGN

provides a consistent but modest improvement in FPR@95 at small-to-moderate ensemble sizes. At $M = 5$, LLE-VGN reduces mean FPR@95 from 0.173 to 0.167; at $M = 10$, from 0.238 to 0.227. While these improvements fall within seed-level variability, the direction is consistent across both ensemble sizes. At $M = 100$, however, VGN degrades both AUC (0.908 *vs.* 0.919) and FPR@95 (0.440 *vs.* 0.348), which may result from the suppression of high-variance predictions during training, reducing ensemble disagreement signals that distinguish ID from OOD at larger ensemble sizes.

Across all configurations, VGMU achieves comparable detection performance while avoiding the $O(M^2C)$ computational cost of pairwise divergence measures, relying instead on ensemble moments with $O(MC)$ complexity.

MCD-LLE: Effect of ensemble configuration. To examine how the source of ensemble diversity affects OOD detection, we evaluate MCD-LLE under varying head (H) and sampling (S) configurations (SI S4.6). VGMU remains remarkably stable across all configurations, maintaining AUC $= 0.934 \pm 0.006$ and FPR@95 ≈ 0.216 regardless of the head–sampling configurations. In contrast, information-theoretic baselines improve as either H or S increases, with mean AUC rising from 0.911 ($M = 5$) to 0.916 ($M = 100$) for EPJS. The AUC gap between VGMU and the baselines ($\approx$ 0.018–0.023) exceeds the observed run-to-run variability across all configurations, representing a consistent separation. This stability arises because VGMU depends only on the top-2 margin and its associated variance, both of which saturate quickly, even with modest ensemble sizes. From a practical perspective, this insensitivity to ensemble configuration is advantageous. VGMU can provide reliable OOD signals without rigorous tuning of MCD-LLE configurations.

Summary. Across ensemble types and configurations, VGMU provides OOD detection performance that is comparable to or exceeds information-theoretic baselines. In the LLE setting, VGMU and EPJS/EU produce similar AUC values with overlapping variability, while VGMU consistently outperforms EPKL in FPR@95 by a margin exceeding run-to-run variation. In the MCD-LLE setting, the separation between VGMU and all baselines is more robust, with AUC differences that consistently exceed the observed standard deviations. VGN provides modest FPR@95 improvements at small ensemble sizes but does not enhance OOD detection at large M. Collectively, these results confirm that variance-gated, margin-based uncertainty preserves sensitivity to distributional shift while remaining computationally efficient.

6 Discussion

6.1 Axiomatic Analysis

We introduced VGE as a computationally efficient alternative to entropy- and divergence-based uncertainty decomposition, with sensitivity to epistemic disagreement. Table 4 positions VGN within the axiomatic framework of Wimmer et al. (2023). These axioms highlight where standard entropy-based decompositions fail under finite ensembles and where disagreement-based alternatives improve epistemic sensitivity (for additional examples and discussion, see SI S6). VGN satisfies non-negativity (A0) and vanishing epistemic uncertainty for identical ensemble members (A1). This behavior is reflected in stable uncertainty estimates on low-noise datasets such as MNIST and in confident predictions on CIFAR-10 when ensemble members agree. In these regions, both the variance-gated decomposition and VGMU assign low uncertainty, consistent with theoretical expectations and the strong rank agreement observed with information-theoretic baselines. Both VGN and pairwise divergence methods satisfy the mean-preserving variance axiom (A3). The margin-variance geometry in Section 5.3 further supports this interpretation. Samples with increased ensemble spread are consistently assigned higher VGMU values, even when mean confidence remains high. Standard entropy-based decompositions violate axioms A2 and A4, relating to the behavior under uniform predictions and injected noise. These violations help explain the weaker alignment and reduced uncertainty mass concentration observed for mutual-information-based epistemic uncertainty in Section 5.2. In practice, entropy-based measures can overestimate uncertainty in diffuse predictions or underestimate uncertainty when confident predictions mask ensemble disagreement. VGN addresses these limitations by suppressing high-variance class probabilities prior to normalization. Rather than relying on entropy identities that may

Table 4: Comparison of uncertainty decomposition frameworks. Each row corresponds to an axiom (A0–A5), and each column indicates whether a given framework satisfies an axiom[1].

Axioms Wimmer et al. (2023)	**Standard Entropy Decomposition** Houlsby et al. (2011)	**Expected Pairwise CE/KL Divergence** Schweighofer et al. (2023)	**Variance-Gated Normalization (proposed)**
A0: TU, AU, EU ≥ 0 (non-negativity)	✓	✓	✓
A1: EU $= 0$ for identical ensemble members.	✓	✓	✓
A2: EU, TU maximal when ensembles are uniform distributions.	✗	◑	◑
A3: EU$^{\uparrow}$, TU$^{\uparrow}$ with mean-preserving variance increases.	✓	✓	✓
A4: AU$^{\uparrow}$, TU$^{\uparrow}$ with addition of uniform noise.	✗	◑	✓
A5: EU invariant to variance-preserving location shifts.	✓	✓	✓

[1] ✓ Satisfied, ✗ violated, ◑ partially satisfied.

not hold under finite sampling, VGN reshapes ensemble predictions using moment-based statistics, providing uncertainty estimates that reflect decision-relevant ambiguity while avoiding quadratic computational costs.

Invariance to variance-preserving location shifts (A5), satisfied by both pairwise divergence methods and VGN, explains why VGMU remains stable under changes in tail-class disagreement. As shown in Section 5.1, this invariance accounts for the divergence between VGMU and full-simplex disagreement measures on CIFAR-100. When ensemble members agree on the most likely classes but differ in low-probability regions, pairwise divergences increase substantially, whereas VGMU remains low. VGMU is decision-focused, quantifying uncertainty through the margin between the top-ranked classes modulated by predictive variance. Disagreement that does not affect the decision boundary is intentionally de-emphasized, aligning uncertainty estimation with selective prediction and human-in-the-loop decision-making.

The axiomatic and results together explain why variance-gated ensembles retain the desirable properties of disagreement-based uncertainty measures while avoiding their principal drawbacks. Pairwise divergence methods scale as $O(M^2C)$, becoming prohibitive for large ensembles or many-class problems. In contrast, VGN operates at $O(MC)$ for uncertainty decomposition and $O(C)$ for VGMU evaluation once ensemble moments are computed, enabling real-time deployment. Despite this efficiency, performance is not compromised. VGMU matches or slightly improves upon information-theoretic baselines in uncertainty ranking and out-of-distribution detection, while VGN improves or preserves calibration across ensemble configurations. The learned sensitivity parameter $\mathbf{k}$ adapts to ensemble diversity and task difficulty, supporting the interpretation of variance-gating as a data-driven mechanism for epistemic control.

6.2 Limitations and Future Work

The proposed framework suggests several directions for future work. First, VGMU focuses on the margin between the top-ranked classes; applications requiring sensitivity to full-simplex disagreement may benefit from hybrid approaches combining margin-based and distributional signals. Second, the current formulation relies on first- and second-order ensemble moments. While this enables linear-time computation, higher-order statistics or alternative measures may capture additional structure in multimodal predictive distributions. Third, the gating function itself is defined using a specific exponential form. This choice was motivated

by smoothness, monotonicity with respect to confidence and variance, and well-behaved gradients for end-to-end optimization. Alternative gating functions or parameterizations may provide different trade-offs between sensitivity and saturation, and exploring such designs remains an open direction. Finally, extending variance-gating to prediction tasks with severe class imbalance remains an open problem. In these settings, decision margins may be ill-defined and ensemble variance may conflate epistemic uncertainty with class-frequency effects. Exploring interactions between VGN and other uncertainty-aware training objectives, such as diversity-promoting loss functions, is another promising direction, as these objectives may interact non-trivially through shared gradient pathways.

7 Conclusion

We introduced VGE, an epistemic-aware uncertainty framework that leverages ensemble disagreement through a signal-to-noise gating principle. VGE unifies two complementary components operating at different stages of the learning pipeline: VGMU, a lightweight margin-based uncertainty score applicable at inference without retraining, and VGN, a differentiable normalization layer that modulates high-variance predictions during training. We derived closed-form vector–Jacobian products that enable end-to-end optimization of VGN through ensemble sample means and variances, allowing uncertainty sensitivity to be learned directly from data. Empirically, across MNIST, SVHN, CIFAR-10, and CIFAR-100, VGMU exhibited strong alignment with information-theoretic uncertainty measures, while revealing a distinct behavior on CIFAR-100. In particular, where pairwise divergence measures emphasize full-simplex disagreement, VGMU prioritizes the decision-relevant margin between top-ranked classes. Despite this difference, VGMU achieved comparable uncertainty ranking and out-of-distribution detection performance at a fraction of the computational cost, yielding a 365× speedup over EPKL on CIFAR-100 and enabling real-time uncertainty estimation. Incorporating VGN during training further suppressed high-variance predictions and reshaped ensemble member distributions, with learned sensitivity parameters adapting to ensemble diversity across models. Overall, VGE provides a computationally efficient approach to epistemic uncertainty estimation, supporting post-hoc evaluation, training-time distribution shaping, or end-to-end integration. These results suggest that decision-focused margin structure offers a practical alternative to pairwise divergence-based uncertainty, particularly in large-class settings.

Broader Impact Statement

Reliable uncertainty estimation is important for deploying machine learning systems in risk-sensitive settings. This work introduces a computationally efficient framework for epistemic-aware uncertainty estimation in ensemble models, enabling real-time uncertainty assessment in large ensembles and many-class problems. As with all uncertainty estimates, the outputs should be used as supporting signals rather than as standalone decision criteria. Overall, this work aims to advance the safe, efficient, and responsible deployment of machine learning models by making epistemic uncertainty estimation more accessible and scalable.

References

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2002.06470.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning (ICML)*, 2018. doi: 10.48550/arXiv.1710.07283.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv*, 2020. doi: 10.48550/arXiv.1912.02757.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016. doi: 10.48550/arXiv.1506.02142.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *arXiv*, 2017. doi: 10.48550/arXiv.1705.07832.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023. doi: 10.1007/s10462-023-10562-9.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv*, 2017. doi: 10.48550/arXiv.1705.08500.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. doi: 10.48550/arXiv.1706.04599.

James Harrison, John Willes, and Jasper Snoek. Variational Bayesian last layers. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2404.11599.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv*, 2011. doi: 10.48550/arXiv.1112.5745.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR)*, 2017. doi: 10.48550/arXiv.1704.00109.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 2021. doi: 10.1007/s10994-021-05946-3.

Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. doi: 10.1109/CVPR.2009.5206627.

Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022. doi: 10.48550/arXiv.2207.08977.

Kaisar Kushibar, Víctor Manuel Campello, Lidia Garrucho Moras, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. *arXiv*, 2022. doi: 10.48550/arXiv.2203.08878.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Conference on Neural Information Processing Systems (NIPS)*, 2017. doi: 10.48550/arXiv.1612.01474.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why M heads are better than one: Training a diverse ensemble of deep networks. *arXiv*, 2015. doi: 10.48550/arXiv.1511.06314.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.48550/arXiv.1802.10501.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. doi: 10.48550/arXiv.1906.02530.

Neal M. Radford. Bayesian learning for neural networks. PhD Thesis, University of Toronto, 1995.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty. In *Neural Information Processing Systems (NeurIPS)*, Mathematics of Modern Machine Learning Workshop, 2023. doi: 10.48550/arXiv.2311.08309.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Conference on Neural Information Processing Systems (NIPS)*, 2018. doi: 10.48550/arXiv.1806.01768.

Sophie Steger, Christian Knoll, Bernhard Klein, Holger Fröning, and Franz Pernkopf. Function space diversity for uncertainty prediction via repulsive last-layer ensembles. In *International Conference on Machine Learning (ICML)*, 2024. doi: 10.48550/arXiv.2412.15758.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations (ICLR)*, 2020. doi: 10.48550/arXiv.2002.06715.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bern Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023. doi: 10.48550/arXiv.2209.03302.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv*, 2017. doi: 10.48550/arXiv.1605.07146.

Supporting Information

Variance-Gated Ensembles: An Epistemic-Aware Framework for Uncertainty Estimation

Table of Contents

S1 Symbols and Abbreviations **S2**

S2 Analytical Derivations and Gradient Expressions **S3**
- S2.1 Gradients through Ensemble Mean S4
- S2.2 Gradients through Ensemble Variance S4
- S2.3 Jacobians and Vector–Jacobian Products for Shared Gates S4
- S2.4 General Reverse-Mode Differentiation for Shared Ensemble Gates S6
- S2.5 Gradients with Respect to Learnable Sensitivity Parameters S6

S3 Implementation Details **S8**
- S3.1 Model Architectures S8
- S3.2 Training Protocols S8
- S3.3 Evaluation and Uncertainty Metrics S8
- S3.4 Variance-Gated Normalization Implementation S9
- S3.5 Numerical Stability S9

S4 Additional Experimental Results **S10**
- S4.1 Rank Consistency and Alignment S10
- S4.2 Uncertainty Mass Concentration S11
- S4.3 Margin–Variance Geometry S12
- S4.4 Effects of Learned **k** S13
- S4.5 Performance and Calibration S14
- S4.6 Out-of-Domain Detection S16
- S4.7 Sensitivity of VGN to Hyperparameter k S17

S5 Risk-Based Interpretation **S19**

S6 Variance-Gated Behavior Across Axioms **S19**

S1 Symbols and Abbreviations

Table S1: Symbols and abbreviations used in the variance-gated normalization framework.

Symbol	Domain or Type	Description
C	$\mathbb{N}$	Number of classes.
M	$\mathbb{N}$	Number of ensemble members.
Δ^{C-1}	simplex	Probability simplex $\{\mathbf{z} \in \mathbb{R}^C \geq 0 : \mathbf{1}^\top \mathbf{z} = 1\}$.
$\mathbf{x}$	input	Input features.
$\mathbf{y}$	categorical	Class label over C classes.
$\mathbf{w}_m$	parameters	Parameters of ensemble member $m \in \{1, \ldots, M\}$.
$\mathbf{1}$	$\mathbb{R}^C$	All-ones vector.
$\mathbf{I}$	$\mathbb{R}^{C \times C}$	Identity matrix.
$\odot$	operator	Hadamard (elementwise) product.
$\mathrm{Diag}(\cdot)$	operator	Diagonal matrix with the given vector on the diagonal.
$\log(\cdot)$	function	Base-2 logarithm (applied elementwise to vectors).
$\mathbf{p}_m$	Δ^{C-1}	Member-m predictive distribution $p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}_m)$.
$p_m(c)$	$[0, 1]$	Probability for class c from member m.
$\bar{\mathbf{z}}$	Δ^{C-1}	Ensemble sample mean (per-class).
$\mathbf{S}$	$\mathbb{R}^C_{\geq 0}$	Per-class ensemble standard deviation (predictive spread).
ε	$\mathbb{R}_{>0}$	Small constant for numerical stability (*e.g.*, 1.0×10^{-8}).
$\mathbf{s}$	$\mathbb{R}^C_{>0}$	Numerically stabilized predictive spread.
k_c	$\mathbb{R}_{>0}$	Classwise sensitivity scalar for gating.
$\mathbf{k}$	$\mathbb{R}^C_{>0}$	Per-class sensitivity vector (learnable) for gating.
$\boldsymbol{\Gamma}$	$[0, 1)^C$	Variance gate with vector $\mathbf{k}$.
Z_m	$\mathbb{R}_{>0}$	Normalization constant for member m.
$\mathbf{q}_m$	Δ^{C-1}	Normalized variance-gated distribution for member m.
$\bar{\mathbf{q}}$	Δ^{C-1}	Variance-gated ensemble mixture (sample mean over members).
H	$[0, 1]$	Normalized Shannon entropy of $\mathbf{z} \in \Delta^{C-1}$.
TU	$[0, 1]$	Total (predictive) uncertainty.
AU	$[0, 1]$	Aleatoric uncertainty (expected per-member entropy).
EU	$[0, 1]$	Epistemic uncertainty (disagreement).
$\mathbf{m}_{ij}$	Δ^{C-1}	Midpoint distribution for members (i, j).
$\mathrm{D}_{\mathrm{KL}}(\mathbf{z}_i \| \mathbf{z}_j)$	$\mathbb{R}_{\geq 0}$	Kullback-Leibler divergence.
$\mathrm{D}_{\mathrm{JS}}(\mathbf{z}_i \| \mathbf{z}_j)$	$[0, 1]$	Jensen-Shannon divergence (normalized).
EPCE	$\mathbb{R}_{\geq 0}$	Expected pairwise cross-entropy.
EPKL	$\mathbb{R}_{\geq 0}$	Expected pairwise KL divergence (disagreement).
EPJS	$[0, 1]$	Expected pairwise Jensen-Shannon divergence (disagreement).
$\hat{y}$	class index or "uncertain"	Predicted label under the margin rule.
SNR	$\mathbb{R}$	Signal-to-noise margin between top-2 classes.
γ	$[0, 1)$	Top-2 variance gate with scalar k.
VGMU	$[0, 1]$	Variance-gated margin uncertainty (lower is more confident).
$\mathcal{L}$	scalar	Training objective (*e.g.*, NLL on mixture distribution $\bar{\mathbf{q}}$).
$\mathbf{u}$	$\mathbb{R}^C$	Upstream gradient at the mixture.
$\boldsymbol{\ell}$	$\mathbb{R}^C$	Unconstrained parameter with $\mathbf{k} = \mathrm{softplus}(\boldsymbol{\ell})$.
$\mathrm{softplus}(\boldsymbol{\ell})$	function	Positive reparameterization for $\mathbf{k}$.
$\sigma(\boldsymbol{\ell})$	function	Logistic function; appears in $\partial\mathcal{L}/\partial\boldsymbol{\ell} = \big(\partial\mathcal{L}/\partial\mathbf{k}\big) \odot \sigma(\boldsymbol{\ell})$.
b	index	Batch/sample index in sums (*e.g.*, $\sum_b$).

S2 Analytical Derivations and Gradient Expressions

This section provides complete analytical derivations of all gradients required for end-to-end training using variance-gated normalization. The results in this section support the summary expressions presented in Section 3 of the main paper. In Table S2, we provide a concise reference for the derived gradients.

Table S2: Summary of analytical gradients for the variance-gated normalization framework. Each gradient describes the partial derivative of the loss $\mathcal{L}$, the gating function $\mathbf{\Gamma}$, or intermediate statistics ($\bar{\mathbf{p}}$, $\mathbf{s}$) and learnable parameter $\mathbf{k}$ (*via* $\boldsymbol{\ell}$) with respect to the quantities that affect variance-aware ensemble training.

Gradient	**Description**
$\dfrac{\partial \bar{\mathbf{q}}}{\partial \mathbf{q}_m} = \dfrac{1}{M}\,\mathbf{I}$	Mixture sensitivity to member distribution.
$\dfrac{\partial \mathcal{L}}{\partial \mathbf{q}_m} = \dfrac{1}{M}\,\mathbf{u}, \;\; \mathbf{u} = \dfrac{\partial \mathcal{L}}{\partial \bar{\mathbf{q}}}$	Upstream gradient distributed equally to members.
$\dfrac{\partial \mathbf{q}_m}{\partial \mathbf{\Gamma}} = \dfrac{1}{Z_m}\left[\mathrm{Diag}(\mathbf{p}_m) - \mathbf{q}_m \mathbf{p}_m^\top\right]$	Jacobian $\mathbf{J}_m$ of normalized gating with respect to the gate $\mathbf{\Gamma}$.
$\dfrac{\partial \mathcal{L}_m}{\partial \mathbf{\Gamma}} = \dfrac{1}{M Z_m}\left[\mathbf{p}_m \odot \mathbf{u} - \mathbf{p}_m(\mathbf{q}_m^\top \mathbf{u})\right]$	Reverse-mode VJP to the gate (per member).
$\dfrac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} = \dfrac{1}{M}\sum_{m=1}^{M} \dfrac{1}{Z_m}\left[\mathbf{p}_m \odot \mathbf{u} - \mathbf{p}_m(\mathbf{q}_m^\top \mathbf{u})\right]$	Total gradient to shared gate across members.
$\dfrac{\partial \mathbf{\Gamma}}{\partial \bar{\mathbf{p}}} = \dfrac{1-\mathbf{\Gamma}}{\mathbf{k}\mathbf{s}}$	Gate sensitivity to mean confidence.
$\dfrac{\partial \mathbf{\Gamma}}{\partial \mathbf{s}} = -\dfrac{(1-\mathbf{\Gamma})\,\bar{\mathbf{p}}}{\mathbf{k}\mathbf{s}^2}$	Gate sensitivity to predictive spread.
$\dfrac{\partial \mathbf{\Gamma}}{\partial \mathbf{k}} = -\dfrac{(1-\mathbf{\Gamma})\,\bar{\mathbf{p}}}{\mathbf{k}^2\mathbf{s}}$	Gate sensitivity scale $\mathbf{k}$.
$\dfrac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}} = \dfrac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \odot \dfrac{1-\mathbf{\Gamma}}{\mathbf{k}\mathbf{s}}$	Backpropagation through gate *via* mean.
$\dfrac{\partial \mathcal{L}}{\partial \mathbf{s}} = -\dfrac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \odot \dfrac{(1-\mathbf{\Gamma})\,\bar{\mathbf{p}}}{\mathbf{k}\mathbf{s}^2}$	Backpropagation through gate *via* spread.
$\dfrac{\partial \mathcal{L}}{\partial \mathbf{k}} = -\sum_b \left(\dfrac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \odot \dfrac{(1-\mathbf{\Gamma})\,\bar{\mathbf{p}}}{\mathbf{k}^2\mathbf{s}}\right)_b$	Backpropagation to $\mathbf{k}$ (sum over classes b).
$\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\ell}} = \dfrac{\partial \mathcal{L}}{\partial \mathbf{k}} \odot \sigma(\boldsymbol{\ell}), \;\; \mathbf{k} = \mathrm{softplus}(\boldsymbol{\ell})$	Through softplus reparameterization of $\mathbf{k}$.
Total per-member gradient contributions:	
$\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m} = \dfrac{1}{M}\left(\left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\mathbf{\Gamma}} + \left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\bar{\mathbf{p}}} + \left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\mathbf{s}}\right)$	Sums direct and two indirect paths.
$\left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\mathbf{\Gamma}} = \dfrac{1}{M Z_m}\left[\mathbf{\Gamma} \odot \mathbf{u} - \mathbf{\Gamma}(\mathbf{q}_m^\top \mathbf{u})\right]$	Direct (local) path through normalization.
$\left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\bar{\mathbf{p}}} = \dfrac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}} \odot \dfrac{1}{M}$	Indirect path *via* ensemble mean.
$\left.\dfrac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right\vert_{\mathbf{s}} = \dfrac{\mathbf{p}_m - \bar{\mathbf{p}}}{M\,\mathbf{s}}$	Indirect path *via* spread.

S2.1 Gradients through Ensemble Mean

This subsection derives gradients that propagate through the ensemble sample mean, which couples the loss to all ensemble members through the shared mixture distribution.

Proposition S2.1 (Through the mean)**.** *Given* $\bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_m$*, each member receives*

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\Big|_{\bar{\mathbf{p}}} = \frac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}}\frac{\partial \bar{\mathbf{p}}}{\partial \mathbf{p}_m} = \frac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}}\frac{1}{M} \tag{S1}$$

Proof. Differentiating $\bar{\mathbf{p}}$ gives $d\bar{\mathbf{p}} = \frac{1}{M}d\mathbf{p}_m$; thus the Jacobian $\partial\bar{\mathbf{p}}/\partial\mathbf{p}_m = \frac{1}{M}\mathbf{I}$. □

Remark S2.1.1. *Every ensemble member contributes equally to the gradient path through the mean.*

S2.2 Gradients through Ensemble Variance

This subsection derives gradients that propagate through the ensemble predictive spread, enabling variance-aware updates that emphasize ensemble members deviating from the mean.

Proposition S2.2 (Through the variance)**.** *Let* $\mathbf{S} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{p}_m - \bar{\mathbf{p}})^2}$ *and* $\mathbf{s} = \mathbf{S} + \varepsilon$ *where* $\varepsilon = 1.0 \times 10^{-8}$ *(for numerical stability), then*

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\Big|_{\mathbf{S}} = \frac{\mathbf{p}_m - \bar{\mathbf{p}}}{M\,\mathbf{S}} \implies \frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\Big|_{\mathbf{s}} \approx \frac{\mathbf{p}_m - \bar{\mathbf{p}}}{M\,\mathbf{s}}. \tag{S2}$$

Proof. Define variance as $\mathbf{v} = \frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{p}_m - \bar{\mathbf{p}})^2$ such that $\mathbf{S} = \sqrt{\mathbf{v}}$. Hence,

$$\frac{\partial \mathbf{S}}{\partial \mathbf{p}_m} = \frac{1}{2\sqrt{\mathbf{v}}}\frac{\partial \mathbf{v}}{\partial \mathbf{p}_m} = \frac{1}{2\mathbf{S}}\frac{\partial \mathbf{v}}{\partial \mathbf{p}_m}. \tag{S3}$$

Using the identity $\mathbf{v} = \mathbb{E}[\mathbf{p}_m^2] - \bar{\mathbf{p}}^2$, we have

$$\frac{\partial \mathbf{v}}{\partial \mathbf{p}_m} = \frac{1}{M}2\mathbf{p}_m - 2\bar{\mathbf{p}}\frac{\partial \bar{\mathbf{p}}}{\partial \mathbf{p}_m}. \tag{S4}$$

Recall $\partial\bar{\mathbf{p}}/\partial\mathbf{p}_m = \frac{1}{M}\mathbf{I}$ therefore,

$$\frac{\partial \mathbf{v}}{\partial \mathbf{p}_m} = \frac{2}{M}(\mathbf{p}_m - \bar{\mathbf{p}}) \implies \frac{\partial \mathbf{S}}{\partial \mathbf{p}_m} = \frac{1}{2\mathbf{S}}\frac{2}{M}(\mathbf{p}_m - \bar{\mathbf{p}}) = \frac{\mathbf{p}_m - \bar{\mathbf{p}}}{M\mathbf{S}} \approx \frac{\mathbf{p}_m - \bar{\mathbf{p}}}{M\mathbf{s}} \tag{S5}$$

□

Remark S2.2.1. *Members farther from the ensemble mean receive proportionally larger gradients, enabling variance awareness in the backward flow of gradients.*

S2.3 Jacobians and Vector–Jacobian Products for Shared Gates

This subsection derives the local Jacobians and corresponding vector-Jacobian products for the variance-gated normalization layer when the gate is shared across ensemble members.

Proposition S2.3. *Let the normalized gated member probabilities be* $\mathbf{q}_m = (\mathbf{p}_m \odot \mathbf{\Gamma})/Z_m$*,* $Z_m = \mathbf{p}_m^\top\mathbf{\Gamma}$*. Then the Jacobian* $\mathbf{J}_m = \partial\mathbf{q}_m/\partial\mathbf{\Gamma} = [\mathrm{Diag}(\mathbf{p}_m) - \mathbf{q}_m\mathbf{p}_m^\top]\,/\,Z_m$*.*

Proof. With $\mathbf{p}_m$ fixed $\partial(\mathbf{p}_m \odot \mathbf{\Gamma}) = \mathbf{p}_m \odot \partial\mathbf{\Gamma}$ and $\partial(\mathbf{p}_m^\top\mathbf{\Gamma}) = \mathbf{p}_m^\top\partial\mathbf{\Gamma}$, we differentiate with respect to $\mathbf{\Gamma}$,

$$d\mathbf{q}_m = \frac{Z_m\partial(\mathbf{p}_m \odot \mathbf{\Gamma}) - (\mathbf{p}_m \odot \mathbf{\Gamma})\partial Z}{Z_m^2} = \frac{Z_m(\mathbf{p}_m \odot \partial\mathbf{\Gamma}) - (\mathbf{p}_m \odot \mathbf{\Gamma})\mathbf{p}_m^\top\partial\mathbf{\Gamma}}{Z_m^2} = \frac{\mathbf{p}_m \odot \partial\mathbf{\Gamma}}{Z_m} - \frac{\mathbf{p}_m \odot \mathbf{\Gamma}}{Z_m^2} \odot \mathbf{p}_m^\top\partial\mathbf{\Gamma} \tag{S6}$$

Substitute $\mathbf{q}_m = (\mathbf{p}_m \odot \mathbf{\Gamma})/\mathbf{p}_m^\top \mathbf{\Gamma}$ and $\mathrm{Diag}(\mathbf{p}_m) \odot \partial \mathbf{\Gamma} = \mathbf{p}_m \partial \odot \mathbf{\Gamma}$,

$$\partial \mathbf{q}_m = \frac{1}{Z_m}\Big[\mathbf{p}_m \odot \partial \mathbf{\Gamma} - \mathbf{q}_m(\mathbf{p}_m^\top \partial \mathbf{\Gamma})\Big] \implies \mathbf{J}_m = \frac{\partial \mathbf{q}_m}{\partial \mathbf{\Gamma}} = \frac{1}{Z_m}\Big[\mathrm{Diag}(\mathbf{p}_m) - \mathbf{q}_m \mathbf{p}_m^\top\Big] \in \mathbb{R}^{C\times C} \tag{S7}$$

□

Corollary S2.3.1 (Jacobian-vector product, forward-mode)**.** *For any direction vector $d\mathbf{\Gamma} \in \mathbb{R}^C$,*

$$d\mathbf{q}_m = \left(\frac{\partial \mathbf{q}_m}{\partial \mathbf{\Gamma}}\right) d\mathbf{\Gamma} = \frac{1}{Z_m}\Big[\mathbf{p}_m \odot d\mathbf{\Gamma} - \mathbf{q}_m(\mathbf{p}_m^\top d\mathbf{\Gamma})\Big]. \tag{S8}$$

This calculates the directional derivative $d\mathbf{q}_m$ for $d\mathbf{\Gamma}$ in $O(C)$ time.

Corollary S2.3.2 (Vector–Jacobian product, reverse mode)**.** *Since the loss is evaluated on the ensemble mixture $\bar{\mathbf{q}}$, each ensemble member receives a scaled upstream gradient $\frac{1}{M}\mathbf{u}$, where $\mathbf{u} = \partial\mathcal{L}/\partial\bar{\mathbf{q}}$. The reverse-mode gradient of $\mathcal{L}_m$ with respect to the gate is therefore*

$$\frac{\partial \mathcal{L}_m}{\partial \mathbf{\Gamma}} = \frac{1}{M}\left(\frac{\partial \mathbf{q}_m}{\partial \mathbf{\Gamma}}\right)^\top \mathbf{u} = \frac{1}{MZ_m}\Big[\mathbf{p}_m \odot \mathbf{u} - \mathbf{p}_m(\mathbf{q}_m^\top \mathbf{u})\Big]. \tag{S9}$$

This calculates the per-member reverse-mode gradient $\partial\mathcal{L}_m/\partial\mathbf{\Gamma}$ in $O(C)$ time through the variance-gated normalization layer.

Remark S2.3.1. *The upstream gradient $\mathbf{u}$ is identical for all ensemble members. Each ensemble member receives a scaled contribution $\frac{1}{M}\mathbf{u}$ when gradients are backpropagated.*

Proposition S2.4. *Let the normalized gated member probabilities be $\mathbf{q}_m = (\mathbf{p}_m \odot \mathbf{\Gamma})/Z_m$, $Z_m = \mathbf{p}_m^\top \mathbf{\Gamma}$. Then the loss with respect to member m probabilities is*

$$\left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\mathbf{\Gamma}} = \frac{1}{MZ_m}\Big[\mathbf{\Gamma} \odot \mathbf{u} - \mathbf{\Gamma}\,(\mathbf{q}_m^\top \mathbf{u})\Big]. \tag{S10}$$

Proof. We define the following, apply the quotient rule on $\mathbf{q}_m$, and substitute values

$$\mathbf{p}_m \odot \mathbf{\Gamma} = Z_m \mathbf{q}_m, \qquad d(\mathbf{p}_m \odot \mathbf{\Gamma}) = \mathrm{Diag}(\mathbf{\Gamma})\, d\mathbf{p}_m, \qquad dZ_m = \mathbf{\Gamma}^\top d\mathbf{p}_m, \qquad \mathbf{\Gamma} \odot \mathbf{u} = \mathrm{Diag}(\mathbf{\Gamma})\,\mathbf{u} \tag{S11}$$

$$d\mathbf{q}_m = \frac{Z_m\, d(\mathbf{p}_m \odot \mathbf{\Gamma}) - (\mathbf{p}_m \odot \mathbf{\Gamma})\, dZ_m}{Z_m^2} \implies d\mathbf{q}_m = \frac{1}{Z_m}\Big[\mathrm{Diag}(\mathbf{\Gamma})\, d\mathbf{p}_m - \mathbf{q}_m\,(\mathbf{\Gamma}^\top d\mathbf{p}_m)\Big]. \tag{S12}$$

Recall $\partial\mathcal{L}/\partial\mathbf{q}_m = \frac{1}{M}\mathbf{u}$, therefore

$$d\mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{q}_m}\right)^\top d\mathbf{q}_m = \left(\frac{1}{M}\mathbf{u}\right)^\top d\mathbf{q}_m \implies d\mathcal{L} = \frac{1}{MZ_m}\,\mathbf{u}^\top\Big[\mathrm{Diag}(\mathbf{\Gamma}) - \mathbf{q}_m \mathbf{\Gamma}^\top\Big] d\mathbf{p}_m, \tag{S13}$$

$$\left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\mathbf{\Gamma}} = \frac{1}{MZ_m}\Big[\mathrm{Diag}(\mathbf{\Gamma}) - \mathbf{\Gamma}\mathbf{q}_m^\top\Big]\mathbf{u} \implies \left.\frac{\partial \mathcal{L}}{\partial \mathbf{p}_m}\right|_{\mathbf{\Gamma}} = \frac{1}{MZ_m}\Big[\mathbf{\Gamma} \odot \mathbf{u} - \mathbf{\Gamma}(\mathbf{q}_m^\top \mathbf{u})\Big]. \tag{S14}$$

□

Remark S2.4.1. *There is no sum over members in $\partial\mathcal{L}/\partial\mathbf{p}_m|_{\mathbf{\Gamma}}$. Each member contributes independently through its own normalization Z_m.*

S2.4 General Reverse-Mode Differentiation for Shared Ensemble Gates

This subsection combines the individual gradient paths into a reverse-mode differentiation rule for ensemble layers with shared gating mechanisms.

Proposition S2.5. *When the ensemble shares a common gate $\mathbf{\Gamma}$, the total gradient accumulates over all members. Let $\mathbf{J}_m = \partial \mathbf{q}_m / \partial \mathbf{\Gamma}$ and $\mathbf{u} = \partial \mathcal{L} / \partial \bar{\mathbf{q}}$, where $\bar{\mathbf{q}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{q}_m$ is the ensemble mixture distribution. Then the total gradient of the loss with respect to $\mathbf{\Gamma}$ is*

$$\frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{J}_m^\top \mathbf{u} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{Z_m} \Big[\mathbf{p}_m \odot \mathbf{u} - \mathbf{p}_m (\mathbf{q}_m^\top \mathbf{u}) \Big], \qquad Z_m = \mathbf{p}_m^\top \mathbf{\Gamma}. \tag{S15}$$

Proof. For a single member $\mathbf{q}_m = (\mathbf{p}_m \odot \mathbf{\Gamma})/Z_m$, the differential is

$$d\mathbf{q}_m = \frac{1}{Z_m} \Big[\mathbf{p}_m \odot d\mathbf{\Gamma} - \mathbf{q}_m (\mathbf{p}_m^\top d\mathbf{\Gamma}) \Big]. \tag{S16}$$

With the upstream gradient $\mathbf{u}$, the change in the loss satisfies

$$d\mathcal{L}_m = \frac{1}{M} \mathbf{u}^\top d\mathbf{q}_m = \frac{1}{M Z_m} \Big[\mathbf{u}^\top (\mathbf{p}_m \odot d\mathbf{\Gamma}) - \mathbf{u}^\top \left[\mathbf{q}_m (\mathbf{p}_m^\top d\mathbf{\Gamma}) \right] \Big] \tag{S17}$$

Since $\mathbf{p}_m \odot d\mathbf{\Gamma} = \operatorname{Diag}(\mathbf{p}_m)\, d\mathbf{\Gamma}$, $\mathbf{u}^\top (\mathbf{p}_m \odot d\mathbf{\Gamma}) = \mathbf{u}^\top \operatorname{Diag}(\mathbf{p}_m)\, d\mathbf{\Gamma}$. Using the product transpose of matrices property $a^\top B x = (B^\top a)^\top x$, we have $\mathbf{u}^\top \operatorname{Diag}(\mathbf{p}_m)\, d\mathbf{\Gamma} = \big(\operatorname{Diag}(\mathbf{p}_m)^\top \mathbf{u} \big)^\top d\mathbf{\Gamma}$. Given a diagonal matrix is symmetric, $\operatorname{Diag}(\mathbf{p}_m)^\top = \operatorname{Diag}(\mathbf{p}_m) \implies \operatorname{Diag}(\mathbf{p}_m)\mathbf{u} = \mathbf{p}_m \odot \mathbf{u}$. Therefore,

$$d\mathcal{L}_m = \frac{1}{M} \mathbf{u}^\top d\mathbf{q}_m = \frac{1}{M Z_m} \Big[(\mathbf{p}_m \odot \mathbf{u})^\top d\mathbf{\Gamma}) - (\mathbf{q}_m^\top \mathbf{u}) \mathbf{p}_m^\top d\mathbf{\Gamma} \Big]. \tag{S18}$$

By identification with the total differential $d\mathcal{L} = (\partial \mathcal{L} / \partial \mathbf{\Gamma})^\top d\mathbf{\Gamma}$, we obtain

$$\frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{Z_m} \Big[\mathbf{p}_m \odot \mathbf{u} - \mathbf{p}_m (\mathbf{q}_m^\top \mathbf{u}) \Big], \tag{S19}$$

□

Remark S2.5.1. *Although the proof begins from the forward differential $d\mathbf{q}_m = (\partial \mathbf{q}_m / \partial \mathbf{\Gamma})\, d\mathbf{\Gamma}$, the final expression corresponds to the reverse-mode vector–Jacobian product used in gradient backpropagation, where $\mathbf{u} = \partial \mathcal{L} / \partial \bar{\mathbf{q}}$.*

Remark S2.5.2. *Since the gate $\mathbf{\Gamma}$ is shared across all ensemble members, the total gradient $\partial \mathcal{L} / \partial \mathbf{\Gamma}$ accumulates contributions from each member distribution $\mathbf{q}_m$. The factor $\frac{1}{M}$ arises from differentiating the loss with respect to the mixture distribution $\bar{\mathbf{q}} = \frac{1}{M} \sum_m \mathbf{q}_m$.*

S2.5 Gradients with Respect to Learnable Sensitivity Parameters

This subsection derives gradients with respect to the learnable sensitivity parameters modulating the strength of variance-gating, including the softplus reparameterization used to enforce positivity.

Proposition S2.6. *For the variance-gating function $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$, the gating scalar $\mathbf{k} > 0$ is learned via a softplus reparameterization $\mathbf{k} = \operatorname{softplus}(\boldsymbol{\ell})$, where $\boldsymbol{\ell} \in \mathbb{R}^C$ are unconstrained learnable parameters. The gating function $\mathbf{\Gamma} \in \mathbb{R}^{B \times C}$ is defined per sample and class, whereas $\mathbf{k}, \mathbf{s} \in \mathbb{R}^C$ are classwise parameters shared across the batch. Then the gradients of the loss are*

$$\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}} = \frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \frac{1 - \mathbf{\Gamma}}{\mathbf{ks}}, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{s}} = -\frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \frac{(1 - \mathbf{\Gamma}) \bar{\mathbf{p}}}{\mathbf{ks}^2}, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{k}} = -\sum_b \left(\frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}} \frac{(1 - \mathbf{\Gamma}) \bar{\mathbf{p}}}{\mathbf{k}^2 \mathbf{s}} \right)_b. \tag{S20}$$

Proof. Recall the partial derivatives

$$\frac{\partial \mathbf{\Gamma}}{\partial \bar{\mathbf{p}}} = \frac{1-\mathbf{\Gamma}}{\mathbf{ks}}, \qquad \frac{\partial \mathbf{\Gamma}}{\partial \mathbf{s}} = -\frac{(1-\mathbf{\Gamma})\bar{\mathbf{p}}}{\mathbf{ks}^2}, \qquad \frac{\partial \mathbf{\Gamma}}{\partial \mathbf{k}} = -\frac{(1-\mathbf{\Gamma})\bar{\mathbf{p}}}{\mathbf{k}^2\mathbf{s}}. \tag{S21}$$

Hence, after applying the chain rule, we obtain

$$\frac{\partial \mathcal{L}}{\partial \bar{\mathbf{p}}} = \frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}}\frac{\partial \mathbf{\Gamma}}{\partial \bar{\mathbf{p}}}, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{s}} = \frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}}\frac{\partial \mathbf{\Gamma}}{\partial \mathbf{s}}, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{k}} = \sum_b \left(\frac{\partial \mathcal{L}}{\partial \mathbf{\Gamma}}\frac{\partial \mathbf{\Gamma}}{\partial \mathbf{k}}\right)_b, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\ell}} = \frac{\partial \mathcal{L}}{\partial \mathbf{k}}\frac{\partial \mathbf{k}}{\partial \boldsymbol{\ell}} = \frac{\partial \mathcal{L}}{\partial \mathbf{k}}\sigma(\boldsymbol{\ell}), \tag{S22}$$

where $\partial\mathcal{L}/\partial\boldsymbol{\ell}$ is the gradient propagated through softplus reparameterization, which enforces positivity of $\mathbf{k}$ and modulates the gradient magnitude by the logistic factor $\sigma(\boldsymbol{\ell}) = 1/(1+e^{-\boldsymbol{\ell}})$. □

Remark S2.6.1. *The partial derivatives reveal a complementary effect of the gating parameters. Increasing* $\bar{\mathbf{p}}$ *results with a corresponding increase in* $\mathbf{\Gamma}$*, reinforcing confident predictions, whereas larger* $\mathbf{s}$ *or* $\mathbf{k}$ *reduce* $\mathbf{\Gamma}$*, suppressing highly variable predictions. Through this mechanism, the variance-gating function adaptively balances aleatoric and epistemic contributions.*

S3 Implementation Details

Reporting and Interpretation All results are presented as mean ± standard deviation over three independent training runs with different random seeds. These statistics are intended to reflect training variability rather than to establish formal statistical significance. We interpret differences conservatively, focusing on (i) consistency of trends across datasets and ensembles, and (ii) differences whose magnitude exceeds the typical run-to-run variability observed across random seeds.

S3.1 Model Architectures

MNIST For the MNIST dataset, we used a LeNet-5 style convolutional neural network (CNN) as the base architecture for DE, MCD, LLE, and MCD–LLE models. The network consists of a shared feature extractor with two convolutional blocks using 5×5 kernels, mapping from 1 to 6 channels in the first block and from 6 to 16 channels in the second. Each block applies a Tanh activation, followed by 2×2 average pooling and spatial dropout ($p = 0.1$). The convolutional blocks ends with an adaptive average pooling layer, producing a fixed 16-dimensional representation that is subsequently flattened. This representation was passed to a multilayer perceptron (MLP) consisting of a linear layer (16 to 120 units), a Tanh activation, and a dropout ($p = 0.1$). Classification is performed by a task-specific head implemented as a two-layer MLP with a $120 \to 84 \to 10$ structure, using Tanh activations and dropout between layers. For last-layer ensemble (LLE) variants, the classifier head H is replaced by a list of 100 independent classifier heads, while all preceding layers are shared.

SVHN, CIFAR-10, and CIFAR-100 For SVHN, CIFAR-10, and CIFAR-100, we employ a ResNet-18 and WideResNet-28-10 (Zagoruyko & Komodakis, 2017) as the base model for DE, MCD, LLE, and MCD-LLE configurations. The network follows the standard BasicBlock for both the WideResNet networks, with spatial dropout applied within convolutional blocks. The networks applies batch normalization, ReLU, and global average pooling. The pooled representation is passed through an additional fully connected block consisting of a linear layer with batch normalization, ReLU activation, and dropout, prior to classification. Dropout is applied with probability $p = 0.1$ for SVHN and $p = 0.3$ for CIFAR-10/100. For last-layer ensemble variants, the classifier module is instantiated as a list of independent classifier heads H, while all preceding layers are shared, analogous to the MNIST setup.

S3.2 Training Protocols

Models were trained using the Adam optimizer with learning rate 1.0×10^{-3} for SVHN, CIFAR-10/100 and 1.0×10^{-4} for MNIST with a batch size 128 and the cross-entropy loss objective. For multi-head architectures, losses were averaged across heads prior to backpropagation. Training was done with early stopping based on the validation loss using a minimum delta of 1.0×10^{-4} and a patience of 5 epochs. Each dataset was normalized using its respective mean and standard deviation prior to training. Training was performed in triplicate using different random seeds under fully deterministic settings, including cuDNN and cuBLAS routines. Experiments were conducted with ensemble sizes of 5, 10, or 100.

S3.3 Evaluation and Uncertainty Metrics

We compare the proposed variance-gated ensemble framework against standard entropy-based uncertainty decompositions and recent information-theoretic approaches (Schweighofer et al., 2023). We report both variance-gated and non-gated variants of entropy- and divergence-based uncertainty measures, together with the proposed variance-gated margin uncertainty. We compare VGMU against: (i) standard entropy-based epistemic uncertainty computed as mutual information Houlsby et al. (2011); (ii) Expected Pairwise KL divergence Schweighofer et al. (2023); and (iii) Expected Pairwise Jensen-Shannon divergence (this work). For variance-gated variants, we report results using learned per-class $\mathbf{k}$.

We include a diversity measure $\mathbb{E}_{i,c}[\mathrm{Var}_M]$, defined as the ensemble variance averaged across samples and classes. In addition to uncertainty and diversity metrics, we evaluate predictive performance and calibration using accuracy, F1-score, and expected calibration error.

We assess rank consistency *via* Spearman's rank correlation ρ and Kendall's τ. Spearman's ρ measures whether samples ranked higher under one measure tend to also be ranked higher under another, capturing overall agreement between rankings. Kendall's τ quantifies pairwise ordering agreement by measuring the fraction of sample pairs whose relative ordering is preserved between two rankings, making it sensitive to local rank inversions. Together, these metrics provide complementary views of global and fine-grained ranking stability across uncertainty scores.

In addition to rank correlation, we report the cumulative area under the curve (AUCc), which summarizes how rapidly scores accumulate when samples are ordered from highest to lowest. AUCc is high when a small number of top-ranked samples account for a large fraction of the total score mass, and low when scores are more evenly distributed. This provides a complementary perspective by characterizing the concentration and sharpness of uncertainty estimates beyond ranking agreement alone.

Monte Carlo Dropout (i.e, MCD and MCD-LLE) inference sampling (S) was done using a dropout rate of $p = 0.1$ with $H = 1$ and $S \in \{5, 10, 100\}$ or $H \in \{1, 10\}$ and $S = 10$. For $H > 1$, feature representations were re-used to improve compute efficiency.

S3.4 Variance-Gated Normalization Implementation

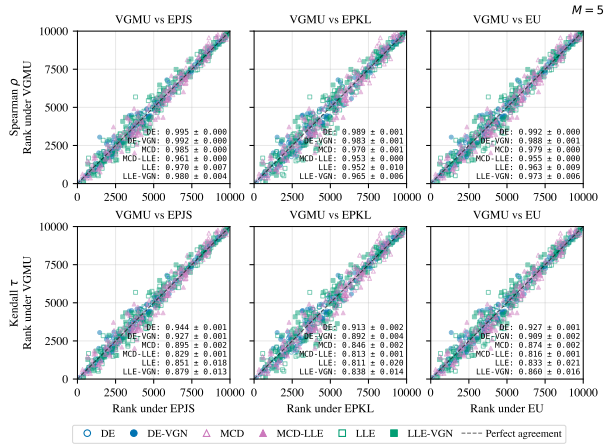
The VGN module is applied on top of ensemble predictions during training. Given an ensemble of M per-member probability distributions $\mathbf{P} \in \mathbb{R}^{B \times M \times C}$, the VGN computes outputs $\mathbf{Q}$ *via* the variance-based gating mechanism. The gated output is computed as $\mathbf{Q} = (\mathbf{P} \odot \mathbf{\Gamma})/\mathbf{Z}$, where $\mathbf{Z}$ normalizes each member's distribution. The final prediction is obtained by averaging the gated distributions: $\bar{\mathbf{Q}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{Q}_m$. Gradients flow through the VGN *via* a custom backward pass that accounts for contributions through the normalization constant $\mathbf{Z}$, the mean $\bar{\mathbf{P}}$, and variance $\mathbf{s}$ pathways. The per-class gate parameters $\mathbf{k}$ are re-parameterized as $\mathbf{k} = \text{softplus}(\boldsymbol{\ell}) + \epsilon$ where $\boldsymbol{\ell}$ is initialized to 0, resulting with $\mathbf{k} \approx 0.693$ at initialization. This ensures strictly positive gate parameters and non-saturating initialization. For sensitivity of VGN to fixed *vs.* learned hyperparameter k, see SI S4.7

S3.5 Numerical Stability

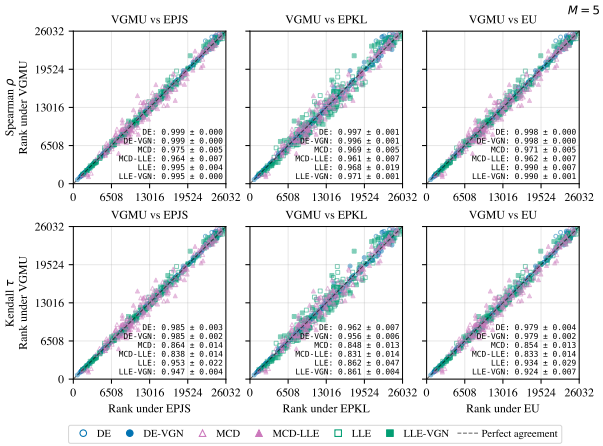
All operations susceptible to numerical instability include a small constant ε (*e.g.*, 1.0×10^{-8}) to prevent underflow, overflow, and division-by-zero errors. Ensemble standard deviations were stabilized as $\mathbf{s} = \mathbf{S} + \varepsilon$. The variance-gating function $\mathbf{\Gamma} = 1 - e^{-\bar{\mathbf{p}}/\mathbf{ks}}$ is clamped below by ε to ensure finite gradients. The per-class scaling parameter $\mathbf{k}$ is learned using a softplus reparameterization and clamped to a minimum value of 1.0×10^{-3} to prevent gate saturation (*i.e.*, excessively large signal-to-noise ratios) and to maintain stability as $\mathbf{k} \to 0$. Normalization constants $Z_m = \mathbf{p}_m^\top \mathbf{\Gamma}$ are likewise lower-bounded by ε; when this bound is active, gradients are passed through unchanged to avoid disrupting backpropagation.

S4 Additional Experimental Results

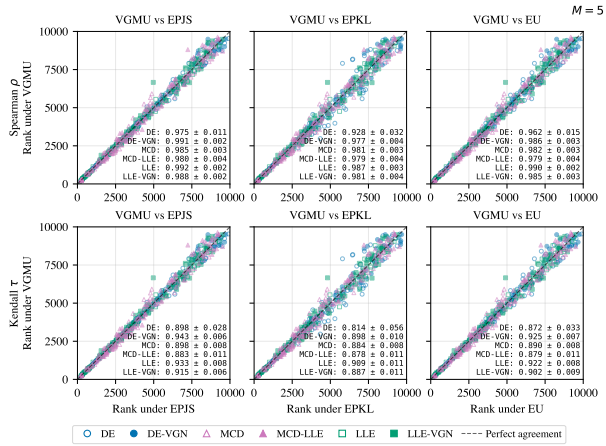
S4.1 Rank Consistency and Alignment



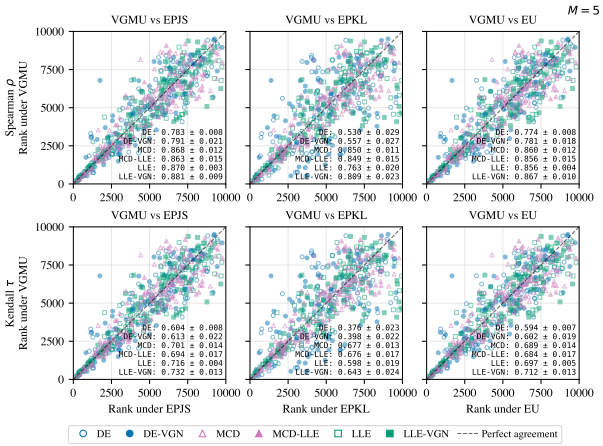
(a) MNIST



(b) SVHN



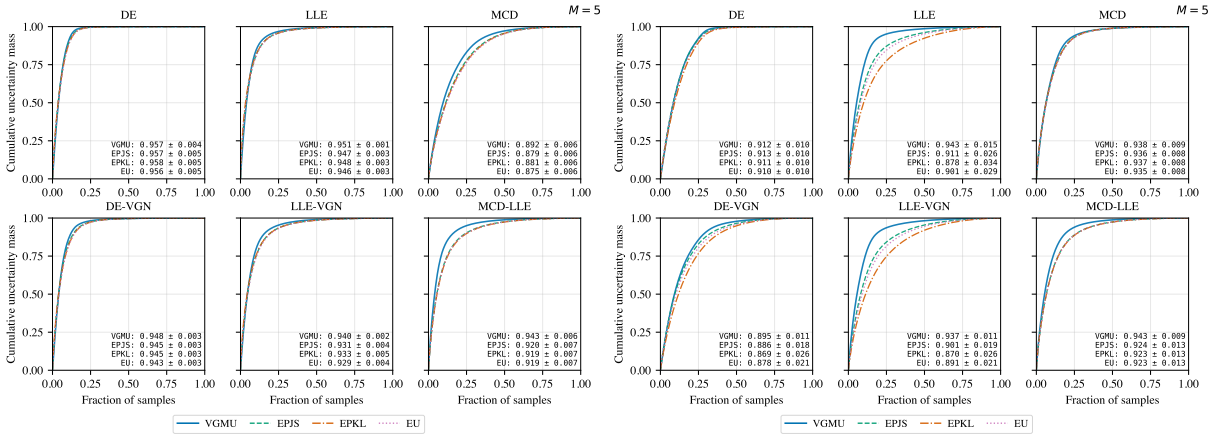
(c) CIFAR-10



(d) CIFAR-100

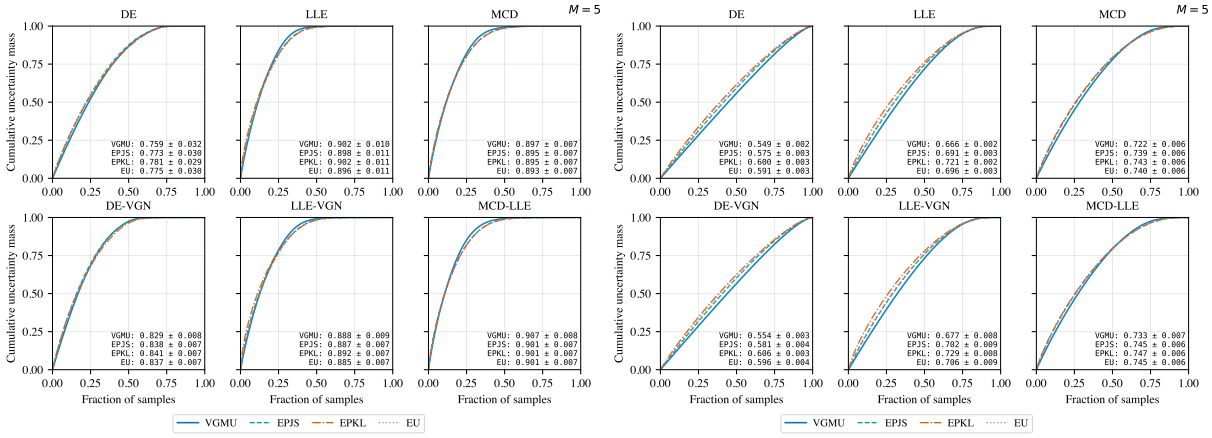
Figure S1: Spearman's ρ and Kendall's τ rank correlations between VGMU and entropy- and divergence-based epistemic uncertainty measures (EPJS, EPKL, and mutual information). Results are reported across datasets and ensemble configurations, illustrating the degree of alignment between margin-based and full-simplex uncertainty rankings.

S4.2 Uncertainty Mass Concentration



(a) MNIST

(b) SVHN

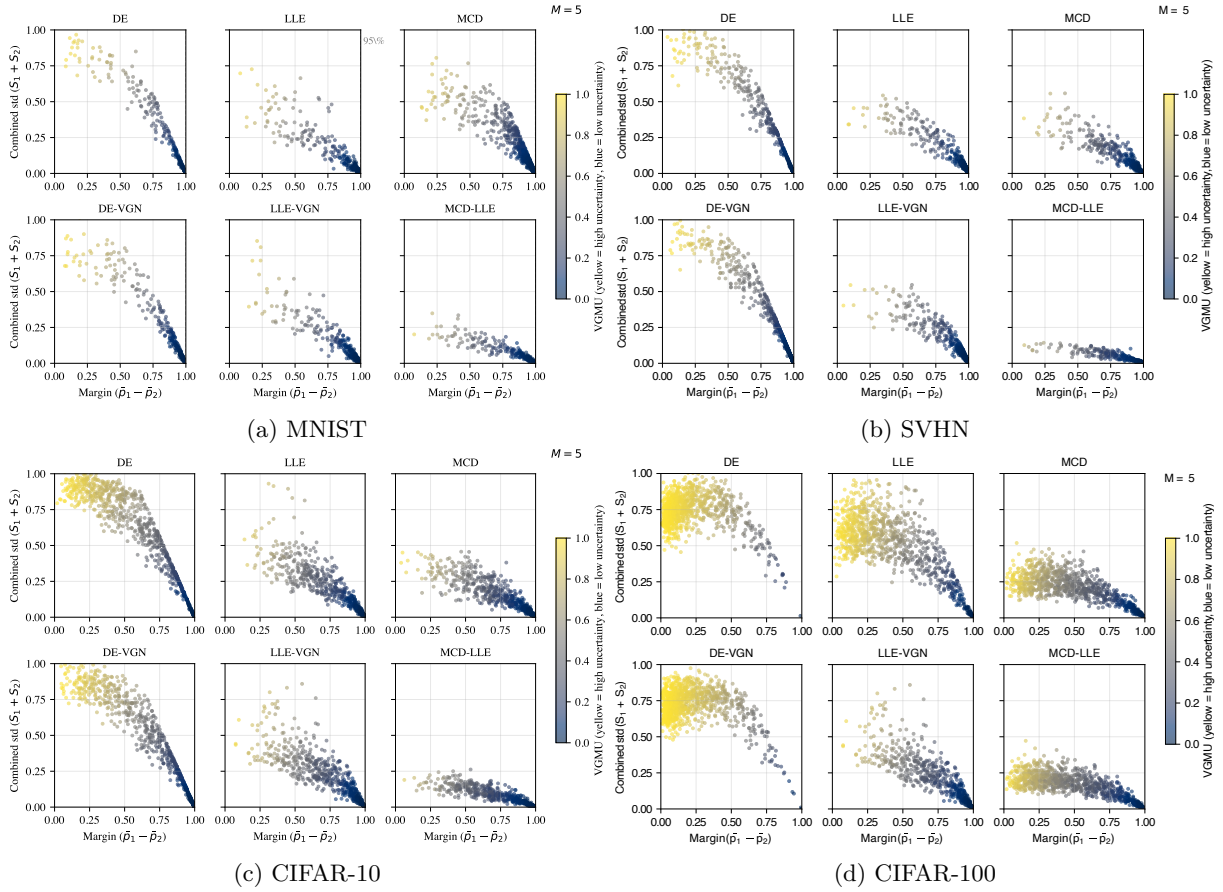


(c) CIFAR-10

(d) CIFAR-100

Figure S2: Cumulative uncertainty mass concentration curves and corresponding AUC_c values for VGMU and information-theoretic baselines. Samples are sorted by descending uncertainty, and higher AUC_c indicates stronger concentration of uncertainty on difficult samples. Results are shown across datasets and ensemble configurations.

S4.3 Margin–Variance Geometry



(a) MNIST (b) SVHN

(c) CIFAR-10 (d) CIFAR-100

Figure S3: Visualization of predictive margin $(\bar{p}_1 - \bar{p}_2)$ and combined predictive spread $(S_1 + S_2)$ for different ensemble methods. Each point corresponds to a test sample and is colored by its VGMU value, illustrating how margin-based uncertainty couples class separability with epistemic disagreement.

S4.4 Effects of Learned k

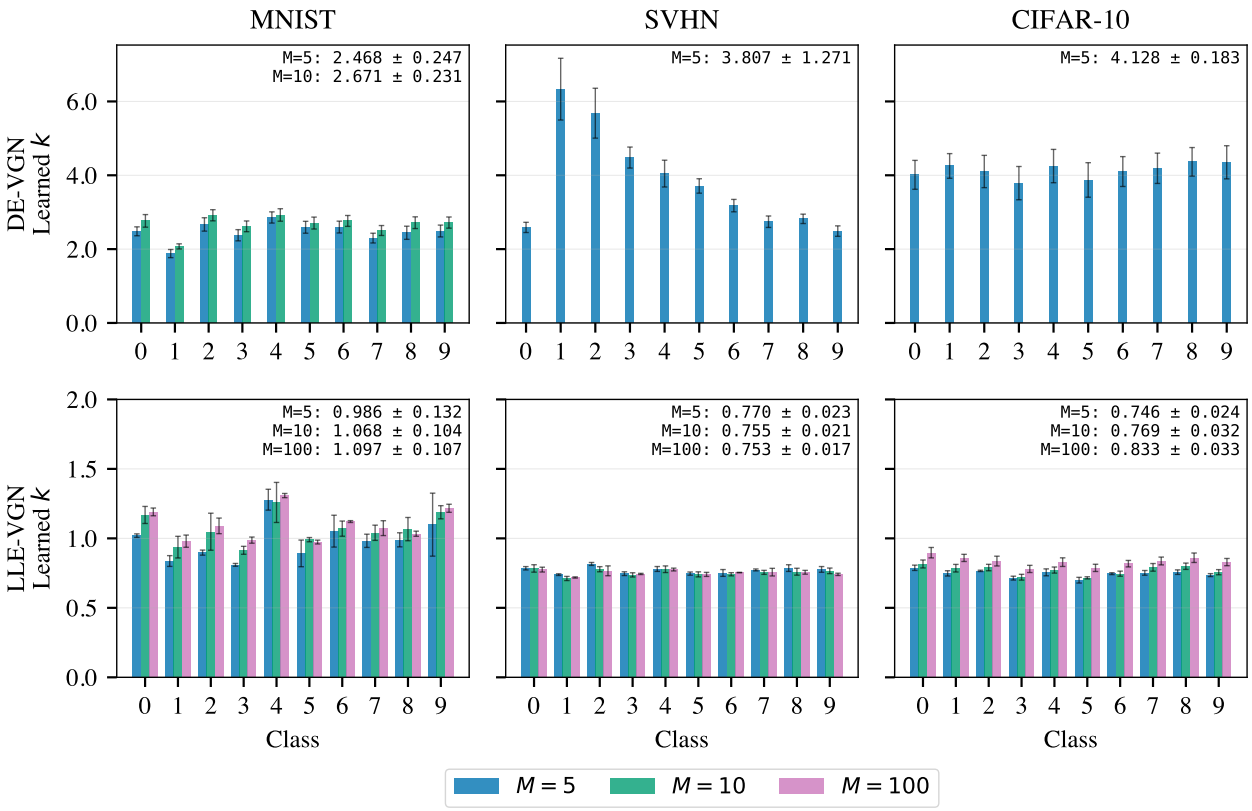


Figure S4: Learned **k**-values.

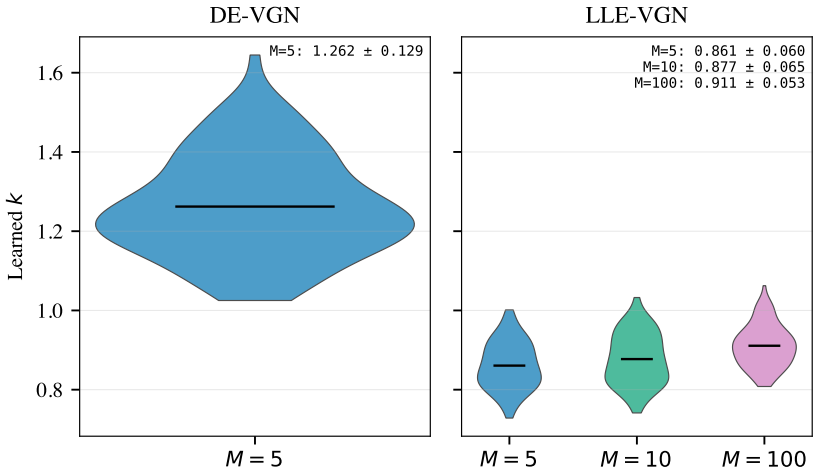


Figure S5: Learned **k**-values for CIFAR-100.

S4.5 Performance and Calibration

Table S3: Performance and calibration metrics for ensemble models and the MNIST dataset.

Network	Accuracy	F1-Score	ECE	Diversity
$M = 1$	0.973 ± 0.001	0.973 ± 0.001	0.004 ± 0.001	—
MCD; $M = 5$	0.971 ± 0.001	0.970 ± 0.001	0.040 ± 0.003	$2.7 \times 10^{-3} \pm 0.1$
MCD; $M = 10$	0.971 ± 0.001	0.971 ± 0.001	0.041 ± 0.005	$2.9 \times 10^{-3} \pm 0.1$
MCD; $M = 100$	0.974 ± 0.001	0.973 ± 0.001	0.041 ± 0.004	$3.0 \times 10^{-3} \pm 0.2$
MCD-LLE; $M = 5$	0.973 ± 0.001	0.973 ± 0.001	0.006 ± 0.002	$1.7 \times 10^{-4} \pm 0.0$
MCD-LLE; $M = 10$	0.973 ± 0.001	0.973 ± 0.001	0.006 ± 0.002	$1.9 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 100$	0.973 ± 0.001	0.973 ± 0.001	0.006 ± 0.002	$2.0 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 1000$	0.973 ± 0.001	0.973 ± 0.001	0.006 ± 0.002	$2.0 \times 10^{-4} \pm 0.1$
DE; $M = 5$	0.981 ± 0.001	0.981 ± 0.001	0.012 ± 0.002	$2.3 \times 10^{-3} \pm 0.2$
DE-VGN; $M = 5$	0.982 ± 0.001	0.982 ± 0.001	0.007 ± 0.002	$1.4 \times 10^{-3} \pm 0.1$
LLE; $M = 5$	0.971 ± 0.003	0.971 ± 0.003	0.005 ± 0.001	$6.0 \times 10^{-4} \pm 0.2$
LLE; $M = 10$	0.972 ± 0.001	0.971 ± 0.001	0.005 ± 0.001	$7.6 \times 10^{-4} \pm 0.9$
LLE; $M = 100$	0.970 ± 0.001	0.970 ± 0.001	0.005 ± 0.002	$1.0 \times 10^{-3} \pm 0.2$
LLE-VGN; $M = 5$	0.970 ± 0.002	0.970 ± 0.002	0.005 ± 0.001	$6.8 \times 10^{-4} \pm 0.4$
LLE-VGN; $M = 10$	0.971 ± 0.002	0.971 ± 0.002	0.005 ± 0.001	$8.9 \times 10^{-4} \pm 0.1$
LLE-VGN; $M = 100$	0.970 ± 0.003	0.969 ± 0.003	0.005 ± 0.001	$1.1 \times 10^{-3} \pm 0.1$

Table S4: Performance and calibration metrics for ensemble models and the SVHN dataset.

Network	Accuracy	F1-Score	ECE	Diversity
$M = 1$	0.948 ± 0.002	0.943 ± 0.003	0.017 ± 0.004	—
MCD; $M = 5$	0.946 ± 0.003	0.942 ± 0.003	0.010 ± 0.004	$7.3 \times 10^{-4} \pm 0.8$
MCD; $M = 10$	0.947 ± 0.003	0.942 ± 0.003	0.009 ± 0.004	$8.1 \times 10^{-4} \pm 1.0$
MCD; $M = 100$	0.947 ± 0.003	0.943 ± 0.003	0.008 ± 0.003	$9.0 \times 10^{-4} \pm 1.1$
MCD-LLE; $M = 5$	0.947 ± 0.003	0.943 ± 0.003	0.016 ± 0.004	$7.3 \times 10^{-5} \pm 1.6$
MCD-LLE; $M = 10$	0.947 ± 0.003	0.943 ± 0.003	0.016 ± 0.004	$8.2 \times 10^{-5} \pm 1.9$
MCD-LLE; $M = 100$	0.948 ± 0.003	0.943 ± 0.003	0.016 ± 0.004	$8.5 \times 10^{-5} \pm 1.9$
MCD-LLE; $M = 1000$	0.947 ± 0.003	0.943 ± 0.003	0.016 ± 0.004	$8.5 \times 10^{-5} \pm 2.0$
DE; $M = 5$	0.955 ± 0.002	0.952 ± 0.002	0.026 ± 0.006	$5.8 \times 10^{-3} \pm 0.6$
DE-VGN; $M = 5$	0.960 ± 0.000	0.957 ± 0.000	0.016 ± 0.003	$4.0 \times 10^{-3} \pm 0.4$
LLE; $M = 5$	0.950 ± 0.002	0.946 ± 0.002	0.013 ± 0.006	$9.7 \times 10^{-4} \pm 1.0$
LLE; $M = 10$	0.949 ± 0.003	0.945 ± 0.003	0.017 ± 0.003	$1.3 \times 10^{-3} \pm 0.1$
LLE; $M = 100$	0.951 ± 0.004	0.947 ± 0.004	0.012 ± 0.002	$3.2 \times 10^{-3} \pm 0.3$
LLE-VGN; $M = 5$	0.953 ± 0.005	0.949 ± 0.006	0.014 ± 0.002	$8.1 \times 10^{-4} \pm 0.2$
LLE-VGN; $M = 10$	0.952 ± 0.002	0.949 ± 0.002	0.012 ± 0.001	$1.1 \times 10^{-3} \pm 0.0$
LLE-VGN; $M = 100$	0.951 ± 0.004	0.947 ± 0.005	0.015 ± 0.004	$2.5 \times 10^{-3} \pm 0.1$

Table S5: Performance and calibration metrics for ensemble models and the CIFAR-10 dataset.

Network	**Accuracy**	**F1-Score**	**ECE**	**Diversity**
$M = 1$	0.853 ± 0.008	0.852 ± 0.008	0.070 ± 0.005	—
MCD; $M = 5$	0.852 ± 0.007	0.851 ± 0.007	0.057 ± 0.002	$1.4 \times 10^{-3} \pm 0.1$
MCD; $M = 10$	0.852 ± 0.007	0.851 ± 0.006	0.055 ± 0.003	$1.5 \times 10^{-3} \pm 0.1$
MCD; $M = 100$	0.853 ± 0.008	0.853 ± 0.007	0.053 ± 0.004	$1.7 \times 10^{-3} \pm 0.1$
MCD-LLE; $M = 5$	0.852 ± 0.009	0.852 ± 0.008	0.067 ± 0.004	$3.1 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 10$	0.853 ± 0.009	0.852 ± 0.008	0.067 ± 0.004	$3.5 \times 10^{-4} \pm 0.2$
MCD-LLE; $M = 100$	0.853 ± 0.008	0.852 ± 0.008	0.067 ± 0.004	$3.6 \times 10^{-4} \pm 0.2$
MCD-LLE; $M = 1000$	0.853 ± 0.008	0.852 ± 0.008	0.067 ± 0.004	$3.7 \times 10^{-4} \pm 0.2$
DE; $M = 5$	0.836 ± 0.012	0.836 ± 0.012	0.049 ± 0.014	$1.9 \times 10^{-2} \pm 0.2$
DE-VGN; $M = 5$	0.875 ± 0.005	0.875 ± 0.006	0.023 ± 0.003	$9.8 \times 10^{-3} \pm 0.5$
LLE; $M = 5$	0.855 ± 0.002	0.855 ± 0.002	0.062 ± 0.012	$2.4 \times 10^{-3} \pm 0.3$
LLE; $M = 10$	0.848 ± 0.003	0.848 ± 0.002	0.058 ± 0.010	$3.6 \times 10^{-3} \pm 0.5$
LLE; $M = 100$	0.851 ± 0.004	0.850 ± 0.005	0.065 ± 0.006	$5.9 \times 10^{-3} \pm 0.8$
LLE-VGN; $M = 5$	0.846 ± 0.008	0.845 ± 0.008	0.067 ± 0.005	$2.3 \times 10^{-3} \pm 0.5$
LLE-VGN; $M = 10$	0.849 ± 0.004	0.849 ± 0.003	0.066 ± 0.003	$3.2 \times 10^{-3} \pm 0.2$
LLE-VGN; $M = 100$	0.847 ± 0.005	0.848 ± 0.004	0.071 ± 0.006	$5.4 \times 10^{-3} \pm 0.8$

Table S6: Performance and calibration metrics for ensemble models and the CIFAR-100 dataset.

Network	**Accuracy**	**F1-Score**	**ECE**	**Diversity**
$M = 1$	0.565 ± 0.012	0.563 ± 0.012	0.122 ± 0.002	—
MCD; $M = 5$	0.564 ± 0.012	0.562 ± 0.012	0.092 ± 0.001	$3.0 \times 10^{-4} \pm 0.2$
MCD; $M = 10$	0.567 ± 0.012	0.565 ± 0.012	0.086 ± 0.001	$3.4 \times 10^{-4} \pm 0.2$
MCD; $M = 100$	0.568 ± 0.012	0.566 ± 0.012	0.082 ± 0.002	$3.7 \times 10^{-4} \pm 0.2$
MCD-LLE; $M = 5$	0.563 ± 0.013	0.562 ± 0.013	0.106 ± 0.001	$1.9 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 10$	0.564 ± 0.014	0.563 ± 0.014	0.103 ± 0.003	$2.2 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 100$	0.565 ± 0.013	0.563 ± 0.013	0.102 ± 0.001	$2.3 \times 10^{-4} \pm 0.1$
MCD-LLE; $M = 1000$	0.565 ± 0.013	0.564 ± 0.013	0.102 ± 0.001	$2.3 \times 10^{-4} \pm 0.1$
DE; $M = 5$	0.487 ± 0.002	0.481 ± 0.003	0.095 ± 0.006	$3.5 \times 10^{-3} \pm 0.0$
DE-VGN; $M = 5$	0.507 ± 0.007	0.504 ± 0.007	0.086 ± 0.006	$3.1 \times 10^{-3} \pm 0.2$
LLE; $M = 5$	0.545 ± 0.002	0.541 ± 0.003	0.074 ± 0.002	$1.6 \times 10^{-3} \pm 0.0$
LLE; $M = 10$	0.543 ± 0.005	0.538 ± 0.005	0.062 ± 0.004	$2.4 \times 10^{-3} \pm 0.1$
LLE; $M = 100$	0.537 ± 0.010	0.533 ± 0.012	0.059 ± 0.019	$4.3 \times 10^{-3} \pm 0.2$
LLE-VGN; $M = 5$	0.548 ± 0.005	0.545 ± 0.007	0.111 ± 0.007	$1.3 \times 10^{-3} \pm 0.1$
LLE-VGN; $M = 10$	0.545 ± 0.014	0.542 ± 0.014	0.095 ± 0.026	$2.0 \times 10^{-3} \pm 0.2$
LLE-VGN; $M = 100$	0.544 ± 0.012	0.542 ± 0.011	0.059 ± 0.009	$4.2 \times 10^{-3} \pm 0.1$

S4.6 Out-of-Domain Detection

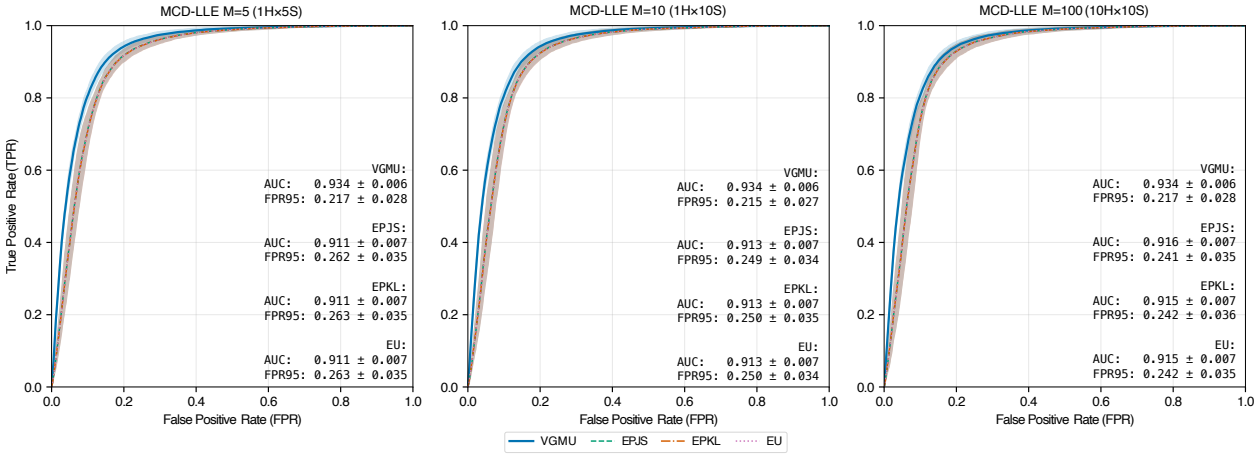


Figure S6: ROC curves for OOD detection (SVHN → CIFAR-10) with MCD-LLE under varying head (H) × sample (S) configurations with increasing stochastic samples: $1H \times 5S$ ($M = 5$), $1H \times 10S$ ($M = 10$), and $10H \times 10S$ ($M = 100$). VGMU curves are nearly identical across all configurations.

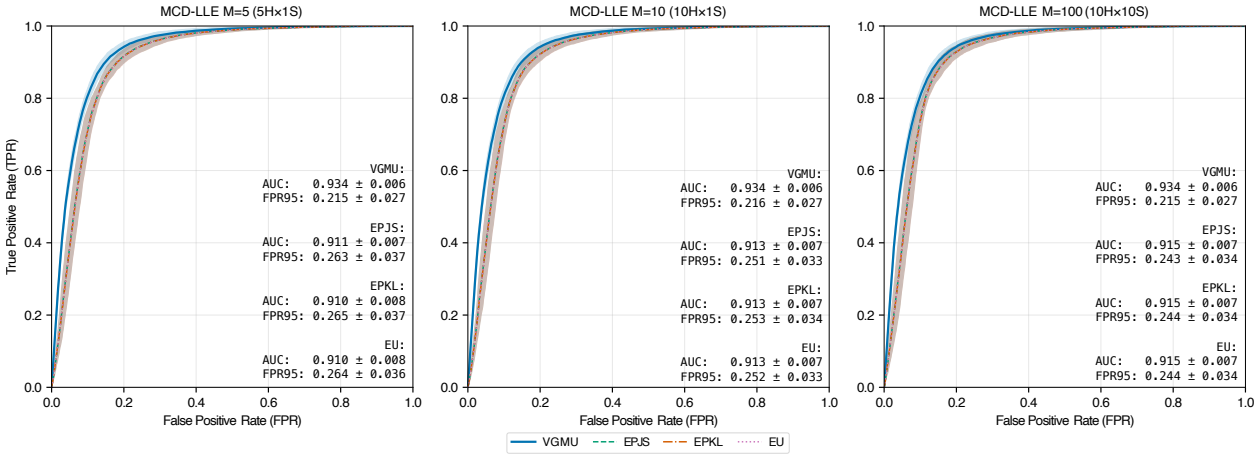


Figure S7: ROC curves for OOD detection (SVHN → CIFAR-10) with MCD-LLE under varying head (H) × sample (S) configurations with increasing classifier heads: $5H \times 1S$ ($M = 5$), $10H \times 1S$ ($M = 10$), and $10H \times 10S$ ($M = 100$). The rightmost panel ($M = 100$) is shared with Figure S6.

S4.7 Sensitivity of VGN to Hyperparameter k

In the main text, the per-class parameter k in VGN is learned end-to-end during training. A natural question is whether this learnable parameterization is necessary, or whether a fixed k suffices. To investigate, we train LLE-VGN on CIFAR-10 (WideResNet-28-10, M=10) with k fixed at three values spanning an order of magnitude: $k \in \{0.127,\ 0.693,\ 4.018\}$, corresponding to $\text{softplus}(\ell) \in \{-2.0,\ 0.0,\ 4.0\}$. All other hyperparameters, including the training protocol and early stopping criteria, are held constant. Results are averaged over three random seeds.

Classification performance. Table S7 reports accuracy, F1-score, ECE, and diversity. Classification accuracy is largely insensitive to the choice of k, with all configurations achieving 0.842–0.849 accuracy. The learnable-k variant (0.849) and fixed $k = 0.127$ (0.849) perform comparably. ECE varies modestly across settings; the non-VGN baseline achieves the lowest ECE (0.058), while fixed $k = 4.018$ resulted with the highest (0.076).

Table S7: Classification performance for LLE and LLE-VGN on CIFAR-10 (WideResNet-28-10, M=10). LLE reports ensemble predictions (P); VGN variants report variance-gated predictions (Q).

Method	k	Accuracy	F1	ECE	Diversity
LLE	–	0.848 ± 0.003	0.848 ± 0.002	0.058 ± 0.010	3.6×10^{-3} ±0.6
LLE-VGN	0.127	0.849 ± 0.006	0.850 ± 0.005	0.065 ± 0.002	3.3×10^{-3} ±0.4
LLE-VGN	0.693	0.845 ± 0.005	0.845 ± 0.005	0.061 ± 0.004	3.6×10^{-3} ±0.4
LLE-VGN	4.018	0.842 ± 0.004	0.843 ± 0.003	0.076 ± 0.009	3.5×10^{-3} ±0.6
LLE-VGN	learned	0.849 ± 0.004	0.849 ± 0.003	0.066 ± 0.003	3.2×10^{-3} ±0.2

Uncertainty quality. Table S8 reports AUCc scores, which measure the concentration of uncertainty mass on difficult samples. Here, the effect of k is more pronounced. Fixed $k = 0.693$ matches the non-VGN baselines, indicating that this value of k effectively removes any benefits of the variance gate. In contrast, fixed $k = 0.127$ and $k = 4.018$ both improve over the baseline, with $k = 4.018$ achieving the highest AUCc across all uncertainty metrics. The learnable-k variants falls between the two best fixed values, indicating that it recovers near-optimal performance without requiring a hyperparameter search.

Table S8: Uncertainty mass concentration (AUCc) for LLE and LLE-VGN on CIFAR-10 (WideResNet-28-10, M=10). Higher values indicate sharper concentration on difficult samples.

Method	k	VGMU	EU	EPJS	EPKL
LLE	–	0.885 ± 0.015	0.868 ± 0.016	0.873 ± 0.016	0.876 ± 0.015
LLE-VGN	0.127	0.894 ± 0.008	0.879 ± 0.009	0.884 ± 0.008	0.885 ± 0.009
LLE-VGN	0.693	0.885 ± 0.006	0.868 ± 0.007	0.873 ± 0.006	0.876 ± 0.006
LLE-VGN	4.018	0.899 ± 0.012	0.886 ± 0.012	0.891 ± 0.012	0.892 ± 0.010
LLE-VGN	learned	0.897 ± 0.009	0.881 ± 0.010	0.886 ± 0.010	0.886 ± 0.010

Rank correlations. Table S9 and Table S10 report Spearman and Kendall rank correlations between VGMU and epistemic uncertainty baselines. The pattern mirrors the AUCc results: fixed $k = 0.693$ matches the baseline, while $k = 4.018$ and learnable k achieve the highest correlations. All VGN variants maintain strong rank agreement ($\rho > 0.986$, $\tau > 0.907$).

Summary. The sensitivity analysis reveals that VGN's benefit depends on the choice of k, with performance varying non-monotonically across the tested range. An unfavorable choice (*e.g.*, $k = 0.693$) can

Table S9: Spearman rank correlation (ρ) between VGMU and epistemic uncertainty baselines on CIFAR-10 (WideResNet-28-10, $M = 10$).

Method	k	**VGMU *vs.* EU**	**VGMU *vs.* EPJS**	**VGMU *vs.* EPKL**
LLE	–	0.988 ± 0.004	0.992 ± 0.002	0.985 ± 0.004
LLE-VGN	0.127	0.991 ± 0.002	0.994 ± 0.001	0.988 ± 0.003
LLE-VGN	0.693	0.989 ± 0.001	0.993 ± 0.001	0.986 ± 0.002
LLE-VGN	4.018	0.993 ± 0.002	0.995 ± 0.001	0.990 ± 0.002
LLE-VGN	learned	0.992 ± 0.002	0.995 ± 0.001	0.990 ± 0.002

Table S10: Kendall rank correlation (τ) between VGMU and epistemic uncertainty baselines on CIFAR-10 (WideResNet-28-10, $M = 10$).

Method	k	**VGMU vs. EU**	**VGMU vs. EPJS**	**VGMU vs. EPKL**
LLE	–	0.918 ± 0.011	0.939 ± 0.009	0.905 ± 0.012
LLE-VGN	0.127	0.928 ± 0.007	0.946 ± 0.005	0.917 ± 0.009
LLE-VGN	0.693	0.920 ± 0.003	0.940 ± 0.003	0.907 ± 0.003
LLE-VGN	4.018	0.934 ± 0.007	0.951 ± 0.006	0.924 ± 0.010
LLE-VGN	learned	0.932 ± 0.008	0.948 ± 0.007	0.923 ± 0.010

remove the effect of variance gating, while a well-chosen fixed k (*e.g.*, 4.018) can slightly outperform the learnable variant. However, the learnable parameterization consistently achieves near-optimal uncertainty calibration without requiring prior knowledge of the appropriate k value, making it the recommended default for practical use.

S5 Risk-Based Interpretation

The term $\mathbf{ks}$ represents a classwise tolerance scale describing the dispersion of ensemble member probabilities around the ensemble mean. Under mild distributional assumptions on ensemble dispersion, $\mathbf{k} \in \{1, 2, 3\}$ corresponds to *ca.* 68%, 95%, or 99.7% of members within one, two, or three standard deviations of the mean, respectively. When ensemble predictions are consistent ($\mathbf{ks} \ll \bar{\mathbf{p}}$), the gate satisfies $\mathbf{\Gamma} \approx 1$, leaving probabilities unchanged ($\mathbf{q}_m \approx \mathbf{p}_m$). When disagreement is significant ($\mathbf{ks} \gg \bar{\mathbf{p}}$), $\mathbf{\Gamma}$ decreases; however, after normalization, probability mass is redistributed such that high-variance (uncertain) classes are suppressed. The quadratic decay of sensitivity with respect to $\mathbf{s}$ and $\mathbf{k}$ ensures that further increases in disagreement produce diminishing changes to the gate. This variance-gating effect is not uniform suppression. The gating selectively down-weights inconsistent ensemble predictions while preserving those supported by ensemble agreement. Therefore, the resulting distribution yields lower uncertainty, translating epistemic disagreement into more conservative, confident predictions. Therefore, variance-gating functions as an epistemic-aware mechanism, suppressing unreliable high-variance predictions while reinforcing consistent ones. Normalization amplifies the relative impact of smaller gates, ensuring that ensemble disagreement yields conservative, bounded, and well-calibrated predictive distributions. This behavior ensures that epistemic disagreement translates into conservative distributions, thereby enforcing the consistency and boundedness properties required by Wimmer's Axioms A0–A5.

S6 Variance-Gated Behavior Across Axioms

Table S11 and Figure S8 provide illustrations of variance-gated ensemble behavior under the axiomatic framework introduced by Wimmer et al. (2023). Table S11 defines ensemble configurations designed to test each axiom (A2–A5), while Figure S8 visualizes the corresponding simplex geometry and uncertainty decompositions as the gating parameter k varies.

A2 (Maximal at Uniform Distribution). When ensemble members span the simplex vertices (P_0), the variance-gated decomposition correctly assigns maximal epistemic uncertainty (EU = 1.0), as members are maximally disagreeing despite the uniform ensemble mean. When members collapse to identical uniform distributions (Q_0), EU vanishes (EU = 0.0) while total and aleatoric uncertainty remain maximal (TU = AU = 1.0). This configuration is invariant to gating since Q_0 has zero variance. The results confirm that VGN distinguishes between distributional ambiguity (high AU) and ensemble disagreement (high EU), consistent with the partial satisfaction noted in Table 4.

A3 (Mean-Preserving Spread). The transformation from P_0 (identical members with $\boldsymbol{\mu} = [0.70, 0.20, 0.10]$, $\boldsymbol{\sigma} = \mathbf{0}$) to Q_0 (spread members with preserved mean, $\boldsymbol{\sigma} = [0.17, 0.10, 0.10]$) demonstrates that VGN correctly increases EU when ensemble disagreement grows without changing the predictive mean. At $k = 0$, EU increases from 0.000 to 0.070. Figure S8 shows that increasing k progressively attenuates this epistemic signal, with EU decreasing from 0.070 ($k = 0$) to 0.055 ($k = 1$) to 0.045 ($k = 2$), illustrating the saturation behavior described in Proposition 3.2. Correspondingly, the VGMU increases from 0.300 to 0.412 ($k = 0$), reflecting decision-relevant uncertainty under ensemble disagreement.

A4 (Center-Shift). Shifting ensemble mass toward the simplex barycenter (from $\boldsymbol{\mu} = [0.70, 0.20, 0.10]$ in P_0 to $\boldsymbol{\mu} = [0.40, 0.35, 0.25]$ in Q_0) while preserving spread increases both TU ($0.730 \to 0.984$) and AU ($0.714 \to 0.971$). EU remains approximately stable across this transformation ($0.016 \to 0.013$ at $k = 0$), demonstrating that VGN correctly attributes the increased uncertainty to aleatoric rather than epistemic sources. The VGMU increases substantially ($0.325 \to 0.887$), reflecting the reduced class separability near the barycenter. This behavior persists across k values, confirming that variance-gating preserves the distinction between location-induced ambiguity and spread-induced disagreement.

A5 (Spread-Preserving Location Shift). Moving ensembles toward a vertex (from $\boldsymbol{\mu} = [0.70, 0.20, 0.10]$ in P_0 to $\boldsymbol{\mu} = [0.90, 0.05, 0.05]$ in Q_0) while preserving spread decreases TU ($0.730 \to 0.359$) and AU ($0.714 \to 0.314$). At $k = 0$, EU shows sensitivity to the location shift ($0.016 \to 0.045$), which represents a deviation from strict axiom compliance. However, as k increases, the gap between P and Q epistemic uncertainty values narrows substantially: from 0.029 at $k = 0$, to 0.015 at $k = 1$, to 0.006 at $k = 2$. This progressive reduction suggests that variance-gating enforces approximate invariance to location shifts, with stronger

Table S11: Illustrative ensembles used to evaluate Wimmer's axiom (A2–A5). Each pair of ensembles P_0 and Q_0 displays a transformation corresponding to a specific axiom. These examples are used to test whether a given uncertainty measure respects the qualitative properties required by each axiom.

Axioms	**Ensemble P_0**	**Ensemble Q_0**
A2: Maximal at Uniform Distribution	$P_0 = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}$	$Q_0 = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$
	$\mu = \begin{bmatrix} 0.33 & 0.33 & 0.33 \end{bmatrix}$	$\mu = \begin{bmatrix} 0.33 & 0.33 & 0.33 \end{bmatrix}$
	$\sigma = \begin{bmatrix} 0.58 & 0.58 & 0.58 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.00 & 0.00 & 0.00 \end{bmatrix}$
A3: Mean-Preserving Spread	$P_0 = \begin{bmatrix} 0.70 & 0.20 & 0.10 \\ 0.70 & 0.20 & 0.10 \\ 0.70 & 0.20 & 0.10 \end{bmatrix}$	$Q_0 = \begin{bmatrix} 0.90 & 0.10 & 0.00 \\ 0.60 & 0.30 & 0.10 \\ 0.60 & 0.20 & 0.20 \end{bmatrix}$
	$\mu = \begin{bmatrix} 0.70 & 0.20 & 0.10 \end{bmatrix}$	$\mu = \begin{bmatrix} 0.70 & 0.20 & 0.10 \end{bmatrix}$
	$\sigma = \begin{bmatrix} 0.00 & 0.00 & 0.00 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.17 & 0.10 & 0.10 \end{bmatrix}$
A4: Center-Shift	$P_0 = \begin{bmatrix} 0.80 & 0.15 & 0.05 \\ 0.70 & 0.20 & 0.10 \\ 0.60 & 0.25 & 0.15 \end{bmatrix}$	$Q_0 = \begin{bmatrix} 0.50 & 0.30 & 0.20 \\ 0.40 & 0.35 & 0.25 \\ 0.30 & 0.40 & 0.30 \end{bmatrix}$
	$\mu = \begin{bmatrix} 0.70 & 0.20 & 0.10 \end{bmatrix}$	$\mu = \begin{bmatrix} 0.40 & 0.35 & 0.25 \end{bmatrix}$
	$\sigma = \begin{bmatrix} 0.10 & 0.05 & 0.05 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.10 & 0.05 & 0.05 \end{bmatrix}$
A5: Spread-Preserving Location Shift	$P_0 = \begin{bmatrix} 0.80 & 0.15 & 0.05 \\ 0.70 & 0.20 & 0.10 \\ 0.60 & 0.25 & 0.15 \end{bmatrix}$	$Q_0 = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.90 & 0.05 & 0.05 \\ 0.80 & 0.10 & 0.10 \end{bmatrix}$
	$\mu = \begin{bmatrix} 0.70 & 0.20 & 0.10 \end{bmatrix}$	$\mu = \begin{bmatrix} 0.90 & 0.05 & 0.05 \end{bmatrix}$
	$\sigma = \begin{bmatrix} 0.10 & 0.05 & 0.05 \end{bmatrix}$	$\sigma = \begin{bmatrix} 0.10 & 0.05 & 0.05 \end{bmatrix}$

gating (larger k) providing closer adherence to axiom A5. The corresponding VGMU values ($0.325 \rightarrow 0.103$ at $k = 0$) reflect the improved class separability near the vertex.

These results provide examples of the axiomatic compliance summarized in Table 4 of the main text. The simplex visualizations in Figure S8 complement the geometric interpretation in Section 3.1.1, illustrating how the four ensemble simple regions (*i.e.*, confident–certain, ambiguous–certain, confident–uncertain, ambiguous–uncertain) manifest under controlled axiom-testing transformations. The progressive reduction in uncertainty measures with increasing k demonstrates the risk-tolerance interpretation discussed in Section S5, where larger k values accommodate greater ensemble disagreement before suppressing predictions. The partial deviations observed, particularly for A5 at low k, reflect the finite-ensemble and exponential gating effects inherent to the VGN formulation, while the convergence with increasing k supports the properties claimed in the main text.

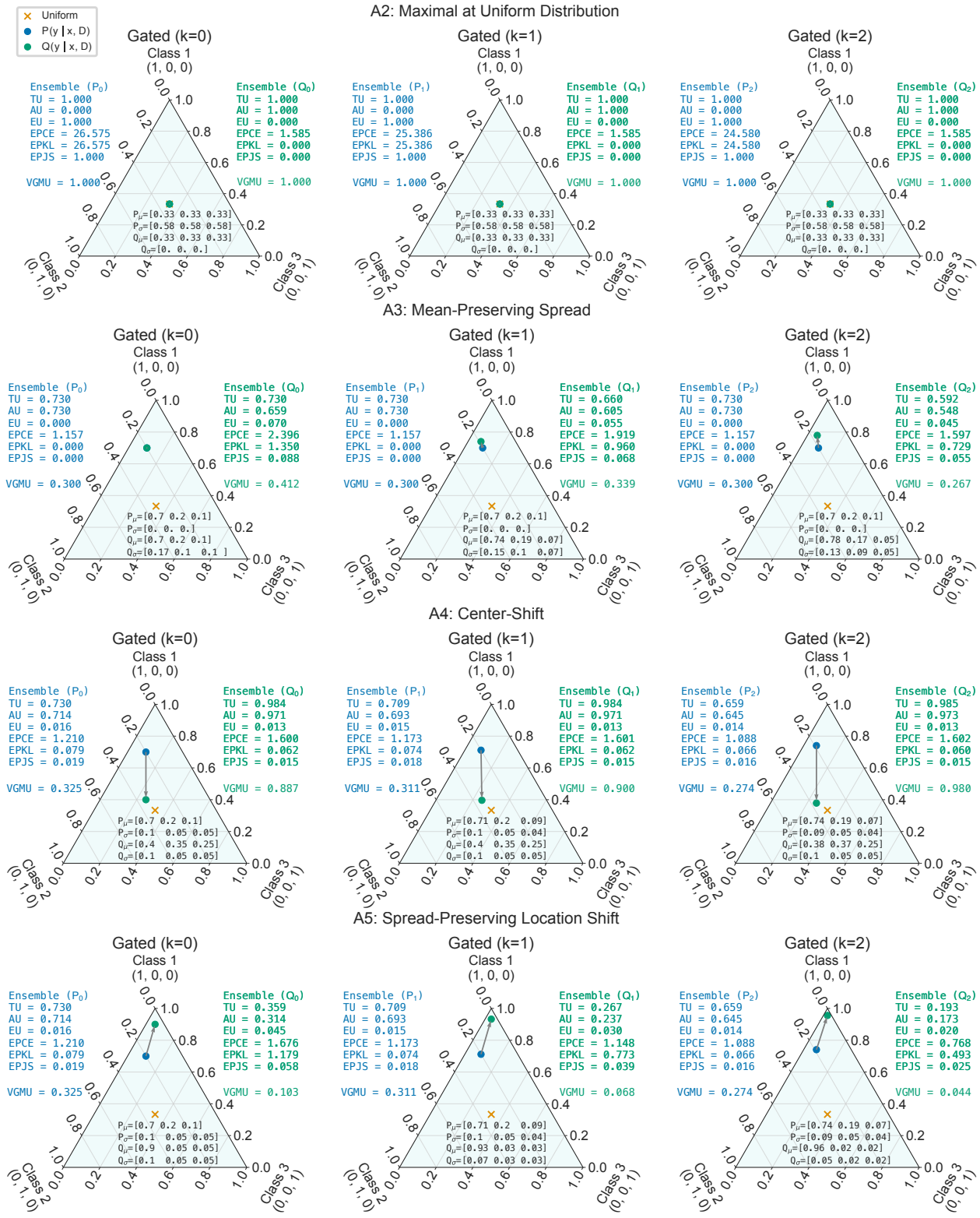


Figure S8: Variance-gated behavior across Wimmer's axioms (A2–A5). Annotations report TU, AU, and EU uncertainty, as well as divergence-based metrics (EPCE, EPKL, EPJS). Gating progressively reduces epistemic contributions, demonstrating compliance with the axioms under controlled variance attenuation. The orange cross denotes the simplex uniform distribution, while circles indicate distributions before (P, blue) and after (Q, green) gating.