WHAT "FORM" OF UNCERTAINTY DOES CONFORMAL PREDICTION QUANTIFY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction (CP) is a post-hoc calibration method that returns prediction sets with distribution-free coverage guarantees, and the size of the output set is often interpreted as a notion of uncertainty. We show that the difference in set size between full and split CP is directly connected to Bayesian posterior variance, which is the epistemic uncertainty of the model. Essentially, this difference in set size is an aggregation of two opposite effects: (1) better model accuracy and therefore smaller set size achieved by training over extra calibration points, and (2) the "formability" of the model at test point towards every possible label. The latter effect vanishes with stronger models or more calibration data (as epistemic uncertainty is known for that behavior). It turns out that the model's formability itself (without the quantization effect of prediction sets, and through estimating the posterior variance) is a useful tool to quantify the epistemic uncertainty. In our empirical evaluation it performs on par with the state of the art in many setups.

1 INTRODUCTION

A model's prediction can be uncertain, either since the data is hard to predict (aleatoric), or because the model is not confident in its prediction due to lack of information (epistemic). Conformal prediction (CP) is a widely used approach in risk-critical applications where quantifying total predictive uncertainty is crucial. Given any model $f(\cdot;\boldsymbol{\theta})$, CP returns a prediction set $\mathcal{C}(\boldsymbol{x}_{n+1})$ for a test point $\boldsymbol{x}_{n+1}$ that is guaranteed to include the true label with user-specified probability $1-\alpha$. This guarantee is distribution-free; namely, it holds regardless of the underlying data distribution, or the model family (Angelopoulos et al., 2024). CP requires an exchangeable set of calibration points, and can be applied in two ways: (i) fine-tuning the model jointly on the calibration set and the test point for every possible pseudo-label (full CP), or (ii) only evaluating the model on the calibration points (split CP). Full CP is known to be more data-efficient, which is particularly appealing in cases with label scarcity (Lei et al., 2018) (e.g. semi-supervised graph node classification). The argument is that in these settings the data is not divided into separate training and calibration sets, which potentially results in better model performance, and smaller sets.

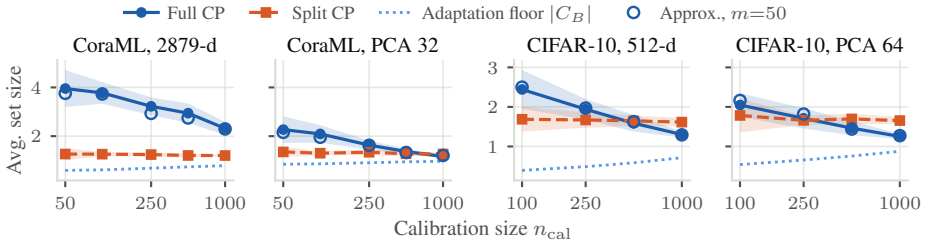

Figure 1: Comparison of full (blue, solid) vs. split (orange, dashed) CP, against the adaptation floor $|\mathcal{C}_B|$ (see § 3), over n_{cal} and feature dimension K, and for [Left] CoraML (transductive node-classification, GCN), and [Right] CIFAR-10. The approximation is with $m = 50$ samples (see § 4).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In contrast, Zong & Xu (2025) show that full CP doesn't always outperform split CP without specifying a definitive reason, leaving this question open: *when and by how much full CP can be more efficient, and what can be concluded from their performance gap?* It turns out (as shown in Fig. 1, and Fig. 2), the performance gap (set size difference) reduces with a larger dataset (calibration, or training), and a lower feature dimension. Intuitively, while in most cases a larger dataset (combination of training and calibration) improves the model's performance, fine-tuning on the test point with a pseudo-label deforms the model towards that label. The strength of this deformation is governed by the calibration density, and the confidence from pretraining both as resisting factors. Strong deformation includes wrong pseudo-labels in the set, resulting in worse overall performance.

In particular, we define *formability* at a test point, as the difference between the predictive probability of any label between two models: one fine-tuned on a calibration set, and the other on the union of the calibration set and the pseudo-labeled test point (on that label). Under Bayesian training, where the epistemic uncertainty is quantified as the disagreement among the plausible (under the posterior) parameters, we show that formability is directly related to the epistemic uncertainty of the predictive distribution at the test point; namely we bound the posterior variance, by a function of formability. This also shows the reason why by increasing calibration set size the difference between full and split CP disappears: the formability (epistemic uncertainty) is known to reduce with more data.

2 Background

Let $f(\cdot;\boldsymbol{\theta})$ be a pretrained model from the family $\mathcal{F}$ specified by the parameters $\boldsymbol{\theta}$. We use $p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ for the predictive label distribution of f at the input $\boldsymbol{x}$. Let $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a holdout calibration set jointly exchangeable with the test point $(\boldsymbol{x}_{n+1}, y_{n+1})$.

Split CP (Vovk et al., 2005) (SCP). With $r_i^{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)$, define the threshold

$$q_{\boldsymbol{\theta}} = Q_{\alpha,n}(r_{1:n}) := \mathbb{Q}_\alpha\left(\sum_{i=1}^{n} \frac{1}{n+1}\delta_{r_i^{\boldsymbol{\theta}}} + \frac{1}{n+1}\delta_{-\infty}\right). \tag{1}$$

and the prediction set $\mathcal{C}^{\mathrm{S}}_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1}) = \{c \in \mathcal{Y} : p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1}) \geq q_{\boldsymbol{\theta}}\}$. This set has a distribution-free coverage guarantee $\Pr[y_{n+1} \in \mathcal{C}^{\mathrm{S}}_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1})] \geq 1-\alpha$. Other scores are discussed in § C.1.6.

Full CP. We define the augmented dataset $\mathcal{D}_{n+1}^{(c)} = \mathcal{D}_n \cup \{(\boldsymbol{x}_{n+1}, c)\}$, for a pseudo-label c. Full conformal prediction (Vovk et al., 2005) (FCP) defines the prediction set $\mathcal{C}^{\mathrm{F}}_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1}) = \{c \in \mathcal{Y} : p^{(c)}_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1}) \geq Q_{\alpha,n}(r^{(c)}_{1:n})\}$ for $r_i^{(c)} = p^{(c)}_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)$, where $p^{(c)}_{\boldsymbol{\theta}}$ is the predictive distribution of the model after fine-tuning on $\mathcal{D}^{(c)}_{n+1}$[1]. This prediction set also comes with the same distribution-free guarantee. In both cases, the guarantee comes from the fact that the true scores $p_\cdot(x_i, y_i)$ are exchangeable for $i \in [n+1]$. The difference is that FCP fine-tunes a single model per test point, per label. The potential use case of this computationally expensive approach is when the labeled dataset is small, and the model can benefit from training on the entire dataset (training and calibration).

Bayesian deep learning. Instead of a point estimate $\boldsymbol{\theta}$ (deterministic value for parameters), Bayesian deep learning (Neal, 1996; Wilson & Izmailov, 2020) defines a distribution $\pi(\boldsymbol{\theta})$ over the model parameters. As a result, the predictive distribution is also defined as $p_\pi(y \mid \boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\theta}\sim\pi}[p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})]$ instead of a single $p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$. Training a Bayesian model is equivalent to finding the posterior parameter distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ given the dataset $\mathcal{D}$ as evidence, and a prior $\pi_0(\boldsymbol{\theta})$ over the parameters (e.g. $\mathcal{N}(\mathbf{0}, \sigma\boldsymbol{I})$ for initialization). Under this definition, the epistemic uncertainty is quantified as the disagreement among parameters that are plausible under the posterior – measured as the variance $\mathrm{Var}_{\boldsymbol{\theta}\sim\pi}[p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})]$ (Sale et al., 2024). Notably, as Jiménez et al. (2025) argue, one cannot quantify the aleatoric uncertainty unless the epistemic uncertainty is sufficiently low.

Basin decomposition. The support of a probability distribution (e.g. $\pi(\boldsymbol{\theta})$) can be decomposed into disjoint basins of attraction $\mathcal{B}_1, \ldots, \mathcal{B}_K$ with $\cup_{k=1}^K \mathcal{B}_k = \mathrm{Support}(\pi)$, and $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ for $k \neq k'$. Each basin is a connected region of the parameter space that contains a local optimum $\hat{\boldsymbol{\theta}}_k$ of the loss function. In § 3.3, we discuss that the posterior predictive distribution can be decomposed into a mixture of distributions over each basin. A better approximation of the variance considers basins other than only the one containing the MAP as well.

[1] Full CP can be applied by both training an initial model, or fine-tuning a pre-trained model.

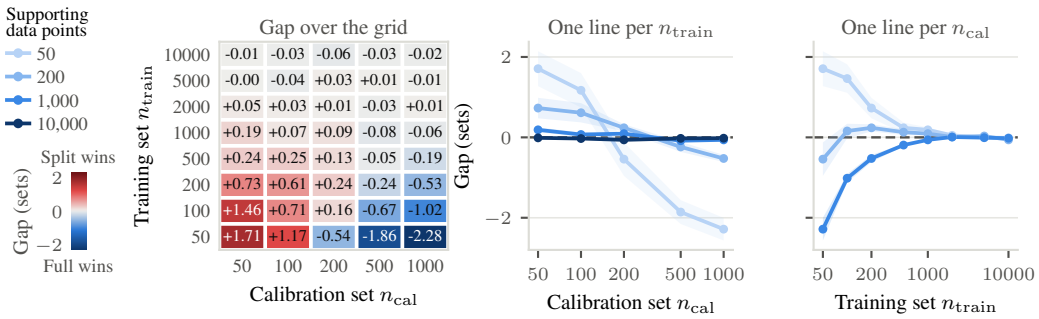

Figure 2: Performance gap between $|C_{\text{full}}|$ and $|C_{\text{split}}|$ over a grid of training and calibration set sizes. This is an extension of the result shown in Fig. 1.

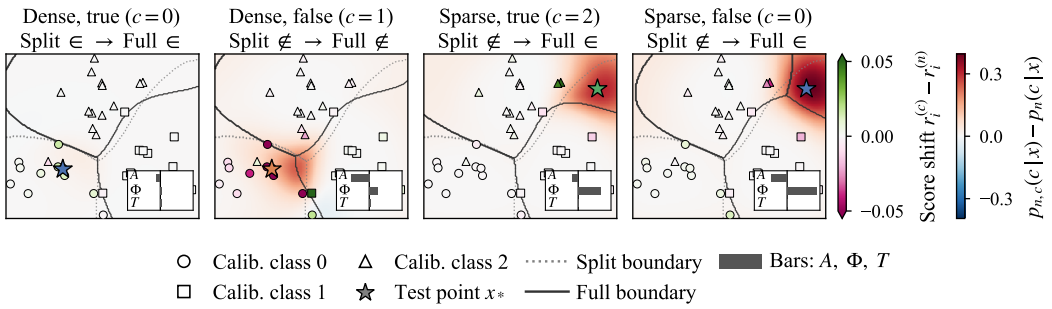

Figure 3: The formability $p_{n,c}(c \mid \boldsymbol{x}) - p_n(c \mid \boldsymbol{x})$ by fine-tuning on the pseudo-label c at the test point (star), in a dense [left pair], and a sparse [right pair] (calibration) neighborhood. Pairs are a correct, and a wrong label c. Calibration points are colored by the shift of their true score (negligible at many points), and the bars report $A_{n,c}$, $\Phi_{n,c}$, and $T_{n,c}$ from Eq. 2.

3 Formability

3.1 Motivating Experiment

First we evaluate the claim that "in label scarcity, full CP outperforms split CP" by comparing full and split conformal prediction on two simple setups with images (`CIFAR-10`; Krizhevsky, 2009), and graphs (`CoraML`; Bojchevski & Günnemann, 2018). It is already mentioned in a couple of works that in some cases full CP is not better than split CP in set size (e.g., Barber et al., 2023; Zong & Xu, 2025). As shown in Fig. 1, however, by increasing the calibration set size, or reducing the feature dimension (using PCA), the performance gap is reduced, until the two methods cross over. These two observations suggest that the performance gap is not governed by the amount of data alone, but by how strongly the model is already determined around the test point. Notably, the argument of "full CP leverages a better model" still serves as a lower bound on the set size, which full CP converges to, by increasing the size of the calibration set.

An intuitive explanation for such behaviour is that when the training and calibration data is scarce, the model can still be deformed towards any label at the test point. Therefore, it is very easy for every (true or false) label of $\boldsymbol{x}_{n+1}$ to pull its own score above the threshold. To observe this mechanism in isolation, we show a synthesized setup where every score is visible on a 2-dim landscape (Fig. 3, details are in § C.1.1). We compare a test point inside a dense calibration neighborhood to one in a sparse support. In all cases, fine-tuning on the pseudo-label deforms the predictive distribution towards the pseudo-label. However, for the one in the dense region, the calibration density resists this deformation, and the self-pull remains small. In the sparse region there are no proximate data to resist. Hence, every pseudo-label pulls its own score above the threshold easily. As a result,

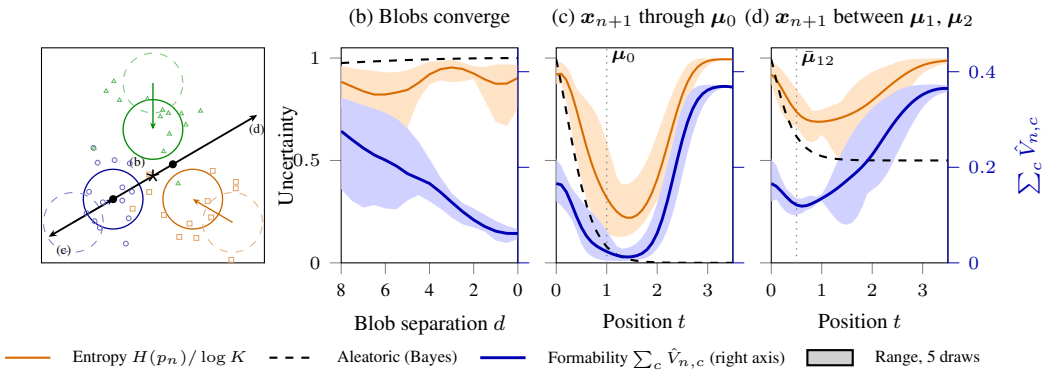

Figure 4: Softmax entropy vs formability variance. The setup is similar to Fig. 3 (see § C.1.2). [Left] Data, and direction of movements. Circles show the mean and standard deviation (radius). We move $\boldsymbol{x}_{n+1}$ from the center to between two blobs (alea ↑, and epis ↓), and from the other side towards the center of the other blob (both ↓). [b] Uncertainty at center (star) when blobs move closer (epis ↓, and alea ↑), *only formability reacts to it*. [c], and [d] show it when moving towards the center of one blob (both ↓), and between two blobs (epis ↓, alea ↑).

the wrong label enters the full set exactly where the model was free to deform. Therefore the performance gap in CPs is governed by the significance of the new test data point w.r.t. the model: how much the model can deform at that point (which is inversely related to the size of $\mathcal{D}_n$, and the strength of the pretrained model as a resisting factor). Namely, this deformation is connected to the epistemic uncertainty – a model with more knowledge (training, and calibration data) is harder to form towards wrong labels.

This direct connection between formability and epistemic uncertainty is particularly important when it compares with methods like softmax entropy. Naturally, the softmax entropy, when evaluated on a single predictor can not distinguish between the two types of uncertainty. Fig. 4 shows that by replicating the setup from Fig. 3 while moving the generative Gaussians close to each other. At the center, the epistemic uncertainty decrease, while aleatoric uncertainty increase. On a single label distribution however, it is presented as $[1/k]_k$ the entire time. Therefore the softmax entropy does not react to this change, while the formability variance does.

3.2 Theory

Let $m_\circ(\boldsymbol{x}, y) := p_\circ(y \mid \boldsymbol{x}) - q_\circ$ be the margin of the label y at $\boldsymbol{x}$ in any of these predictive regimes: (i) split CP with the pretrained predictive distribution $p_{\boldsymbol{\theta}}$ (call it m_0), (ii) full CP fine-tuned on the augmented set $\mathcal{D}_{n+1}^{(c)}$ with predictive distribution $p_{n,c}$ (call it $m_{n,c}$), and (iii) CP fine-tuned only on the calibration set $\mathcal{D}_n$ with predictive distribution p_n (call it m_n). Note that (iii) is not a valid CP, as the calibration and test points are not exchangeable, but it represents the effect of "better model from more data". In all cases, a label is in the prediction set if its margin is non-negative. We decompose the full margin and define three new variables $A_{n,c}$, $\Phi_{n,c}$, and $T_{n,c}$ as

$$\begin{aligned} m_{n,c}(\boldsymbol{x}, c) &= m_0(\boldsymbol{x}, c) + (m_n(\boldsymbol{x}, c) - m_0(\boldsymbol{x}, c)) + (m_{n,c}(\boldsymbol{x}, c) - m_n(\boldsymbol{x}, c)) \\ &= m_0(\boldsymbol{x}, c) + \underbrace{\left[p_{\boldsymbol{\theta}}^{(n)}(c \mid \boldsymbol{x}_{n+1}) - p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\right] - \left(q_{\boldsymbol{\theta}}^{(n)} - q_{\boldsymbol{\theta}}\right)}_{A_{n,c}} \\ &\quad + \underbrace{p_{\boldsymbol{\theta}}^{(c)}(c \mid \boldsymbol{x}_{n+1}) - p_{\boldsymbol{\theta}}^{(n)}(c \mid \boldsymbol{x}_{n+1})}_{\Phi_{n,c}} - \underbrace{\left(q_{\boldsymbol{\theta}}^{(c)} - q_{\boldsymbol{\theta}}^{(n)}\right)}_{T_{n,c}}. \end{aligned} \tag{2}$$

The above decomposition is trivial, but each of the introduced terms has an interpretation: $A_{n,c}$ captures the adaptation-driven efficiency which is the result of fine-tuning on the calibration set. It simulates the effect from which we expected full CP to be more efficient. $\Phi_{n,c}$, and $T_{n,c}$ are the effects of training on the additional pseudo-labeled test point (we call the total $\bar{\Phi}_{n,c} := \Phi_{n,c} - T_{n,c}$ the net formability). Adding $(\boldsymbol{x}_{n+1}, c)$ to the training set pulls the model towards that label

(increasing the margin), but at the same time, it affects the other calibration points, which can change the quantile threshold. Ultimately, the full and split CP set size is only a quantization of the margins; i.e.

$$|\mathcal{C}^F(\boldsymbol{x}_{n+1})| - |\mathcal{C}^S(\boldsymbol{x}_{n+1})| = \sum_{c\in\mathcal{Y}}\Big[\mathbf{1}\{m_0 + A_{n,c} + \Phi_{n,c} - T_{n,c} \geq 0\} - \mathbf{1}\{m_0 \geq 0\}\Big].$$

Under Bayesian Learning. Discussing this decomposition under Bayesian learning has two advantages: (i) Bayesian learning comes with a well-defined notion of epistemic uncertainty; i.e. the variance of the posterior predictive distribution, and (ii) in the Bayesian framework, we can write fine–tunings as a sequence of Bayesian updates, which decouples the effect of the additional pseudo-label from the calibration data. Let $\pi_0(\boldsymbol{\theta})$ be the prior parameter distribution centered around $\boldsymbol{\theta}_0$ (initial or pretrained model), $\pi_n(\boldsymbol{\theta})$ be the posterior distribution after observing the calibration set $\mathcal{D}_n$ (we call it "calibration posterior"), and $\pi_{n,c}(\boldsymbol{\theta})$ be the posterior distribution with $\mathcal{D}_{n+1}^{(c)} = \mathcal{D}_n \cup (\boldsymbol{x}_{n+1}, c)$ as evidence (we call it "augmented posterior"). By sequential Bayesian updating (Lemma B.1), the posterior distribution $\pi_{n,c}$ can be computed using Bayes' rule with $\{(\boldsymbol{x}_{n+1}, c)\}$ as a single evidence, and π_n as the prior. For ease of notation, let

$$w_{n,c} := p_{\pi_n}(c \mid \boldsymbol{x}_{n+1}) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})], \tag{3}$$

and assume $w_{n,c} > 0$. In Bayesian learning, we define the formability on $(\boldsymbol{x}_{n+1}, c)$ as

$$\Phi_{n,c}(\boldsymbol{x}_{n+1}) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_{n,c}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] - \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]. \tag{4}$$

The following lemma shows that the formability is directly related to the epistemic uncertainty of the predictive distribution at $\boldsymbol{x}_{n+1}$.

Lemma 3.1. *The (per-label) variance of the calibration posterior predictive distribution at $\boldsymbol{x}_{n+1}$ is proportional to the formability at $\boldsymbol{x}_{n+1}$:*

$$\mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] = w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}), \quad \textit{and} \tag{5}$$

therefore, $0 \leq \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] \leq \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big)$.

As a result, for a finite label set, the total posterior variance (epistemic uncertainty) satisfies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\Big[\|p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}_{n+1}) - p_{\pi_n}(\cdot \mid \boldsymbol{x}_{n+1})\|_2^2\Big] &= \sum_{c\in\mathcal{Y}} w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}) \\ &\leq \sum_{c\in\mathcal{Y}} \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big).\end{aligned} \tag{6}$$

The two terms $\Phi_{n,c}$, and $T_{n,c}$ in Eq. 2 are governed by the same posterior on the pseudo-label. As both variance, and $w_{n,c}$ are positive, the formability $\Phi_{n,c}$ is positive. This means that training on the pseudo-label necessarily increases its score. At the same time, it also moves every calibration score, and therefore the threshold. To study all of the above, we define the *forming kernel* as the covariance kernel over the posterior distribution:

$$\mathcal{K}_n(z, z') := \mathrm{Cov}_{\boldsymbol{\theta}\sim\pi_n}\big(p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}),\, p_{\boldsymbol{\theta}}(y' \mid \boldsymbol{x}')\big), \qquad z = (\boldsymbol{x}, y),\ z' = (\boldsymbol{x}', y'), \tag{7}$$

and use the shorthands $C_{i,c} := \mathcal{K}_n\big((\boldsymbol{x}_i, y_i), (\boldsymbol{x}_{n+1}, c)\big)$ for the kernel between calibration points, and the pseudo-label hypothesis, $V_i := \mathcal{K}_n\big((\boldsymbol{x}_i, y_i), (\boldsymbol{x}_i, y_i)\big) = \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)]$, and the calibration-posterior variance at the test point, $V_{n,c} := \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] = \Phi_{n,c}(\boldsymbol{x}_{n+1})w_{n,c}$ (Lemma 3.1).

Theorem 3.2. *Under Bayesian training, with scores $r_i^{(n)} = p_{\pi_n}(y_i \mid \boldsymbol{x}_i)$, $r_i^{(c)} = p_{\pi_{n,c}}(y_i \mid \boldsymbol{x}_i)$ and $q_{\boldsymbol{\theta}}^{(n)} = Q_{\alpha,n}(r_{1:n}^{(n)})$, $q_{\boldsymbol{\theta}}^{(c)} = Q_{\alpha,n}(r_{1:n}^{(c)})$, the following statements hold: (i) the update in each calibration score r_i is exactly $r_i^{(c)} - r_i^{(n)} = \frac{C_{i,c}}{w_{n,c}}$ and bounded in absolute value by $\frac{\sqrt{V_i\, V_{n,c}}}{w_{n,c}}$. As a result, with $[u \pm v]$ denoting the closed interval $[u - v,\, u + v]$, we have (ii - Decomposition):*

$$m_{n,c} \in m_0 + A_{n,c} + \left[\frac{V_{n,c} - \max_i C_{i,c}}{w_{n,c}}, \frac{V_{n,c} - \min_i C_{i,c}}{w_{n,c}}\right] \tag{8}$$

$$\in \underbrace{m_0 + A_{n,c}}_{\textit{defined as } m_n} + \underbrace{\left[\frac{\sqrt{V_{n,c}}(\sqrt{V_{n,c}} \pm \max_i \sqrt{V_i})}{w_{n,c}}\right]}_{\textit{net formability (epistemic) } \bar{\Phi}_{n,c}}. \tag{9}$$

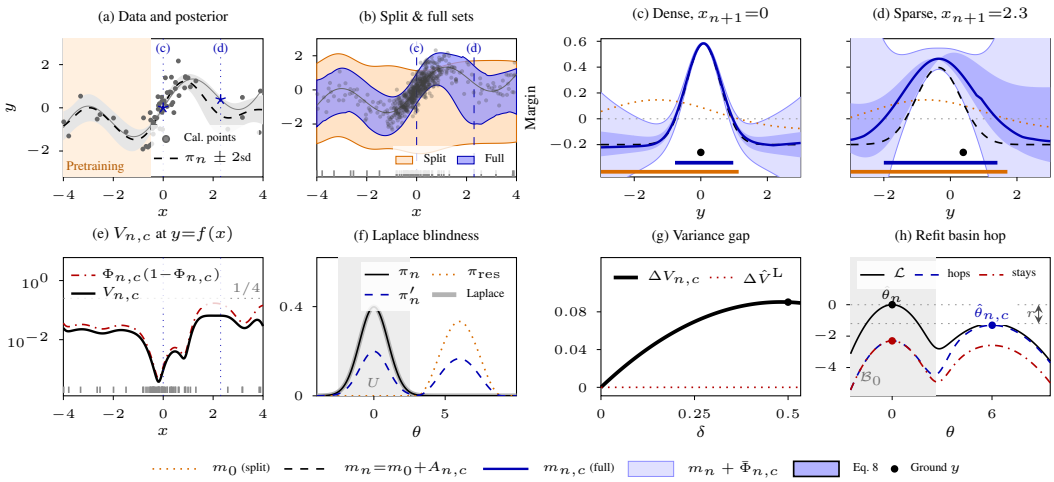

Figure 5: [a - e] Thm. 3.2 for Bayesian linear regression: [a] data, training covariate, and calibration points; [b] split and full sets along x; [c,d] the margins $m_\circ$ from Thm. 3.2 in dense and sparse area; [e] posterior variance, and its upper bound from Eq. 6. Details are in § C.1.3. [f-h] Illustration of Prop. 3.3, and Prop. 3.4: [f] same Laplace approximation on two different distributions: π_n (solid), and a second distribution π'_n (dashed); [g] the variance difference by increasing the impact of π_{res}. [h] The refit can jump to a different basin, i.e. $\pi_{n,c}$ can be outside the basin of the original MAP $\hat{\theta}_n$.

Thm. 3.2 shows that under Bayesian training, the set size for full CP can be bounded using only π_0, and π_n without recomputing $\pi_{n,c}$ over every label c. Fig. 5 evaluates this section exactly on a conjugate Bayesian linear regression (setup in § C.1.3), where all three posteriors are Gaussian and every term of Eq. 9 is a closed-form moment: the identities of Lemma 3.1 and Thm. 3.2(i) hold to machine precision, the full margin computed from the actual augmented posterior stays inside the band for every candidate, and the band is tight exactly where the calibration neighborhood is dense.

Consistent Bayesian learner. By increasing n (calibration set size) the posterior π_n concentrates. Lemma 3.1, and Thm. 3.2 tie formability, and quantile change to exactly the disagreement between plausible parameters from π_n. The concentration shrinks $\Phi_{n,c}$ and $T_{n,c}$ and the full margin $m_{n,c}$ converges to the calibration-posterior margin $m_{n,c} \to m_0 + A_{n,c}$ which is also observed in Fig. 1.

3.3 Deterministic Fine-Tuning

Thm. 3.2, and Lemma 3.1 are written for Bayesian learning (given posterior π_n), which is intractable to compute for deep models. A common regime for deterministic fine-tuning finds the maximizer of the (log-) posterior, i.e., $\hat{\boldsymbol{\theta}}_n := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$, for $\mathcal{L}(\boldsymbol{\theta}) := \log \pi_0(\boldsymbol{\theta}) + \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)$; e.g. the common ℓ_2-regularized objective assigns a Gaussian prior $\pi_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \tau^2 \boldsymbol{I})$. Daxberger et al. (2021) approximate π_n with a Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Sigma}}_n)$ centered at the MAP $\hat{\boldsymbol{\theta}}_n$, and covariance given by the inverse Hessian of the log-posterior; i.e. $\boldsymbol{\Sigma} = -\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}_n)^{-1}|_{\hat{\theta}}$, which approximates the variance of the *predictive* distribution as $V_{n,c} \approx \hat{V}^{\text{L}} := \nabla_{\boldsymbol{\theta}} p_{\hat{\boldsymbol{\theta}}_n}(c \mid \boldsymbol{x}_{n+1})^\top \hat{\boldsymbol{\Sigma}}_n \nabla_{\boldsymbol{\theta}} p_{\hat{\boldsymbol{\theta}}_n}(c \mid \boldsymbol{x}_{n+1})$. This approximation with a Gaussian distribution is the result of a second-order Taylor expansion of $\mathcal{L}$ around $\hat{\boldsymbol{\theta}}_n$. By construction, a shortcoming is that this approximation is always uni-modal (Gaussian), regardless of the shape of the true posterior (which can be multi-modal, skewed, etc.). In other words, there are dissimilar distributions (with different variances) that are all approximated by the same Gaussian distribution under this approach. Aware of this approximation scheme, Prop. 3.3 designs such dissimilar distributions.

Proposition 3.3. *Let π_n be twice differentiable at $\hat{\boldsymbol{\theta}}_n$, and U be an open neighborhood of $\hat{\boldsymbol{\theta}}_n$. Define $\eta := \pi_n(\hat{\boldsymbol{\theta}}_n) - \sup_{\boldsymbol{\theta} \notin U} \pi_n(\boldsymbol{\theta}) > 0$. Let π_{res} be a probability density that vanishes on U, and for $\delta \in (0,1)$ satisfy $\delta \|\pi_{\text{res}}\|_\infty < (1-\delta)\eta$, and set $\pi'_n := (1-\delta)\pi_n + \delta \pi_{\text{res}}$. Then:*

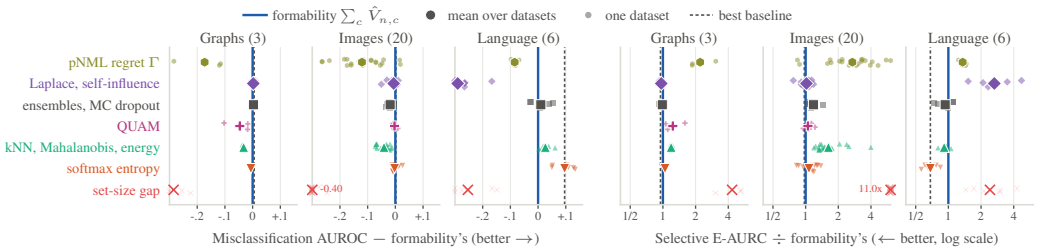

Figure 6: Formability against families of uncertainty measures, for in-distribution data, on graphs, images and language. For each modality the relative performance is gathered from different datasets and models. Each family of methods includes several variations (e.g. Laplace is last-layer and all-parameter linearised, last-layer sampled; self-influence). A detailed comparison is in Fig. 13.

(i) Every maximizer of π'_n lies in U and is a maximizer of π_n. In particular $\hat{\boldsymbol{\theta}}_n$ maximizes π'_n, and $\hat{V}^{\mathrm{L}}(\pi'_n) = \hat{V}^{\mathrm{L}}(\pi_n)$.

(ii) With $w_{\mathrm{res},c} := \mathbb{E}_{\boldsymbol{\theta}\sim\pi_{\mathrm{res}}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]$ *and* $V_{\mathrm{res},c} := \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_{\mathrm{res}}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]$,

$$V_{n,c}(\pi'_n) - V_{n,c}(\pi_n) \geq \delta\big[(1-\delta)(w_{\mathrm{res},c} - w_{n,c})^2 - V_{n,c}(\pi_n)\big]. \tag{10}$$

Part (i) shows that the Laplace approximation of the different distributions remains the same while (ii) the variance of the predictive distribution can differ. This is well known, and it is what motivates mixing several Laplace fits (Eschenhagen et al., 2021), ensembling across basins (Wilson & Izmailov, 2020), or searching directly for plausible parameters that disagree about the prediction (Schweighofer et al., 2023). Basically, the issue stems from the fact that the Bayesian posterior distribution can have multiple basins, and basins other than the one with $\hat{\theta}_n$ can contribute meaningfully to the predictive variance. Formability retrains the model over each candidate label, meaning that it can jump to the basin that is more likely to predict the candidate label. The following proposition shows that the refit can result in a maximum a posteriori estimate outside of the basin of the original MAP.

Proposition 3.4. *Let $\mathcal{B}_0$ be a basin containing $\hat{\boldsymbol{\theta}}_n$, and suppose that for some $r \geq 0$ and $0 < u < v \leq 1$, and some $\theta' \notin \mathcal{B}_0$ outside of the basin containing the MAP estimate, we have*

$$\sup_{\theta\in\mathcal{B}_0} p_\theta(c \mid \boldsymbol{x}_{n+1}) \leq u < v \leq p_{\theta'}(c \mid \boldsymbol{x}_{n+1}), \quad \textit{and} \quad \mathcal{L}(\hat{\theta}_n) - \mathcal{L}(\theta') \leq r.$$

Then, if $\log(v/u) > r$, *the augmented MAP* $\hat{\theta}_{n,c} \notin \mathcal{B}_0$. *In either case* $p_{\hat{\theta}_{n,c}}(c \mid \boldsymbol{x}_{n+1}) \geq e^{-r}v$.

4 Experiments

With our experiments we aim to show that (i) the performance gap between full and split conformal prediction is governed by the size of the calibration set, training set, and the dimension of the feature space; and that (ii) it stems from the model's formability which is directly connected to epistemic uncertainty (synthetic experiments). (iii) The formability itself can serve as a tool to quantify epistemic uncertainty, through the variance as computed in Eq. 6, $\hat{V}_{n,c} = \hat{w}_{n,c}\hat{\Phi}_{n,c}$. We compare formability against several existing uncertainty quantification methods: softmax entropy (Hendrycks & Gimpel, 2017), deep ensembles (Lakshminarayanan et al., 2017), MC-dropout (Gal & Ghahramani, 2016), Laplace approximation (Daxberger et al., 2021), and DeepKNN (Papernot & McDaniel, 2018). The closest methods to formability are pNML regret (Bibas et al., 2019), and the set size gap, against which we specifically compare, together with the best of the aforementioned methods in Fig. 7. Reporting against the best does not mean that there is a single best method at every level, but rather every method has its own sweet spot. Overall relative performance is reported in Fig. 6, and detailed reports are in § C.

For language models, refitting together over a calibration set and test point for each pseudo-label becomes unrealistically expensive. Therefore we approximate the refit with a single step of LoRA

(Hu et al., 2022) fine-tuning. Our updates are done through damped curvature-informed steps, and for calibration points we cache the gradients. We compare methods on six four-option question-answering setups (MMLU, MedQA on small Qwen, and Llama models). The specifics of the setup are in § C.2. Note that the language modeling setup is defined in a very simplistic way. The model is simply asked to choose one of the A–D answers and the score is the softmax over those specific tokens (see § C.2). The model updates are over a LoRA that is applied to the query and value matrices of the attention layers. These are specifically chosen for the computation to fit on a single GPU. The results are also not comparable to the image and graph setups. Defining a clear formability signal for language models is an open question, and we leave it for future work.

Throughout this paper we assumed the TPS score function (Sadinle et al., 2019) $s(\boldsymbol{x}, y) = p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$, and $1 - \alpha = 0.9$ unless otherwise specified. While CP is commonly phrased with non-conformity (disagreement) scores, we use agreement as it simplifies the derivations, and interpretation. Setups are equivalent through a change in sign. For uncertainty quantification, we use two main metrics: (i) the misclassification AUROC (Hendrycks & Gimpel, 2017) which is the area under the curve of true positive rate vs false positive rate over rejection according to the uncertainty score function; 0.5 is chance and 1.0 is the perfect score. (ii) The selective E-AURC (Geifman et al., 2019) is the difference in AURC between the uncertainty score, and the ideal uncertainty score (the oracle). In this score lower is better, and 0 is the best score.

Full-vs-Split CP. In Fig. 1 we compare full and split conformal prediction on node and image-classification tasks. For dimensionality reduction, we apply PCA on input features. Specifically on `CIFAR-10`, the features are the embeddings from an `ImageNet`-pretrained ResNet-18 (He et al., 2016) backbone, (training is done on the two-layer MLP as classification head). Since full CP is computationally expensive to examine, we report these results on a random subset of the test set containing $300 \sim 400$ instances. Prior to observing test examples, we can approximate the full CP set size by a small (e.g. $m = 50$) number of calibration points over leave-one-out refits (keeping $n-1$ points as calibration set and the left out point as test set). We run ablation study and show in Fig. 2, that by increasing calibration size, training size, or decreasing the dimension of the feature space, the full CP outperforms split CP, and it converges to the adaptation lower bound $|\mathcal{C}_B|$ (consistent with Thm. 3.2). With a larger training set, the effect of calibration set shows a diminishing return.

Synthetic experiment. In Fig. 3 we visualize the concept discussed in § 3. If a point is far from the calibration set it can deform the model more easily, its score can increase more, and full conformal set size would also inflate. It also shows the three variables defined in Eq. 2. To synthesize the data, we sample three classes from overlapping Gaussians, and use a multinomial logistic regression over fixed RBF features as the model. The pretrained model $\boldsymbol{\theta}_0$ is trained on only 3 points per class, and every fine-tuning minimizes ℓ_2-regularized cross-entropy towards $\boldsymbol{\theta}_0$ (i.e. the MAP objective with the prior $\mathcal{N}(\boldsymbol{\theta}_0, \tau^2 \boldsymbol{I})$). We fine-tune once on the calibration set ($n = 42$) to obtain p_n, and once per candidate label c on $\mathcal{D}_{n+1}^{(c)}$ for $p_{n,c}$. The complete anatomy for every candidate label is in § C.1.1. An important extension of this experiment is shown in Fig. 4, where we move the Gaussian generators towards the center which reduces epistemic uncertainty, and increases aleatoric uncertainty at the center. Here the softmax entropy remains steady, while the formability variance decreases. This shows that formability directly distinguishes the epistemic uncertainty, while the softmax entropy is agnostic to this decomposition.

In Fig. 5 we evaluate Thm. 3.2 in a setup where every posterior can be computed exactly. We sample from $y = \sin(1.5x) + 0.3x + \varepsilon$, with Gaussian noise of standard deviation 0.5, and use Bayesian linear regression over fixed RBF features. The model is pretrained on 40 points in $[-4, -0.5]$, while the 60 calibration points concentrate around $x = 0$. The full sets are narrower than split, especially where calibration data is dense. In sparse regions, the posterior variance and formability are larger, and the full set expands relative to the model fitted on calibration data alone. The full margins remain within the theoretical bounds at both test points. Details are in § C.1.3.

Formability as a quantification of epistemic uncertainty. While refitting on every label as an uncertainty quantification is not an entirely new concept (Bibas et al. (2019) define regret based on refitting to every candidate label), we connect the refit through formability to Bayesian posterior variance, and the uncertainty captured in full conformal prediction. While both we and Bibas et al. (2019) take information from models fitted on pseudo-labels, they take the logarithm-sum of the candidate probabilities, namely, $\log\big(1 + \sum_{c \in \mathcal{Y}} \Phi_{n,c}\big)$, while we directly estimate the posterior variance $\Phi_{n,c} \cdot w_{n,c}$. This is derived directly from CP's uncertainty. Interestingly, as shown in Fig. 7, this

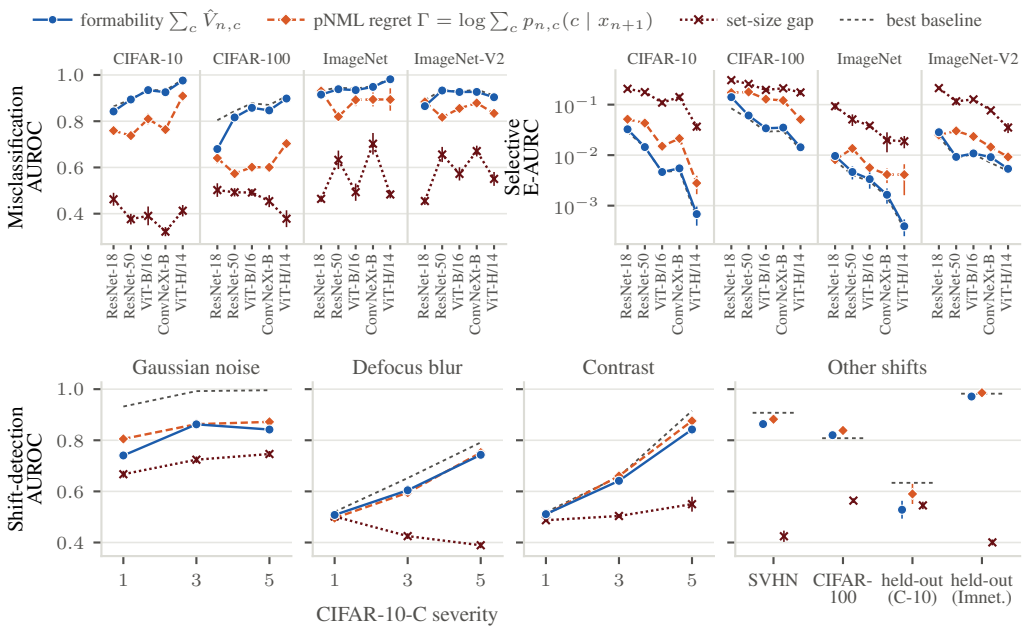

Figure 7: Comparison of formability, regret and conformal set size gap over image benchmarks.

minor change in the definition can result in a different evaluation on the data. Formability performs significantly better on in-domain error detection, while the regret-based uncertainty performs better on out-of-distribution data. We compare all methods briefly in Fig. 6. The CP set size gap (full - split) itself is not a good uncertainty quantification; mainly because it quantizes all the information captured in Eq. 6, which is a continuous value, into a binary decision per label.

5 Conclusion

We investigated the performance gap between full, and split conformal prediction. We showed that the gap is the resultant of two opposite effects: adapting the model on the calibration set (more data resulting in better model, and hence smaller sets), and forming it towards each pseudo-label at the test point. Under Bayesian learning, the second effect (formability) is directly tied to posterior predictive variance, and fades as the calibration posterior concentrates. This explains why full CP can become more efficient when the model is already well determined around the test point, and in sparse regions it performs poorly. However, this gap suggest to use formability as a measure of epistemic uncertainty that carries much more information compared to set size (due to the 0/1 binarization). Our experiments on images and graphs show its value for detecting prediction errors; and to some degree, ood detection. In many benchmarks, formability is on par with the best of the existing uncertainty quantification methods, and in some cases it outperforms them.

Limitations. The most important limitation is the computational cost of refitting the model on every candidate label. Additionally our analysis, and definition of formability is based on the softmax score function. CP itself is agnostic to the choice of score function, and scores like APS also can lead to meaningful interpretations of the full-vs-split gap. We leave this for future work. Additionally, conformal prediction assumes exchangeability of the data, which is not always satisfied in practice, specifically for uncertainty quantification use cases. Formability implicitly assume that refits are able to reach the MAP estimate. This could challenge the quality in overparameterized models, or in wrong early stopping scenarios. Potentially our language model experiments are affected by this.

AI USE STATEMENT

We used generative AI tools for implementing the experiments (we designed the experimental setup ourselves and asked AI to implement and run them. We also inspected the results for potential issues). We also used AI for rephrasing parts of the paper, from grammar and choice of words, to better presenting the proofs. The proofs are ultimately rewritten by the authors, but AI was used to assist in the process. The research idea was originally proposed by the authors, including the main idea, theory, and the primary design of the experiments.

ETHICS STATEMENT

This work does not raise any ethical concerns, as it is a general study, and a theoretical contribution to uncertainty quantification. The experiments are conducted on publicly available datasets, and the models are trained on these datasets. The authors have no conflicts of interest to disclose.

REPRODUCIBILITY STATEMENT

All datasets, and models used in the experiments are publicly available. The code for reproducing the experiments is also shared in the supplementary material of this paper.

REFERENCES

Javier Abad Martinez, Umang Bhatt, Adrian Weller, and Giovanni Cherubin. Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6631–6639, 2023. doi: 10.1609/aaai.v37i6.25814. URL https://ojs.aaai.org/index.php/AAAI/article/view/25814.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024. URL https://arxiv.org/abs/2411.11824.

Anastasios Nikolas Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *International Conference on Learning Representations*, 2021. URL https://api.semanticscholar.org/CorpusID:221995507.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965. URL https://arxiv.org/abs/1905.02928.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Umang Bhatt, Adrian Weller, and Giovanni Cherubin. Fast conformal classification using influence functions. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 303–305, 2021. URL https://proceedings.mlr.press/v152/bhatt21a.html.

Koby Bibas, Yaniv Fogel, and Meir Feder. Deep pNML: Predictive normalized maximum likelihood for deep neural networks. *arXiv preprint arXiv:1904.12286*, 2019. URL https://arxiv.org/abs/1904.12286.

Aleksandar Bojchevski and Stephan Günnemann. Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1ZdKJ-0W.

Siu Lun Chau, Soroush H. Zargarbashi, Yusuf Sale, and Michele Caprio. Quantifying epistemic predictive uncertainty in conformal prediction. *arXiv preprint arXiv:2602.01667*, 2026. URL https://arxiv.org/abs/2602.01667.

Alvaro H. C. Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/b6fa3ed9624c184bd73e435123bd576a-Abstract-Conference.html.

Luben Miguel Cruz Cabezas, Vagner Silva Santos, Thiago Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 443–470. PMLR, 2025. URL https://proceedings.mlr.press/v286/cruz-cabezas25a.html.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/a7c9585703d275249f30a088cebba0ad-Abstract.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Runa Eschenhagen, Erik Daxberger, Philipp Hennig, and Agustinus Kristiadi. Mixtures of Laplace approximations for improved post-hoc uncertainty in deep learning. *Bayesian Deep Learning Workshop, NeurIPS*, 2021. URL https://arxiv.org/abs/2111.03577.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:235458009.

Alireza Javanmardi, Soroush H. Zargarbashi, Santo M. A. R. Thies, Willem Waegeman, Aleksandar Bojchevski, and Eyke Hüllermeier. Optimal conformal prediction under epistemic uncertainty. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 2461–2479. PMLR, 2026. URL https://proceedings.mlr.press/v337/javanmardi26a.html.

Sebastián Jiménez, Mira Jürgens, and Willem Waegeman. Why machine learning models fail to fully capture epistemic uncertainty. *arXiv preprint arXiv:2505.23506*, 2025. URL https://arxiv.org/abs/2505.23506v1.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.nips.cc/paper_files/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1612.01474.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://arxiv.org/abs/1604.04173.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996. URL https://link.springer.com/book/10.1007/978-1-4612-0745-0.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018. URL https://arxiv.org/abs/1803.04765.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. URL https://arxiv.org/abs/1609.00451.

Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 3159–3179. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/sale24a.html.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/3e0b96206965f5f05b0b4550c0e73ff0-Abstract-Conference.html.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 804–814, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. URL https://alrw.net/.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/322f62469c5e3c7dc3e58f5a4d1ea399-Abstract.html.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12292–12318, 2023. URL https://proceedings.mlr.press/v202/h-zargarbashi23a.html.

Yuwei Zong and Yanwen Xu. Comparative analysis of conformal prediction: Split, full, and adaptive approaches for statistical and neural network models. *Proceedings of the Design Society*, 5:981–990, 2025. doi: 10.1017/pds.2025.10112. URL https://doi.org/10.1017/pds.2025.10112.

A Related Works

Conformal prediction. CP (Vovk et al., 2005; Angelopoulos et al., 2024) is considered a distribution-free uncertainty quantification technique, where the prediction sets are guaranteed to cover the true label with high probability. With that, many consider the size of the prediction set to represent how uncertain the model is on a specific input. While the method works with any model, and any score function, there are many works designing various scores to improve the set size efficiency, adaptivity, and approximation of conditional coverage (Romano et al., 2020; Angelopoulos et al., 2021; Zargarbashi et al., 2023). We use the basic threshold prediction sets (Sadinle et al., 2019) since it directly connects to the predictive distribution of the model which can connect perfectly with our decomposition in § 3, and through which we bound the epistemic uncertainty. In § C.1.6 we repeat our experiments under an adaptive (rank-based) score, where the full–split gap is wider, and never closes. While full and split conformal prediction are two endpoints of a computational cost spectrum, there are other variants such as Jackknife+ (Barber et al., 2021) in between. There are techniques that approximate full CP with lower computational cost (Bhatt et al., 2021; Abad Martinez et al., 2023), through influence functions (Koh & Liang, 2017). Both directions are outside of our focus, as we address the fundamental difference between full and split CP.

CP and epistemic uncertainty. Javanmardi et al. (2026); Cruz Cabezas et al. (2025) incorporate a second-order distribution (which contains explicit epistemic uncertainty) into conformal prediction. In the reverse direction, Chau et al. (2026); Correia et al. (2024) provide interpretations of split conformal prediction in uncertainty quantification; one through connecting split CP to imprecise integral probability metrics, and the other through information theory. In contrast, we provide a direct connection between the difference of full and split CP set size and the epistemic uncertainty, by bounding the posterior variance of the predictive distribution.

Our work connects three arguments: (i) When does full CP outperform split CP? Inconsistencies in comparison are mentioned by Barber et al. (2023) and extensively studied by Zong & Xu (2025). However, neither work discusses what the underlying factor is that governs this difference. We show that the unexpected inefficiency in full CP set size stems from the formability of the model at the test point. (ii) We show that, in a Bayesian framework, the formability is directly related to the posterior variance (epistemic uncertainty). The concept of *deterministic* formability itself is not entirely new. Bibas et al. (2019) define a learnability measure (regret) as the log sum of the predictive probabilities $\hat{\theta}_{n,c}$ normalized by the probabilities from $\hat{\theta}_n$. However, its connection with Bayesian posterior variance, comparison with deterministic Laplace approximation through basin decomposition (Daxberger et al., 2021), or its influence on full conformal prediction do not follow automatically. (iii) We show that on larger models and various tasks this notion of formability still can be used as a practical tool to quantify epistemic uncertainty. Our work also connects with many other uncertainty quantification approaches. While the posterior distribution of parameters is a canonical way to quantify epistemic uncertainty, on larger models it is computationally near-impossible to compute. Laplace Redux (Daxberger et al., 2021) approximates this distribution as a Gaussian around the MAP estimate. However, as we show in Prop. 3.3 it is agnostic to other basins of attraction, and therefore miscalculates the epistemic uncertainty where other highly plausible basins disagree with the current predictive distribution. A remedy to that is MC-dropout (Gal & Ghahramani, 2016), and ensemble training (Lakshminarayanan et al., 2017), which are parameter centric approximations. Formability however is a data-centric approximation. On computation, the comparison depends on whether multiple model inference or test-time training is more favorable. Schweighofer et al. (2023) instead searches for adversarial models with high disagreement to the baseline, while attaining high posterior plausibility. In comparison, formability is a local measure, searching for such models through retraining. The comparison results as shown in Fig. 17 are very close. Our ultimate goal is also not to come up with a new notion of uncertainty, but to showcase what is the underlying uncertainty quantification in conformal prediction.

B Supplementary to Theory

We first prove the sequential updating property used in § 3.2.

Lemma B.1 (Sequential Bayesian updating)**.** *Let π_n be the posterior from the prior π_0 and evidence $\mathcal{D}_n$, and $\pi_{n,c}$ be the posterior from the prior π_0 and evidence $\mathcal{D}_{n+1}^{(c)}$. For a fixed $\boldsymbol{x}_{n+1}$, assume its*

label to be independent of $\mathcal{D}_n$ given $\boldsymbol{\theta}$, and $w_{n,c} = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] > 0$ (recalling Eq. 3). Then $\pi_{n,c}$ equals the posterior computed from the prior π_n on the single observation $\{(\boldsymbol{x}_{n+1}, c)\}$; i.e.,

$$\pi_{n,c}(\boldsymbol{\theta}) = \frac{p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\pi_n(\boldsymbol{\theta})}{\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]} = \frac{p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\pi_n(\boldsymbol{\theta})}{w_{n,c}}. \tag{11}$$

Proof. For two sets of evidence E_1 and E_2, Bayes' rule with the joint evidence gives

$$p(\boldsymbol{\theta} \mid E_1, E_2) = \frac{p(E_1, E_2 \mid \boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta})}{p(E_1, E_2)} = \frac{p(E_2 \mid \boldsymbol{\theta}, E_1)\,p(E_1 \mid \boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta})}{p(E_2 \mid E_1)\,p(E_1)},$$

Bayes' rule on E_1 alone yields $p(E_1 \mid \boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid E_1)\,p(E_1)$, and the law of total probability expands $p(E_2 \mid E_1) = \int p(E_2 \mid \vartheta, E_1)\,p(\vartheta \mid E_1)\,d\vartheta$. Substituting both,

$$p(\boldsymbol{\theta} \mid E_1, E_2) = \frac{p(E_2 \mid \boldsymbol{\theta}, E_1)\,p(\boldsymbol{\theta} \mid E_1)}{\int p(E_2 \mid \vartheta, E_1)\,p(\vartheta \mid E_1)\,d\vartheta}. \tag{12}$$

This means that the posterior after observing E_1 can be used as the prior when conditioning on E_2, and it results in the same posterior as conditioning jointly on both sets of evidence.

In our case, $E_1 = \mathcal{D}_n$ and $E_2 = \{(\boldsymbol{x}_{n+1}, c)\}$, therefore $p(\boldsymbol{\theta} \mid E_1) = \pi_n(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid E_1, E_2) = \pi_{n,c}(\boldsymbol{\theta})$. With $\boldsymbol{x}_{n+1}$ fixed, we have $p(E_2 \mid \boldsymbol{\theta}, E_1) = p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})$, and the denominator of Eq. 12 becomes $\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] = w_{n,c}$, which is positive by assumption. □

Lemma 3.1. *The (per-label) variance of the calibration posterior predictive distribution at $\boldsymbol{x}_{n+1}$ is proportional to the formability at $\boldsymbol{x}_{n+1}$:*

$$\mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] = w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}), \quad \textit{and} \tag{5}$$

therefore, $0 \le \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] \le \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big)$.

As a result, for a finite label set, the total posterior variance (epistemic uncertainty) satisfies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\Big[\|p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}_{n+1}) - p_{\pi_n}(\cdot \mid \boldsymbol{x}_{n+1})\|_2^2\Big] &= \sum_{c\in\mathcal{Y}} w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}) \\ &\le \sum_{c\in\mathcal{Y}} \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big).\end{aligned} \tag{6}$$

Proof. Using Eq. 11,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\pi_{n,c}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] &= \int p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\pi_{n,c}(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= \int p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\frac{p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\pi_n(\boldsymbol{\theta})}{w_{n,c}}\,d\boldsymbol{\theta} = \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})^2]}{w_{n,c}}.\end{aligned}$$

Substituting it into Eq. 4 yields

$$\begin{aligned}\Phi_{n,c}(\boldsymbol{x}_{n+1}) &= \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})^2]}{w_{n,c}} - w_{n,c} \\ &= \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})^2] - w_{n,c}^2}{w_{n,c}} = \frac{\mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]}{w_{n,c}}.\end{aligned}$$

Multiplying both sides by $w_{n,c}$ results in Eq. 5.

For the remainder of the proof, we will use the following rule: for any $X \in [0, 1]$ we have

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \le \mathbb{E}[X] - (\mathbb{E}[X])^2 = \mathbb{E}[X](1 - \mathbb{E}[X]).$$

As $0 \le p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1}) \le 1$, its posterior variance is bounded by

$$\mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] \le w_{n,c}(1 - w_{n,c}).$$

Replacing the variance in Eq. 5 with this inequality gives $\Phi_{n,c}(\boldsymbol{x}_{n+1}) \leq 1 - w_{n,c}$, and hence $w_{n,c} \leq 1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})$. Therefore,

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] &= w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}) \\ &\leq \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big).\end{aligned}$$

For a finite number of labels, the expected squared Euclidean norm is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\Big[\|p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}_{n+1}) - p_{\pi_n}(\cdot \mid \boldsymbol{x}_{n+1})\|_2^2\Big] &= \sum_{c\in\mathcal{Y}} \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\Big[\big(p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1}) - w_{n,c}\big)^2\Big] \\ &= \sum_{c\in\mathcal{Y}} \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] \\ &= \sum_{c\in\mathcal{Y}} w_{n,c}\Phi_{n,c}(\boldsymbol{x}_{n+1}) \\ &\leq \sum_{c\in\mathcal{Y}} \Phi_{n,c}(\boldsymbol{x}_{n+1})\big(1 - \Phi_{n,c}(\boldsymbol{x}_{n+1})\big),\end{aligned}$$

which proves Eq. 6.

□

B.1 Proof of Thm. 3.2

Before proving Thm. 3.2, we first prove two lemmas: (i) Lemma B.2 shows how conditioning on the pseudo-labeled test point moves the posterior expectation w.r.t. the expectation under the prior. In our case the posterior is the augmented posterior, and the prior is the calibration posterior. (ii) Lemma B.3 records the stability of the conformal quantile under coordinatewise shifts of the scores. The proof relies on two auxiliary lemmas: the first shows how conditioning on the pseudo-labeled test point moves *any* posterior expectation, and the second records the stability of the conformal quantile under coordinatewise shifts of the scores.

Lemma B.2. *For any function $h : \Theta \to \mathbb{R}$ with $\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}|h(\boldsymbol{\theta})| < \infty$,*

$$\mathbb{E}_{\boldsymbol{\theta}\sim\pi_{n,c}}[h(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[h(\boldsymbol{\theta})] + \frac{\mathrm{Cov}_{\boldsymbol{\theta}\sim\pi_n}\big(h(\boldsymbol{\theta}),\, p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\big)}{w_{n,c}}. \tag{13}$$

Proof. Using Eq. 11,

$$\mathbb{E}_{\boldsymbol{\theta}\sim\pi_{n,c}}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta})\frac{p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\pi_n(\boldsymbol{\theta})}{w_{n,c}}\, d\boldsymbol{\theta} = \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\big[h(\boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\big]}{w_{n,c}}.$$

By the definition of covariance,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}\big[h(\boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\big] &= \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[h(\boldsymbol{\theta})]\, \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] \\ &\quad + \mathrm{Cov}_{\boldsymbol{\theta}\sim\pi_n}\big(h(\boldsymbol{\theta}),\, p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\big),\end{aligned}$$

and $\mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})] = w_{n,c}$. Dividing by $w_{n,c}$ completes the proof. □

Note that Eq. 5 in Lemma 3.1 is the special case $h(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})$, since the covariance of a variable with itself is its variance.

Lemma B.3 (Quantile stability under score shifts)**.** *The map $r_{1:n} \mapsto Q_{\alpha,n}(r_{1:n})$ defined in Eq. 1 is non-decreasing in every coordinate, and translation equivariant whenever the quantile is finite, i.e., $Q_{\alpha,n}(r_{1:n} + t\mathbf{1}) = Q_{\alpha,n}(r_{1:n}) + t$ for every $t \in \mathbb{R}$. For any shift vector $\delta_{1:n}$,*

$$Q_{\alpha,n}(r_{1:n}) + \min_i \delta_i \;\leq\; Q_{\alpha,n}(r_{1:n} + \delta_{1:n}) \;\leq\; Q_{\alpha,n}(r_{1:n}) + \max_i \delta_i. \tag{14}$$

Proof. $Q_{\alpha,n}(r_{1:n})$ is the α-quantile of the empirical distribution over the multiset $\{r_1, \ldots, r_n, \infty\}$, and hence returns the k-th smallest element of this multiset, where the rank k is fixed by α and n (with $Q_{\alpha,n} = \infty$ when $k = n+1$; we assume α and n are such that the quantile is finite). Increasing

any coordinate cannot decrease any order statistic, which gives monotonicity, and adding the same t to every score adds t to every finite order statistic, which gives translation equivariance. For the sandwich, note that coordinatewise $r_i + \min_j \delta_j \le r_i + \delta_i \le r_i + \max_j \delta_j$. Applying monotonicity and then translation equivariance to both sides yields Eq. 14. □

Theorem 3.2. *Under Bayesian training, with scores $r_i^{(n)} = p_{\pi_n}(y_i \mid \boldsymbol{x}_i)$, $r_i^{(c)} = p_{\pi_{n,c}}(y_i \mid \boldsymbol{x}_i)$ and $q_{\boldsymbol{\theta}}^{(n)} = Q_{\alpha,n}(r_{1:n}^{(n)})$, $q_{\boldsymbol{\theta}}^{(c)} = Q_{\alpha,n}(r_{1:n}^{(c)})$, the following statements hold: (i) the update in each calibration score r_i is exactly $r_i^{(c)} - r_i^{(n)} = \frac{C_{i,c}}{w_{n,c}}$ and bounded in absolute value by $\frac{\sqrt{V_i\, V_{n,c}}}{w_{n,c}}$. As a result, with $[u \pm v]$ denoting the closed interval $[u - v,\, u + v]$, we have (ii - Decomposition):*

$$m_{n,c} \in m_0 + A_{n,c} + \left[\frac{V_{n,c} - \max_i C_{i,c}}{w_{n,c}},\, \frac{V_{n,c} - \min_i C_{i,c}}{w_{n,c}}\right] \tag{8}$$

$$\in \underbrace{m_0 + A_{n,c}}_{\text{defined as } m_n} + \underbrace{\left[\frac{\sqrt{V_{n,c}}\big(\sqrt{V_{n,c}} \pm \max_i \sqrt{V_i}\big)}{w_{n,c}}\right]}_{\text{net formability (epistemic) } \bar{\Phi}_{n,c}}. \tag{9}$$

Proof. **Part (i).** Under Bayesian training, the calibration scores are posterior predictive probabilities: $r_i^{(n)} = p_{\pi_n}(y_i \mid \boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_n}[p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)]$ and $r_i^{(c)} = p_{\pi_{n,c}}(y_i \mid \boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_{n,c}}[p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)]$. Applying Lemma B.2 with $h(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i)$ gives

$$r_i^{(c)} - r_i^{(n)} = \frac{\mathrm{Cov}_{\boldsymbol{\theta}\sim\pi_n}\big(p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i),\, p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})\big)}{w_{n,c}} = \frac{C_{i,c}}{w_{n,c}}.$$

The bound $|C_{i,c}| \le \sqrt{V_i V_{n,c}}$ is the Cauchy–Schwarz inequality for the covariance.

Part (ii). We first bound the threshold drift. By definition, $T_{n,c} = q_{\boldsymbol{\theta}}^{(c)} - q_{\boldsymbol{\theta}}^{(n)}$ with $q_{\boldsymbol{\theta}}^{(c)} = Q_{\alpha,n}(r_{1:n}^{(c)})$ and $q_{\boldsymbol{\theta}}^{(n)} = Q_{\alpha,n}(r_{1:n}^{(n)})$. Defining δ_i as the difference in the scores after the update, we have $r_{1:n}^{(c)} = r_{1:n}^{(n)} + \delta_{1:n}$ with $\delta_i = C_{i,c}/w_{n,c}$, so Lemma B.3 gives

$$\frac{\min_i C_{i,c}}{w_{n,c}} \le T_{n,c} \le \frac{\max_i C_{i,c}}{w_{n,c}}, \qquad \text{hence} \qquad |T_{n,c}| \le \frac{\max_i |C_{i,c}|}{w_{n,c}} \le \frac{\sqrt{V_{n,c}}\max_i \sqrt{V_i}}{w_{n,c}}, \tag{15}$$

where the last step is the Cauchy–Schwarz bound of part (i).

Next, Eq. 2 is an algebraic identity that holds for any three fits: $m_{n,c} = m_0 + A_{n,c} + \Phi_{n,c} - T_{n,c}$. Under Bayesian refitting where the formability is defined as Eq. 4, $\Phi_{n,c} = p_{\pi_{n,c}}(c \mid \boldsymbol{x}_{n+1}) - p_{\pi_n}(c \mid \boldsymbol{x}_{n+1})$, we write $\Phi_{n,c} - T_{n,c} = \frac{V_{n,c}}{w_{n,c}} - T_{n,c}$ (given $\Phi_{n,c} = V_{n,c}/w_{n,c}$). Therefore, $m_{n,c} = m_0 + A_{n,c} + V_{n,c}/w_{n,c} - T_{n,c}$. By Eq. 15, we have

$$\begin{aligned}\frac{V_{n,c}}{w_{n,c}} - T_{n,c} &\in \left[\frac{V_{n,c} - \sqrt{V_{n,c}}\max_i \sqrt{V_i}}{w_{n,c}},\ \frac{V_{n,c} + \sqrt{V_{n,c}}\max_i \sqrt{V_i}}{w_{n,c}}\right] \\ &= \left[\frac{\sqrt{V_{n,c}}\big(\sqrt{V_{n,c}} \pm \max_i \sqrt{V_i}\big)}{w_{n,c}}\right] = \bar{\Phi}_{n,c},\end{aligned}$$

which proves Eq. 9. □

B.2 Proof of Prop. 3.3

Proposition 3.3. *Let π_n be twice differentiable at $\hat{\boldsymbol{\theta}}_n$, and U be an open neighborhood of $\hat{\boldsymbol{\theta}}_n$. Define $\eta := \pi_n(\hat{\boldsymbol{\theta}}_n) - \sup_{\boldsymbol{\theta}\notin U} \pi_n(\boldsymbol{\theta}) > 0$. Let π_{res} be a probability density that vanishes on U, and for $\delta \in (0,1)$ satisfy $\delta\|\pi_{\mathrm{res}}\|_\infty < (1-\delta)\eta$, and set $\pi_n' := (1-\delta)\pi_n + \delta\pi_{\mathrm{res}}$. Then:*

(i) Every maximizer of π_n' lies in U and is a maximizer of π_n. In particular $\hat{\boldsymbol{\theta}}_n$ maximizes π_n', and $\hat{V}^{\mathrm{L}}(\pi_n') = \hat{V}^{\mathrm{L}}(\pi_n)$.

(ii) With $w_{\mathrm{res},c} := \mathbb{E}_{\boldsymbol{\theta}\sim\pi_{\mathrm{res}}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]$ and $V_{\mathrm{res},c} := \mathrm{Var}_{\boldsymbol{\theta}\sim\pi_{\mathrm{res}}}[p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})]$,

$$V_{n,c}(\pi_n') - V_{n,c}(\pi_n) \ge \delta\big[(1-\delta)(w_{\mathrm{res},c} - w_{n,c})^2 - V_{n,c}(\pi_n)\big]. \tag{10}$$

Proof. (i) On U, $\pi_{\text{res}} = 0$, so $\pi'_n = (1-\delta)\pi_n \leq (1-\delta)\pi_n(\hat{\boldsymbol{\theta}}_n) = \pi'_n(\hat{\boldsymbol{\theta}}_n)$, which means that π'_n is maximized at the same points as π_n in U. Outside U,

$$\pi'_n(\boldsymbol{\theta}) \leq (1-\delta)\big(\pi_n(\hat{\boldsymbol{\theta}}_n) - \eta\big) + \delta\|\pi_{\text{res}}\|_\infty < (1-\delta)\,\pi_n(\hat{\boldsymbol{\theta}}_n) = \pi'_n(\hat{\boldsymbol{\theta}}_n).$$

Hence, the maximizers of π'_n, and π_n are the same in U, and $\hat{\boldsymbol{\theta}}_n$ is one of them. Since $\log \pi'_n = \log(1-\delta) + \log \pi_n$ on the open set U (meaning that the distribution is a scaled version of π_n in that region), the two log-densities have the same Hessian at $\hat{\boldsymbol{\theta}}_n$, hence the same $\hat{\boldsymbol{\Sigma}}_n$ (the $\log(1-\delta)$ disappears when differentiating). The gradient $\nabla_{\boldsymbol{\theta}} p_{\hat{\boldsymbol{\theta}}_n}(c \mid \boldsymbol{x}_{n+1})$ depends only on the model and on $\hat{\boldsymbol{\theta}}_n$. Therefore $\hat{V}^{\text{L}}(\pi'_n) = \hat{V}^{\text{L}}(\pi_n)$.

(ii) Write $f(\boldsymbol{\theta}) := p_{\boldsymbol{\theta}}(c \mid \boldsymbol{x}_{n+1})$. Expectations are linear in the density, so

$$\mathbb{E}_{\pi'_n}[f] = (1-\delta)\, w_{n,c} + \delta\, w_{\text{res},c}, \qquad \mathbb{E}_{\pi'_n}[f^2] = (1-\delta)\big(V_{n,c}(\pi_n) + w_{n,c}^2\big) + \delta\big(V_{\text{res},c} + w_{\text{res},c}^2\big).$$

Subtracting the square of the first from the second,

$$\begin{aligned} V_{n,c}(\pi'_n) &= (1-\delta)\, V_{n,c}(\pi_n) + \delta\, V_{\text{res},c} + (1-\delta)\, w_{n,c}^2 + \delta\, w_{\text{res},c}^2 - \big((1-\delta)\, w_{n,c} + \delta\, w_{\text{res},c}\big)^2, \\ &= (1-\delta)\, V_{n,c}(\pi_n) + \delta\, V_{\text{res},c} + \delta(1-\delta)(w_{\text{res},c} - w_{n,c})^2. \end{aligned}$$

where the last three terms equal $\delta(1-\delta)\, w_{n,c}^2 + \delta(1-\delta)\, w_{\text{res},c}^2 - 2\delta(1-\delta)\, w_{n,c} w_{\text{res},c} = \delta(1-\delta)(w_{\text{res},c} - w_{n,c})^2$. Subtracting $V_{n,c}(\pi_n)$ from both sides results in Eq. 10, and the inequality follows from $V_{\text{res},c} \geq 0$. □

B.3 Proof of Prop. 3.4

Proposition 3.4. *Let $\mathcal{B}_0$ be a basin containing $\hat{\boldsymbol{\theta}}_n$, and suppose that for some $r \geq 0$ and $0 < u < v \leq 1$, and some $\theta' \notin \mathcal{B}_0$ outside of the basin containing the MAP estimate, we have*

$$\sup_{\theta \in \mathcal{B}_0} p_\theta(c \mid \boldsymbol{x}_{n+1}) \leq u < v \leq p_{\theta'}(c \mid \boldsymbol{x}_{n+1}), \quad \textit{and} \quad \mathcal{L}(\hat{\theta}_n) - \mathcal{L}(\theta') \leq r.$$

Then, if $\log(v/u) > r$, the augmented MAP $\hat{\theta}_{n,c} \notin \mathcal{B}_0$. In either case $p_{\hat{\theta}_{n,c}}(c \mid \boldsymbol{x}_{n+1}) \geq e^{-r} v$.

Proof. Adding the candidate $(\boldsymbol{x}_{n+1}, c)$ to the calibration set adds $\log p_\theta(c \mid \boldsymbol{x}_{n+1})$ to the log posterior. Thus the augmented MAP $\hat{\theta}_{n,c}$ globally maximizes $\mathcal{L}(\theta) + \log p_\theta(c \mid \boldsymbol{x}_{n+1})$. With $\mathcal{L}(\theta') \geq \mathcal{L}(\hat{\theta}_n) - r$, and $\log p_{\theta'}(c \mid \boldsymbol{x}_{n+1}) \geq \log v$, we have:

$$\begin{aligned} \mathcal{L}(\hat{\theta}_{n,c}) + \log p_{\hat{\theta}_{n,c}}(c \mid \boldsymbol{x}_{n+1}) &\geq \mathcal{L}(\theta') + \log p_{\theta'}(c \mid \boldsymbol{x}_{n+1}) \\ &\geq \mathcal{L}(\hat{\theta}_n) - r + \log v. \end{aligned}$$

Since $\hat{\theta}_n$ maximizes the calibration log posterior, $\mathcal{L}(\hat{\theta}_{n,c}) \leq \mathcal{L}(\hat{\theta}_n)$. Subtracting $\mathcal{L}(\hat{\theta}_{n,c})$ from the preceding inequality therefore yields $\log p_{\hat{\theta}_{n,c}}(c \mid \boldsymbol{x}_{n+1}) \geq \log v - r$, or $p_{\hat{\theta}_{n,c}}(c \mid \boldsymbol{x}_{n+1}) \geq e^{-r} v$. This bound holds whether or not $\log(v/u) > r$. If that condition holds, then $e^{-r} v > u$, whereas every model in $\mathcal{B}_0$ assigns c probability at most u. Hence $\hat{\theta}_{n,c} \notin \mathcal{B}_0$. □

C Supplementary Results, and Experiments

C.1 Synthetic Experiments

C.1.1 Anatomy of the Split-to-Full CP

This experiment replicates the transition from split to full conformal prediction in a two-dimensional setup where every quantity of Eq. 2 is directly visible. Fig. 3 in § 3.1 shows a compact presentation of the result (one correct, and one wrong candidate per test point). Fig. 8 reports the complete anatomy for every candidate label at both test points, in addition to the score function at each stage.

Setup. We sample three classes from overlapping Gaussians (means on a triangle with pairwise distance 4, standard deviation 1.5). The model is a multinomial logistic regression over fixed RBF

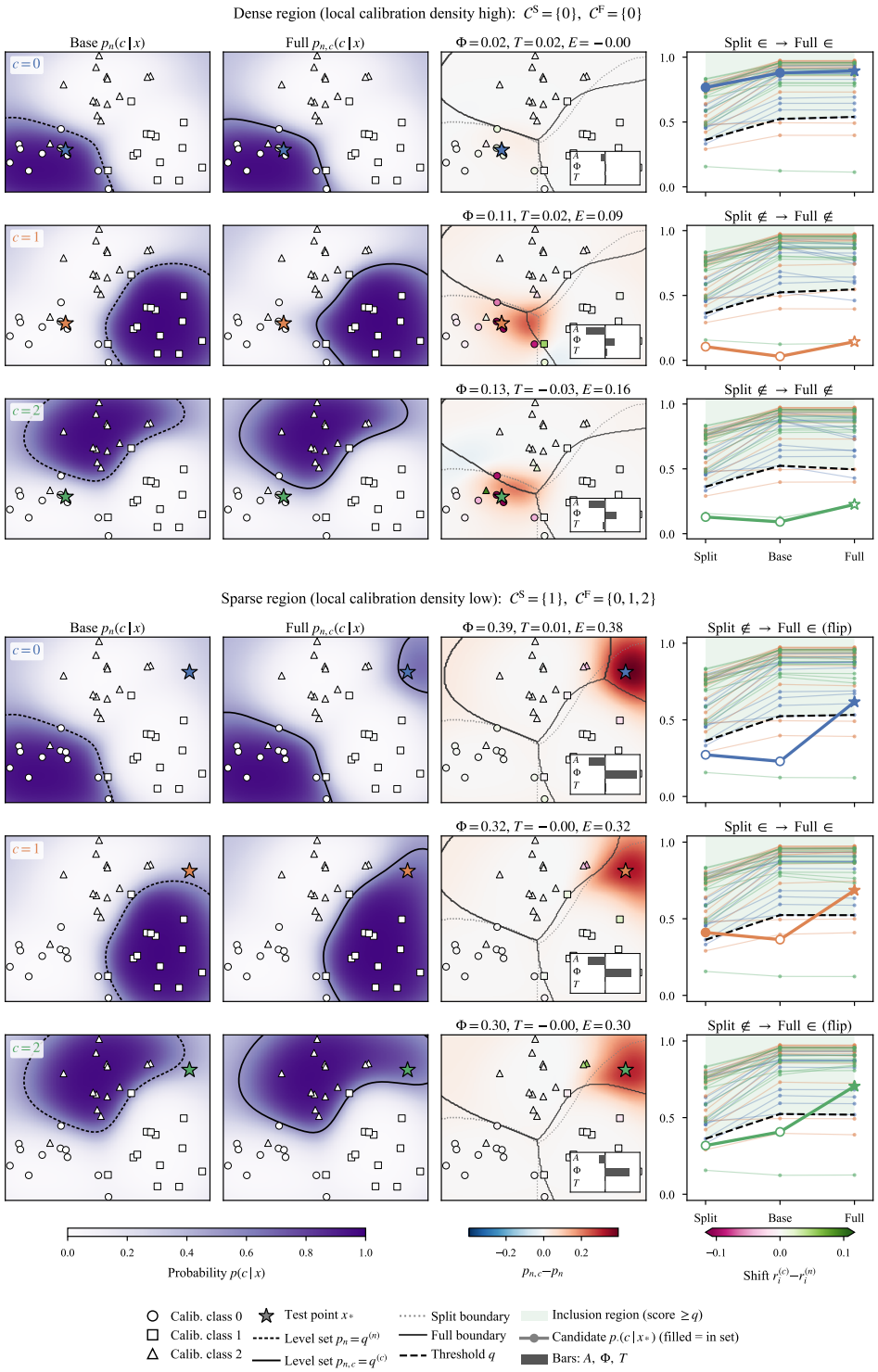

Figure 8: Complete anatomy of the split-to-full margin.

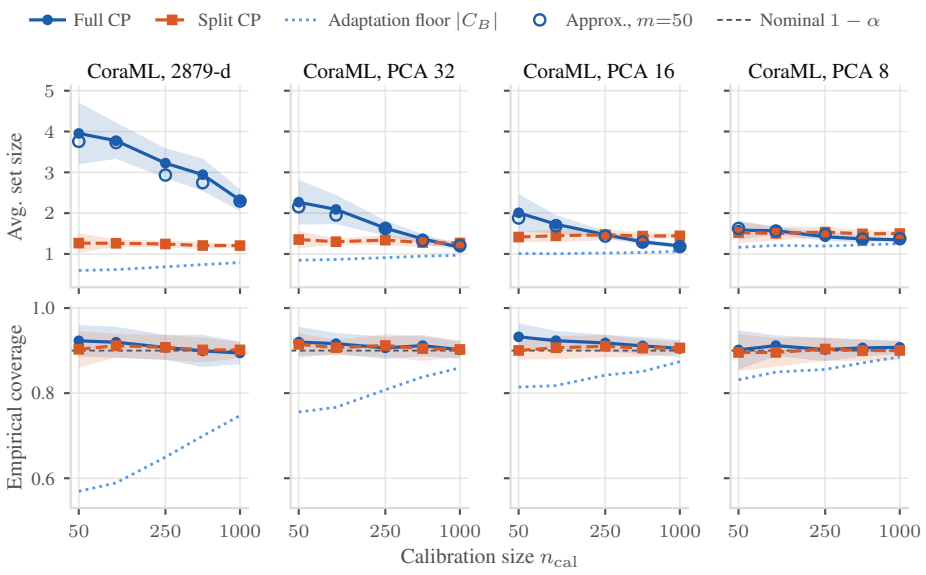

Figure 9: Full (blue, solid) vs. split (orange, dashed) CP on `CoraML` (transductive node-classification, GCN) over the calibration set size n_{cal}, with the 2879-d bag-of-words features [Left], and their PCA reduction to $K = 32$, 16, and 8 [Middle Left to Right]. [Top] Average set size against the adaptation floor $|\mathcal{C}_B|$ (dotted). Hollow markers are the approximate full CP with $m = 50$. [Bottom] Empirical coverage of the same sets against the nominal $1 - \alpha = 0.9$ (black, dashed). Bands are ± 1 standard deviation across calibration splits.

features, so that every refit is an exact convex MAP problem. The pretrained model $\boldsymbol{\theta}_0$ is trained on 3 points per class (with a prior centered at zero), and every fine-tuning minimizes the cross-entropy plus the L_2 pull $\frac{1}{2\tau^2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ with $\tau^2 = 1$; i.e. the MAP objective under the prior $\mathcal{N}(\boldsymbol{\theta}_0, \tau^2 \boldsymbol{I})$. The calibration set has $n = 42$ points, the score is the predictive probability, and $1 - \alpha = 0.9$. We fine-tune once on $\mathcal{D}_n$ (the base model p_n), and once per candidate on $\mathcal{D}_{n+1}^{(c)}$ (the full models $p_{n,c}$). We examine two test points that differ only in their local calibration density: one is placed inside the class-0 cluster, the other away from every calibration point.

Description for Fig. 8. Rows iterate over the candidate labels $c \in \{0, 1, 2\}$. There are two blocks, one for the test point in the dense area (top), and the other (bottom) the sparse one. Columns from left to right are (i) the base predictive field $p_n(c \mid \boldsymbol{x})$ together with its level set at the base threshold $q_{\boldsymbol{\theta}}^{(n)}$ (dashed): the region inside the contour is where the label c would enter the prediction set under the base model. (ii) The same field after fine-tuning on the pseudo-label, $p_{n,c}(c \mid \boldsymbol{x})$, with the level set at $q_{\boldsymbol{\theta}}^{(n,c)}$ (solid). (iii) Deformation $p_{n,c}(c \mid \boldsymbol{x}) - p_n(c \mid \boldsymbol{x})$. The calibration markers are colored by the shift of their own score, $r_i^{(c)} - r_i^{(n)} = C_{i,c}/w_{n,c}$ from Thm. 3.2. The figure also shows the quantities $A_{n,c}$, $\Phi_{n,c}$, and $T_{n,c}$ on one scale shared across all plots. (iv) Shows every score through the three regimes (split, base, full): thin lines are the calibration scores, the dashed line is the threshold, the shaded area is where the label is included in the set, and the bold line is the candidate itself, with filled markers indicating membership at that stage.

As expected, at the dense region the self-pull is resisted by the surrounding calibration points (and the training) with $\Phi_{n,c} \leq 0.13$ for the wrong candidates. In other words, the deformation is small, and sharply localized. The threshold also barely moves, and the two level sets remain relatively steady. Conversely, at the sparse test point where there is much less resistance from other points, fine-tuning on *any* pseudo-label bends the model freely ($\Phi_{n,c} \approx 0.3$–0.4 for every candidate), no calibration score reacts ($C_{i,c} \approx 0$, hence $T_{n,c} \approx 0$), and as a result, every candidate – the wrong ones included – flips into the full set. The takeaway is that the full CP loses exactly where the formability $\Phi_{n,c} = V_{n,c}/w_{n,c}$ is large.

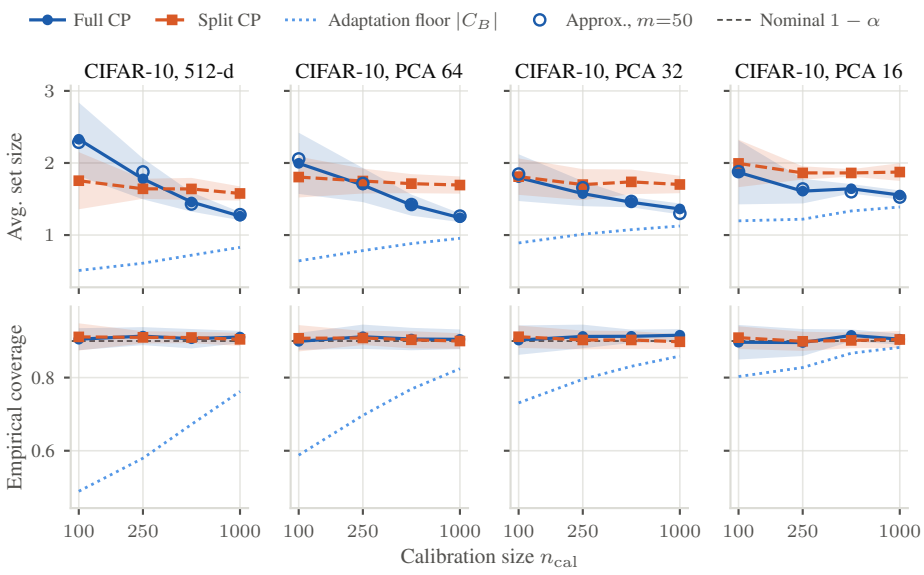

Figure 10: Full (blue, solid) vs. split (orange, dashed) CP on `CIFAR-10` (two-layer MLP over a frozen ResNet-18) over the calibration set size n_{cal}, with the 512-d embedding [Left], and its PCA reduction to $K = 64$, 32, and 16 [Middle Left to Right]. [Top] Average set size against the adaptation floor $|\mathcal{C}_B|$ (dotted). Hollow markers are the approximate full CP with $m = 50$. [Bottom] Empirical coverage of the same sets against the nominal $1 - \alpha = 0.9$ (black, dashed). Bands are ± 1 standard deviation across calibration splits.

C.1.2 Aleatoric against Epistemic Uncertainty

Fig. 4 in § 3.1 reads two uncertainty measures on the setup of § C.1.1 while the aleatoric part of the uncertainty is moved on purpose. The generating law is known, so the aleatoric part is known in closed form: with $\boldsymbol{\mu}_k$ the class means (at pairwise distance d), and the common standard deviation 1.5, the Bayes posterior at $\boldsymbol{x}$ is $\eta_k(\boldsymbol{x}) \propto \exp\big(-\|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2/(2 \cdot 1.5^2)\big)$, and we report $1 - \frac{K}{K-1}\mathrm{TV}\big(\eta(\boldsymbol{x}), \frac{1}{K}\big)$, with TV the total variation distance to the uniform law over the $K = 3$ classes. It is 1 where the label is a coin flip among the three classes, $1/2$ where it is a coin flip between two of them, 0 where it is determined, and it depends neither on the model nor on the sample. The two measures are read from the model fine-tuned on $\mathcal{D}_n$: the normalized softmax entropy $H(p_n(\cdot \mid \boldsymbol{x}_{n+1}))/\log K$, and the formability variance $\sum_c \hat{V}_{n,c} = \sum_c \hat{w}_{n,c}\hat{\Phi}_{n,c}$ of Eq. 6, with $\hat{w}_{n,c} = p_n(c \mid \boldsymbol{x}_{n+1})$, and $\hat{\Phi}_{n,c} = p_{n,c}(c \mid \boldsymbol{x}_{n+1}) - p_n(c \mid \boldsymbol{x}_{n+1})$ from one MAP refit per pseudo-label, exactly as in § C.1.1.

Setup. Model, pretraining (3 points per class), calibration ($n = 42$), prior, and α are those of § C.1.1. The grid of RBF centers is the one of that experiment extended by the same spacing (to 15×14 centers), so that it covers the wider configurations, and the longer paths below. The data are drawn once as standard normal offsets $\boldsymbol{z}_i$, and placed as $\boldsymbol{x}_i = \boldsymbol{\mu}_{y_i}(d) + 1.5\,\boldsymbol{z}_i$, so that the sample moves continuously with d. The means $\boldsymbol{\mu}_k(d)$ are the means of § C.1.1 scaled about their centroid $\bar{\boldsymbol{\mu}}$ to pairwise distance d ($d = 4$ is the configuration of Fig. 3), and the pretrained model, and p_n are refitted at every d. In Fig. 4(b), d runs from 8 to 0, and $\boldsymbol{x}_{n+1} = \bar{\boldsymbol{\mu}}$. In (c), and (d), $d = 4$, and $\boldsymbol{x}_{n+1}(t) = \bar{\boldsymbol{\mu}} + t\,\|\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}\|\,\boldsymbol{u}$ for $t \in [0, 3.5]$ along a unit direction $\boldsymbol{u}$: in (c) towards the center of class 0, which the path reaches at $t = 1$; in (d) towards the midpoint $\bar{\boldsymbol{\mu}}_{12}$ of classes 1, and 2, which it reaches at $t = 0.5$. The path of (d) lies, up to 0.01, on the perpendicular bisector of $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$, so the Bayes posterior stays a two-way coin flip between the two classes along the rest of it. Every curve is the mean over 5 draws of the data, and the band is their range. The aleatoric reference does not depend on the draw.

Reading. In (b), the label at $\boldsymbol{x}_{n+1}$ is a three-way coin flip at every d (the centroid is, up to 0.02, equidistant from the three means), so the aleatoric reference is at least 0.976 throughout. The soft-

max entropy reads this: its mean stays between 0.82 and 0.92 (the range over the draws is 0.59–0.98). The labeled points, however, pile up around $\boldsymbol{x}_{n+1}$ as the classes converge – the Gaussian kernel density of $\mathcal{D}_n$ at $\boldsymbol{x}_{n+1}$ (bandwidth 1.1) grows from 0.002 to 0.046 – and the formability variance reads this: it falls in every draw, from 0.276 to 0.061 on average ($\sum_c \hat{\Phi}_{n,c}$ from 0.94 to 0.20). The two measures move independently: the entropy is flat, while the formability variance falls by a factor of 4.5. In (c), the aleatoric reference falls from 0.994 to 0.079 at $\boldsymbol{\mu}_0$, and to 0.002 at $t = 2$. Both measures fall towards $\boldsymbol{\mu}_0$, the entropy from 0.92 to 0.32, and the formability variance from 0.166 to 0.024, as they must: with $p_n(0 \mid \boldsymbol{x}_{n+1})$ near one every plausible model agrees, and $V_{n,c} \leq w_{n,c}(1 - w_{n,c})$ (Lemma 3.1) leaves no room for disagreement. Both reach their minimum at $t = 1.4$ (0.22, and 0.012), where the labeled points start to thin out (the density is 0.018 there, 0.009 at $t = 2$, 0.003 at $t = 2.5$, and below 10^{-3} from $t = 3$ on), and then both rise: the formability variance to 0.37, since a model that no data resist is free to deform towards any label ($\sum_c \hat{\Phi}_{n,c} = 1.1$; every refit lifts its pseudo-label to about 0.7), and the entropy to 0.99, since the RBF logits decay to the bias away from the data, and the softmax returns to the near-uniform bias. The entropy therefore reads the same value, its maximum, at the centroid and at the far end of the path, where the aleatoric reference is 0.994, and 0: the first reading is aleatoric, the second is entirely epistemic, and the entropy cannot tell the two apart. In (d), the aleatoric reference falls from 0.994 to 0.62 at the midpoint $\bar{\boldsymbol{\mu}}_{12}$, and then settles at 0.50: the label remains a coin flip between classes 1, and 2 along the rest of the path. The entropy reads 0.74 at the midpoint, its minimum 0.69 at $t = 0.9$, and then climbs back to 0.99. The formability variance reads 0.12 at the midpoint, and rises to 0.37 as the density falls from 0.014 to 0. Between $t = 1$ and $t = 3.5$ the aleatoric part is constant, so the rise of the entropy from 0.69 to 0.99 is entirely epistemic; the entropy does not say so, and the formability variance does. Wherever the entropy is high, then, the formability variance separates the two cases: it is small where the entropy is high because the labels are mixed and the labeled points are many (the centroid at small d, the neighborhood of $\bar{\boldsymbol{\mu}}_{12}$), and large where the entropy is high because nothing constrains the model (the centroid at $d = 8$, the far ends of both paths).

C.1.3 Exact Evaluation on Bayesian Linear Regression

The margin decomposition for full - split CP in § 3.2 is general. However, using the Bayesian learning regime we can show that $\Phi_{n,c}$ is a function of epistemic uncertainty. Such a setup is unrealistic for large-scale deep learning, for which we discuss the same concept under deterministic fine-tuning in § 3.3. Meanwhile, the Bayesian setup is exactly realizable in a small Gaussian-linear model, where every quantity of § 3.2 can be evaluated in closed form. The synthetic dataset is drawn from $f^\star(x) = \sin(1.5x) + 0.3x$ plus noise: 40 pretraining points uniform on $[-4, -0.5]$, and the calibration set is drawn from the mixture $x \sim \frac{3}{4}\mathcal{N}(0, 0.5^2) + \frac{1}{4}\,\mathrm{U}[-4, 4]$, therefore its density is high around $x = 0$ and low elsewhere ($n = 60$). The model is $y = \boldsymbol{\theta}^\top \boldsymbol{\psi}(x) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 0.5$, and $\boldsymbol{\psi}(x) \in \mathbb{R}^{13}$ a bias plus 12 Gaussian bumps of width 0.7 centered on a grid over $[-4, 4]$. The score is the likelihood $p_{\boldsymbol{\theta}}(y \mid x) = \mathcal{N}(y; \boldsymbol{\theta}^\top \boldsymbol{\psi}(x), \sigma^2)$, which takes values in $[0,\, 1/(\sigma\sqrt{2\pi})] = [0, 0.8]$. The bound $V_{n,c} \leq \Phi_{n,c}(1 - \Phi_{n,c})$ of Lemma 3.1 therefore applies to it as it does to a class probability, and the candidates c range over a grid on $[-3, 3]$. The pretrained model $\boldsymbol{\theta}_0$ is trained with prior $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $1 - \alpha = 0.9$.

Closed forms. For $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ write $m(x) := \boldsymbol{\mu}^\top \boldsymbol{\psi}(x)$ and $s^2(x) := \boldsymbol{\psi}(x)^\top \boldsymbol{S} \boldsymbol{\psi}(x)$. The calibration posterior is $\pi_n = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{S}_n)$ with $\boldsymbol{S}_n^{-1} = \boldsymbol{I} + \sigma^{-2} \sum_{i=1}^n \boldsymbol{\psi}(x_i)\boldsymbol{\psi}(x_i)^\top$ and $\boldsymbol{\mu}_n = \boldsymbol{S}_n\big(\boldsymbol{\theta}_0 + \sigma^{-2} \sum_{i=1}^n y_i \boldsymbol{\psi}(x_i)\big)$, and $\pi_{n,c}$ is the same update with the pair $(\boldsymbol{x}_{n+1}, c)$ added. This update is the refit that produces $m_{n,c}$, $\Phi_{n,c}$, and $T_{n,c}$ in Fig. 5. The moments that enter Lemma 3.1 and Thm. 3.2 are, with $\boldsymbol{\Psi} := [\boldsymbol{\psi}(x), \boldsymbol{\psi}(x')]$,

$$\mathbb{E}_{\boldsymbol{\theta}}[p_{\boldsymbol{\theta}}(y \mid x)] = \mathcal{N}\big(y;\, m(x),\, \sigma^2 + s^2(x)\big),$$

$$\mathbb{E}_{\boldsymbol{\theta}}[p_{\boldsymbol{\theta}}(y \mid x)^2] = \frac{1}{2\sigma\sqrt{\pi}}\, \mathcal{N}\big(y;\, m(x),\, \sigma^2/2 + s^2(x)\big),$$

$$\mathbb{E}_{\boldsymbol{\theta}}[p_{\boldsymbol{\theta}}(y \mid x)\, p_{\boldsymbol{\theta}}(y' \mid x')] = \mathcal{N}\big((y, y');\, (m(x), m(x')),\, \sigma^2 \boldsymbol{I} + \boldsymbol{\Psi}^\top \boldsymbol{S} \boldsymbol{\Psi}\big),$$

which are used to compute the scores r_i and the candidate mass $w_{n,c}$, the variances V_i and $V_{n,c}$, and the forming kernel $C_{i,c}$ of Eq. 7. Each identity of § 3.2 is then compared against the quantity it predicts, computed independently from the refit $\pi_{n,c}$.

Conclusion. Our sanity check (which is also shown in Fig. 5) succeeds as over all 481×3 candidate–test pairs, the two independent computations are close up to the numerical instability; i.e.

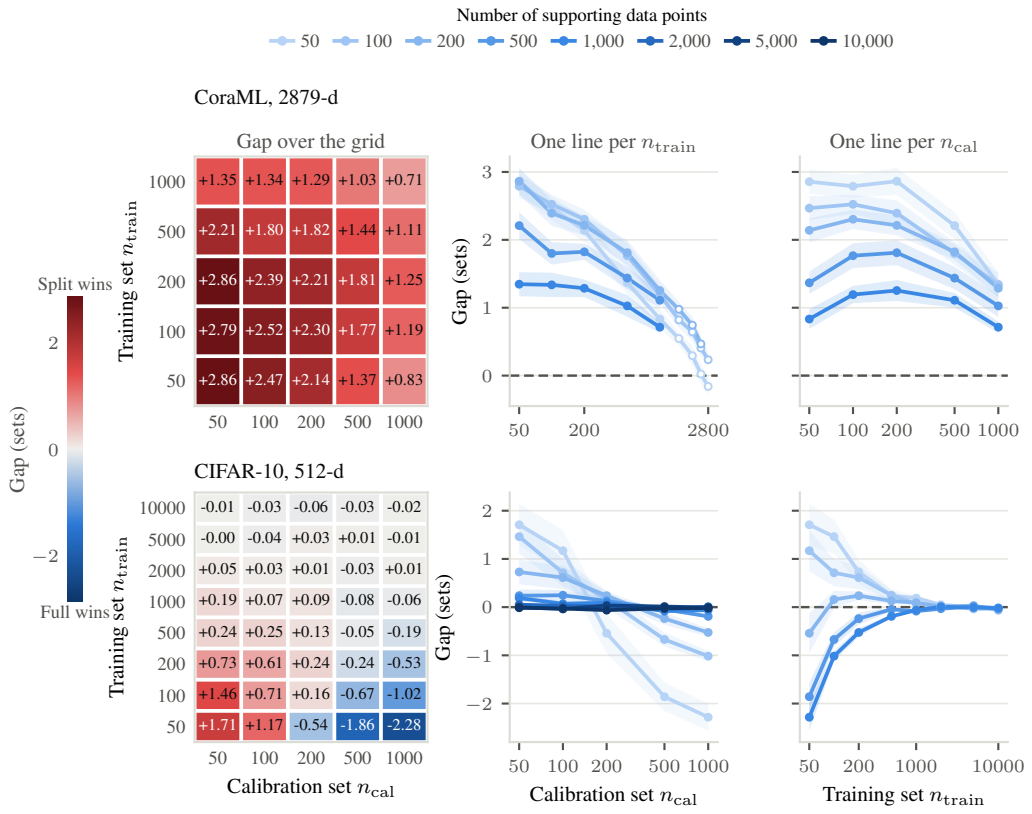

Figure 11: The gap $|\mathcal{C}^F| - |\mathcal{C}^S|$ in average set size over the training set size n_{train}, and the calibration set size n_{cal} (positive: split CP is smaller, negative: full CP is smaller), on `CoraML` with 2879-d features [Top], and `CIFAR-10` with the full 512-d embedding [Bottom]. [Left] The gap over the grid. [Middle] The gap against n_{cal}, one line per n_{train}. Hollow markers are the `CoraML` extension arm up to $n_{\text{cal}} = 2800$. [Right] The gap against n_{train}, one line per n_{cal}. In the line panels, the shade encodes the size of the other set (light to dark = small to large), and bands are ± 1 standard error of the paired difference across calibration splits. $1 - \alpha = 0.9$.

$\max_{i,c} |r_i^{(c)} - r_i^{(n)} - C_{i,c}/w_{n,c}| = 7 \cdot 10^{-15}$, $\max_c |V_{n,c} - w_{n,c}\Phi_{n,c}| = 3 \cdot 10^{-16}$, $|C_{i,c}| \leq \sqrt{V_i V_{n,c}}$ throughout, and $m_{n,c}$ lies inside the band of Eq. 9 with a smallest slack of $6.5 \cdot 10^{-3}$ (inside the kernel band of Eq. 15, $2.4 \cdot 10^{-4}$). The formability is the dominant term wherever the neighborhood is sparse ($\max_c \Phi_{n,c} = 0.24$ against $\max_c |T_{n,c}| = 0.03$ at the sparse test point), which is why the full set inflates there and not at the dense point, where $\max_c \Phi_{n,c} = 0.02$.

C.1.4 The Frozen Image Backbones

Every image experiment in this paper trains a small head on top of a frozen `ImageNet` (Deng et al., 2009)-pretrained backbone, and the feature space from which the formability is defined is the output of this backbone (this especially matters for CIFAR-10 experiments with feature reduction). The `ImageNet` output layer is removed, the remaining network embeds each image once, and only the head is trained, or refitted. The backbone itself is never updated, which helps us to reduce the cost of per-candidate refit from the entire model to only the head. Table 1 lists the specifications of the ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), and ConvNeXt (Liu et al., 2022) backbones. ViT-H/14 uses the SWAG weights of Singh et al. (2022).

Table 1: The frozen `ImageNet`-pretrained backbones over which the image experiments are run. *Weights* is the `torchvision` checkpoint. *Original* counts the parameters of the classifier as loaded, and *fixed* the ones that remain once its `ImageNet` output layer is removed, which is the form in which we load it. d is the dimension of the embedding the head sees, and it is also the dimension the PCA reduction of § C.1.5 starts from. The top-1 accuracy is the reference number reported for the same checkpoint on `ImageNet`-1k.

		Parameters		Feature	`ImageNet`-1k
Backbone	Weights	Original	Fixed	dim. d	top-1 (%)
ResNet-18	`IMAGENET1K_V1`	11.7M	11.2M	512	69.8
ResNet-50	`IMAGENET1K_V2`	25.6M	23.5M	2048	80.9
ViT-B/16	`IMAGENET1K_V1`	86.6M	85.8M	768	81.1
ConvNeXt-B	`IMAGENET1K_V1`	88.6M	87.6M	1024	84.1
ViT-H/14	`IMAGENET1K_SWAG_E2E_V1`	633.5M	632.2M	1280	88.6

C.1.5 Full vs. Split CP over the Feature Dimension

Setup. We use a two-layer GCN (Kipf & Welling, 2017) over the 2879-dimensional bag-of-words node features of `CoraML`, and a two-layer MLP over the 512-dimensional embedding of a frozen `ImageNet`-pretrained ResNet-18 for `CIFAR-10`. Full CP trains the model from initialization on the training set, the calibration set, and the pseudo-labeled test point. We reduce the features by PCA to $K \in \{32, 16, 8\}$ for `CoraML`, and $K \in \{64, 32, 16\}$ for `CIFAR-10`, and sweep n_{cal}, and n_{tr}. The score is the predictive probability of § 2, and $1 - \alpha = 0.9$. The adaptation floor $|\mathcal{C}_B|$ is the size of the posterior conformal set $\mathcal{C}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_{n+1})$ i.e. split CP on the model trained on the training, and the calibration set together, which needs a single fit per calibration set, and no per-candidate refit.

Results. In Fig. 9, and Fig. 10 the plots go from the full feature space to the smallest K. The top row reports the average set size against n_{cal}, where the hollow markers are the approximate full CP with $m = 50$, and the bottom row reports the empirical coverage of the same sets against the nominal $1 - \alpha$. Full, and split CP are at the nominal level in every panel which shows the sanity of the result. The adaptation floor is not (as expected): it under-covers by a wide margin in the full feature space (≈ 0.57 on `CoraML` at $n_{\text{cal}} = 50$, and ≈ 0.49 on `CIFAR-10` at $n_{\text{cal}} = 100$), and it approaches $1 - \alpha$ as n_{cal} grows, or K shrinks. This is expected, since $\mathcal{C}_B$ scores the calibration points it was trained on, and the amount of this optimism is again a formability – now read at the calibration points.

Similarly in set size, we observe that by increasing the calibration points, the difference between the two approaches decreases, until full CP crosses the split CP. Notably this effect is more visible in `CIFAR-10` than in `CoraML`. Mainly this is since the training size of the `CIFAR-10` is larger than that of `CoraML`. We also observe that reducing the feature dimension closes the performance gap.

An interesting side observation is that the question of whether the full CP is more efficient than the split CP can be answered without the need to compute the full CP. One can simply estimate it with a small sample rate ($m = 50$) by computing a leave-one-out refit on the calibration set, and that point alone. This is a relatively cheap operation to check the performance gap, without the need for expensive computation. The method turns out to be a good approximation as shown in both figures (the hollow markers are the approximate full CP with $m = 50$).

We report the extended experiment on the gap $|\mathcal{C}^F(\boldsymbol{x}_{n+1})| - |\mathcal{C}^S(\boldsymbol{x}_{n+1})|$ over a grid of n_{train}, and n_{cal}, on `CoraML` (2879-d), and on `CIFAR-10` (the 512-d embedding). The results are shown in Fig. 11. This is also same as our expectation. Another takeaway from this result is that the training set size also decreases the performance, and as the training size increases, the effect of the calibration set size becomes less significant.

C.1.6 Adaptive Prediction Sets

In this work we focus on threshold prediction sets as the score function (Sadinle et al., 2019): it directly sets the conformity score to the predictive probability of the candidate label; i.e. $s(\boldsymbol{x}, y) = p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$. That object is very easy to work with and it directly connects to the decomposition

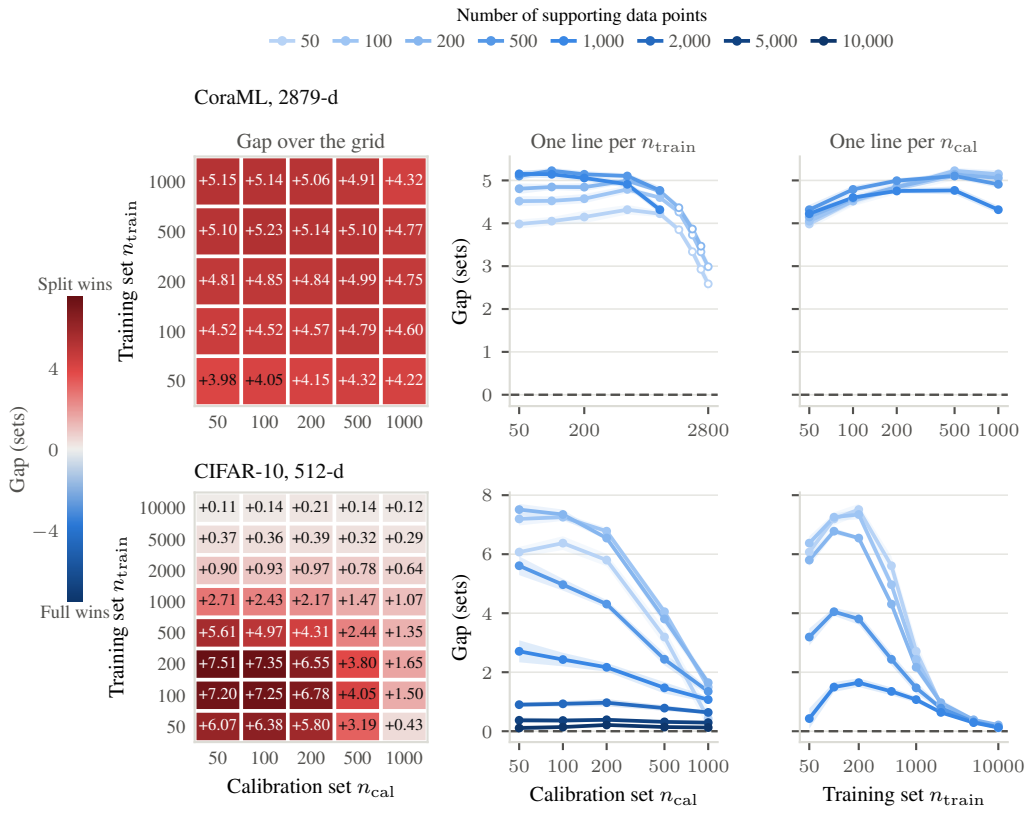

Figure 12: The grid of Fig. 11 read under the randomized APS score (Romano et al., 2020), off the same refits: the gap $|\mathcal{C}^F| - |\mathcal{C}^S|$ in average set size over the training set size n_{train}, and the calibration set size n_{cal} (positive: split CP is smaller), on `CoraML` with 2879-d features [Top], and `CIFAR-10` with the full 512-d embedding [Bottom]. [Left] The gap over the grid. [Middle] The gap against n_{cal}, one line per n_{train}. Hollow markers are the `CoraML` extension arm up to $n_{\text{cal}} = 2800$. [Right] The gap against n_{train}, one line per n_{cal}. In the line panels, the shade encodes the size of the other set (light to dark = small to large), and bands are ± 1 standard error of the paired difference across calibration splits. The color scale is centered at zero as in Fig. 11, but it spans a range about two and a half times wider. No cell of either grid is negative. $1 - \alpha = 0.9$.

of Eq. 2 to the updates of the predictive distribution after refits. However, conformal prediction is agnostic to the choice of score, and the same margin-decomposition can apply on other score functions as well. It is an important question to ask what interpretation does the full-split gap (and in particular the margin gap) has under that different score.

We repeat the experiment shown in Fig. 10, this time under the adaptive prediction sets (APS) score (Romano et al., 2020). Result is shown in Fig. 12. This ablation does not need to train a new model since we have already cached the refits from the previous experiment.

APS applies the same conformal prediction as shown in § 2, but instead on the following score function:

$$s(\boldsymbol{x}, y) = 1 - \sum_{y' : p_{\boldsymbol{\theta}}(y'|\boldsymbol{x}) > p_{\boldsymbol{\theta}}(y|\boldsymbol{x})} p_{\boldsymbol{\theta}}(y' \mid \boldsymbol{x}) - u\, p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}),$$

and we use the randomized variant, with one $u \sim \text{Unif}(0, 1)$ drawn per (point, rank) once per split, and reused by every refit. The score therefore remains a deterministic function of the predictive distribution, and the exchangeability argument is untouched. Everything else – the grid, the test points, the models, and $1 - \alpha = 0.9$ – is as in Fig. 11.

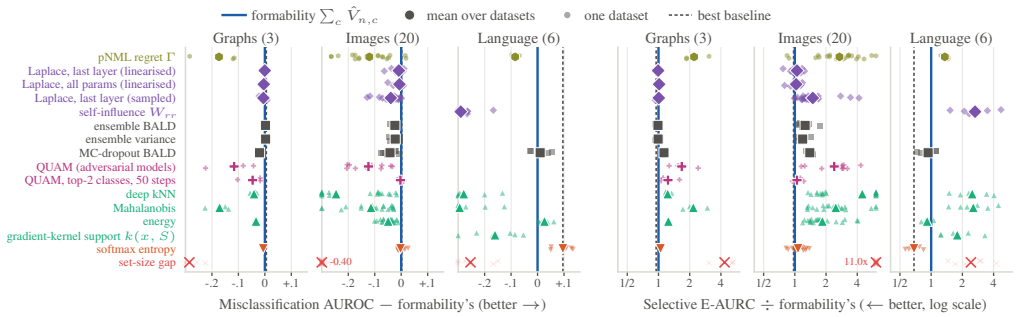

Figure 13: Formability against families of uncertainty measures, for in-distribution data, on graphs, images and language. For each modality the relative performance is gathered from different datasets and models.

Result. Split CP is smaller in every cell of both grids, and by a wide margin, and by increasing the training points or calibration points the gap converges to zero. APS can be read as a fractional knapsack problem over the probabilty distribution: take the smallest number of labels (all with equal weight), until $1-\beta$ cumulative probability is captured. Then conformalization (computing the quantile) is an answer to: what cumulative $1-\beta$ in model's predictive probability can capture the $1-\alpha$ probability in the true distribution. The gap is therefore a measure of how much the model's predictive distribution is misaligned with the true distribution, and the gap converges to zero as the model sees more labels and becomes better aligned with the true distribution. However deriving the same formability decomposition as in § 3.2 is not straightforward, specially since the score involves the ranking of the labels, and the refit can change this ranking. We leave this as an open question for future work.

C.1.7 Breakdown of Uncertainty Quantification

In Fig. 13 we report the detailed comparison against other uncertainty measures. Formability is close to the best methods on graphs and images, while on language softmax entropy performs better. The full–split set-size gap performs poorly across all three modalities, as it quantizes the continuous formability signal into a binary signal.

Fig. 14 extends the image comparison of Fig. 6, and Fig. 7 across different backbones and distribution shifts. Formability stays close to the best uncertainty measures for ranking errors on clean images, while for shift detection the result depends on the type of shift and is weaker on Gaussian noise.

Fig. 15 further extends the image comparison of Fig. 6 by grouping `CIFAR-10` images according to human-label disagreement (which can be considered a proxy for aleatoric uncertainty). Formability remains close to the best methods consistently within these groups, while ranking the errors generally becomes harder as the disagreement increases.

Ablation on training set size. In addition we also stress-test the measures against the number of labels the model has seen during training. Fig. 16 shows this comparison across various training budgets.

The models are the ones of § C.1.5. We vary the number of labels the model has seen, $N = n_{\text{train}} + n_{\text{cal}}$, from 50 to 1600 on `CoraML`, and from 50 to 3000 on `CIFAR-10` – with the calibration set held at $n_{\text{cal}} = 20$. Note that here our main argument is not to show a powerful conformal prediction, but to show the quality of formability in capturing epistemic uncertainty. Therefore the small calibration size is justified for computational efficiency. Columns are the two datasets, with the accuracy of the base model at each rung along the top. [Top] The misclassification AUROC (higher is better; the dotted rule is chance). [Bottom] The selective E-AURC (lower is better). In all training budgets, formability remains among the best methods.

Cost against quality. The refits are the most expensive reading of the epistemic uncertainty by construction: one fine-tuning per pseudo-label, per test point. We evaluate the cost, and (AUROC,

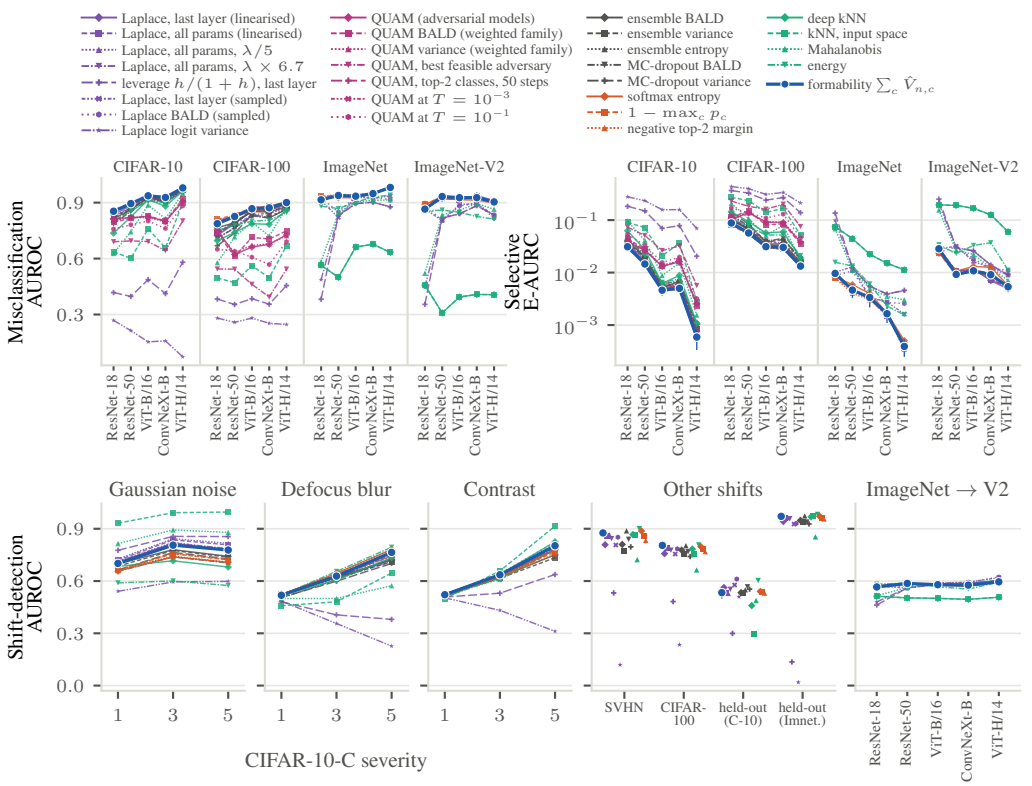

Figure 14: Formability against uncertainty measures across frozen image backbones and distribution shifts. [Top] Misclassification AUROC (higher is better) and selective E-AURC (lower is better, log scale) on clean `CIFAR-10`, `CIFAR-100`, `ImageNet`, and `ImageNet-V2`, with five backbones in each block. [Bottom] Shift-detection AUROC for `CIFAR-10-C` at three severities, Street View House Numbers (`SVHN`), `CIFAR-100`, held-out classes, and `ImageNet-V2` against `ImageNet` validation images. The blue line is $\sum_c \hat{V}_{n,c}$. Bars are ± 1 standard error over splits. Other colors distinguish Laplace, ensemble and MC-dropout, QUAM, feature-based, and softmax measures.

E-AURC) on different measures for five datasets: `CoraML`, `CiteSeer`, and `PubMed` with the two-layer GCN of § 4, and `CIFAR-10`, and `ImageNet-100` (its validation images) with the two-layer MLP over the 512-d embedding of the frozen ResNet-18 (Table 1). The refits, the measures, and the two metrics are as in § 4. Every measure is timed per evaluation point with CUDA synchronization. The fit of the base model is charged to no one, since every measure needs it, and the softmax controls are a forward pass that is already made. The refits are fitted in a batch over the candidate labels, which is how full CP is run throughout the paper. One at a time, they would cost one to two orders of magnitude more. QUAM is timed at two budgets: every class attacked for 100 steps, and the two most probable classes for 50.

Fig. 17 shows the result: one column per dataset. The horizontal axis is the wall-clock seconds per evaluation point (log). The vertical axis is the misclassification AUROC [Top], and the selective E-AURC [Bottom]. Markers are means over the calibration splits, and bars are ± 1 standard error on both axes. The two QUAM budgets are joined by a line, from the cheap one to the full one.

C.2 Formability on Language Models

For language model tasks, we use four-option question answering: each model assigns probabilities to the answer letters A–D, and the conformal set contains a subset of these letters. We do not evaluate free-form generations. Even keeping most of the parameters frozen, running the same pipeline as the images, and graphs can be significantly more expensive. Therefore we approximate formability

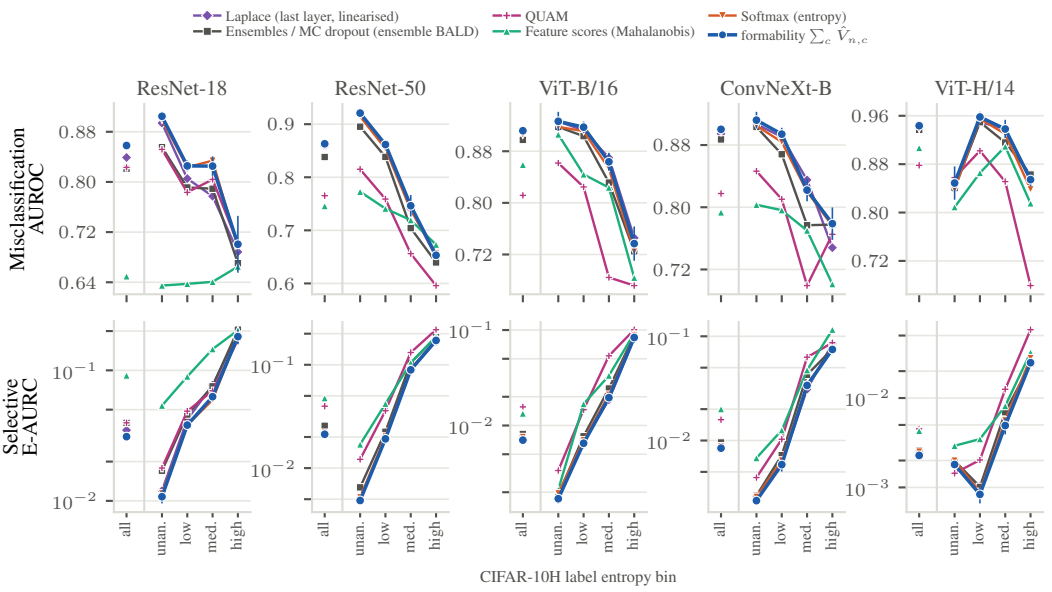

Figure 15: Error ranking within bins of human-label disagreement on `CIFAR-10H`. The 1000 `CIFAR-10` test images are grouped by the entropy H of about 51 human labels per image: unanimous ($H = 0$), low ($0 < H < 0.2$), medium ($0.2 \leq H < 0.5$), and high ($H \geq 0.5$). "All" is the pooled result. Each column is a frozen backbone. [Top] Misclassification AUROC (higher is better). [Bottom] Selective E-AURC (lower is better, log scale). Blue: $\sum_c \hat{V}_{n,c}$ with ± 1 standard error over splits. Other lines: last-layer Laplace, ensemble BALD, QUAM, Mahalanobis, and softmax entropy.

by a local update of a LoRA adapter. In general we follow the full conformal approximation from Abad Martinez et al. (2023), but this time over the LoRA adapter parameters, for the classification task.

Setup. We use six model–dataset combinations: Llama-1B, Llama-3B (Grattafiori et al., 2024), Qwen-1.5B, and Qwen-3B (Qwen Team, 2024) on MMLU (Hendrycks et al., 2021) humanities and business, and Llama-1B on MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). The pretrained language model is frozen, and a LoRA adapter is fitted on 2000 labeled training questions. The LoRA adapter is applied with rank $r = 8$, and scaling $\alpha = 16$, dropout 0, on query and value matrices of the attention layers. The same setup is also used for the ensemble model. From separate pools of 6000 calibration questions, and 800 evaluation questions, we vary the calibration set size from 50 to 2000, with 20 calibration draws. The score is the probability assigned to an answer letter. Every uncertainty measure ranks the same evaluation questions, using the misclassification AUROC, and selective E-AURC as mentioned in § 4. Softmax entropy, kNN, Mahalanobis and the ensemble are not computed from $p_{\boldsymbol{\theta}}$. They come from a separate cached model fitted on the training questions plus one fixed set of calibration questions. Formability is however computed from damped influence steps that start at that same anchor adapter. The comparison is not entirely fair, but it is the best we can do given the expensive computational cost of the refits on LLMs.

Approximate refits. Let $\tilde{\boldsymbol{\theta}}_n$ denote the adapter fitted on the training questions, and updated by a single damped influence step over calibration points, and let $\tilde{\boldsymbol{\theta}}_{n,c}$ denote the adapter fitted on those same questions plus $(\boldsymbol{x}_{n+1}, c)$. These are the base, and refit whose difference defines formability. The pseudo-labeled test question is the only additional observation. We approximate the two fits by damped influence steps in the adapter parameters (Koh & Liang, 2017; Abad Martinez et al., 2023), rather than optimizing a new adapter for every candidate. In particular, the additional step from the base has the form

$$\tilde{\boldsymbol{\theta}}_{n,c} - \tilde{\boldsymbol{\theta}}_n \;\approx\; -\tfrac{1}{N+1}(H + \lambda \boldsymbol{I})^{-1} g_c, \tag{16}$$

where $g_c = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{x}_{n+1}, c; \boldsymbol{\theta}_{\text{ref}})$, and $\boldsymbol{\theta}_{\text{ref}}$ is the adapter at which the gradients are cached. Here H is the *empirical Fisher* of the existing training questions (note that a preconditioner containing the

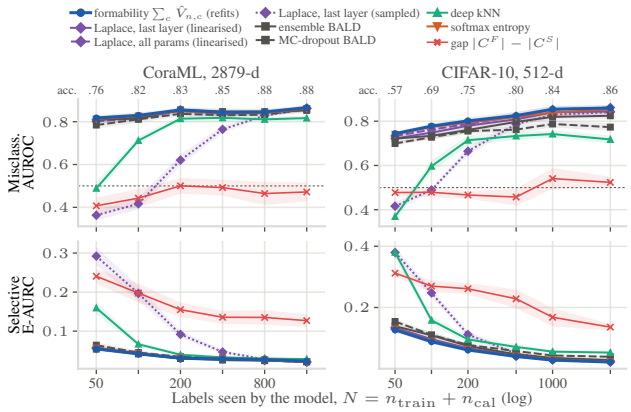

Figure 16: The epistemic measures of § 4 against the number of labels the model has seen, $N = n_{\text{train}} + n_{\text{cal}}$ (log), at the end of training (epoch 300), on `CoraML` with the 2879-d node features [Left], and `CIFAR-10` with the 512-d embedding [Right]. The accuracy of the base model at each rung is printed along the top. [Top] Misclassification AUROC (higher is better; dotted: chance). [Bottom] Selective E-AURC (lower is better). Every measure ranks the same base model on the same 200 evaluation points of a split, with $n_{\text{cal}} = 20$ held fixed. Blue circles: $\sum_c \hat{V}_{n,c}$ from the refits. Purple diamonds: the linearized last-layer Laplace (solid), the all-parameter empirical-Fisher form (dashed), and the sampled last-layer Laplace (dotted). Grey squares: ensemble BALD (solid), and MC-dropout BALD (dashed). Green triangles: deep kNN. Orange: softmax entropy. Red crosses: the set-size gap $|\mathcal{C}^F| - |\mathcal{C}^S|$. Lines are means over 8 calibration splits, bands ± 1 standard error.

calibration questions, but not the test question would break their exchangeability). Writing N for the combined size of the training, and calibration questions, and $g_i = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}_{\text{ref}})$, we use

$$H = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} g_i g_i^\top. \tag{17}$$

Each g_i uses the question's observed label y_i. This differs from the generalized Gauss–Newton matrix (the model Fisher for softmax cross-entropy), which averages $g_{i,c} g_{i,c}^\top$ over all four labels with weights $p(c \mid \boldsymbol{x}_i)$. This observed-label outer product is not, in general, a reliable approximation to the model Fisher or the Hessian (Kunstner et al., 2019). Cached gradients, and their bilinear products give the changes of the candidate probability, and the calibration scores without running every refit in the main sweep. We vary the λ independently of the calibration size (every calibration size is run at each λ). Fig. 18 shows the results at $\lambda/\text{ref} = 1.5$, with $\text{ref} = \text{tr}(H)/n_{\text{train}}$. With $\tilde{p}_n$, and $\tilde{p}_{n,c}$ denoting the predictions of the two approximate fits, we compute $\hat{w}_{n,c} = \tilde{p}_n(c \mid \boldsymbol{x}_{n+1})$, $\hat{\Phi}_{n,c} = \tilde{p}_{n,c}(c \mid \boldsymbol{x}_{n+1}) - \tilde{p}_n(c \mid \boldsymbol{x}_{n+1})$, and $\sum_c \hat{V}_{n,c} = \sum_c \hat{w}_{n,c} \hat{\Phi}_{n,c}$. The same approximate refit changes the calibration scores, and hence the full-conformal threshold. This is a local approximation to adapter refitting, and does not update the frozen language-model weights.

Result. Similar to the image and graph experiments, the approximate formability shows a performance that is competitive with the best method and performs better than many other approaches. However, the performance of the method is not as strong as the softmax entropy, and in general the results are not at all comparable with the image baseline. Our intuition is that with all the compromises we made to make the experiments feasible (and the training to fit on a single GPU), it is very much possible that the refit approximation is not as accurate as it should be. Finding a more efficient way to compute the refit, or a better approximation is an interesting direction for future work. The results are shown in Fig. 18.

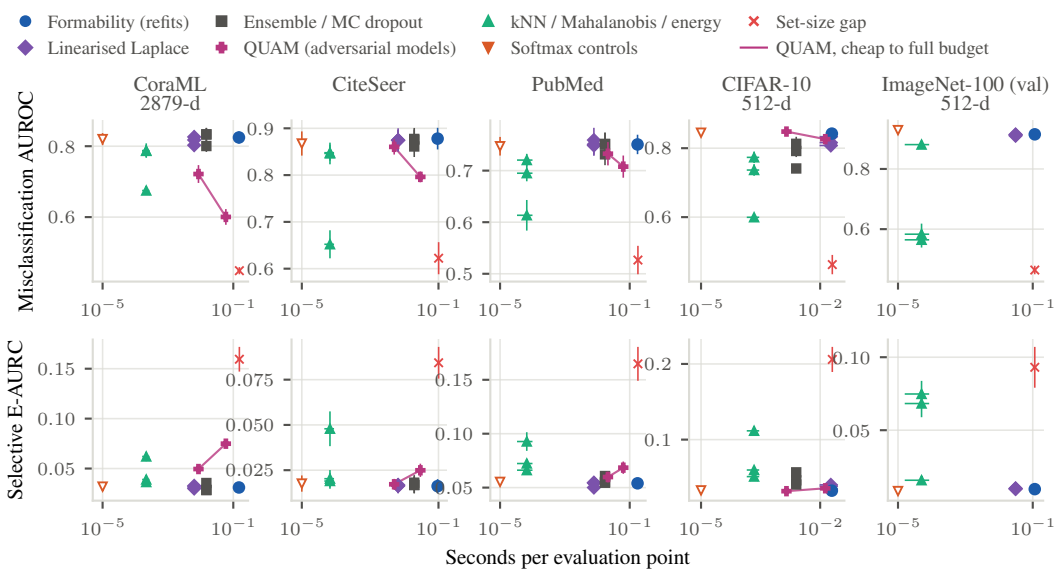

Figure 17: Cost against quality, one marker per measure, one column per dataset. [Horizontal] Wall-clock seconds per evaluation point (log), timed with CUDA synchronization. The softmax controls are a forward pass that is already made, and are drawn hollow at the left edge. [Vertical] Misclassification AUROC [Top, higher is better], and selective E-AURC [Bottom, lower is better]. Means over calibration splits, bars ± 1 standard error on both axes. QUAM (Schweighofer et al., 2023) is drawn at two budgets – the two most probable classes attacked for 50 steps, and every class for 100 – joined by a line from the cheap budget to the full one.

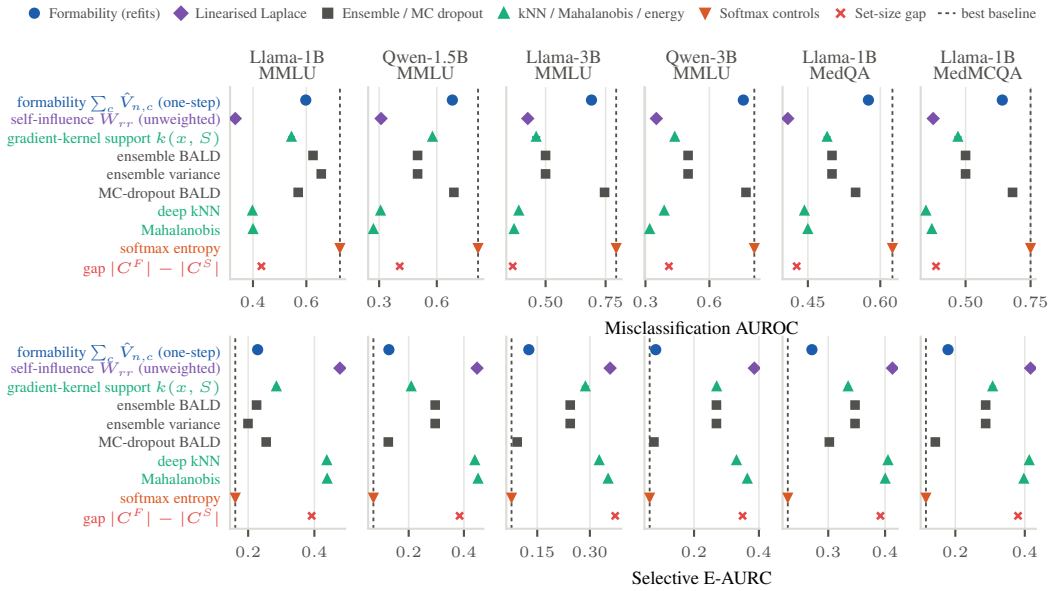

Figure 18: Formability on four-option language-model tasks. Each column ranks errors on the same evaluation questions. [Top] Misclassification AUROC (higher is better). [Bottom] Selective E-AURC (lower is better).