CONFORMAL PREDICTION SETS QUANTIFY INFORMATION GAIN: A THEORETICAL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction is a popular tool for uncertainty quantification that outputs prediction sets with finite-sample coverage guarantees. While prediction set size is commonly used as a heuristic measure of uncertainty, the information-theoretic basis for this interpretation remains poorly understood. In this work, we provide such a foundation using a decision-theoretic generalization of entropy tailored to set-valued prediction. In particular, we introduce a family of generalized information measures based on the size and coverage of conformal prediction sets. Notably, Shannon mutual information admits an exact integral representation in terms of these measures. We then show that, in standard classification settings, the reduction in conformal set size from additional information (i) is sandwiched between calibration-dependent members of this family and (ii) obeys a data processing inequality, both up to finite-sample calibration and model error terms. Together, our results formally relate conformal prediction to classical information-theoretic quantities and justify using set-size reduction as an information gain metric. Empirically, we validate our theory across 11 classification settings and show that set-size reduction and Shannon mutual information can rank features differently in a greedy feature selection experiment.

1 INTRODUCTION

Conformal prediction is a general procedure for constructing prediction sets with finite-sample and distribution-free coverage guarantees under minimal assumptions (Angelopoulos & Bates, 2021). Recently, the *split conformal prediction* variant has gained popularity as a lightweight and scalable calibration procedure, with applications in healthcare (Lu et al., 2022), dynamic control (Lindemann et al., 2023), and language modeling (Lu et al., 2023; Quach et al., 2024), among other settings.

Beyond coverage, conformal predictors are commonly evaluated by their inefficiency, or expected set size (Vovk et al., 2016; Sadinle et al., 2019). At the instance level, set size also provides an easy-to-compute and operationally meaningful heuristic measure of uncertainty (Cresswell et al., 2024), since more ambiguous inputs typically produce larger sets (Romano et al., 2019). Motivated by this interpretation, we examine conformal set-size reduction as a natural notion of information gain.

The relationship between conformal set size and information-theoretic quantities has been studied both empirically (e.g. Hagos & Lundström, 2026) and theoretically by using list-decoding arguments to upper bound the conditional entropy in terms of set size and coverage (Correia et al., 2024). However, these one-sided bounds do not establish whether set-size reduction itself behaves as a principled information measure. In this work, we instead analyze conformal prediction using a decision-theoretic notion of entropy (DeGroot, 1962). This perspective provides new insights into expected conformal set size and its connection to Shannon entropy and mutual information.

Our primary contributions are as follows: (i) we introduce a family of generalized information measures for set-valued prediction that exactly recovers Shannon mutual information via an integral representation; (ii) we prove that, in standard classification settings, conformal set-size reduction approximately lies between calibration-dependent members of this family and obeys a data processing inequality; and (iii) we empirically validate our theory across 11 classification settings and verify that set-size reduction and Shannon mutual information can induce different feature selection orderings in practice. Our results further the theoretical understanding of uncertainty quantification with conformal prediction and motivate set-size reduction as a principled information gain metric.

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.

2 PRELIMINARIES

We introduce the notation and relevant background on information theory and conformal prediction. In this work, we consider a classification setting in which $X \in \mathcal{X}$ is a generic feature vector and $Y \in \mathcal{Y} \coloneqq \{1, \ldots, K\}$ is a categorical prediction target.

Generalized entropy and information. Our analysis relies on a decision-theoretic generalization of entropy and information, in which uncertainty is represented using Bayes risk (DeGroot, 1962; Grünwald & Dawid, 2004). Suppose we fix an action space $\mathcal{A}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. The generalized entropy induced by ℓ is defined as

$$H_\ell(Y \mid X) = \mathbb{E}\left[\inf_{a \in \mathcal{A}} \mathbb{E}[\ell(Y, a) \mid X]\right]. \tag{1}$$

Given side information Z, the generalized mutual information between X and Y quantifies the predictive value of X by the reduction in Bayes risk under the chosen loss.

$$I_\ell(X; Y \mid Z) = H_\ell(Y \mid Z) - H_\ell(Y \mid X, Z) \geq 0. \tag{2}$$

Generalized mutual information retains several natural information-theoretic properties, including non-negativity and a data processing inequality (Duchi et al., 2018). The marginal versions $H_\ell(Y)$ and $I_\ell(X; Y)$ are defined by taking the respective conditioning variables to generate the trivial σ-algebra (e.g., if X and Z are almost surely constant, respectively).

Example. Shannon entropy $H(Y) = -\mathbb{E}[\log P_Y(Y)]$ is a special case of the generalized entropy by taking the action space $\mathcal{A} = \Delta^{K-1}$ to be the probability simplex over $\mathcal{Y}$ and using the log loss function $\ell(y, a) = -\log a(y)$. The Bayes-optimal action is then $a = P_Y$, giving $H_\ell(Y) = H(Y)$.

Conformal prediction. Conformal prediction provides a general framework for converting the output of a predictive model into a prediction set with finite-sample coverage guarantees (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos et al., 2025). In this work, we focus on *split conformal prediction* (Papadopoulos et al., 2002; Lei et al., 2018), which we briefly review next.

Let f be a pretrained model and $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ be a calibration dataset. Given a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$, define the calibration scores $S_i = s(X_i, Y_i)$ for $i = 1, \ldots, n$. Fix a nominal target coverage level $1 - \alpha \in (0, 1)$, and define the calibrated conformal threshold $\hat{\tau}_\alpha = \text{Quantile}\big((1 + 1/n)(1 - \alpha); \{S_i\}_{i=1}^n\big)$. At inference time, the conformal prediction set for a new test point is $\mathcal{C}_\alpha(X_{\text{test}}) = \{y \in \mathcal{Y} : s(X_{\text{test}}, y) \leq \hat{\tau}_\alpha\}$.

A key appeal of split conformal prediction is that the calibration procedure is lightweight and yields a finite-sample marginal coverage guarantee.

Proposition 1 (Conformal coverage guarantee)**.** *Suppose that the calibration observations $\mathcal{D}_{\text{cal}}$ and test point $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable and that the scores S_i are almost surely distinct. Then,*

$$1 - \alpha \leq \mathbb{P}\big(Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\big) \leq 1 - \alpha + \tfrac{1}{n+1}, \tag{3}$$

where the probability is taken jointly over the calibration data and the test point.

Proposition 1 holds for *any* model f and score function s. Nonetheless, the score function determines the size and structure of the conformal sets (Angelopoulos et al., 2021). We review two standard choices for classification. Let $q(\cdot \mid x) \in \Delta^{K-1}$ denote the predicted distribution for an input x.

Under the *probability score* (Sadinle et al., 2019), $s(x, y) = -q(y \mid x)$ is the negative predicted probability of the label. The resulting conformal prediction set $\mathcal{C}_\alpha(x) = \{y : q(y \mid x) \geq -\hat{\tau}_\alpha\}$ is therefore a level set of the predicted class probabilities determined by the calibrated cutoff.

The *Adaptive Prediction Sets (APS) score* (Romano et al., 2020) instead computes the score by accumulating the predicted probability mass of labels that are at least as likely as the true label. Formally, let $q(y^{(1)} \mid x) > q(y^{(2)} \mid x) > \cdots > q(y^{(K)} \mid x)$ denote the ordered predicted class probabilities (assume for simplicity that there are no ties). Let $r_x(y)$ be the rank of the label y, i.e., $y = y^{(r_x(y))}$. The randomized APS score function is

$$s(x, y) = \sum_{\ell=1}^{r_x(y)-1} q(y^{(\ell)} \mid x) + U \cdot q(y \mid x), \tag{4}$$

where $U \sim \text{Unif}[0, 1]$ is drawn independently to randomize inclusion of the boundary label. If there are ties among the predicted probabilities, we break them at random following Romano et al. (2020).

3 CONFORMAL ENTROPY AND MUTUAL INFORMATION

We introduce a family of generalized entropy and mutual information measures motivated by set-valued prediction. Our main result then establishes a two-sided bound showing that information gain from conformal prediction is bracketed by calibration-dependent members of this family, up to calibration and model error terms. Lastly, we extend the construction to the APS score. All proofs are deferred to Appendix A.

3.1 GENERALIZED INFORMATION MEASURES FOR THE PROBABILITY SCORE

We begin with a family of loss functions indexed by $\lambda \in (0, 1]$. For a set-valued prediction $\Gamma \subseteq \mathcal{Y}$, define the loss function

$$\ell_\lambda(Y, \Gamma) = |\Gamma| + \tfrac{1}{\lambda}\mathbf{1}(Y \notin \Gamma), \tag{5}$$

which penalizes both the size of Γ and its miscoverage of the true label Y. We use ℓ_λ to define a generalized entropy with action space $\mathcal{A} = 2^{\mathcal{Y}}$, the collection of all subsets of $\mathcal{Y}$.

Definition 1 (Conformal entropy and mutual information)**.** *For a fixed $\lambda \in (0, 1]$, the conformal entropy is defined as*

$$H_\lambda(Y \mid X) = \mathbb{E}\Big[\inf_{\Gamma \subseteq \mathcal{Y}} \mathbb{E}\big[\ell_\lambda(Y, \Gamma) \mid X\big]\Big]. \tag{6}$$

The corresponding conformal mutual information is $I_\lambda(X; Y \mid Z) = H_\lambda(Y \mid Z) - H_\lambda(Y \mid X, Z)$.

Since the action space is finite, the conformal entropy can equivalently be expressed as an infimum over (measurable) decision rules $\Gamma : \mathcal{X} \to \mathcal{A}$, so that $H_\lambda(Y \mid X) = \inf_{\Gamma: \mathcal{X} \to \mathcal{A}} \mathbb{E}\big[\ell_\lambda(Y, \Gamma(X))\big]$. We next characterize the Bayes-optimal actions under ℓ_λ.

Proposition 2 (Optimal set-valued predictors)**.** *Fix $\lambda \in (0, 1]$ and let $p(\cdot \mid X)$ be the conditional distribution of Y given X. The decision rule $\Gamma^*_\lambda(X; p) = \{y \in \mathcal{Y} : p(y \mid X) \geq \lambda\}$ minimizes $\mathbb{E}\big[\ell_\lambda(Y, \Gamma(X))\big]$ over all $\Gamma : \mathcal{X} \to 2^{\mathcal{Y}}$.*

Proposition 2 explains the conformal entropy terminology: the Bayes-optimal action $\Gamma^*_\lambda(X; p)$ is exactly the conformal prediction set from an oracle model $f(X) = p(\cdot \mid X)$ using the probability score at the calibrated threshold $\hat{\tau}_\alpha = -\lambda$. This connection should be unsurprising since $\mathbb{E}[\ell_\lambda(Y, \Gamma)]$ is the Lagrangian relaxation of minimizing expected set size subject to a coverage constraint, which is precisely the tradeoff underlying the probability score (Sadinle et al., 2019).

In practice, the model predictions may not coincide with the true conditional distribution of Y. The resulting model imperfection can be quantified using a generalized information divergence[1].

Definition 2 (Conformal information divergence)**.** *For a fixed $\lambda \in (0, 1]$, the conformal information divergence between conditional distributions $p(\cdot \mid X)$ and $q(\cdot \mid X)$ is defined as*

$$D_\lambda(p(\cdot \mid X) \parallel q(\cdot \mid X)) = \mathbb{E}_X\big[\mathbb{E}_{Y \sim p(\cdot|X)}[\ell_\lambda(Y, \Gamma^*_\lambda(X; q)) - \ell_\lambda(Y, \Gamma^*_\lambda(X; p))]\big]. \tag{7}$$

Intuitively, $D_\lambda(p \parallel q)$ is the excess risk incurred by acting under q when the true distribution is p. The divergence is non-negative by the optimality of $\Gamma^*_\lambda(X; p)$ and vanishes when $p = q$.

Using our construction, we can decompose Shannon mutual information into a uniform average of conformal mutual information terms over λ, analogous to the classical decomposition of the log loss into cost-weighted binary losses (Schervish, 1989).

Theorem 1 (Integral representation of Shannon information)**.** *Let I denote the Shannon mutual information (in nats). For any random variables X and Z,*

$$I(X; Y \mid Z) = \int_0^1 I_\lambda(X; Y \mid Z)\, d\lambda. \tag{8}$$

This identity connects information gain in conformal prediction to classical information-theoretic quantities. At a high level, Proposition 2 expresses conformal mutual information in terms of differences in set size and miscoverage. Standard coverage guarantees then allow us to bound set-size reduction in terms of I_λ at calibrated values of λ, each of which corresponds to a slice of Shannon mutual information by Theorem 1. The following section makes this intuition precise.

[1] More precisely, D_λ is a pseudo-divergence, since $D_\lambda(p \parallel q) = 0 \implies p = q$.

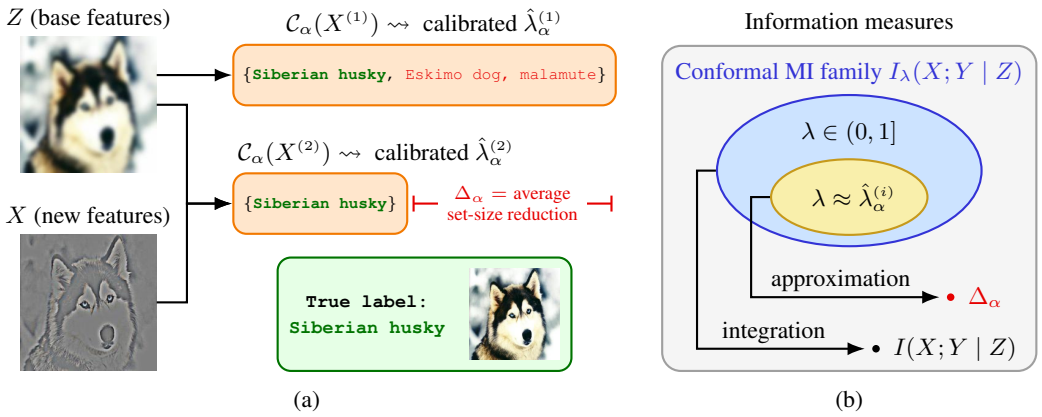

Figure 1: We illustrate an example of our setting in (1a), where Z is a low-resolution image and X is its residual. The information gain of X beyond Z can be quantified by Δ_α. We develop the information hierarchy in (1b), showing that conformal mutual information near the calibrated λ approximates Δ_α, while Shannon mutual information aggregates the entire family (Theorem 1).

3.2 Approximate information gain using conformal set-size reduction

We first formalize the information gained from additional features X beyond base features Z with conformal prediction. For notational convenience, let $X^{(1)} = Z$ and $X^{(2)} = (X, Z)$.

For $i \in \{1, 2\}$, let $p(\cdot \mid X^{(i)})$ and $q(\cdot \mid X^{(i)})$ denote the true and predicted conditional distributions, respectively. Let $\hat{\tau}_\alpha^{(i)}$ be the calibrated probability-score threshold using $X^{(i)}$ at nominal coverage level $1 - \alpha$, obtained from a shared calibration dataset $\mathcal{D}_{\text{cal}}$ of size n. Let $\mathcal{C}_\alpha(X^{(i)})$ denote the corresponding conformal prediction set. The (training-conditional) expected set-size reduction is

$$\Delta_\alpha = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(1)})|\big] - \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big], \tag{9}$$

where the expectations are over a new test input. This setting is illustrated in Figure 1a.

We emphasize that Δ_α depends on $\mathcal{D}_{\text{cal}}$. One might expect set-size reduction to be non-negative on average over draws of the calibration data (i.e., $\mathbb{E}[\Delta_\alpha] \geq 0$). In the following analysis, we show that an approximate version of this intuition holds.

Theorem 2 (Probability score sandwich bound). *Fix a calibration dataset $\mathcal{D}_{\text{cal}}$ and nominal coverage level $1 - \alpha$. For $i \in \{1, 2\}$, let $\hat{\lambda}_\alpha^{(i)} = -\hat{\tau}_\alpha^{(i)}$ and assume that $\hat{\lambda}_\alpha^{(i)} > 0$. Moreover, define*

$$I^{(i)} = I_{\hat{\lambda}_\alpha^{(i)}}(X; Y \mid Z), \qquad D^{(i)} = D_{\hat{\lambda}_\alpha^{(i)}}(p(\cdot \mid X^{(i)}) \,\|\, q(\cdot \mid X^{(i)})). \tag{10}$$

Let the training-conditional miscoverage be $\varepsilon^{(i)} = \mathbb{P}(Y \notin \mathcal{C}_\alpha(X^{(i)}); \mathcal{D}_{\text{cal}})$, where the probability is over a fresh test point $(X^{(i)}, Y)$, and define $\varepsilon^\Delta = \varepsilon^{(2)} - \varepsilon^{(1)}$. Then,

$$I^{(2)} - D^{(2)} + \varepsilon^\Delta / \hat{\lambda}_\alpha^{(2)} \;\leq\; \Delta_\alpha \;\leq\; I^{(1)} + D^{(1)} + \varepsilon^\Delta / \hat{\lambda}_\alpha^{(1)}. \tag{11}$$

Theorem 2 states that conformal set-size reduction is bracketed by conformal mutual information terms at two different values of λ, up to calibration terms and model error. By standard conformal arguments, the training-conditional miscoverage concentrates around α as n grows, which allows us to control the calibration term in expectation under suitable regularity conditions.

Corollary 1 (Probability score sandwich bound in expectation). *Fix a nominal coverage level $1 - \alpha$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. For $i \in \{1, 2\}$, define $\hat{\lambda}_\alpha^{(i)}$, $I^{(i)}$, and $D^{(i)}$ as in Theorem 2, and suppose that $s^{(i)} = \text{Var}(1/\hat{\lambda}_\alpha^{(i)})^{1/2} < \infty$. Then,*

$$\mathbb{E}\big[I^{(2)}\big] - \mathbb{E}\big[D^{(2)}\big] - s^{(2)}/\sqrt{n} \;\leq\; \mathbb{E}[\Delta_\alpha] \;\leq\; \mathbb{E}\big[I^{(1)}\big] + \mathbb{E}\big[D^{(1)}\big] + s^{(1)}/\sqrt{n}, \tag{12}$$

where all expectations are over the calibration dataset $\mathcal{D}_{\text{cal}}$.

Since $I_\lambda \geq 0$, Corollary 1 implies that on average over draws of the calibration data, the expected set-size reduction from adding X is asymptotically non-negative when $\mathbb{E}[D^{(2)}] = 0$ and $s^{(2)}$ remains bounded as n increases.

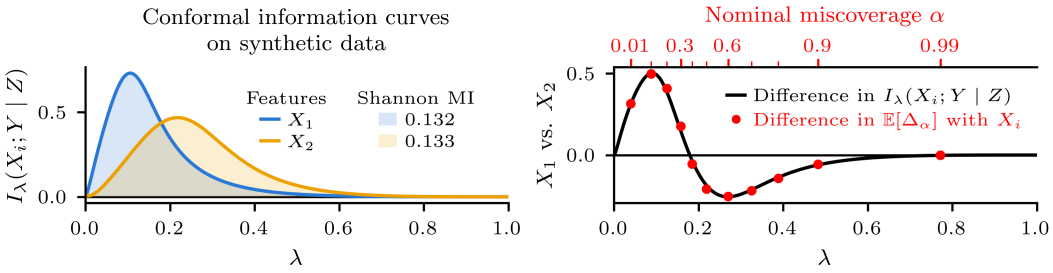

Figure 2: Conformal mutual information accurately reflects feature utility for reducing set size, whereas Shannon mutual information may not. We show information profiles $\lambda \mapsto I_\lambda(X_i; Y \mid Z)$ for two candidate features X_1 and X_2 with similar Shannon mutual information (left), along with their differences in I_λ and $\mathbb{E}[\Delta_\alpha]$ using the probability score (right). Nominal miscoverage levels $\alpha \in \{0.01, 0.1, \ldots, 0.9, 0.99\}$ are indicated along the top axis. Despite having nearly identical Shannon mutual information, X_1 yields a larger conformal set-size reduction at high coverage ($1 - \alpha > 0.6$), whereas X_2 reduces set size more at low coverage levels. Each difference in $\mathbb{E}[\Delta_\alpha]$ is positioned at the geometric mean of the two calibrated thresholds.

The variance condition in Corollary 1 is mild, requiring only that small predicted probabilities of the true label $q(Y \mid X^{(i)})$ are sufficiently rare. In Appendix A, we further relax this to almost-sure positivity using an alternative bound that holds with high probability over the calibration data.

We now revisit the integral representation in Theorem 1. Shannon mutual information uniformly averages I_λ over all $\lambda \in (0, 1]$, whereas Δ_α is approximated by members near the calibrated $\hat{\lambda}_\alpha^{(i)}$. We summarize this relationship in Figure 1b. Since these two quantities operate across different regions of λ, Shannon mutual information can misrepresent the utility of new features for conformal prediction. We demonstrate this behavior in Figure 2 using a long-tailed synthetic setting with two candidate features X_1 and X_2. Details on the data-generating process are described in Appendix B.

3.3 Data processing inequality for conformal set-size reduction

Conformal set-size reduction enjoys additional properties that make it attractive as an approximate notion of mutual information. Here, we highlight one such property by deriving a data processing inequality (DPI) for conformal prediction sets.

Suppose that for each data point we also observe a *single realization* of a degraded feature $\tilde{X}$ obtained from a (possibly noisy) channel, so that $Y \to (X, Z) \to (\tilde{X}, Z)$ forms a Markov chain. Let $\tilde{X}^{(2)} = (\tilde{X}, Z)$ and, similarly to Δ_α, define the set-size reduction with the degraded features as

$$\tilde{\Delta}_\alpha = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(1)})|\big] - \mathbb{E}\big[|\mathcal{C}_\alpha(\tilde{X}^{(2)})|\big]. \tag{13}$$

The following analysis shows that conformal set-size reduction satisfies an approximate version of the DPI by comparing Δ_α and $\tilde{\Delta}_\alpha$.

Theorem 3 (Probability score DPI). *Fix a calibration dataset $\mathcal{D}_{\text{cal}}$ and nominal coverage level $1 - \alpha$. Let $\hat{\lambda}_\alpha^{(2)}$, $D^{(2)}$, and $\varepsilon^{(2)}$ be defined as in Theorem 2 and assume $\hat{\lambda}_\alpha^{(2)} > 0$. Define the training-conditional miscoverage for the degraded features as $\tilde{\varepsilon}^{(2)} = \mathbb{P}\big(Y \notin \mathcal{C}_\alpha(\tilde{X}^{(2)}); \mathcal{D}_{\text{cal}}\big)$. Then,*

$$\tilde{\Delta}_\alpha \le \Delta_\alpha + D^{(2)} + (\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)})/\hat{\lambda}_\alpha^{(2)}. \tag{14}$$

As before, a suitable regularity condition controls the finite-sample calibration term in expectation.

Corollary 2 (Probability score DPI in expectation). *Fix a nominal coverage level $1-\alpha$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Let $\hat{\lambda}_\alpha^{(2)}$ and $D^{(2)}$ be defined as in Theorem 2, and suppose that $s^{(2)} = \mathrm{Var}(1/\hat{\lambda}_\alpha^{(2)})^{1/2} < \infty$. Then,*

$$\mathbb{E}[\tilde{\Delta}_\alpha] \le \mathbb{E}[\Delta_\alpha] + \mathbb{E}[D^{(2)}] + s^{(2)}/\sqrt{n}. \tag{15}$$

Corollary 2 establishes the approximate DPI for the conformal set-size reduction: on average over draws of the calibration data, the degraded feature $\tilde{X}$ reduces the set size by no more than X does, up to model imperfection and an $O(n^{-1/2})$ calibration term. However, in practice, these slack terms could make the degraded feature spuriously appear more informative.

3.4 EXTENSION OF CONFORMAL MUTUAL INFORMATION FOR APS SCORE

We extend our results to conformal prediction sets with the APS score. Unlike the probability score, whose calibrated threshold acts as a single global cutoff on $q(\cdot \mid x)$, the APS score effectively applies an input-dependent threshold with randomization at the boundary. Hence, we make two key extensions in our analysis. First, we index the conformal mutual information by a random $\Lambda \in (0, 1]$ in place of the fixed λ. Second, we introduce $U \sim \mathrm{Unif}[0, 1]$ independent of (Λ, X, Y) to randomize the set-valued action. Let Ω denote the space of measurable functions $[0, 1] \to 2^{\mathcal{Y}}$.

Definition 3 (Conformal entropy and mutual information for random Λ)**.** *Let $\Lambda \in (0, 1]$ be a random index variable and let $\ell_\Lambda(Y, \Gamma)$ denote the loss in Equation 5 evaluated using the value of Λ. Define the conformal entropy under Λ as*

$$H_\Lambda(Y \mid X) = \mathbb{E}\Big[\inf_{\Gamma \in \Omega} \mathbb{E}\big[\ell_\Lambda(Y, \Gamma(U)) \mid X\big]\Big], \tag{16}$$

where the inner expectation is taken over Y, U, and Λ. The corresponding conformal mutual information is $I_\Lambda(X; Y \mid Z) = H_\Lambda(Y \mid Z) - H_\Lambda(Y \mid X, Z)$.

The random Λ allows the loss index to depend on the input. Since the action space is finite, we again express the conformal entropy as an infimum over measurable decision rules $\Gamma : \mathcal{X} \times [0, 1] \to \mathcal{A}$ and show that conformal prediction sets using the APS score achieve H_Λ for suitable choices of Λ.

Proposition 3 (Optimal set-valued predictors for random Λ)**.** *Suppose that $\Lambda = \xi(X)$ for some fixed measurable function $\xi : \mathcal{X} \to (0, 1]$, and let $p(\cdot \mid X)$ be the conditional distribution of Y given X. Let $\Gamma^*_\Lambda(X, U; p)$ be any decision rule such that, for all $u \in [0, 1]$,*

$$\big\{y : p(y \mid X) > \xi(X)\big\} \subseteq \Gamma^*_\Lambda(X, u; p) \subseteq \big\{y : p(y \mid X) \geq \xi(X)\big\}. \tag{17}$$

Then, $\Gamma^*_\Lambda(X, U; p)$ minimizes $\mathbb{E}\left[\ell_\Lambda(Y, \Gamma(X, U))\right]$ over all measurable $\Gamma : \mathcal{X} \times [0, 1] \to 2^{\mathcal{Y}}$.

The conformal prediction set using an oracle model and the APS score is a Bayes-optimal action $\Gamma^*_\Lambda(X, U; p)$, where Λ is the *induced* input-dependent probability threshold equal to the first predicted probability (in decreasing order) for which the cumulative mass exceeds $\hat{\tau}_\alpha$. When $\hat{\tau}_\alpha < 1$, this quantity is well-defined and satisfies $\Lambda = \xi(X) \in (0, 1]$ for some measurable function ξ.

We define Δ_α as in Equation 9, with the expectation now also taken over U. Furthermore, we overload $\hat{\tau}^{(i)}_\alpha$ to denote the calibrated APS-score threshold using $X^{(i)}$ and $\varepsilon^{(i)}$ to be the $X^{(2)}$-conditional miscoverage. We next formalize the conformal information divergence under a random Λ.

Definition 4 (Conformal information divergence for random Λ)**.** *Suppose that $\Lambda = \xi(X)$ for some fixed measurable function $\xi : \mathcal{X} \to (0, 1]$. Define the conformal information divergence under Λ between conditional distributions $p(\cdot \mid X)$ and $q(\cdot \mid X)$ as*

$$D_\Lambda(p(\cdot \mid X) \,\|\, q(\cdot \mid X)) = \mathbb{E}_X\big[\mathbb{E}_{Y \sim p(\cdot|X),\, U}[\ell_\Lambda(Y, \Gamma^*_\Lambda(X, U; q)) - \ell_\Lambda(Y, \Gamma^*_\Lambda(X, U; p))]\big]. \tag{18}$$

In Definition 4, the value of D_Λ depends on the boundary randomization choice for $\Gamma^*_\Lambda(X, U; q)$. When Λ is the probability threshold induced by the APS score of $q(\cdot \mid X)$ at some $\tau < 1$, we use the convention that $\Gamma^*_\Lambda(X, U; q)$ is the corresponding conformal prediction set. We are now ready to extend Theorem 2 for the APS score.

Theorem 4 (APS score sandwich bound)**.** *Fix a calibration dataset $\mathcal{D}_{\mathrm{cal}}$ and nominal coverage level $1 - \alpha$. For $i \in \{1, 2\}$, assume that $\hat{\tau}^{(i)}_\alpha < 1$, and let $\hat{\Lambda}^{(i)}_\alpha$ be the input-dependent random probability threshold induced by $\hat{\tau}^{(i)}_\alpha$. Moreover, define*

$$I^{(i)} = I_{\hat{\Lambda}^{(i)}_\alpha}(X; Y \mid Z), \qquad D^{(i)} = D_{\hat{\Lambda}^{(i)}_\alpha}(p(\cdot \mid X^{(i)}) \,\|\, q(\cdot \mid X^{(i)})). \tag{19}$$

Let $\varepsilon^{(i)} = \mathbb{P}(Y \notin \mathcal{C}_\alpha(X^{(i)}) \mid X^{(2)}; \mathcal{D}_{\mathrm{cal}})$, where the probability is taken jointly over a fresh test label Y given $X^{(2)}$ and the APS randomization U. Define $\varepsilon^\Delta = \varepsilon^{(2)} - \varepsilon^{(1)}$. Then,

$$I^{(2)} - D^{(2)} + \mathbb{E}\big[\varepsilon^\Delta / \hat{\Lambda}^{(2)}_\alpha\big] \;\leq\; \Delta_\alpha \;\leq\; I^{(1)} + D^{(1)} + \mathbb{E}\big[\varepsilon^\Delta / \hat{\Lambda}^{(1)}_\alpha\big], \tag{20}$$

where the expectations are over $X^{(2)}$, and both bounds are finite.

Theorem 4 reduces to Theorem 2 when $\hat{\Lambda}$ is constant. For a random $\hat{\Lambda}$, the ε^Δ terms depend on the $X^{(2)}$-conditional miscoverage, rather than the training-conditional miscoverage. In Appendix A, we derive corresponding bounds in expectation and with high probability over the calibration data, as well as a corresponding data processing inequality. In general, the APS results incur additional slack on both sides of the bound, because the conditional miscoverage typically does not concentrate as n grows. For oracle models, this slack vanishes in the upper bound.

4 Related Work

We build on the decision-theoretic notion of information introduced by DeGroot (1962). A long line of work has studied generalized entropies in relation to proper scoring rules (Gneiting & Raftery, 2007) and information-theoretic divergences (Reid & Williamson, 2011; García-García & Williamson, 2012). We instead focus on a particular family of generalized information measures tailored to set-valued prediction. The loss in Equation 5 is related to class-selective rejection (Ha, 1996) and classification with rejection (Herbei & Wegkamp, 2006), though these ideas have largely been developed independently of conformal prediction (Chzhen et al., 2021).

A growing body of empirical work studies conformal set size as an operational measure of predictive uncertainty (Fayyad et al., 2024; Portela et al., 2025; Hagos & Lundström, 2026). On the theoretical side, several works characterize expected set size as a measure of predictive inefficiency using finite-sample estimates (Dhillon et al., 2024) and model-dependent bounds (Zecchin et al., 2024). However, these works do not formally motivate set size as a measure of uncertainty or information.

The work closest to ours is Correia et al. (2024), who use list-decoding arguments to upper bound the conditional entropy in terms of set size and coverage. In contrast, our decision-theoretic approach uses generalized entropy to exactly characterize set size and miscoverage. This formulation yields two-sided bounds on set-size reduction, which we further connect to Shannon mutual information via an integral representation. These results provide a direct information-theoretic interpretation of set-size reduction that is not captured by existing work.

Information gain metrics are widely used in practice for active learning (Houlsby et al., 2011; Kirsch et al., 2019; Bickford Smith et al., 2023) and feature acquisition (Covert et al., 2023). While these approaches typically rely on classical information-theoretic quantities, Neiswanger et al. (2022) study decision-theoretic information to optimize downstream utility. Separately, conformal prediction has been used to design acquisition policies based on prediction set size (Kharazian et al., 2024). Our main contribution is identifying theoretical connections between these notions of information gain.

5 Experiments

We empirically examine our theoretical results and then probe set-size reduction and Shannon mutual information in a feature selection experiment. Motivated by the behavior in Figure 2, we show that these two criteria can rank features differently in practice. For all experiments, we fix the nominal coverage to be $1 - \alpha = 0.9$, unless noted otherwise.

Datasets. We consider 11 datasets spanning synthetic, image, and tabular domains. We use three synthetic settings. The first (S1) defines the true conditional distribution using a two-layer ReLU network, while the second (S2) uses a Gaussian mixture model with each mixture component representing a class. For the feature selection experiments, we additionally introduce a third setting (S3), where the conditional distribution has a long low-probability tail. All synthetic datasets provide access to the true conditional probabilities, allowing us to evaluate our results using an oracle model.

For tabular domains, we use the Wine Quality (Wine) (Cortez et al., 2009), Human Activity Recognition (HAR) (Reyes-Ortiz et al., 2013), Letter Recognition (Letter) (Slate, 1991), and Covertype (Blackard, 1998) datasets. For image classification, we use FMNIST (Xiao et al., 2017), CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and ImageNet-1k (Deng et al., 2009). Dataset sizes, dimensions, numbers of classes, and preprocessing details are provided in Appendix B.

Feature construction. We partition coordinates of the synthetic data and features of the tabular datasets into disjoint subsets Z and X. For images, we take Z to be a Gaussian-blurred version of the original data and X to be its residual (see Figure 1a). The DPI experiments construct $\tilde{X}$ by either adding Gaussian noise or applying Gaussian blur to X. Candidate features in the selection experiment are partitions of the input vector for synthetic and tabular datasets or a pretrained dense representation for image datasets. Appendix B provides further details on feature construction.

Models. We train a model for each feature set on the real-world datasets and hold it fixed across subsequent calibration procedures. The tabular datasets use MLPs with two or three hidden layers, ReLU activations, and dropout. For FMNIST, CIFAR-10, and CIFAR-100, we train CNNs with max pooling, while for ImageNet-1k we fine-tune a pretrained ResNet-18 (He et al., 2016).

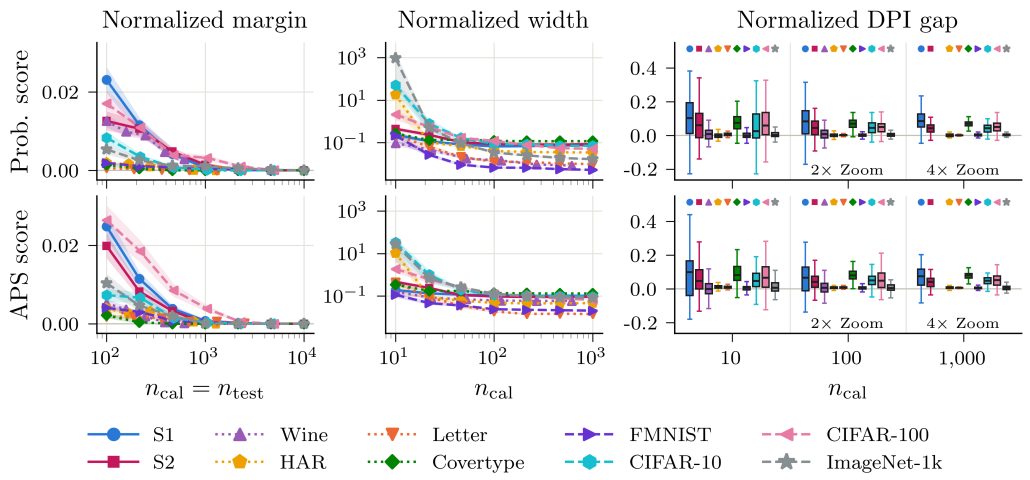

Figure 3: We show the normalized margin, width, and DPI gap based on 100 random draws of the calibration and test sets from the fixed evaluation pool. The margin and width panels report the mean $\pm$ standard error, while the DPI gap panel summarizes the distribution using boxplots. Datasets whose evaluation pool is too small for a given calibration size are omitted.

Training. We use the oracle model for the synthetic settings and generate a fixed evaluation pool of 200,000 samples. For the tabular datasets, we allocate 60% of the data to training, 15% to validation, and the remaining 25% to an evaluation pool. The image datasets have dedicated evaluation pools, and the remaining data is split into 80% for training and 20% for validation.

Across all training procedures, we use early stopping based on validation loss and select the learning rate, weight decay, and dropout rate by grid search. In Appendix C, we provide additional details on model architectures, training procedures, hyperparameters, and predictive accuracy.

5.1 Sandwich and DPI validation

We first evaluate the sandwich bounds in Theorems 2 and 4. From each evaluation pool, we sample separate calibration and test sets to estimate Δ_α and $\varepsilon^{(i)}$. In the synthetic settings, the true conditional distribution is known, so $I^{(i)}$ is computable and $D^{(i)} = 0$. For the real-world datasets, $I^{(i)}$ and $D^{(i)}$ are inaccessible, so we use the plug-in model estimates, which set $D^{(i)} = 0$. Moreover, we use a k-NN approximation for $I^{(2)}$ in the APS lower bound since $\hat{\Lambda}_\alpha^{(2)}$ depends on X (see Appendix D). The real-data results are thus an empirical corroboration rather than strict validation of the theory.

Letting L and U denote the lower and upper bounds, we measure the magnitude of violations using the normalized margin $d(\Delta_\alpha, [L, U])/K$, where $d(x, [L, U])$ is the distance from x to $[L, U]$. We separately quantify the informativeness of the bound using the normalized width $(U - L)/K$.

The left column of Figure 3 reports the normalized margin as we vary $n_{\text{cal}} = n_{\text{test}}$. We limit the synthetic sample sizes to match the real datasets for a controlled comparison. Even with oracle models, the average margin is positive since finite samples introduce Monte Carlo error, but it approaches zero as sample sizes increase. The margins on the real data exhibit similar behavior. The middle column reports the normalized width for a fixed n_{test}. The intervals narrow with more calibration data, but the width remains nonzero due to the gap between $I^{(1)}$ and $I^{(2)}$. At very small n_{cal} and large K, the $1/\hat{\lambda}_\alpha^{(i)}$ factor makes the bounds vacuous since the calibrated threshold is typically small.

We next empirically analyze Theorem 3 and its APS analogue (Theorem 5 in Appendix A). The right column of Figure 3 reports the normalized DPI gap $(\Delta_\alpha - \tilde{\Delta}_\alpha)/K$ as we vary n_{cal} with n_{test} fixed. At small calibration sizes, finite-sample effects can produce negative DPI gaps, meaning that the degraded features $\tilde{X}$ appear to reduce set size more than the original features X. As n_{cal} increases, however, these violations shrink in magnitude and the gap concentrates around non-negative values.

We include additional experimental analyses in Appendix E, including violation rates for Theorems 2 and 4, results across nominal coverage levels, and DPI behavior under varying noise levels.

Table 1: We evaluate feature selection performance using conformal set-size reduction. Each row represents one selection criterion and values are reported as mean ± standard error over 10 random splits of the evaluation pool. Arrows indicate the direction of improvement, and the best-performing method for each metric is bolded. Empirical coverage is reported for reference without bolding.

Dataset	Criterion	Prob.		APS		Acc.	
		$\mathbb{E}[\lvert\mathcal{C}\rvert]$ (↓)	Cov.	$\mathbb{E}[\lvert\mathcal{C}\rvert]$ (↓)	Cov.	Top-1 (↑)	Top-5 (↑)
S3 ($K = 40$)	Δ_α (Prob.)	$\mathbf{2.239 \pm 0.002}$	0.900 ± 0.000	2.629 ± 0.002	0.900 ± 0.000	0.750 ± 0.001	$\mathbf{0.952 \pm 0.000}$
	Δ_α (APS)	2.243 ± 0.002	0.900 ± 0.000	$\mathbf{2.611 \pm 0.002}$	0.900 ± 0.000	0.752 ± 0.000	$\mathbf{0.952 \pm 0.000}$
	Shannon MI	2.447 ± 0.005	0.900 ± 0.000	2.876 ± 0.006	0.900 ± 0.000	0.767 ± 0.000	$\mathbf{0.952 \pm 0.000}$
	Accuracy	3.613 ± 0.020	0.900 ± 0.000	4.487 ± 0.029	0.900 ± 0.000	$\mathbf{0.768 \pm 0.001}$	0.914 ± 0.001
	Random	3.046 ± 0.100	0.900 ± 0.000	3.591 ± 0.116	0.900 ± 0.000	0.672 ± 0.009	0.931 ± 0.004
CIFAR ($K = 100$)	Δ_α (Prob.)	12.683 ± 0.100	0.904 ± 0.002	15.661 ± 0.127	0.902 ± 0.002	0.466 ± 0.002	$\mathbf{0.753 \pm 0.002}$
	Δ_α (APS)	$\mathbf{12.594 \pm 0.088}$	0.902 ± 0.002	$\mathbf{15.472 \pm 0.129}$	0.902 ± 0.002	$\mathbf{0.467 \pm 0.002}$	$\mathbf{0.753 \pm 0.002}$
	Shannon MI	13.763 ± 0.115	0.902 ± 0.002	16.597 ± 0.145	0.902 ± 0.002	0.458 ± 0.003	0.741 ± 0.002
	Accuracy	12.872 ± 0.130	0.902 ± 0.002	15.877 ± 0.158	0.901 ± 0.002	$\mathbf{0.467 \pm 0.003}$	0.752 ± 0.002
	Random	14.243 ± 0.253	0.902 ± 0.003	17.189 ± 0.224	0.902 ± 0.003	0.454 ± 0.002	0.737 ± 0.002
Letter ($K = 26$)	Δ_α (Prob.)	$\mathbf{2.753 \pm 0.016}$	0.898 ± 0.003	$\mathbf{3.358 \pm 0.019}$	0.900 ± 0.001	0.756 ± 0.001	$\mathbf{0.926 \pm 0.000}$
	Δ_α (APS)	2.835 ± 0.038	0.898 ± 0.002	3.398 ± 0.032	0.900 ± 0.001	0.758 ± 0.002	0.924 ± 0.001
	Shannon MI	2.801 ± 0.019	0.899 ± 0.003	3.399 ± 0.026	0.901 ± 0.001	0.758 ± 0.001	0.924 ± 0.001
	Accuracy	2.879 ± 0.020	0.899 ± 0.002	3.415 ± 0.022	0.901 ± 0.001	$\mathbf{0.763 \pm 0.001}$	0.922 ± 0.001
	Random	4.459 ± 0.184	0.897 ± 0.002	5.054 ± 0.174	0.900 ± 0.001	0.642 ± 0.011	0.871 ± 0.007

5.2 Feature selection

We now investigate conformal set-size reduction as an information gain metric in a sequential feature selection setting with candidate features $\{X_1, \ldots, X_m\}$. At each step, we select the subsequent feature by maximizing the conformal set-size reduction or Shannon mutual information (MI), with top-1 accuracy gain and random selection as baselines. The purpose of this experiment is to empirically probe the theoretical differences between these two criteria, as highlighted in Figure 2, on real-world datasets, rather than to propose a competitive active feature acquisition method.

Evaluating the selection criteria requires estimating the label distribution for many subsets of candidate features. We amortize this computation by training a single mask-conditioned classifier that can make predictions from arbitrary subsets of observed features (Yoon et al., 2019; Covert et al., 2021). The evaluation pool is split into two halves to determine the feature ordering and evaluate its performance on held-out data. Further details on the analysis procedure are provided in Appendix D.

We measure the resulting conformal prediction set size and coverage using the probability and APS scores, as well as top-1 and top-5 accuracy. To summarize performance across the selection trajectory, we average each metric over all non-empty proper prefixes of the feature ordering. Table 1 summarizes results for one dataset from each domain, with the remaining results in Appendix E.

Selecting features according to conformal set-size reduction consistently produces the smallest sets, although the relative performance of the probability-score and APS criteria varies. This advantage is partly expected since Δ_α directly targets set size, but the gap to Shannon MI shows that it can induce different feature orderings on real data. In an ablation on CIFAR-100, this gap persists when we retrain a separate model for each selected feature subset, indicating that it is not solely an artifact of the shared mask-conditioned model. Accuracy-based selection achieves the highest top-1 accuracy, while Δ_α matches it on top-5 accuracy. We visualize metric trajectories in Appendix E.

6 Discussion

In this work, we provide new theoretical insights relating conformal prediction and information theory. Our main results show that in standard classification settings, Δ_α is bracketed by calibration-dependent members of the conformal mutual information family, which exactly recovers Shannon mutual information via an integral representation. Moreover, set-size reduction satisfies an approximate data processing inequality, further motivating it as a principled measure of information gain.

Limitations. Our theoretical analysis is limited to classification with the probability and APS scores, while our empirical evaluation considers sequential feature selection with a globally fixed ordering. Future work could extend these connections to other conformal scores and regression settings or develop active feature acquisition methods using Δ_α as a proxy objective.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing, improve the presentation and clarity of proofs, and assist with experimental code. The main theoretical results were developed by the authors, and all subsequent AI-assisted edits were reviewed by the authors. Any AI-written code was manually reviewed and subject to automated validation checks. We did not use generative AI tools to generate synthetic data or draft sections of the paper. We take responsibility for the final content of this work, including any text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2025.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, 2023.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *CoRR*, abs/1707.08819, 2017.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification – overview via a unified framework. *CoRR*, abs/2102.12318, 2021.

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.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Wine quality. UCI Machine Learning Repository, 2009.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22, 2021.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 6424–6447. PMLR, 2023.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Morris H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33, 1962.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Guneet S. Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, 2024.

John C. Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence, and surrogate risk. *The Annals of Statistics*, 46, 2018.

Jamil Fayyad, Shadi Alijani, and Homayoun Najjaran. Empirical validation of conformal prediction for trustworthy skin lesions classification. *Computer Methods and Programs in Biomedicine*, 253, 2024.

Dario García-García and Robert C. Williamson. Divergences and risks for multiclass experiments. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, 2012.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 2007.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32, 2004.

Thien M. Ha. An optimum class-selective rejection rule for pattern recognition. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 2. IEEE, 1996.

Misgina Tsighe Hagos and Claes Lundström. Performance of conformal prediction in capturing aleatoric uncertainty. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

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*, 2016.

Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34, 2006.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011.

Zahra Kharazian, Tony Lindgren, Sindri Magnusson, and Henrik Boström. Copal: Conformal prediction in active learning an algorithm for enhancing remaining useful life estimation in predictive maintenance. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230. PMLR, 2024.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

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, 2018.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8, 2023.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022.

Charles Lu, Bhawesh Kumar, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. In *ICML 2023 Workshop on Neural Conversational AI: What's Left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) Chatbots?*, 2023.

Olivier Marchal and Julyan Arbel. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22, 2017.

Willie Neiswanger, Lantao Yu, Shengjia Zhao, Chenlin Meng, and Stefano Ermon. Generalizing bayesian optimization with decision-theoretic entropies. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*. Springer, 2002.

Alberto Portela, Julio R. Banga, and Marcos Matabuena. Conformal prediction for uncertainty quantification in dynamic biological systems. *PLOS Computational Biology*, 21, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12, 2011.

Jorge Luis Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114, 2019.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17, 1989.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 2008.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Slate. Letter recognition. UCI Machine Learning Repository, 1991.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*, volume 29. Springer, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications*. Springer, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019.

Matteo Zecchin, Sangwoo Park, Osvaldo Simeone, and Fredrik Hellström. Generalization and informativeness of conformal prediction. In *2024 IEEE International Symposium on Information Theory*. IEEE, 2024.

SUPPLEMENTARY MATERIAL FOR CONFORMAL PREDICTION SETS QUANTIFY INFORMATION GAIN: A THEORETICAL PERSPECTIVE

A PROOFS

We first establish several useful results about conformal entropy and mutual information, followed by concentration results for the finite-sample miscoverage terms. We then prove our main results on conformal prediction with the probability score in Appendix A.2, and extend these results to the APS score in Appendix A.3.

A.1 PROPERTIES OF CONFORMAL ENTROPY AND MUTUAL INFORMATION

In this section, we first prove our results on optimal set-valued predictors for conformal entropy and derive the integral representation of Shannon mutual information. We also establish a data processing inequality for generalized mutual information.

A.1.1 PROPOSITION 2: OPTIMAL SET-VALUED PREDICTORS

Consider the risk for a fixed $\lambda \in (0, 1]$. Using Equation 5,

$$\mathbb{E}\left[\ell_\lambda(Y, \Gamma(X))\right] = \mathbb{E}[|\Gamma(X)|] + \tfrac{1}{\lambda}\,\mathbb{P}(Y \notin \Gamma(X)). \tag{21}$$

We can decompose the risk into a sum over the elements of $\Gamma(X)$ and its complement. Since $p(\cdot \mid X)$ is the true conditional probability of Y given X,

$$\mathbb{E}[|\Gamma(X)|] + \tfrac{1}{\lambda}\,\mathbb{P}(Y \notin \Gamma(X)) = \mathbb{E}\left[\sum_{y \in \Gamma(X)} 1 + \tfrac{1}{\lambda} \sum_{y \notin \Gamma(X)} p(y \mid X)\right]. \tag{22}$$

Using $\sum_{y \in \mathcal{Y}} p(y \mid X) = 1$, we have

$$\mathbb{E}_X\left[\mathbb{E}_{p(\cdot|X)}[\ell_\lambda(Y, \Gamma(X))]\right] = \tfrac{1}{\lambda} + \mathbb{E}\left[\sum_{y \in \Gamma(X)} \left(1 - \tfrac{1}{\lambda}\, p(y \mid X)\right)\right]. \tag{23}$$

Hence, $\Gamma(X)$ minimizes the risk pointwise in x if and only if it includes every y with $p(y \mid x) > \lambda$ and excludes every y with $p(y \mid x) < \lambda$. The decision rule that outputs probability superlevel sets

$$\Gamma^*_\lambda(X; p) = \{y \in \mathcal{Y} : p(y \mid X) \geq \lambda\} \tag{24}$$

is one such minimizer for every $x \in \mathcal{X}$ and thus achieves the conformal entropy $H_\lambda(Y \mid X)$.

A.1.2 THEOREM 1: INTEGRAL REPRESENTATION OF SHANNON INFORMATION

We first establish the corresponding integral representation for the Shannon entropy.

Lemma 1 (Integral representation of Shannon entropy)**.** *Let H denote the Shannon entropy (in nats). For any random variable X,*

$$1 + H(Y \mid X) = \int_0^1 H_\lambda(Y \mid X)\, d\lambda. \tag{25}$$

Proof. Recall that the decision rule $\Gamma^*_\lambda(X; p) = \{y : p(y \mid X) \geq \lambda\}$ achieves $H_\lambda(Y \mid X)$. At this optimum value, each $y \in \mathcal{Y}$ either increases the set size by one when $p(y \mid x) \geq \lambda$ or increases the miscoverage penalty term by $p(y \mid x)/\lambda$ otherwise. Hence,

$$H_\lambda(Y \mid X) = \mathbb{E}_X\left[\sum_{y \in \mathcal{Y}} \min\left\{1, \frac{p(y \mid X)}{\lambda}\right\}\right], \tag{26}$$

and therefore by Tonelli's theorem,

$$\int_0^1 H_\lambda(Y \mid X)\, d\lambda = \mathbb{E}_X\left[\sum_{y \in \mathcal{Y}} \int_0^1 \min\left\{1, \frac{p(y \mid X)}{\lambda}\right\} d\lambda\right]. \tag{27}$$

Evaluating the inner integral,

$$\int_0^1 \min\left\{1, \frac{p(y \mid X)}{\lambda}\right\} d\lambda = \underbrace{\int_0^{p(y|X)} d\lambda}_{p(y \mid X)} + \underbrace{\int_{p(y|X)}^1 \frac{p(y \mid X)}{\lambda}\, d\lambda}_{-p(y \mid X)\log p(y \mid X)}. \tag{28}$$

Substituting Equation 28 into Equation 27,

$$\int_0^1 H_\lambda(Y \mid X)\, d\lambda = \mathbb{E}_X\left[\sum_{y\in\mathcal{Y}} p(y \mid X) - p(y \mid X)\log p(y \mid X)\right] = 1 + H(Y \mid X). \tag{29}$$

□

The integral representation of Shannon information now follows immediately from Lemma 1.

$$H(Y \mid Z) - H(Y \mid X, Z) = \int_0^1 H_\lambda(Y \mid Z)\, d\lambda - \int_0^1 H_\lambda(Y \mid X, Z)\, d\lambda. \tag{30}$$

Thus,

$$I(X; Y \mid Z) = \int_0^1 I_\lambda(X; Y \mid Z)\, d\lambda. \tag{31}$$

A.1.3 Proposition 3: Optimal set-valued predictors for random Λ

Consider the risk when $\Lambda = \xi(X)$, where $\xi : \mathcal{X} \to (0, 1]$ is a fixed measurable function. Using Equation 5 with a fixed value of $u \in [0, 1]$,

$$\mathbb{E}\left[\ell_\Lambda(Y, \Gamma(X, u))\right] = \mathbb{E}_X[|\Gamma(X, u)|] + \mathbb{E}_X\left[\tfrac{1}{\xi(X)}\mathbb{P}(Y \notin \Gamma(X, u) \mid X)\right]. \tag{32}$$

Since $p(\cdot \mid X)$ is the true conditional probability of Y given X, the risk decomposition in Equations 22 and 23 gives that

$$\mathbb{E}\left[\ell_\Lambda(Y, \Gamma(X, u))\right] = \mathbb{E}_X\left[\tfrac{1}{\xi(X)} + \sum_{y\in\Gamma(X,u)}\left(1 - \tfrac{1}{\xi(X)}p(y \mid X)\right)\right]. \tag{33}$$

Hence, $\Gamma(X, u)$ minimizes the risk pointwise in x if and only if it includes every y with $p(y \mid x) > \xi(x)$ and excludes every y with $p(y \mid x) < \xi(x)$. Thus, any decision rule $\Gamma^*_\Lambda(X, U; p)$ such that, for all $u \in [0, 1]$,

$$\big\{y : p(y \mid X) > \xi(X)\big\} \subseteq \Gamma^*_\Lambda(X, u; p) \subseteq \big\{y : p(y \mid X) \geq \xi(X)\big\}, \tag{34}$$

minimizes $\mathbb{E}\left[\ell_\Lambda(Y, \Gamma(X, U))\right]$ and thus achieves the conformal entropy $H_\Lambda(Y \mid X)$. The decision rule may choose to include or exclude labels at the boundary $p(y \mid X) = \xi(X)$ depending on U.

A.1.4 Lemma 2: Generalized information DPI

The data processing inequality for conformal prediction sets follows from a more general result for generalized mutual information. For completeness, we include the result here. Let H_ℓ and I_ℓ denote the generalized entropy and mutual information under a loss $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, as defined in Section 2.

Lemma 2 (Generalized information DPI)**.** *Let ℓ be any loss with finite action space $\mathcal{A}$, and suppose $Y \to X \to \tilde{X}$ forms a Markov chain. Then,*

$$H_\ell(Y \mid X) \leq H_\ell(Y \mid \tilde{X}). \tag{35}$$

Equivalently, $I_\ell(\tilde{X}; Y) \leq I_\ell(X; Y)$.

Proof. Fix any decision rule g for the $\tilde{X}$-problem. Since $Y \perp \tilde{X} \mid X$,

$$\mathbb{E}\big[\ell(Y, g(\tilde{X})) \mid X, \tilde{X} = \tilde{x}\big] = \mathbb{E}\big[\ell(Y, g(\tilde{x})) \mid X\big] \geq \inf_{a\in\mathcal{A}} \mathbb{E}\big[\ell(Y, a) \mid X\big]. \tag{36}$$

Taking the joint expectation over $(X, \tilde{X})$,

$$\mathbb{E}\left[\inf_{a\in\mathcal{A}} \mathbb{E}\big[\ell(Y,a) \mid X\big]\right] \le \mathbb{E}\big[\ell(Y, g(\tilde{X}))\big]. \tag{37}$$

Therefore,

$$H_\ell(Y \mid X) \le \mathbb{E}\big[\ell(Y, g(\tilde{X}))\big]. \tag{38}$$

Minimizing the right-hand side over g gives $H_\ell(Y \mid X) \le H_\ell(Y \mid \tilde{X})$. The equivalent form expressed in terms of mutual information follows by subtracting both sides from $H_\ell(Y)$. □

A.2 Probability score results

We now prove our main results for the probability score in Section 3.2.

A.2.1 Theorem 2: Probability score sandwich bound

We wish to bound the expected conformal set-size reduction

$$\Delta_\alpha = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(1)})|\big] - \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big] \tag{39}$$

in terms of the conformal mutual information I_λ, for a fixed realization of the calibration data $\mathcal{D}_{\text{cal}}$. Since $\mathcal{D}_{\text{cal}}$ is fixed, we treat the calibrated thresholds $\hat{\lambda}_\alpha^{(1)}, \hat{\lambda}_\alpha^{(2)}$ and the training-conditional miscoverage rates $\varepsilon^{(1)}, \varepsilon^{(2)}$ as constants. A key intermediate quantity in our analysis is

$$\rho \coloneqq H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)}), \tag{40}$$

which compares the conformal entropies at the two different calibrated thresholds. We first establish a sandwich bound on ρ, then convert it into a bound on Δ_α using the relationship between conformal prediction sets and Bayes-optimal actions.

Lemma 3 (Probability score sandwich bound on ρ)**.** *Fix a calibration dataset $\mathcal{D}_{\text{cal}}$ and nominal coverage level $1-\alpha$. For $i \in \{1,2\}$, define $\hat{\lambda}_\alpha^{(i)}$, $I^{(i)}$, $D^{(i)}$, and $\varepsilon^{(i)}$ as in Theorem 2, and assume that $\hat{\lambda}_\alpha^{(i)} > 0$. Then,*

$$I^{(2)} - D^{(1)} + \varepsilon^{(1)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right) \;\le\; \rho \;\le\; I^{(1)} + D^{(2)} + \varepsilon^{(2)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right). \tag{41}$$

Proof. In both bounds we insert an intermediate conformal entropy term that isolates I_λ and leaves a cross-threshold conformal entropy difference that we control with an appropriately chosen predictor. Let $p(\cdot \mid X)$ be the true conditional distribution and $q(\cdot \mid X)$ be the predicted conditional distribution. We start with the upper bound.

Upper bound: Adding and subtracting $H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(2)})$ decomposes ρ as

$$\rho = I^{(1)} + H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)}). \tag{42}$$

It remains to upper bound the cross-threshold conformal entropy difference. By the definition of the conformal information divergence,

$$H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) = \mathbb{E}\big[|\Gamma^*_{\hat{\lambda}_\alpha^{(2)}}(X^{(2)}; q)|\big] + \frac{1}{\hat{\lambda}_\alpha^{(2)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\lambda}_\alpha^{(2)}}(X^{(2)}; q)\big) - D^{(2)}. \tag{43}$$

Under the $\ell_{\hat{\lambda}_\alpha^{(1)}}$ loss, the same predictor has risk no smaller than $H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(2)})$. Thus,

$$H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) \le \mathbb{E}\big[|\Gamma^*_{\hat{\lambda}_\alpha^{(2)}}(X^{(2)}; q)|\big] + \frac{1}{\hat{\lambda}_\alpha^{(1)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\lambda}_\alpha^{(2)}}(X^{(2)}; q)\big). \tag{44}$$

By Proposition 2, we have $\Gamma^*_{\hat{\lambda}_\alpha^{(2)}}(X^{(2)}; q) = \mathcal{C}_\alpha(X^{(2)})$, so the miscoverage probabilities in Equations 44 and 43 are both $\varepsilon^{(2)}$. Subtracting Equation 43 from Equation 44 gives

$$H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) \le D^{(2)} + \varepsilon^{(2)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right). \tag{45}$$

Substituting Equation 45 into 42 establishes the upper bound.

Lower bound: Adding and subtracting $H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(1)})$ instead decomposes ρ as

$$\rho = I^{(2)} + H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(1)}). \tag{46}$$

It remains to lower bound the cross-threshold conformal entropy difference. By the definition of the conformal information divergence,

$$H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) = \mathbb{E}\big[|\Gamma^*_{\hat{\lambda}_\alpha^{(1)}}(X^{(1)}; q)|\big] + \frac{1}{\hat{\lambda}_\alpha^{(1)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\lambda}_\alpha^{(1)}}(X^{(1)}; q)\big) - D^{(1)}. \tag{47}$$

Under the $\ell_{\hat{\lambda}_\alpha^{(2)}}$ loss, the same predictor has risk no smaller than $H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(1)})$. Thus,

$$H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(1)}) \le \mathbb{E}\big[|\Gamma^*_{\hat{\lambda}_\alpha^{(1)}}(X^{(1)}; q)|\big] + \frac{1}{\hat{\lambda}_\alpha^{(2)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\lambda}_\alpha^{(1)}}(X^{(1)}; q)\big). \tag{48}$$

By Proposition 2, we have $\Gamma^*_{\hat{\lambda}_\alpha^{(1)}}(X^{(1)}; q) = \mathcal{C}_\alpha(X^{(1)})$, so the miscoverage probabilities in Equations 48 and 47 are both $\varepsilon^{(1)}$. Subtracting Equation 48 from Equation 47 gives

$$H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(1)}) \ge \varepsilon^{(1)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right) - D^{(1)}. \tag{49}$$

Substituting Equation 49 into Equation 46 establishes the lower bound. □

We now convert the two-sided bound on ρ in Lemma 3 into the corresponding bound on Δ_α in Theorem 2. Using Proposition 2, we can write the conformal entropy at $\hat{\lambda}_\alpha^{(1)}$ and $\hat{\lambda}_\alpha^{(2)}$ as

$$H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(1)})|\big] + \frac{1}{\hat{\lambda}_\alpha^{(1)}}\,\mathbb{P}\big(Y \notin \mathcal{C}_\alpha(X^{(1)}); \mathcal{D}_{\text{cal}}\big) - D^{(1)}, \tag{50}$$

and

$$H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big] + \frac{1}{\hat{\lambda}_\alpha^{(2)}}\,\mathbb{P}\big(Y \notin \mathcal{C}_\alpha(X^{(2)}); \mathcal{D}_{\text{cal}}\big) - D^{(2)}. \tag{51}$$

Recall $\rho = H_{\hat{\lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\lambda}_\alpha^{(2)}}(Y \mid X^{(2)})$, and the miscoverage probabilities in Equations 50 and 51 are $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$, respectively. Subtracting the two identities,

$$\Delta_\alpha = \rho - \frac{\varepsilon^{(1)}}{\hat{\lambda}_\alpha^{(1)}} + \frac{\varepsilon^{(2)}}{\hat{\lambda}_\alpha^{(2)}} + D^{(1)} - D^{(2)}. \tag{52}$$

We combine Lemma 3 with the identity in Equation 52 to obtain the corresponding bounds on Δ_α.

Upper bound: Substituting the upper bound from Equation 41,

$$\Delta_\alpha \le I^{(1)} + D^{(2)} + \varepsilon^{(2)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right) - \frac{\varepsilon^{(1)}}{\hat{\lambda}_\alpha^{(1)}} + \frac{\varepsilon^{(2)}}{\hat{\lambda}_\alpha^{(2)}} + D^{(1)} - D^{(2)}. \tag{53}$$

Therefore,

$$\Delta_\alpha \le I^{(1)} + D^{(1)} + \frac{\varepsilon^{(2)} - \varepsilon^{(1)}}{\hat{\lambda}_\alpha^{(1)}}. \tag{54}$$

Lower bound: Substituting the lower bound from Equation 41,

$$\Delta_\alpha \ge I^{(2)} - D^{(1)} + \varepsilon^{(1)}\left(\frac{1}{\hat{\lambda}_\alpha^{(1)}} - \frac{1}{\hat{\lambda}_\alpha^{(2)}}\right) - \frac{\varepsilon^{(1)}}{\hat{\lambda}_\alpha^{(1)}} + \frac{\varepsilon^{(2)}}{\hat{\lambda}_\alpha^{(2)}} + D^{(1)} - D^{(2)}. \tag{55}$$

Therefore,

$$\Delta_\alpha \ge I^{(2)} - D^{(2)} + \frac{\varepsilon^{(2)} - \varepsilon^{(1)}}{\hat{\lambda}_\alpha^{(2)}}. \tag{56}$$

This completes the proof.

A.2.2 CONCENTRATION OF FINITE-SAMPLE MISCOVERAGE DIFFERENCE

The bounds in Equations 56 and 54 contain the finite-sample calibration terms $\varepsilon^{\Delta}/\hat{\lambda}_{\alpha}^{(i)}$ for $i \in \{1,2\}$ involving the difference in miscoverage rates. We now turn our attention to these terms under a random calibration dataset. Using standard arguments from conformal prediction, the marginal distribution of the coverage $1-\varepsilon^{(i)}$ is that of an order statistic of a uniform sample. We make this result precise below and include the proof for completeness.[2]

Lemma 4 (Training-conditional coverage)**.** *Fix a nominal coverage level $1-\alpha \in (0,1)$ such that $k = \lceil (1-\alpha)(n+1) \rceil \leq n$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Then,*

$$1-\varepsilon^{(i)} \sim \mathrm{Beta}(k,\, n+1-k), \qquad \textit{for } i \in \{1,2\}. \tag{57}$$

Proof. The argument is identical for $i \in \{1,2\}$, so we suppress the superscript. Let $S_1, \ldots, S_n$ be the non-conformity scores on the calibration data and S_{n+1} be the score of an independent test point. Let G denote the CDF of the score distribution. The test point is covered if and only if $S_{n+1} \leq \hat{\tau}_{\alpha} = S_{(k)}$. Hence, the training-conditional coverage is

$$1-\varepsilon = \mathbb{P}\big(S_{n+1} \leq \hat{\tau}_{\alpha}\ ;\ \mathcal{D}_{\mathrm{cal}}\big) = G\big(S_{(k)}\big). \tag{58}$$

If the scores are almost surely distinct, then G is continuous. Thus, $G(S_j) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$, so $G(S_{(k)})$ is the k-th order statistic of n uniform samples, which follows $\mathrm{Beta}(k, n+1-k)$. □

Because Lemma 4 applies for $i \in \{1,2\}$, the two miscoverages share the same mean. Thus,

$$\mathbb{E}\big[\varepsilon^{(1)}\big] = \mathbb{E}\big[\varepsilon^{(2)}\big] = 1 - \frac{k}{n+1} = \frac{n+1-k}{n+1} \leq \alpha. \tag{59}$$

Since $\mathbb{E}[\varepsilon^{\Delta}] = 0$, one might hope that the calibration term $\varepsilon^{\Delta}/\hat{\lambda}_{\alpha}^{(i)}$ also vanishes in expectation. Unfortunately, this is not the case. Taking expectations gives

$$\mathbb{E}\left[\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(i)}}\right] = \mathrm{Cov}\left(\frac{1}{\hat{\lambda}_{\alpha}^{(i)}},\ \varepsilon^{\Delta}\right), \tag{60}$$

which is generally nonzero because the calibrated thresholds are correlated with the miscoverages. We provide two approaches to control $\varepsilon^{\Delta}/\hat{\lambda}_{\alpha}^{(i)}$ without placing assumptions on the *joint* distribution of $(\varepsilon^{(1)}, \varepsilon^{(2)})$. First, we analyze its expectation when the reciprocal of the calibrated threshold has finite variance. We then derive a high-probability guarantee that only requires the calibrated threshold to be almost surely positive.

Lemma 5 (Expectation bound on the calibration term)**.** *Fix a nominal coverage level $1-\alpha$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. For $i \in \{1,2\}$, if $s^{(i)} = \mathrm{Var}(1/\hat{\lambda}_{\alpha}^{(i)})^{1/2} < \infty$, then*

$$\left|\mathbb{E}\left[\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(i)}}\right]\right| \leq \frac{s^{(i)}}{\sqrt{n}}. \tag{61}$$

Proof. Fix $i \in \{1,2\}$. Let $\mathrm{Var}(\varepsilon^{(i)}) = \sigma^2$. By Lemma 4,

$$\sigma^2 = \frac{(n+1-k)\,k}{(n+1)^2(n+2)} \leq \frac{1}{4(n+2)} \leq \frac{1}{4n}. \tag{62}$$

Expanding the variance of $\varepsilon^{\Delta} = \varepsilon^{(2)} - \varepsilon^{(1)}$ and applying Cauchy–Schwarz to the cross term,

$$\mathrm{Var}(\varepsilon^{\Delta}) = 2\sigma^2 - 2\,\mathrm{Cov}\big(\varepsilon^{(1)}, \varepsilon^{(2)}\big) \leq 2\sigma^2 + 2\sqrt{\mathrm{Var}(\varepsilon^{(1)})\,\mathrm{Var}(\varepsilon^{(2)})} = 4\sigma^2 \leq 1/n. \tag{63}$$

Since $\mathbb{E}[\varepsilon^{\Delta}] = 0$, Equation 60 and the Cauchy–Schwarz inequality give

$$\left|\mathbb{E}\left[\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(i)}}\right]\right| = \left|\mathrm{Cov}\left(\frac{1}{\hat{\lambda}_{\alpha}^{(i)}},\ \varepsilon^{\Delta}\right)\right| \leq \sqrt{\mathrm{Var}(1/\hat{\lambda}_{\alpha}^{(i)})\mathrm{Var}(\varepsilon^{\Delta})} \leq \frac{s^{(i)}}{\sqrt{n}}. \tag{64}$$

□

[2] We follow the proof given in Angelopoulos et al. (2025); see also Vovk et al. (2005).

Lemma 5 only requires $1/\hat{\lambda}_\alpha^{(i)}$ to have finite variance. Under additional smoothness assumptions (e.g., Lipschitz continuity of the reciprocal quantile function of the scores), the expectation can instead be bounded at a faster $O(1/n)$ rate. Alternatively, we can relax the condition to almost-sure positivity and control the calibration term with high probability.

Lemma 6 (High-probability bound on the calibration term)**.** *Fix a nominal coverage level $1-\alpha$ and $\delta \in (0,1)$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Suppose that there exists $\lambda_{\delta/2} > 0$ such that*

$$\mathbb{P}\big(\min\{\hat{\lambda}_\alpha^{(1)}, \hat{\lambda}_\alpha^{(2)}\} \geq \lambda_{\delta/2}\big) \geq 1 - \tfrac{\delta}{2}. \tag{65}$$

Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\mathrm{cal}}$,

$$\left|\frac{\varepsilon^\Delta}{\hat{\lambda}_\alpha^{(i)}}\right| \leq \frac{1}{\lambda_{\delta/2}}\sqrt{\frac{2\log(8/\delta)}{n}}, \qquad \textit{for both } i \in \{1,2\}. \tag{66}$$

Proof. By Lemma 4, we have $\varepsilon^{(i)} \sim \mathrm{Beta}(n+1-k,\, k)$, and let μ denote their common mean. The $\mathrm{Beta}(\alpha,\beta)$ distribution is sub-Gaussian with proxy variance $\sigma_p^2 = 1/(4(\alpha+\beta+1))$ (Marchal & Arbel, 2017). Hence, by the sub-Gaussian tail bound,

$$\mathbb{P}\big(|\varepsilon^{(i)} - \mu| \geq t\big) \leq 2\exp\big(-t^2/(2\sigma_p^2)\big) = 2\exp\big(-2(n+2)t^2\big) \leq 2\exp\big(-2nt^2\big). \tag{67}$$

Setting the right-hand side equal to $\delta/4$ gives $t_\delta = \sqrt{\log(8/\delta)/(2n)}$. By a union bound over $i \in \{1,2\}$, with probability at least $1-\frac{\delta}{2}$ we have $|\varepsilon^{(i)} - \mu| \leq t_\delta$, and therefore

$$|\varepsilon^\Delta| = \big|(\varepsilon^{(2)} - \mu) - (\varepsilon^{(1)} - \mu)\big| \leq 2t_\delta = \sqrt{\frac{2\log(8/\delta)}{n}}. \tag{68}$$

By assumption, with probability at least $1-\frac{\delta}{2}$ we have $\hat{\lambda}_\alpha^{(i)} \geq \lambda_{\delta/2}$ for both i. By a union bound, with probability at least $1-\delta$,

$$\left|\frac{\varepsilon^\Delta}{\hat{\lambda}_\alpha^{(i)}}\right| \leq \frac{|\varepsilon^\Delta|}{\lambda_{\delta/2}} \leq \frac{1}{\lambda_{\delta/2}}\sqrt{\frac{2\log(8/\delta)}{n}}. \tag{69}$$

□

A.2.3 Corollary 1: Probability score sandwich bound in expectation

The result follows by applying Lemma 5 to Theorem 2. For $i \in \{1,2\}$,

$$-\frac{s^{(i)}}{\sqrt{n}} \leq \mathbb{E}\left[\frac{\varepsilon^\Delta}{\hat{\lambda}_\alpha^{(i)}}\right] \leq \frac{s^{(i)}}{\sqrt{n}}. \tag{70}$$

Taking expectations over the calibration data in Theorem 2 gives

$$\mathbb{E}[I^{(2)}] - \mathbb{E}[D^{(2)}] + \mathbb{E}\left[\frac{\varepsilon^\Delta}{\hat{\lambda}_\alpha^{(2)}}\right] \;\leq\; \mathbb{E}[\Delta_\alpha] \;\leq\; \mathbb{E}[I^{(1)}] + \mathbb{E}[D^{(1)}] + \mathbb{E}\left[\frac{\varepsilon^\Delta}{\hat{\lambda}_\alpha^{(1)}}\right]. \tag{71}$$

Substituting the appropriate lower and upper bounds from Equation 70 into 71 completes the proof.

A.2.4 Corollary 3: Probability score sandwich bound with high probability

Corollary 3 (Probability score sandwich bound with high probability)**.** *Fix a nominal coverage level $1-\alpha$ and $\delta \in (0,1)$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Suppose that there exists $\lambda_{\delta/2} > 0$ such that*

$$\mathbb{P}(\min\{\hat{\lambda}_\alpha^{(1)}, \hat{\lambda}_\alpha^{(2)}\} \geq \lambda_{\delta/2}) \geq 1 - \tfrac{\delta}{2}. \tag{72}$$

Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\mathrm{cal}}$,

$$I^{(2)} - D^{(2)} - \sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^2}} \;\leq\; \Delta_\alpha \;\leq\; I^{(1)} + D^{(1)} + \sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^2}}. \tag{73}$$

The result follows by applying Lemma 6 to Theorem 2. By Lemma 6, with probability at least $1-\delta$,

$$-\sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^2}} \le \frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(i)}} \le \sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^2}}. \tag{74}$$

Since Theorem 2 holds for every fixed calibration dataset with $\hat{\lambda}_{\alpha}^{(i)} > 0$, on this event

$$I^{(2)} - D^{(2)} + \frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(2)}} \;\le\; \Delta_{\alpha} \;\le\; I^{(1)} + D^{(1)} + \frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(1)}}. \tag{75}$$

Substituting the appropriate lower and upper bounds from Equation 74 into 75 completes the proof.

A.2.5 Theorem 3: Probability score DPI

We convert the data processing inequality for generalized mutual information into a corresponding approximate result for conformal prediction sets for a fixed calibration dataset. Since Δ_{α} and $\tilde{\Delta}_{\alpha}$ share the term $\mathbb{E}[|\mathcal{C}_{\alpha}(X^{(1)})|]$,

$$\tilde{\Delta}_{\alpha} - \Delta_{\alpha} = \mathbb{E}\big[|\mathcal{C}_{\alpha}(X^{(2)})|\big] - \mathbb{E}\big[|\mathcal{C}_{\alpha}(\tilde{X}^{(2)})|\big]. \tag{76}$$

Using Proposition 2 and the definition of the conformal information divergence, we can write

$$\mathbb{E}\big[|\mathcal{C}_{\alpha}(X^{(2)})|\big] = H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid X^{(2)}) - \frac{\varepsilon^{(2)}}{\hat{\lambda}_{\alpha}^{(2)}} + D^{(2)}. \tag{77}$$

For the degraded features, $\mathcal{C}_{\alpha}(\tilde{X}^{(2)})$ is a feasible decision rule under the $\ell_{\hat{\lambda}_{\alpha}^{(2)}}$ loss, so its risk is no smaller than $H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid \tilde{X}^{(2)})$. Evaluating the risk gives

$$H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid \tilde{X}^{(2)}) \le \mathbb{E}\big[|\mathcal{C}_{\alpha}(\tilde{X}^{(2)})|\big] + \frac{\tilde{\varepsilon}^{(2)}}{\hat{\lambda}_{\alpha}^{(2)}}. \tag{78}$$

Combining Equations 78 and 77,

$$\tilde{\Delta}_{\alpha} - \Delta_{\alpha} \le H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid X^{(2)}) - H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid \tilde{X}^{(2)}) + D^{(2)} + \frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\lambda}_{\alpha}^{(2)}}. \tag{79}$$

By Lemma 2, $H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid X^{(2)}) \le H_{\hat{\lambda}_{\alpha}^{(2)}}(Y \mid \tilde{X}^{(2)})$. Thus,

$$\tilde{\Delta}_{\alpha} - \Delta_{\alpha} \le D^{(2)} + \frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\lambda}_{\alpha}^{(2)}}. \tag{80}$$

This completes the proof.

A.2.6 Corollary 2: Probability score DPI in expectation

Since $\tilde{\varepsilon}^{(2)}$ and $\varepsilon^{(2)}$ are both training-conditional miscoverages at coverage level $1-\alpha$, they share the same marginal distribution by Lemma 4. The result follows by applying Lemma 5 to Theorem 3 with $\varepsilon^{\Delta} = \tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}$. Thus, we get

$$\mathbb{E}\left[\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(2)}}\right] \le \frac{s^{(2)}}{\sqrt{n}}. \tag{81}$$

Taking expectations over the calibration data in Theorem 3 gives

$$\mathbb{E}[\tilde{\Delta}_{\alpha}] \;\le\; \mathbb{E}[\Delta_{\alpha}] + \mathbb{E}[D^{(2)}] + \mathbb{E}\left[\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(2)}}\right]. \tag{82}$$

Substituting Equation 81 into Equation 82 completes the proof.

A.2.7 Corollary 4: Probability score DPI with high probability

Corollary 4 (Probability score DPI with high probability)**.** *Fix a nominal coverage level $1-\alpha$ and $\delta \in (0,1)$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Suppose that there exists $\lambda_{\delta/2} > 0$ such that*

$$\mathbb{P}(\hat{\lambda}_{\alpha}^{(2)} \ge \lambda_{\delta/2}) \ge 1 - \tfrac{\delta}{2}. \tag{83}$$

Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\mathrm{cal}}$,

$$\tilde{\Delta}_{\alpha} \;\le\; \Delta_{\alpha} + D^{(2)} + \sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^2}}. \tag{84}$$

The result follows by applying the proof in Lemma 6 to Theorem 3 with $\varepsilon^{\Delta} = \tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}$. Since $\tilde{\varepsilon}^{(2)}$ and $\varepsilon^{(2)}$ are both training-conditional miscoverages at level $1-\alpha$, they share the same marginal distribution by Lemma 4. Then, Lemma 6 gives, with probability at least $1-\delta$,

$$\frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(2)}} \leq \sqrt{\frac{2\log(8/\delta)}{n\lambda_{\delta/2}^{2}}}. \tag{85}$$

Since Theorem 3 holds for every fixed calibration dataset with $\hat{\lambda}_{\alpha}^{(2)} > 0$, on this event

$$\tilde{\Delta}_{\alpha} \;\leq\; \Delta_{\alpha} + D^{(2)} + \frac{\varepsilon^{\Delta}}{\hat{\lambda}_{\alpha}^{(2)}}. \tag{86}$$

Substituting the upper bound from Equation 85 into Equation 86 completes the proof.

A.3 APS score results

We now prove our results using the APS score from Section 3.4. The arguments mirror the proofs in Appendix A.2, with the fixed calibrated thresholds replaced by the input-dependent thresholds induced by the APS score. As in Section 2, we assume without loss of generality that there are no ties among the predicted probabilities, since any ties can be broken at random following Romano et al. (2020).

A.3.1 Regularity of the induced APS threshold

We first show that $\hat{\tau}_{\alpha} < 1$ ensures $\hat{\Lambda}_{\alpha} \in (0, 1]$ as well as boundedness of $\mathbb{E}[1/\hat{\Lambda}_{\alpha}]$ and $D_{\hat{\Lambda}_{\alpha}}$.

Lemma 7 (APS threshold and divergence bounds)**.** *Fix a calibration dataset $\mathcal{D}_{\mathrm{cal}}$ and nominal coverage level $1-\alpha$. Let $\hat{\tau}_{\alpha}$ be the calibrated APS threshold and let $\hat{\Lambda}_{\alpha} = \xi(X)$ be the induced probability threshold. If $\hat{\tau}_{\alpha} < 1$, then for every $x \in \mathcal{X}$,*

$$\xi(x) \geq \frac{1-\hat{\tau}_{\alpha}}{K}. \tag{87}$$

Consequently, $\hat{\Lambda}_{\alpha} \in (0, 1]$ and $\mathbb{E}[1/\hat{\Lambda}_{\alpha}] \leq K/(1-\hat{\tau}_{\alpha})$. Moreover,

$$D_{\hat{\Lambda}_{\alpha}}\big(p(\cdot \mid X) \,\|\, q(\cdot \mid X)\big) \leq K + \frac{K}{1-\hat{\tau}_{\alpha}}. \tag{88}$$

Proof. Fix x and order the classes as in Section 2. Define $P_j = \sum_{\ell \leq j} q(y^{(\ell)} \mid x)$ with $P_0 = 0$.

Since $P_K = 1 > \hat{\tau}_{\alpha}$, there is a first index j with $P_j > \hat{\tau}_{\alpha}$, and $\xi(x) = q(y^{(j)} \mid x)$ by definition. Furthermore, we have $P_{j-1} \leq \hat{\tau}_{\alpha}$, so $P_j = P_{j-1} + \xi(x) \leq \hat{\tau}_{\alpha} + \xi(x)$. Each of the $K-j$ remaining labels has predicted probability at most $\xi(x)$, so

$$1 - \hat{\tau}_{\alpha} - \xi(x) \;\leq\; 1 - P_j \;=\; \sum_{\ell > j} q(y^{(\ell)} \mid x) \;\leq\; (K-j)\,\xi(x). \tag{89}$$

Rearranging gives $1 - \hat{\tau}_{\alpha} \leq (K-j+1)\,\xi(x) \leq K\xi(x)$, which is Equation 87. Taking expectations over X gives the bound on $\mathbb{E}[1/\hat{\Lambda}_{\alpha}]$.

For the divergence bound, the decision rule $\Gamma^{*}_{\hat{\Lambda}_{\alpha}}(X, U; q)$ has risk at most $K + \mathbb{E}[1/\hat{\Lambda}_{\alpha}]$ when the conditional distribution is $p(\cdot \mid X)$. Equation 88 then follows from the definition of $D_{\hat{\Lambda}_{\alpha}}$. □

A.3.2 Theorem 4: APS score sandwich bound

We wish to bound the expected conformal set-size reduction

$$\Delta_{\alpha} = \mathbb{E}\big[|\mathcal{C}_{\alpha}(X^{(1)})|\big] - \mathbb{E}\big[|\mathcal{C}_{\alpha}(X^{(2)})|\big] \tag{90}$$

in terms of the conformal mutual information I_{Λ} under a random index, for a fixed realization of the calibration data $\mathcal{D}_{\mathrm{cal}}$. For $i \in \{1, 2\}$, let $\xi^{(i)}$ be the fixed measurable map sending an input to its induced probability threshold, so that $\hat{\Lambda}_{\alpha}^{(i)} = \xi^{(i)}(X^{(i)})$.

Since $\mathcal{D}_{\text{cal}}$ is fixed, the conditional miscoverages $\varepsilon^{(1)}, \varepsilon^{(2)}$ are deterministic functions of $X^{(2)}$, where $\varepsilon^{(i)} = \mathbb{P}(Y \notin \mathcal{C}_\alpha(X^{(i)}) \mid X^{(2)}; \mathcal{D}_{\text{cal}})$ with the conditional probability taken over the fresh test label $Y \sim p(\cdot \mid X^{(2)})$ and the independent randomization $U \sim \text{Unif}[0,1]$.

A key intermediate quantity in our analysis is

$$\rho \coloneqq H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}), \tag{91}$$

which compares the conformal entropies at the two different calibrated thresholds. We first establish a sandwich bound on ρ, then convert it into a bound on Δ_α by making use of the fact that conformal prediction sets using the APS score satisfy the Bayes-optimality criterion in Proposition 3.

Lemma 8 (APS score sandwich bound on ρ)**.** *Fix a calibration dataset $\mathcal{D}_{\text{cal}}$ and nominal coverage level $1-\alpha$. For $i \in \{1,2\}$, assume that $\hat{\tau}_\alpha^{(i)} < 1$, and define $\hat{\Lambda}_\alpha^{(i)}$, $I^{(i)}$, $D^{(i)}$, and $\varepsilon^{(i)}$ as in Theorem 4. Then,*

$$I^{(2)} - D^{(1)} + \mathbb{E}\left[\varepsilon^{(1)}\left(\frac{1}{\hat{\Lambda}_\alpha^{(1)}} - \frac{1}{\hat{\Lambda}_\alpha^{(2)}}\right)\right] \;\le\; \rho \;\le\; I^{(1)} + D^{(2)} + \mathbb{E}\left[\varepsilon^{(2)}\left(\frac{1}{\hat{\Lambda}_\alpha^{(1)}} - \frac{1}{\hat{\Lambda}_\alpha^{(2)}}\right)\right], \tag{92}$$

where the expectations are over $X^{(2)}$, and both bounds are finite.

Proof. The proof closely mirrors the argument in Lemma 3. Let $p(\cdot \mid X)$ be the true conditional distribution of Y given X and $q(\cdot \mid X)$ be the predicted conditional distribution. By Lemma 7, we have $\hat{\Lambda}_\alpha^{(i)} \in (0,1]$ with $\mathbb{E}[1/\hat{\Lambda}_\alpha^{(i)}]$ and $D^{(i)}$ finite, so we can make use of Proposition 3 and every term in our analysis is finite.

Upper bound: Adding and subtracting $H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(2)})$ decomposes ρ as

$$\rho = I^{(1)} + H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}). \tag{93}$$

It remains to upper bound the cross-threshold conformal entropy difference. By the definition of the conformal information divergence,

$$H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) = \mathbb{E}\big[|\Gamma^*_{\hat{\Lambda}_\alpha^{(2)}}(X^{(2)}, U; q)|\big] + \mathbb{E}\left[\frac{1}{\hat{\Lambda}_\alpha^{(2)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\Lambda}_\alpha^{(2)}}(X^{(2)}, U; q) \mid X^{(2)}\big)\right] - D^{(2)}. \tag{94}$$

Under the $\ell_{\hat{\Lambda}_\alpha^{(1)}}$ loss, the same predictor has risk no smaller than $H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(2)})$. Thus,

$$H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) \le \mathbb{E}\big[|\Gamma^*_{\hat{\Lambda}_\alpha^{(2)}}(X^{(2)}, U; q)|\big] + \mathbb{E}\left[\frac{1}{\hat{\Lambda}_\alpha^{(1)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\Lambda}_\alpha^{(2)}}(X^{(2)}, U; q) \mid X^{(2)}\big)\right]. \tag{95}$$

By Proposition 3, we can choose $\Gamma^*_{\hat{\Lambda}_\alpha^{(2)}}(X^{(2)}, U; q) = \mathcal{C}_\alpha(X^{(2)})$ since $\hat{\Lambda}_\alpha^{(2)} = \xi^{(2)}(X^{(2)})$.

Hence, the miscoverage terms in Equations 94 and 95 become $\mathbb{E}[\varepsilon^{(2)}/\hat{\Lambda}_\alpha^{(2)}]$ and $\mathbb{E}[\varepsilon^{(2)}/\hat{\Lambda}_\alpha^{(1)}]$, respectively. Subtracting Equation 94 from Equation 95 gives

$$H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(2)}) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) \le D^{(2)} + \mathbb{E}\left[\varepsilon^{(2)}\left(\frac{1}{\hat{\Lambda}_\alpha^{(1)}} - \frac{1}{\hat{\Lambda}_\alpha^{(2)}}\right)\right]. \tag{96}$$

Substituting Equation 96 into 93 establishes the upper bound.

Lower bound: Adding and subtracting $H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(1)})$ instead decomposes ρ as

$$\rho = I^{(2)} + H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(1)}). \tag{97}$$

It remains to lower bound the cross-threshold conformal entropy difference. By the definition of the conformal information divergence and the law of iterated expectation,

$$H_{\hat{\Lambda}_\alpha^{(1)}}(Y \mid X^{(1)}) = \mathbb{E}\big[|\Gamma^*_{\hat{\Lambda}_\alpha^{(1)}}(X^{(1)}, U; q)|\big] + \mathbb{E}\left[\frac{1}{\hat{\Lambda}_\alpha^{(1)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\Lambda}_\alpha^{(1)}}(X^{(1)}, U; q) \mid X^{(2)}\big)\right] - D^{(1)}. \tag{98}$$

Under the $\ell_{\hat{\Lambda}_\alpha^{(2)}}$ loss, the same predictor has risk no smaller than $H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(1)})$. Thus,

$$H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(1)}) \le \mathbb{E}\big[|\Gamma^*_{\hat{\Lambda}_\alpha^{(1)}}(X^{(1)}, U; q)|\big] + \mathbb{E}\left[\frac{1}{\hat{\Lambda}_\alpha^{(2)}}\,\mathbb{P}\big(Y \notin \Gamma^*_{\hat{\Lambda}_\alpha^{(1)}}(X^{(1)}, U; q) \mid X^{(2)}\big)\right]. \tag{99}$$

By Proposition 3, we can choose $\Gamma^*_{\hat{\Lambda}^{(1)}_\alpha}(X^{(1)}, U; q) = \mathcal{C}_\alpha(X^{(1)})$ since $\hat{\Lambda}^{(1)}_\alpha = \xi^{(1)}(X^{(1)})$.

Hence, the miscoverage terms in Equations 98 and 99 become $\mathbb{E}[\varepsilon^{(1)}/\hat{\Lambda}^{(1)}_\alpha]$ and $\mathbb{E}[\varepsilon^{(1)}/\hat{\Lambda}^{(2)}_\alpha]$, respectively. Subtracting Equation 99 from Equation 98 gives

$$H_{\hat{\Lambda}^{(1)}_\alpha}(Y \mid X^{(1)}) - H_{\hat{\Lambda}^{(2)}_\alpha}(Y \mid X^{(1)}) \geq \mathbb{E}\left[\varepsilon^{(1)}\left(\frac{1}{\hat{\Lambda}^{(1)}_\alpha} - \frac{1}{\hat{\Lambda}^{(2)}_\alpha}\right)\right] - D^{(1)}. \tag{100}$$

Substituting Equation 100 into Equation 97 establishes the lower bound. □

We now convert the two-sided bound on ρ in Lemma 8 into the corresponding bound on Δ_α in Theorem 4. Because $\hat{\Lambda}^{(i)}_\alpha = \xi^{(i)}(X^{(i)})$ and $\hat{\tau}^{(i)}_\alpha < 1$, Proposition 3 lets us write the conformal entropy at $\hat{\Lambda}^{(1)}_\alpha$ and $\hat{\Lambda}^{(2)}_\alpha$ as

$$H_{\hat{\Lambda}^{(1)}_\alpha}(Y \mid X^{(1)}) = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(1)})|\big] + \mathbb{E}\left[\frac{\varepsilon^{(1)}}{\hat{\Lambda}^{(1)}_\alpha}\right] - D^{(1)}, \tag{101}$$

and

$$H_{\hat{\Lambda}^{(2)}_\alpha}(Y \mid X^{(2)}) = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big] + \mathbb{E}\left[\frac{\varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] - D^{(2)}, \tag{102}$$

where each term is finite by Lemma 7. Recall $\rho = H_{\hat{\Lambda}^{(1)}_\alpha}(Y \mid X^{(1)}) - H_{\hat{\Lambda}^{(2)}_\alpha}(Y \mid X^{(2)})$. Subtracting the two identities,

$$\Delta_\alpha = \rho - \mathbb{E}\left[\frac{\varepsilon^{(1)}}{\hat{\Lambda}^{(1)}_\alpha}\right] + \mathbb{E}\left[\frac{\varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] + D^{(1)} - D^{(2)}. \tag{103}$$

We combine Lemma 8 with the identity in Equation 103 to obtain the corresponding bounds on Δ_α.

Upper bound: Substituting the upper bound from Equation 92,

$$\Delta_\alpha \leq I^{(1)} + D^{(2)} + \mathbb{E}\left[\varepsilon^{(2)}\left(\frac{1}{\hat{\Lambda}^{(1)}_\alpha} - \frac{1}{\hat{\Lambda}^{(2)}_\alpha}\right)\right] - \mathbb{E}\left[\frac{\varepsilon^{(1)}}{\hat{\Lambda}^{(1)}_\alpha}\right] + \mathbb{E}\left[\frac{\varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] + D^{(1)} - D^{(2)}. \tag{104}$$

Therefore,

$$\Delta_\alpha \leq I^{(1)} + D^{(1)} + \mathbb{E}\left[\frac{\varepsilon^{(2)} - \varepsilon^{(1)}}{\hat{\Lambda}^{(1)}_\alpha}\right]. \tag{105}$$

Lower bound: Substituting the lower bound from Equation 92,

$$\Delta_\alpha \geq I^{(2)} - D^{(1)} + \mathbb{E}\left[\varepsilon^{(1)}\left(\frac{1}{\hat{\Lambda}^{(1)}_\alpha} - \frac{1}{\hat{\Lambda}^{(2)}_\alpha}\right)\right] - \mathbb{E}\left[\frac{\varepsilon^{(1)}}{\hat{\Lambda}^{(1)}_\alpha}\right] + \mathbb{E}\left[\frac{\varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] + D^{(1)} - D^{(2)}. \tag{106}$$

Therefore,

$$\Delta_\alpha \geq I^{(2)} - D^{(2)} + \mathbb{E}\left[\frac{\varepsilon^{(2)} - \varepsilon^{(1)}}{\hat{\Lambda}^{(2)}_\alpha}\right]. \tag{107}$$

This completes the proof.

A.3.3 Decomposition of the APS calibration term

Unlike for the probability score, the calibration terms involve a threshold $\hat{\Lambda}^{(i)}_\alpha$ and conditional miscoverage $\varepsilon^{(i)}$ which vary with the test input $X^{(2)}$. We next isolate the part of the calibration term that concentrates with n from an input-dependent covariance that does not.

Lemma 9 (Decomposition of the APS calibration term)**.** *Fix a calibration dataset $\mathcal{D}_{\text{cal}}$ and nominal coverage level $1-\alpha$. For $i \in \{1, 2\}$, assume $\hat{\tau}^{(i)}_\alpha < 1$ and define*

$$\bar{\varepsilon}^{(i)} = \mathbb{P}\big(Y \notin \mathcal{C}_\alpha(X^{(i)}); \mathcal{D}_{\text{cal}}\big), \qquad M^{(i)} = \mathbb{E}\left[\frac{1}{\hat{\Lambda}^{(i)}_\alpha}\right], \tag{108}$$

and let $\bar{\varepsilon}^{\Delta} = \bar{\varepsilon}^{(2)} - \bar{\varepsilon}^{(1)}$. Then,

$$\mathbb{E}_{X^{(2)}}\left[\frac{\varepsilon^{\Delta}}{\hat{\Lambda}^{(i)}_\alpha}\right] = \bar{\varepsilon}^{\Delta}\, M^{(i)} + R^{(i)}, \qquad \textit{where } R^{(i)} \coloneqq \mathrm{Cov}_{X^{(2)}}\left(\varepsilon^{\Delta},\, \frac{1}{\hat{\Lambda}^{(i)}_\alpha}\right). \tag{109}$$

Moreover, let $\bar{\varepsilon}_Z^{(i)} = \mathbb{P}(Y \notin \mathcal{C}_\alpha(X^{(i)}) \mid X^{(1)}; \mathcal{D}_{\text{cal}})$ *denote the* $X^{(1)}$*-conditional miscoverage. Define*

$$\kappa^{(2)} = \sqrt{\mathrm{Var}_{X^{(2)}}(\varepsilon^{(2)} - \varepsilon^{(1)})}, \qquad \bar{\kappa}^{(1)} = \sqrt{\mathrm{Var}_{X^{(1)}}(\bar{\varepsilon}_Z^{(2)} - \bar{\varepsilon}_Z^{(1)})} \tag{110}$$

and

$$v^{(2)} = \sqrt{\mathrm{Var}_{X^{(2)}}(1/\hat{\Lambda}_\alpha^{(2)})}, \qquad \bar{v}^{(1)} = \sqrt{\mathrm{Var}_{X^{(1)}}(1/\hat{\Lambda}_\alpha^{(1)})}. \tag{111}$$

Then, $|R^{(1)}| \le \bar{\kappa}^{(1)}\, \bar{v}^{(1)}$ *and* $|R^{(2)}| \le \kappa^{(2)}\, v^{(2)}$.

Proof. By the law of iterated expectation, $\bar{\varepsilon}^{(i)} = \mathbb{E}_{X^{(2)}}[\varepsilon^{(i)}]$. Thus, Equation 109 follows from the identity $\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B] + \mathrm{Cov}(A, B)$ with $A = \varepsilon^\Delta$ and $B = 1/\hat{\Lambda}_\alpha^{(i)}$, together with Lemma 7.

By the law of total covariance,

$$\mathrm{Cov}_{X^{(2)}}\left(\varepsilon^\Delta,\, \frac{1}{\hat{\Lambda}_\alpha^{(1)}}\right) = \mathrm{Cov}_Z\left(\mathbb{E}_X[\varepsilon^\Delta \mid Z],\, \frac{1}{\hat{\Lambda}_\alpha^{(1)}}\right). \tag{112}$$

By the law of iterated expectation, $\mathbb{E}_X[\varepsilon^{(i)} \mid Z] = \bar{\varepsilon}_Z^{(i)}$. The bounds on $R^{(i)}$ now follow directly from the Cauchy–Schwarz inequality for covariance, $|\mathrm{Cov}(A, B)| \le \sqrt{\mathrm{Var}(A)\,\mathrm{Var}(B)}$, together with Lemma 7 to ensure that $\mathrm{Var}(1/\hat{\Lambda}_\alpha^{(i)})$ is finite. □

The κ factors quantify the spread of the coverage gap across inputs, while the v factors measure the spread of the induced APS threshold. These typically will not shrink with n. However, under an oracle model, $R^{(1)}$ vanishes since the randomized APS score achieves exact conditional coverage. We formalize this next.

Lemma 10 (Oracle conditional coverage)**.** *Assume the oracle model* $q(\cdot \mid X^{(i)}) = p(\cdot \mid X^{(i)})$ *for* $i \in \{1, 2\}$*. Then, for each* $i \in \{1, 2\}$ *and every realization of* $\mathcal{D}_{\text{cal}}$*,*

$$\mathbb{P}\big(Y \notin \mathcal{C}_\alpha(X^{(i)}) \;\big|\; X^{(i)}; \mathcal{D}_{\text{cal}}\big) = 1 - \hat{\tau}_\alpha^{(i)} \qquad \textit{almost surely,} \tag{113}$$

where the conditional probability is over a fresh label $Y \sim p(\cdot \mid X^{(i)})$ *and* $U \sim \mathrm{Unif}[0, 1]$.

Proof. Fix x and order the classes so that $p(y^{(1)} \mid x) > \cdots > p(y^{(K)} \mid x)$. Write

$$P_j = \sum_{\ell \le j} p(y^{(\ell)} \mid x), \tag{114}$$

with $P_0 = 0$. If $Y = y^{(j)}$, the randomized APS score is $s(x, Y) = P_{j-1} + U\, p(y^{(j)} \mid x)$, which is uniform on $(P_{j-1}, P_j]$. Since $\mathbb{P}(Y = y^{(j)} \mid X = x) = P_j - P_{j-1}$, mixing over all j gives $s(x, Y) \mid X = x \sim \mathrm{Unif}(0, 1]$. Hence,

$$\mathbb{P}\big(Y \notin \mathcal{C}_\alpha(x) \mid X = x\big) = \mathbb{P}\big(s(x, Y) > \hat{\tau}_\alpha^{(i)} \mid X = x\big) = 1 - \hat{\tau}_\alpha^{(i)}. \tag{115}$$

Thus, Equation 113 holds almost surely. □

Using Lemma 10 with $X^{(1)}$, we get $\bar{\varepsilon}_Z^{(1)} = 1 - \hat{\tau}_\alpha^{(1)}$. With $X^{(2)}$ and the law of total probability,

$$\bar{\varepsilon}_Z^{(2)} = \mathbb{E}\big[\varepsilon^{(2)} \mid X^{(1)}\big] = \mathbb{E}\big[1 - \hat{\tau}_\alpha^{(2)} \mid X^{(1)}\big] = 1 - \hat{\tau}_\alpha^{(2)}. \tag{116}$$

Given a fixed calibration dataset, $\bar{\varepsilon}_Z^{(2)} - \bar{\varepsilon}_Z^{(1)} = \hat{\tau}_\alpha^{(1)} - \hat{\tau}_\alpha^{(2)}$ has zero variance and $\bar{\kappa}^{(1)} = 0$. Thus, $R^{(1)} = 0$ under an oracle model. The same argument does not apply to $R^{(2)}$, since $\varepsilon^{(1)}$ is conditioned on $X^{(2)}$, whereas the corresponding predictor only observes $X^{(1)}$.

A.3.4 Corollary 5: APS score sandwich bound in expectation

Corollary 5 (APS score sandwich bound in expectation)**.** *Fix a nominal coverage level* $1 - \alpha$*. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. For* $i \in \{1, 2\}$*, define* $I^{(i)}$ *and* $D^{(i)}$ *as in Theorem 4 and* $\kappa^{(2)}$*,* $\bar{\kappa}^{(1)}$*,* $v^{(2)}$*,* $\bar{v}^{(1)}$*, and* $M^{(i)}$ *as in Lemma 9. Furthermore, suppose that* $\mathbb{E}[(1 - \hat{\tau}_\alpha^{(i)})^{-2}] < \infty$ *and let* $m^{(i)} = \mathrm{Var}(M^{(i)})^{1/2}$*. Then,*

$$\mathbb{E}[I^{(2)} - D^{(2)} - \kappa^{(2)}\, v^{(2)}] - \frac{m^{(2)}}{\sqrt{n}} \;\le\; \mathbb{E}[\Delta_\alpha] \;\le\; \mathbb{E}[I^{(1)} + D^{(1)} + \bar{\kappa}^{(1)}\, \bar{v}^{(1)}] + \frac{m^{(1)}}{\sqrt{n}}, \tag{117}$$

where all expectations are over the calibration dataset $\mathcal{D}_{\text{cal}}$*, and both bounds are finite.*

Proof. By Lemma 9,

$$\mathbb{E}_{X^{(2)}}\left[\frac{\varepsilon^{\Delta}}{\hat{\Lambda}_{\alpha}^{(i)}}\right] = \bar{\varepsilon}^{\Delta}\, M^{(i)} + R^{(i)}. \tag{118}$$

By Lemma 4, we have $1 - \bar{\varepsilon}^{(i)} \sim \mathrm{Beta}(k, n+1-k)$, where $k = \lceil (1-\alpha)(n+1) \rceil$. Hence, $\mathbb{E}[\bar{\varepsilon}^{\Delta}] = 0$ and, following the proof of Lemma 5, we have $\mathrm{Var}(\bar{\varepsilon}^{\Delta}) \le 1/n$. Lemma 7 also gives that $M^{(i)} \le K/(1-\hat{\tau}_{\alpha}^{(i)})$, so $m^{(i)}$ is finite. By the Cauchy–Schwarz inequality,

$$\left|\mathbb{E}[\bar{\varepsilon}^{\Delta}\, M^{(i)}]\right| = \left|\mathrm{Cov}(\bar{\varepsilon}^{\Delta}, M^{(i)})\right| \le \frac{m^{(i)}}{\sqrt{n}}. \tag{119}$$

We also have $\mathbb{E}[R^{(2)}] \ge -\mathbb{E}[\kappa^{(2)}\, v^{(2)}]$ and $\mathbb{E}[R^{(1)}] \le \mathbb{E}[\bar{\kappa}^{(1)}\, \bar{v}^{(1)}]$ by Lemma 9. Taking expectations over $\mathcal{D}_{\mathrm{cal}}$ in Theorem 4 and then substituting Equation 118 together with the bounds on $\mathbb{E}[R^{(i)}]$ gives Equation 117.

It remains to show that all other terms are finite. Since $\mathbb{E}[(1-\hat{\tau}_{\alpha}^{(i)})^{-2}] < \infty$, we have $\hat{\tau}_{\alpha}^{(i)} < 1$ almost surely, so Theorem 4 applies. Lemma 7 then gives

$$\frac{1}{\hat{\Lambda}_{\alpha}^{(i)}} \le \frac{K}{1-\hat{\tau}_{\alpha}^{(i)}}, \qquad D^{(i)} \le \frac{2K}{1-\hat{\tau}_{\alpha}^{(i)}},$$

where the right-hand sides have finite expectation since $\mathbb{E}[(1-\hat{\tau}_{\alpha}^{(i)})^{-2}] < \infty$. Since $v^{(2)}$ and $\bar{v}^{(1)}$ are standard deviations of $1/\hat{\Lambda}_{\alpha}^{(i)}$ over the test input, they are also bounded by $K/(1-\hat{\tau}_{\alpha}^{(i)})$. Moreover, $\kappa^{(2)}, \bar{\kappa}^{(1)} \le 1$. Thus, both bounds are finite. □

A.3.5 Corollary 6: APS score sandwich bound with high probability

Corollary 6 (APS score sandwich bound with high probability)**.** *Fix a nominal coverage level $1-\alpha$ and $\delta \in (0,1)$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Suppose that there exists a constant $\eta_{\delta/2} < 1$ such that*

$$\mathbb{P}\big(\max\{\hat{\tau}_{\alpha}^{(1)},\, \hat{\tau}_{\alpha}^{(2)}\} \le \eta_{\delta/2}\big) \ge 1 - \tfrac{\delta}{2}. \tag{120}$$

Let $c_{\delta/2} = \frac{K}{1-\eta_{\delta/2}}$. Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\mathrm{cal}}$,

$$I^{(2)} - D^{(2)} - \sqrt{\frac{2c_{\delta/2}^2 \log(8/\delta)}{n}} - \kappa^{(2)}\, v^{(2)} \;\le\; \Delta_{\alpha} \;\le\; I^{(1)} + D^{(1)} + \sqrt{\frac{2c_{\delta/2}^2 \log(8/\delta)}{n}} + \bar{\kappa}^{(1)}\, \bar{v}^{(1)}, \tag{121}$$

where $I^{(i)}$ and $D^{(i)}$ are defined as in Theorem 4 and $\bar{\kappa}^{(1)}$, $\kappa^{(2)}$, $\bar{v}^{(1)}$, and $v^{(2)}$ are defined as in Lemma 9.

Proof. By Lemma 9, if $\hat{\tau}_{\alpha}^{(i)} < 1$, then

$$\mathbb{E}_{X^{(2)}}\left[\frac{\varepsilon^{\Delta}}{\hat{\Lambda}_{\alpha}^{(i)}}\right] = \bar{\varepsilon}^{\Delta}\, M^{(i)} + R^{(i)}. \tag{122}$$

By Lemma 4, $\bar{\varepsilon}^{(i)} \sim \mathrm{Beta}(n+1-k, k)$. Following the proof of Lemma 6, each $\bar{\varepsilon}^{(i)}$ is sub-Gaussian with proxy variance $1/(4(n+2))$, so with probability at least $1 - \frac{\delta}{2}$,

$$|\bar{\varepsilon}^{\Delta}| \le \sqrt{\frac{2\log(8/\delta)}{n}}. \tag{123}$$

By assumption, with probability at least $1 - \frac{\delta}{2}$ we have $\hat{\tau}_{\alpha}^{(i)} \le \eta_{\delta/2}$ for both i, so Lemma 7 gives $M^{(i)} \le c_{\delta/2}$. By a union bound, with probability at least $1-\delta$,

$$|\bar{\varepsilon}^{\Delta}\, M^{(i)}| \le c_{\delta/2}\sqrt{\frac{2\log(8/\delta)}{n}} = \sqrt{\frac{2c_{\delta/2}^2 \log(8/\delta)}{n}}. \tag{124}$$

Moreover, $R^{(2)} \ge -\kappa^{(2)}\, v^{(2)}$ and $R^{(1)} \le \bar{\kappa}^{(1)}\, \bar{v}^{(1)}$ by Lemma 9. Since Theorem 4 holds for every fixed calibration dataset with $\hat{\tau}_{\alpha}^{(i)} < 1$, substituting these bounds completes the proof. □

A.3.6 THEOREM 5: APS SCORE DPI

We extend the data processing inequality to the APS score. As in Section 3.3, suppose that for each data point we also observe a *single realization* of a degraded feature $\tilde{X}$ obtained from a (possibly noisy) channel, so that $Y \to (X, Z) \to (\tilde{X}, Z)$ forms a Markov chain. Let $\tilde{X}^{(2)} = (\tilde{X}, Z)$ and define the set-size reduction with the degraded features as

$$\tilde{\Delta}_\alpha = \mathbb{E}[|\mathcal{C}_\alpha(X^{(1)})|] - \mathbb{E}[|\mathcal{C}_\alpha(\tilde{X}^{(2)})|]. \tag{125}$$

Theorem 5 (APS score DPI)**.** *Fix a calibration dataset $\mathcal{D}_{\mathrm{cal}}$ and nominal coverage level $1-\alpha$. Define the conditional miscoverage $\tilde{\varepsilon}^{(2)} = \mathbb{P}(Y \notin \mathcal{C}_\alpha(\tilde{X}^{(2)}) \mid X^{(2)}; \mathcal{D}_{\mathrm{cal}})$, where the probability is taken jointly over a noisy realization of $\tilde{X}$ and fresh test label Y given $X^{(2)}$ and the APS randomization U. Moreover, let $\hat{\Lambda}_\alpha^{(2)}$, $D^{(2)}$, and $\varepsilon^{(2)}$ be defined as in Theorem 4, and assume that $\hat{\tau}_\alpha^{(2)} < 1$. Then,*

$$\tilde{\Delta}_\alpha \le \Delta_\alpha + D^{(2)} + \mathbb{E}\big[(\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)})/\hat{\Lambda}_\alpha^{(2)}\big], \tag{126}$$

where the expectation is over $X^{(2)}$, and the right-hand side is finite.

Proof. We convert the data processing inequality for generalized mutual information into a corresponding approximate result for conformal prediction sets for a fixed calibration dataset. Since Δ_α and $\tilde{\Delta}_\alpha$ share the term $\mathbb{E}[|\mathcal{C}_\alpha(X^{(1)})|]$,

$$\tilde{\Delta}_\alpha - \Delta_\alpha = \mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big] - \mathbb{E}\big[|\mathcal{C}_\alpha(\tilde{X}^{(2)})|\big]. \tag{127}$$

Because $\hat{\Lambda}_\alpha^{(2)} = \xi^{(2)}(X^{(2)})$ is a function of $X^{(2)}$, using Proposition 3 gives

$$\mathbb{E}\big[|\mathcal{C}_\alpha(X^{(2)})|\big] = H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) - \mathbb{E}\left[\frac{\varepsilon^{(2)}}{\hat{\Lambda}_\alpha^{(2)}}\right] + D^{(2)}, \tag{128}$$

where each term is finite by Lemma 7. For the degraded features, $\mathcal{C}_\alpha(\tilde{X}^{(2)})$ is a feasible decision rule under the $\ell_{\hat{\Lambda}_\alpha^{(2)}}$ loss, so its risk is no smaller than $H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid \tilde{X}^{(2)})$. Evaluating the risk and using the law of iterated expectation gives

$$H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid \tilde{X}^{(2)}) \le \mathbb{E}\big[|\mathcal{C}_\alpha(\tilde{X}^{(2)})|\big] + \mathbb{E}\left[\frac{\tilde{\varepsilon}^{(2)}}{\hat{\Lambda}_\alpha^{(2)}}\right]. \tag{129}$$

Combining Equations 129 and 128,

$$\tilde{\Delta}_\alpha - \Delta_\alpha \le H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid \tilde{X}^{(2)}) + D^{(2)} + \mathbb{E}\left[\frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\Lambda}_\alpha^{(2)}}\right]. \tag{130}$$

Since $\hat{\Lambda}_\alpha^{(2)} = \xi^{(2)}(X^{(2)})$ is a function of $X^{(2)}$, the loss index is fixed given $X^{(2)}$ and the proof of Lemma 2 applies to $\ell_{\hat{\Lambda}_\alpha^{(2)}}$, so $H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X^{(2)}) \le H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid \tilde{X}^{(2)})$. Thus,

$$\tilde{\Delta}_\alpha - \Delta_\alpha \le D^{(2)} + \mathbb{E}\left[\frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\Lambda}_\alpha^{(2)}}\right]. \tag{131}$$

Finiteness of each term on the right-hand side follows from Lemma 7. □

As with the sandwich bound in Theorem 4, we provide versions of the DPI either in expectation or with high probability over the calibration data.

A.3.7 COROLLARY 7: APS SCORE DPI IN EXPECTATION

Corollary 7 (APS score DPI in expectation)**.** *Fix a nominal coverage level $1-\alpha$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Define $D^{(2)}$ as in Theorem 4 and $v^{(2)}$ as in Lemma 9, and let $\tilde{\kappa}^{(2)} = \sqrt{\mathrm{Var}_{X^{(2)}}(\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)})}$. Moreover, suppose $\hat{\tau}_\alpha^{(2)}$ satisfies the conditions of Corollary 5, with $m^{(2)}$ defined accordingly. Then,*

$$\mathbb{E}[\tilde{\Delta}_\alpha] \;\le\; \mathbb{E}[\Delta_\alpha] + \mathbb{E}[D^{(2)}] + \frac{m^{(2)}}{\sqrt{n}} + \mathbb{E}\big[\tilde{\kappa}^{(2)}\, v^{(2)}\big], \tag{132}$$

where all expectations are over the calibration dataset $\mathcal{D}_{\mathrm{cal}}$, and the right-hand side is finite.

Proof. The proof follows the same argument as Corollary 5. Define $\bar{\tilde{\varepsilon}}^{(2)} \coloneqq \mathbb{P}(Y \notin \mathcal{C}_\alpha(\tilde{X}^{(2)}); \mathcal{D}_{\text{cal}})$ and $\bar{\varepsilon}^{\Delta}_{\text{DPI}} \coloneqq \bar{\tilde{\varepsilon}}^{(2)} - \bar{\varepsilon}^{(2)}$. By Lemma 9 with $\bar{\varepsilon}^{\Delta}_{\text{DPI}}$,

$$\mathbb{E}_{X^{(2)}}\left[\frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] = \bar{\varepsilon}^{\Delta}_{\text{DPI}}\, M^{(2)} + R^{(2)}_{\text{DPI}}, \qquad |R^{(2)}_{\text{DPI}}| \le \tilde{\kappa}^{(2)}\, v^{(2)}. \tag{133}$$

Since $\bar{\tilde{\varepsilon}}^{(2)}$ and $\bar{\varepsilon}^{(2)}$ are both training-conditional miscoverages at coverage level $1-\alpha$, Lemma 4 gives that they share the same marginal distribution. Following the argument for Equation 119 with $\bar{\varepsilon}^{\Delta}_{\text{DPI}}$ in place of $\bar{\varepsilon}^{\Delta}$,

$$\left|\mathbb{E}[\bar{\varepsilon}^{\Delta}_{\text{DPI}}\, M^{(2)}]\right| \le \frac{m^{(2)}}{\sqrt{n}}. \tag{134}$$

Taking expectations over $\mathcal{D}_{\text{cal}}$ in Theorem 5 and substituting Equations 133 and 134 completes the proof. All terms on the right-hand side are finite using the same argument as in Corollary 5. □

A.3.8 Corollary 8: APS score DPI with high probability

Corollary 8 (APS score DPI with high probability)**.** *Fix a nominal coverage level $1-\alpha$ and $\delta \in (0,1)$. Assume that the calibration scores are almost surely distinct and that the calibration data and future test samples are i.i.d. Suppose that there exists a constant $\eta_{\delta/2} < 1$ such that*

$$\mathbb{P}\big(\hat{\tau}^{(2)}_\alpha \le \eta_{\delta/2}\big) \ge 1 - \tfrac{\delta}{2}. \tag{135}$$

Let $c_{\delta/2} = \frac{K}{1-\eta_{\delta/2}}$. Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\text{cal}}$,

$$\tilde{\Delta}_\alpha \;\le\; \Delta_\alpha + D^{(2)} + \sqrt{\frac{2c^2_{\delta/2}\log(8/\delta)}{n}} + \tilde{\kappa}^{(2)}\, v^{(2)}, \tag{136}$$

where $D^{(2)}$ is defined as in Theorem 4 and $\tilde{\kappa}^{(2)}$ and $v^{(2)}$ are defined as in Corollary 7.

Proof. The proof follows the same argument as Corollary 6. Define $\bar{\tilde{\varepsilon}}^{(2)} \coloneqq \mathbb{P}(Y \notin \mathcal{C}_\alpha(\tilde{X}^{(2)}); \mathcal{D}_{\text{cal}})$ and $\bar{\varepsilon}^{\Delta}_{\text{DPI}} \coloneqq \bar{\tilde{\varepsilon}}^{(2)} - \bar{\varepsilon}^{(2)}$ as in Corollary 7. By Lemma 9, if $\hat{\tau}^{(2)}_\alpha < 1$, then

$$\mathbb{E}_{X^{(2)}}\left[\frac{\tilde{\varepsilon}^{(2)} - \varepsilon^{(2)}}{\hat{\Lambda}^{(2)}_\alpha}\right] = \bar{\varepsilon}^{\Delta}_{\text{DPI}}\, M^{(2)} + R^{(2)}_{\text{DPI}}, \qquad |R^{(2)}_{\text{DPI}}| \le \tilde{\kappa}^{(2)}\, v^{(2)}. \tag{137}$$

Since $\bar{\tilde{\varepsilon}}^{(2)}$ and $\bar{\varepsilon}^{(2)}$ are both training-conditional miscoverages at coverage level $1-\alpha$, Lemma 4 gives that they share the same marginal distribution. Following the proof of Lemma 6, with probability at least $1-\frac{\delta}{2}$,

$$|\bar{\varepsilon}^{\Delta}_{\text{DPI}}| \le \sqrt{\frac{2\log(8/\delta)}{n}}. \tag{138}$$

By assumption, with probability at least $1-\frac{\delta}{2}$ we have $\hat{\tau}^{(2)}_\alpha \le \eta_{\delta/2}$, so Lemma 7 gives $M^{(2)} \le c_{\delta/2}$. By a union bound, with probability at least $1-\delta$,

$$|\bar{\varepsilon}^{\Delta}_{\text{DPI}}\, M^{(2)}| \le c_{\delta/2}\sqrt{\frac{2\log(8/\delta)}{n}} = \sqrt{\frac{2c^2_{\delta/2}\log(8/\delta)}{n}}. \tag{139}$$

Since Theorem 5 holds for every fixed calibration dataset with $\hat{\tau}^{(2)}_\alpha < 1$, substituting Equation 139 together with $R^{(2)}_{\text{DPI}} \le \tilde{\kappa}^{(2)}\, v^{(2)}$ completes the proof. □

B DATASETS

We provide additional details on the synthetic and real datasets used in our experiments. The dataset statistics, such as sample sizes and feature dimensions, are summarized in Table 2.

B.1 DATASET DETAILS

We provide a brief overview of the datasets. Throughout, we will use $\oplus$ as a concatenation operator.

S1. The first synthetic dataset samples the features $Z \oplus X \sim \mathcal{N}(0, I_d)$, where $d = 4$. The first two coordinates are the base features Z and the remaining two coordinates are the additional features X. The label Y is sampled as

$$Y \mid X, Z \sim \mathrm{Categorical}\big(\mathrm{softmax}(W_2\,\mathrm{ReLU}(W_1(Z \oplus X) + b_1) + b_2)\big), \tag{140}$$

with $K = 10$ classes and hidden width $h = 16$.

The weights $W_1 \in \mathbb{R}^{h\times d}$, $W_2 \in \mathbb{R}^{K\times h}$ and biases $b_1 \in \mathbb{R}^h$, $b_2 \in \mathbb{R}^K$ are drawn once, with i.i.d. entries $W_1 \sim \mathcal{N}(0, 1/d)$, $W_2 \sim \mathcal{N}(0, 1/h)$, and $b_1, b_2 \sim \mathcal{N}(0, 0.01)$. Columns of W_1 corresponding to X are scaled by a factor of 4. We compute the marginal oracle models $p(\cdot \mid Z)$ by Gauss–Hermite quadrature, while $p(\cdot \mid X, Z)$ is available in closed form.

S2. The second synthetic dataset samples a label $Y \sim \mathrm{Unif}\{0, \ldots, K-1\}$ with $K = 50$ classes and features $Z \oplus X \mid Y = k \sim \mathcal{N}(\mu_k, I_d)$, where $d = 20$. The first 10 coordinates are the base features Z and the remaining 10 coordinates are the additional features X.

Each class mean $\mu_k = \mu_k^Z \oplus \mu_k^X$ is drawn once by sampling a vector of i.i.d. standard normal entries in each block, then rescaling it to a fixed norm where $\|\mu_k^Z\| \approx \sqrt{2}$ and $\|\mu_k^X\| \approx 2$. Since Y is uniform and the features are Gaussian, the oracle posteriors $p(\cdot \mid Z)$ and $p(\cdot \mid X, Z)$ are available in closed form by Bayes' rule.

S3. The third synthetic dataset is designed for the feature selection experiment. We use $K = 40$ classes separated into a small set of 8 head classes containing 70% of the prior mass and a large set of 32 tail classes containing the remaining 30% of the prior mass. The base feature is

$$Z \mid Y = k \sim \mathcal{N}(c_k + \delta_k,\, I_{d_Z}), \tag{141}$$

with $d_Z = 8$, where $c_k = c_{\mathrm{head}}$ for head classes and $c_k = c_{\mathrm{tail}}$ for tail classes. Each center c_k has norm 2 and δ_k is a smaller per-class offset of norm 0.5.

A pool of $m = 50$ scalar candidate features $X_1, \ldots, X_m$ is observed independently given Y,

$$X_j \mid Y = k \sim \mathcal{N}(a_j\,\mathrm{dir}_j[k],\, 1), \tag{142}$$

where $a_j \in \mathbb{R}$ is a per-candidate amplitude and $\mathrm{dir}_j \in \mathbb{R}^K$ is a per-candidate direction vector over the classes. Candidates are split across three roles, which set a_j and dir_j as follows:

- 18 *head* detectors: the 8 head classes are randomly partitioned into two groups of 4. The direction dir_j is zero for all non-head classes and takes values $\pm 1/\sqrt{8}$ on the two head groups, with opposite signs for each group, so that dir_j has unit norm. The amplitude is $a_j = 2$;
- 18 *tail* detectors: the same construction, splitting the 32 tail classes into two groups of 16;
- 14 *filler* features: either dir_j is a random unit vector with $a_j = 0.4$ or $\mathrm{dir}_j = 0$.

The tail detector amplitude is calibrated by bisection so that the average Shannon information $I(X_j; Y \mid Z)$ across tail detectors matches that across head detectors. Since the base and candidate features are conditionally independent given Y, the posterior $p(y \mid Z, X_S)$ is available in closed form via Bayes' rule for any subset $S \subseteq \{1, \ldots, m\}$.

Note that S3 is relatively complex because it is designed to exhibit clear cases where conformal set-size reduction and Shannon mutual information disagree in feature selection. We observe similar behavior on several real-world datasets, albeit to a lesser extent.

An adapted version of S3 is used for Figure 2, with $K = 24$ classes split into head, tail, and filler groups and $X_1, X_2 \in \mathbb{R}^8$. The class-conditional mean of X_1 is a distinct random vector for each of the 8 tail classes and zero otherwise, and X_2 is defined analogously for the 4 head classes.

Table 2: Sample sizes, feature dimensions, and number of classes for all datasets. For the synthetic datasets S1–S3, the conditional distributions are known and no training data are required. Here, n_{eval} denotes the number of samples available for calibration and evaluation. For tabular datasets, the input features are partitioned into X and Z. For image datasets, X and Z correspond to separate image representations with the indicated channel dimensions. The number of candidate features in the feature selection experiments is denoted by m.

Dataset	n_{total}	n_{eval}	d_{total}	d_X	d_Z	m	K
S1	200,000	200,000	4	2	2	—	10
S2	200,000	200,000	20	10	10	10	50
S3	200,000	200,000	58	50	8	50	40
Wine	4,898	1,224	11	6	5	11	7
HAR	10,299	2,575	561	348	213	51	6
Letter	20,000	5,000	16	9	7	16	26
Covertype	581,012	145,253	54	10	44	54	7
FMNIST	70,000	10,000	$2 \times 28 \times 28$	$1 \times 28 \times 28$	$1 \times 28 \times 28$	64	10
CIFAR-10	60,000	10,000	$6 \times 32 \times 32$	$3 \times 32 \times 32$	$3 \times 32 \times 32$	64	10
CIFAR-100	60,000	10,000	$6 \times 32 \times 32$	$3 \times 32 \times 32$	$3 \times 32 \times 32$	64	100
ImageNet-1k	1,331,167	50,000	$6 \times 64 \times 64$	$3 \times 64 \times 64$	$3 \times 64 \times 64$	64	1000

Wine Quality (Wine). The Wine Quality dataset (Cortez et al., 2009) contains physicochemical measurements and quality ratings for red and white wines. We use the white wine subset, which contains 4,898 samples with 11 physicochemical features. The prediction target is the wine quality rating ranging from 3 to 9.

Human Activity Recognition (HAR). The Human Activity Recognition dataset (Reyes-Ortiz et al., 2013) contains smartphone sensor measurements collected from participants performing six activities. We use the preprocessed feature representation, consisting of 10,299 samples with 561 features. The prediction target is one of six activities: walking, walking upstairs, walking downstairs, sitting, standing, or laying.

Letter Recognition (Letter). The Letter Recognition dataset (Slate, 1991) contains 20,000 samples of capital English letters, each represented by 16 numerical features extracted from the corresponding character image. The prediction target is one of the 26 letters of the English alphabet.

Covertype. The Covertype dataset (Blackard, 1998) contains 581,012 samples describing forested areas using 54 cartographic features, comprising 10 quantitative features and 44 binary features encoding wilderness area and soil type information. The prediction target is one of seven forest cover types.

Fashion-MNIST (FMNIST). Fashion-MNIST (Xiao et al., 2017) consists of 70,000 grayscale 28×28 images of fashion items from 10 classes. The classes are T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The dataset is split into 60,000 training samples and 10,000 test samples.

CIFAR-10. CIFAR-10 (Krizhevsky, 2009) consists of 60,000 RGB 32×32 images from 10 object classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The dataset contains 50,000 training samples and 10,000 test samples.

CIFAR-100. CIFAR-100 (Krizhevsky, 2009) consists of 60,000 RGB 32×32 images from 100 object classes. The dataset contains 50,000 training samples and 10,000 test samples.

ImageNet-1k. ImageNet-1k (Deng et al., 2009) is a large-scale image classification dataset containing images from 1,000 object classes. We use the 64×64 downsampled version of the dataset (Chrabaszcz et al., 2017) containing 1,281,167 training samples and 50,000 test samples.

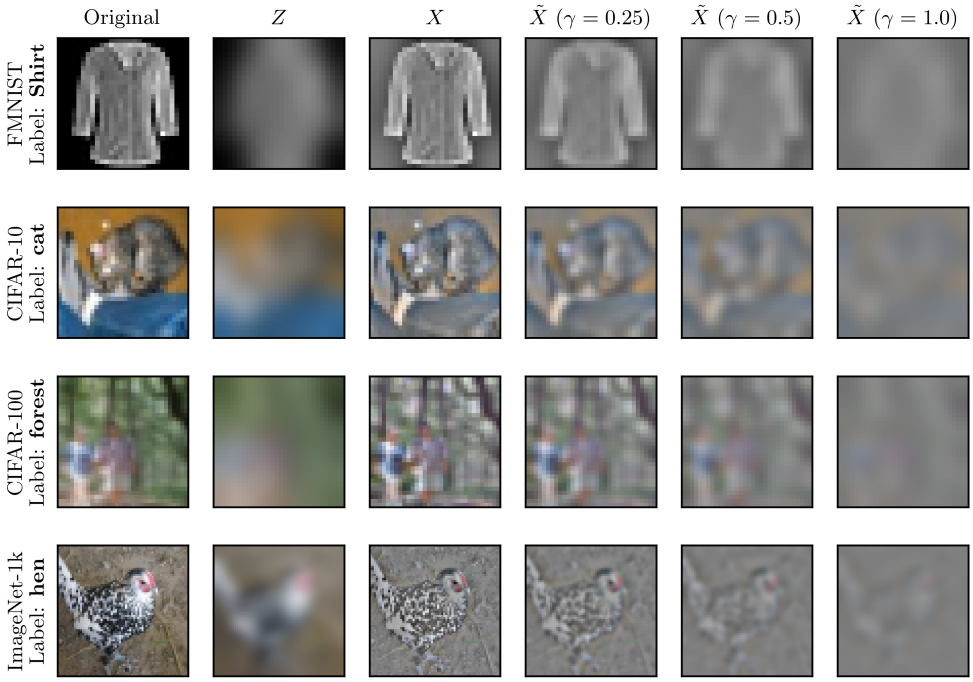

Figure 4: Examples of the original image, base, new, and degraded features for each image dataset. The degraded features apply an additional Gaussian blur with standard deviation $\gamma\sigma$ to X.

B.2 Feature construction details

We provide additional details on the construction of the base features Z and additional features X for each dataset. For the synthetic datasets, Z consists of the first d_Z feature coordinates and X consists of the remaining d_X coordinates.

For the tabular datasets, we partition the input features into X and Z based on natural groupings of the feature types, as detailed below.

Wine Quality (Wine). The base features Z consist of five physical properties: chlorides, density, residual sugar, alcohol content, and pH. The additional features X consist of the remaining six acidity- and sulfur-related measurements.

Human Activity Recognition (HAR). The base features Z consist of the 213 gyroscope-derived measurements, including time- and frequency-domain statistics computed from the gyroscope signals. The additional features X consist of the remaining 348 features, including accelerometer-derived measurements and the three angle-between-vectors features.

Letter Recognition (Letter). The base features Z consist of seven statistical moments of the x- and y-pixel positions. The additional features X consist of the remaining nine features: five bounding-box features describing position, width, height, and pixel count, and four edge-count and edge-correlation features.

Covertype. The base features Z consist of the 44 binary features encoding wilderness area and soil type. The additional features X consist of the 10 quantitative cartographic variables, including elevation, slope, and distances to hydrology.

For all image datasets, we construct Z by applying a Gaussian blur with standard deviation $\sigma = 3$ to the original image, with the Gaussian kernel truncated at four standard deviations on each side. The additional features X are given by the residual between the original and blurred images. An example of the features for each image dataset is shown in Figure 4.

B.3 DEGRADED FEATURE CONSTRUCTION DETAILS

We provide additional details on the construction of the degraded feature $\tilde{X}$ used in the data processing inequality experiments (Section 3.3). We apply degradation only to the additional features X, leaving the base features Z unchanged. We consider three degradation levels, *low*, *medium*, and *high*, corresponding to degradation multipliers $\gamma \in \{0.25, 0.5, 1.0\}$, respectively. We report results for all three degradation levels in Appendix E. In the main text, we report results only for the medium degradation level, corresponding to $\gamma = 0.5$.

For the synthetic and tabular datasets, we degrade X by adding independent Gaussian noise to each coordinate. For a degradation level γ, the noise standard deviation for each coordinate is γ times its empirical standard deviation, estimated using the training data. The degradation noise is sampled once for each dataset and held fixed across all noise levels and subsequent experimental replicates in the analysis.

For the image datasets, we degrade X by applying an additional Gaussian blur to the residual features, while leaving the already-blurred base features Z unchanged. For a degradation level γ, we use a blur standard deviation of $\gamma\sigma$, where $\sigma = 3$ is the blur scale used to construct Z. An example of the degraded features for each image dataset is shown in Figure 4.

The synthetic datasets (S1 and S2) have known oracle distributions, so we compute the degraded posterior in closed form or using Gauss–Hermite quadrature. For the real-world datasets, we train a separate model on $(\tilde{X}, Z)$ for each degradation level.

B.4 CANDIDATE FEATURE DETAILS

We describe the candidate features $\{X_1, \ldots, X_m\}$ used in the feature selection experiments. In the synthetic settings, the base features Z are always observed, and each coordinate of the additional features X forms a separate candidate feature. This yields 10 and 50 candidate features for S2 and S3, respectively, for which we can compute oracle conditional probabilities given any subset of features. We exclude S1 from the feature selection experiments due to the low dimensionality of X.

For the tabular datasets, no features are observed initially and each input feature (including those previously in Z) is treated as a separate candidate. The only exception is HAR, which contains 561 input features. To keep the number of candidates manageable, we randomly partition the features into 51 groups of 11, with each group forming a candidate feature.

The image candidates are constructed from the final convolutional representation of the frozen backbone trained on (X, Z). As with the tabular datasets, no features are observed initially. Since the purpose of the feature selection experiment is to empirically probe the differences between set-size reduction and Shannon information, this representation serves to provide a collection of meaningful candidates, rather than physically acquirable features.

On FMNIST, CIFAR-10, and CIFAR-100, we partition the channels of the final convolutional feature map into 64 contiguous groups. Each group contains 2 channels of size 3×3 on FMNIST and 4 channels of size 4×4 on CIFAR-10 and CIFAR-100. The ResNet backbone on ImageNet-1k instead produces 512 globally average-pooled channels, which we partition into 64 groups of 8.

The number of candidate features m for each dataset is summarized in Table 2.

Table 3: Selected hyperparameters and training configurations for the classification models. The learning rate, weight decay, and dropout rate are selected via grid search. The maximum number of epochs and patience correspond to the full training procedure after hyperparameter selection.

Dataset	Params.	Max. epochs	Patience	Learning rate	Weight decay	Dropout
Wine	600	1,000	100	4×10^{-4}	1×10^{-4}	0.3
HAR	178K	1,000	100	4×10^{-4}	1×10^{-4}	0.1
Letter	106K	1,000	100	1×10^{-3}	1×10^{-4}	0.2
Covertype	424K	500	100	4×10^{-4}	1×10^{-5}	0.1
FMNIST	105K	1,000	100	1×10^{-3}	4×10^{-5}	0.3
CIFAR-10	414K	500	100	1×10^{-3}	1×10^{-4}	0.1
CIFAR-100	783K	500	100	1×10^{-3}	1×10^{-4}	0.3
ImageNet-1k	11.7M	25	5	4×10^{-4}	4×10^{-5}	0.1

C TRAINING DETAILS

We provide details on the model architectures, training procedures, hyperparameter selection, and final pre-calibration accuracy of our classification models. The synthetic datasets use oracle probabilities and therefore do not require model training.

C.1 MODEL ARCHITECTURES

For the tabular datasets, we use multi-layer perceptrons (MLPs) with dropout after each hidden layer. The hidden-layer widths are $(16, 16)$ for Wine, $(256, 128)$ for HAR, $(256, 256, 128)$ for Letter, and $(512, 512, 256)$ for Covertype. All non-binary features are independently standardized to zero mean and unit variance using statistics computed from the training data.

For FMNIST, CIFAR-10, and CIFAR-100, we use three-layer CNNs with 3×3 convolutional kernels, batch normalization, and ReLU activations. The convolutional layers have channel widths $(32, 64, 128)$ for FMNIST and $(64, 128, 256)$ for CIFAR-10 and CIFAR-100. Each convolutional layer is followed by 2×2 max pooling. After the final layer, we flatten the spatial representation, apply dropout, and use a linear classification head to produce the predicted probabilities.

For ImageNet-1k, we use the ResNet-18 architecture. We directly fine-tune the pretrained network on the base features Z using standard ImageNet normalization and modify the input layer to accommodate the additional channels when using both X and Z. For all ResNet models, the final convolutional representation is passed through global adaptive pooling and dropout, followed by a linear classification head.

C.2 TRAINING OPTIMIZATION

We train all models on the real datasets using the Adam optimizer with weight decay to minimize the standard cross-entropy loss. All newly initialized model parameters use the default PyTorch initialization. For the image datasets, we additionally apply data augmentation consisting of horizontal flips with probability 0.5 and random crops after zero-padding the images by 2 pixels for FMNIST and 4 pixels for CIFAR-10, CIFAR-100, and ImageNet-1k. We also use cosine annealing to decay the learning rate from its initial value to zero over the maximum number of training epochs.

Training uses a batch size of 256 and early stopping based on validation loss, with the checkpoint attaining the lowest validation loss restored at the end of training. Separate models are trained for the base, degraded, and augmented feature sets. We set the maximum number of training epochs to 1,000 for Wine, HAR, Letter, and FMNIST; 500 for CIFAR-10, CIFAR-100, and Covertype; and 25 for ImageNet-1k. We use an early-stopping patience of 100 epochs for all datasets except ImageNet-1k, for which we use a patience of 5 epochs.

Table 4: Top-1 classification accuracy (%) on the held-out evaluation pool for the augmented, degraded, and base feature sets, and for the mask-conditioned model with all candidate features observed. For the synthetic datasets S1 and S2, we report the accuracy of the oracle model.

Dataset	(X, Z)	$(\tilde{X}, Z)$ low noise	$(\tilde{X}, Z)$ medium noise	$(\tilde{X}, Z)$ high noise	Z	Mask-conditioned (all)
S1	56.52	54.86	51.83	46.69	39.33	—
S2	50.83	48.84	44.00	34.21	18.43	—
Wine	55.15	54.98	53.43	52.78	50.33	55.80
HAR	98.29	97.40	96.54	93.86	88.08	98.72
Letter	97.68	95.52	92.00	86.76	86.52	97.36
Covertype	96.24	83.11	74.93	68.85	65.16	94.12
FMNIST	93.06	92.77	91.43	89.92	89.85	93.11
CIFAR-10	86.40	84.10	77.55	70.31	68.80	86.30
CIFAR-100	60.97	56.77	50.14	42.94	40.36	56.79
ImageNet-1k	58.79	57.84	54.57	47.06	43.70	57.99

C.3 Mask-conditioned model training

The feature selection experiments require predicted conditional probabilities given arbitrary subsets of the candidate features. We therefore modify the model architectures to accommodate arbitrary conditioning masks. For the tabular datasets, we modify the MLP to take inputs of the form $[x \odot M, M]$, where x is the full feature vector and M is a binary mask indicating which features are observed, following the masked predictor construction in Yoon et al. (2019).

For the image datasets, we use the pretrained CNN or ResNet model previously trained on the full set of features (X, Z). We freeze the backbone through the final convolutional layer, fixing the representation from which the candidate features are constructed. For FMNIST, CIFAR-10, and CIFAR-100, we train a masked MLP on this representation using the same masking construction as for the tabular datasets. We instead train a linear classification head on the masked representation on ImageNet-1k for scalability.

During training, we independently sample a conditioning mask for each example. Specifically, we first sample $u \sim \mathrm{Unif}[0, 1]$ and then independently mask each candidate feature with probability u. To ensure sufficient exposure to the two extreme cases, we additionally mask all candidate features with probability 0.05 and observe all candidate features with probability 0.15.

We double both the maximum number of training epochs and the patience relative to the full-feature training procedure described in Appendix C.2. For early stopping, we average the validation loss across masks generated with fixed masking probabilities $u \in \{0, 0.25, 0.5, 0.75, 1\}$.

C.4 Hyperparameter selection

We select the learning rate, weight decay, and dropout rate via grid search. We consider learning rates in $(10^{-3}, 4 \times 10^{-4}, 10^{-4})$, weight decay values in $(10^{-4}, 4 \times 10^{-5}, 10^{-5})$, and dropout rates in $(0.1, 0.2, 0.3)$, yielding 27 configurations in total. Hyperparameters are selected using the augmented model with both X and Z. The selected configuration is then shared across the augmented model, the base model using only Z, and the models using degraded features $\tilde{X}$ and Z.

For each configuration, we train the augmented model using 20% of both the maximum epoch budget and early-stopping patience specified above. The image models use the same cosine annealing schedule during hyperparameter selection. For ImageNet-1k, each configuration is trained for the reduced epoch budget without early stopping.

Among the 27 configurations, we select the one achieving the highest validation accuracy after early stopping. We then train all models using this configuration with their respective epoch budget and early-stopping patience. The selected hyperparameters for each dataset are summarized in Table 3.

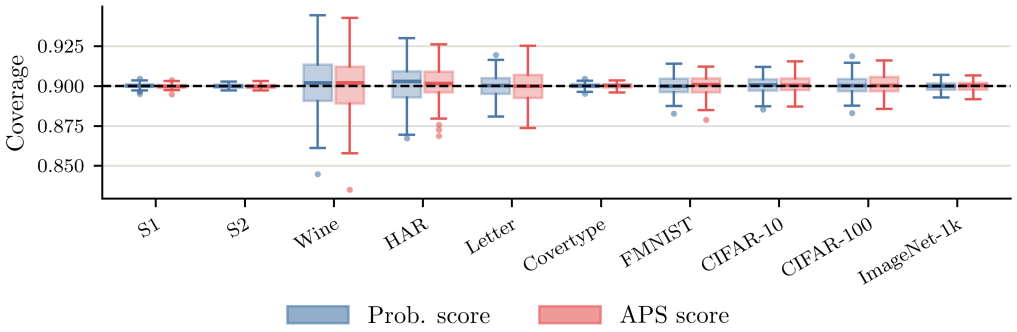

Figure 5: Empirical marginal coverage of the conformal predictors using (X, Z) over 100 random equal calibration and test splits of the evaluation pool. The dashed line indicates the nominal coverage level $1 - \alpha = 0.9$.

C.5 Accuracy and coverage metrics

We report the top-1 classification accuracy of all models on the held-out evaluation pool. We evaluate models using the augmented features (X, Z), the degraded features $(\tilde{X}, Z)$ at each noise level, the base features Z, as well as the mask-conditioned model with all candidate features observed. The results are summarized in Table 4.

As expected, classification accuracy generally decreases as features are degraded or removed. The mask-conditioned models also obtain comparable performance to models trained on the full feature set (X, Z), despite supporting predictions given arbitrary subsets of features. The models using the augmented features do not achieve state-of-the-art performance on these classification tasks, since we use simplified architectures and training procedures. Our primary goal is not to maximize predictive accuracy, but to evaluate our theoretical results under realistic levels of model imperfection.

We additionally evaluate the marginal coverage of the conformal predictors using (X, Z) at the nominal coverage level $1 - \alpha = 0.9$, which is used throughout most of our experimental results. For each dataset, we randomly split the evaluation pool in half into calibration and test sets over 100 replicates. As shown in Figure 5, both the probability and APS scores achieve empirical coverage close to the nominal level on all datasets, with larger variability on the datasets with smaller evaluation pools, such as Wine and HAR.

D ANALYSIS DETAILS

We provide additional details on the estimation of the bounds, the evaluation protocols, and the analysis metrics used in our validation and feature selection experiments.

D.1 ESTIMATION OF THE BOUNDS

To evaluate the conformal mutual information terms in Theorem 2, we estimate $I_\lambda(X;Y \mid Z)$ with the plug-in (or oracle) probability distributions on the test set at the calibrated $\hat{\lambda}_\alpha^{(i)}$. For Theorem 4, we compute $I^{(1)}$ in the same way, since $\hat{\Lambda}_\alpha^{(1)}$ depends only on Z. However, the term $I^{(2)}$ in the lower bound cannot be computed using a plug-in estimate. Recall that

$$I^{(2)} = H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid Z) - H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid X, Z). \tag{143}$$

The second term can be evaluated separately for each test point, but the first term is problematic because $\hat{\Lambda}_\alpha^{(2)}$ depends on both X and Z. In particular,

$$H_{\hat{\Lambda}_\alpha^{(2)}}(Y \mid Z) = \mathbb{E}_Z\Big[\sum\nolimits_y \min\Big\{1,\, \mathbb{E}\Big[\tfrac{p(y|X,Z)}{\hat{\Lambda}_\alpha^{(2)}} \;\Big|\; Z\Big]\Big\}\Big]. \tag{144}$$

Since the conditional expectation over $X \mid Z$ lies inside the minimum, computing it requires multiple draws of X for each Z. In the synthetic settings, $p(x \mid z)$ is known, so we estimate it by Monte Carlo sampling with $M = 256$ draws of $X \mid Z$ for each test point. For the real-world datasets, we instead average $q(y \mid x_j, z_j)/\hat{\Lambda}_\alpha^{(2)}(x_j, z_j)$ over the $k = 10$ nearest other neighbors j of each test point in the evaluation pool. Neighbors are computed using Euclidean distance on the base features Z for tabular data and cosine distance on the learned embedding of Z for image data.

In the synthetic settings, we use the oracle conditional distributions, so $D^{(i)} = 0$ exactly. For the real-world datasets, estimating $D^{(i)}$ requires the true conditional distribution, so we set $D^{(i)} = 0$, which is consistent with the plug-in estimates of $I^{(i)}$. We emphasize that the resulting bounds for the real-world datasets are therefore approximate. These approximations do not undermine our validation, since the synthetic settings evaluate the bounds using the true conditional distributions, while the real-world datasets complement these results under realistic model imperfections.

The terms involving miscoverage rates are estimated by the test-set average of

$$\frac{\mathbf{1}[Y \notin \mathcal{C}_\alpha(X^{(2)})] - \mathbf{1}[Y \notin \mathcal{C}_\alpha(X^{(1)})]}{\hat{\lambda}_\alpha^{(i)}}. \tag{145}$$

For the APS bounds, we analogously replace $\hat{\lambda}_\alpha^{(i)}$ with $\hat{\Lambda}_\alpha^{(i)}$. Since the random threshold is a function of the input, this estimator is unbiased for the corresponding calibration terms in the bounds, conditional on the calibration set.

D.2 EVALUATION PROTOCOLS

As described in Section 5, we reserve a dedicated pool of n_{eval} samples from each synthetic and real-world dataset for calibration and testing. We next describe how this evaluation pool is used in each experiment.

D.2.1 VALIDATION EXPERIMENTS

We first describe the evaluation protocol for the validation experiments. For each experimental replicate, we sample a calibration dataset of size n_{cal} and a test dataset of size n_{test} without replacement from the evaluation pool. The calibration and test datasets are disjoint but do not necessarily partition the entire pool. For a given random seed, the sampled datasets are nested across sample sizes, while different seeds yield independent draws.

For each metric, we consider three regimes: (i) jointly varying $n_{\text{cal}} = n_{\text{test}}$, (ii) varying n_{cal} with n_{test} fixed, and (iii) varying n_{test} with n_{cal} fixed. For the normalized margin and violation rate, we use a logarithmically spaced grid of five to seven points ranging from 100 to $\min(10{,}000, n_{\text{eval}}/2)$, inclusive. Because the bound width becomes visually uninformative at larger sample sizes, we instead use a logarithmically spaced grid from 10 to $\min(1{,}000, n_{\text{eval}}/2)$. In regimes (ii) and (iii), the fixed sample size is set to the maximum of this grid.

Calibration follows the standard split conformal prediction procedure using either the probability score or APS score. For APS, we use the same score randomization draw $U \sim \mathrm{Unif}[0,1]$ between the procedures using $X^{(1)}$ and $X^{(2)}$. Both procedures are calibrated on the same calibration set and evaluated on the same test set. We estimate Δ_α by the difference in average prediction set size on the test set under $X^{(1)}$ and $X^{(2)}$.

We additionally vary the nominal coverage level over $1-\alpha \in \{0.8, 0.85, 0.9, 0.95, 0.99\}$. For these experiments, n_{cal} and n_{test} are held fixed at the respective maximum sample sizes used for each metric, i.e., $\min(10{,}000, n_{\mathrm{eval}}/2)$ for the normalized margin and violation rate and $\min(1{,}000, n_{\mathrm{eval}}/2)$ for the bound width.

For the DPI validation experiments, we follow a similar calibration and evaluation procedure by sampling from the same evaluation pool. We fix the test set size at $n_{\mathrm{test}} = \min(10{,}000, n_{\mathrm{eval}}/2)$ and vary $n_{\mathrm{cal}} \in \{10, 100, 1{,}000\}$, omitting values of n_{cal} that exceed $n_{\mathrm{eval}}/2$. Within each replicate, the procedures using the augmented and degraded features share the same calibration and test sets. We intentionally consider small calibration sample sizes to illustrate how the normalized DPI gap concentrates as n_{cal} increases and to demonstrate that finite-sample calibration variability can result in a negative gap at sufficiently small n_{cal}.

D.2.2 Feature selection experiments

We next describe the evaluation protocol for the feature selection experiments. For each of 10 random seeds, we randomly partition the evaluation pool into three disjoint subsets, using 50% of the samples as a selection set, 25% as a calibration set, and the remaining 25% as a test set. The selection set is used to compute the criteria and determine the feature ordering, while the calibration and test sets are used only to evaluate the resulting sequence. In the synthetic settings, the base features Z are always observed, whereas no features are known initially for the real-world datasets.

Note that this setting contrasts with the active feature acquisition (AFA) literature, which typically selects features to acquire for each individual instance and trains dedicated acquisition policies to optimize predictive performance under a feature budget (Shim et al., 2018). We use a different evaluation protocol because our goal is not to propose a competitive AFA method. Instead, we use a simple protocol to empirically probe how the criteria from our theory rank features in practice.

Let S denote the set of candidate features, initialized as $S = \emptyset$. At each step, we evaluate a selection criterion for every remaining candidate X_j using the predicted probabilities given $X_{S\cup\{j\}}$, add the highest-scoring candidate to S, and repeat until all candidates are exhausted. This yields a single global feature ordering for each criterion and seed, which is shared across all test samples. For each prefix of the ordering, we then calibrate the conformal predictor on the calibration set at nominal coverage $1-\alpha = 0.9$ and evaluate set size, coverage, and accuracy on the test set.

We compare the following selection criteria, each computed using the predicted probabilities of the oracle or mask-conditioned model as a plug-in estimate.

- Δ_α (Prob.): the reduction in the average set size using probability score with $X_{S\cup\{j\}}$;
- Δ_α (APS): the reduction in the average set size using APS score with $X_{S\cup\{j\}}$;
- I_λ (Prob.): the conformal mutual information $I_\lambda(X_j; Y \mid X_S)$ evaluated at the geometric mean $\lambda = (\hat{\lambda}_\alpha^{(1)} \hat{\lambda}_\alpha^{(2)})^{1/2}$ of the calibrated thresholds using X_S and $X_{S\cup\{j\}}$;
- Shannon MI: the plug-in estimate $\hat{I}(X_j; Y \mid X_S) = \hat{H}(Y \mid X_S) - \hat{H}(Y \mid X_{S\cup\{j\}})$;
- Accuracy: the empirical top-1 accuracy gain with $X_{S\cup\{j\}}$;
- Random: a uniformly random ordering of the candidate features.

Since the conformal-based criteria require separate calibration data, we use 5-fold cross-fitting within the selection set. We calibrate the thresholds on four folds and evaluate the criterion on the held-out fold, then average the resulting values across folds. The Shannon mutual information and accuracy criteria do not require calibration and are computed directly on the full selection set.

We do not include an analogous I_Λ criterion for the APS score since it requires the same k-NN approximation of the conditional expectation over $X_j \mid X_S$ as described above for $I^{(2)}$. Repeating

this approximation for every candidate at every step is computationally expensive and introduces additional approximation error, so we omit this criterion.

D.3 METRICS

We next formalize the evaluation metrics in further detail.

D.3.1 VALIDATION EXPERIMENTS

In the validation experiments, we consider four main metrics: normalized margin, violation rate, normalized width, and normalized DPI gap. Except for the violation rate, we normalize all metrics by the number of classes K in order to compare across datasets. We next define each metric formally.

Let $[L, U]$ denote the estimated bounds from Theorem 2 for the probability score or Theorem 4 for APS. When computing the normalized margin, violation rate, and normalized width, we discard replicates with invalid bounds ($U < L$), which occur in roughly 3% of cases, and average over the remaining random splits of the data. This exclusion does not change our conclusions.

The *normalized margin* quantifies the magnitude by which the empirical Δ_α falls outside the theoretical bounds for a particular realization of the calibration and test data. Although the theoretical bounds hold exactly for the population quantities under the assumptions of our results, finite-sample Monte Carlo estimation and model approximation on the real-world datasets can lead to empirical violations. We define

$$\text{Normalized Margin} = \frac{d(\Delta_\alpha, [L, U])}{K} = \frac{\max\{L - \Delta_\alpha,\, \Delta_\alpha - U,\, 0\}}{K}, \tag{146}$$

where $d(x, \mathcal{I})$ denotes the distance from x to the interval $\mathcal{I}$, with $d(x, \mathcal{I}) = 0$ whenever $x \in \mathcal{I}$.

The *violation rate* examines how frequently the empirical Δ_α falls outside the theoretical bounds, providing a binary measure of empirical validity that does not account for the magnitude of a violation. We define

$$\text{Violation Rate} = \mathbb{P}\big(\Delta_\alpha \notin [L, U]\big), \tag{147}$$

where the probability is estimated over random replicates of the calibration and test data. In practice, we expand the lower and upper bounds by a small slack of 10^{-4} to account for floating-point error when computing this metric.

The *normalized width* measures the informativeness of the bounds through the size of the interval $[L, U]$. We define

$$\text{Normalized Width} = \frac{U - L}{K}. \tag{148}$$

Since we exclude replicates with $U < L$, the normalized width is non-negative for every replicate.

Finally, the *normalized DPI gap* compares the conformal set-size reduction obtained by augmenting the base features Z with either X or its degraded counterpart $\tilde{X}$. For each replicate, we use identical test sets to estimate the expected set-size reductions Δ_α and $\tilde{\Delta}_\alpha$ and define

$$\text{Normalized DPI Gap} = \frac{\Delta_\alpha - \tilde{\Delta}_\alpha}{K}. \tag{149}$$

We examine the distribution of this gap across random draws of the calibration and test data.

D.3.2 FEATURE SELECTION EXPERIMENTS

In the feature selection experiments, we evaluate each criterion by the quality of the conformal predictors obtained along its selection trajectory. Let S_j denote the first j candidates acquired under a given criterion. For each j, we compute the average prediction set size and empirical coverage on the test set using X_{S_j} for both the probability score and the APS score. We additionally report the top-1 and top-5 classification accuracy of the model given X_{S_j} on the test set.

Let $\mathcal{E}_j$ denote any one of these metrics evaluated using the first j features. To summarize performance across the entire selection trajectory, we report the average

$$\bar{\mathcal{E}} = \frac{1}{m-1} \sum_{j=1}^{m-1} \mathcal{E}_j, \tag{150}$$

where we exclude the cases $j = 0$ and $j = m$ since all criteria observe identical features at the endpoints. Intuitively, $\bar{\mathcal{E}}$ approximates the area under the curve of each metric normalized by m. This summarizes the performance across all steps rather than at a single fixed budget.

D.4 Implementation details

The validation experiments require evaluating a large number of calibration and test draws across sample sizes, random seeds, nominal coverage levels, feature sets, and conformity scores. To reduce the computational cost, we presort the predicted class probabilities for the evaluation pool and use binary search to efficiently compute the quantities required for each replicate. This optimization is an exact reformulation of the standard computation and does not affect any reported results.

For an evaluation pool of N samples and K classes, presorting requires $O(NK \log K)$ time. Each subsequent replicate requires $O(n_{\text{cal}} \log n_{\text{cal}} + n_{\text{test}} \log K)$ time. This provides a substantial computational speedup over directly recomputing the conformal prediction sets, particularly for datasets with large K such as ImageNet-1k.

The presorting approach does not apply to the feature selection experiments, since the evaluated subsets depend on the selection criterion and the number of possible subsets grows exponentially. Instead, we naively compute the predicted probabilities for each candidate subset at each step.

E ADDITIONAL EXPERIMENTS

In this section, we further investigate the empirical behavior of our sandwich bounds for set-size reduction using the probability and APS scores. We then provide additional results for the DPI under different noise levels and report feature selection results across the remaining datasets. Details of the evaluation procedures and metrics are described in Appendix D.

E.1 SANDWICH VALIDATION RESULTS

We provide additional results demonstrating the empirical behavior of Theorems 2 and 4. We first investigate the violation rate of the bounds as we jointly vary n_{cal} and n_{test}. We show these results in Figure 6, together with the normalized margin and width results from Figure 3 for comparison.

Consistent with the normalized margin, the violation rate is nonzero at smaller calibration and test set sizes but decreases toward zero as the sample size increases. The main outlier is the Wine dataset using the probability score, for which the violation rate remains high over the evaluated range. This behavior largely reflects the binary nature of the violation rate. The corresponding normalized margin behaves similarly to those of the other datasets, indicating that the violations themselves are small in magnitude.

Comparing the probability and APS scores, we find that the violation rate is typically lower for APS across datasets and sample sizes. The lower violation rate for APS is consistent with its wider bounds, which are easier to satisfy and less sensitive to finite-sample estimation error.

We next examine the normalized margin, violation rate, and normalized width as we independently and jointly vary n_{cal} and n_{test} in Figures 7–9. Additionally, we investigate the three metrics across different nominal coverage levels in Figure 10.

Normalized margin (Figure 7). The magnitude of the empirical violations is driven primarily by the size of the test set. As n_{test} increases with n_{cal} fixed, the normalized margin decreases toward zero. In contrast, with n_{test} fixed, the margin is largely insensitive to n_{cal}. This suggests that the observed violations primarily arise from finite test-set estimation error.

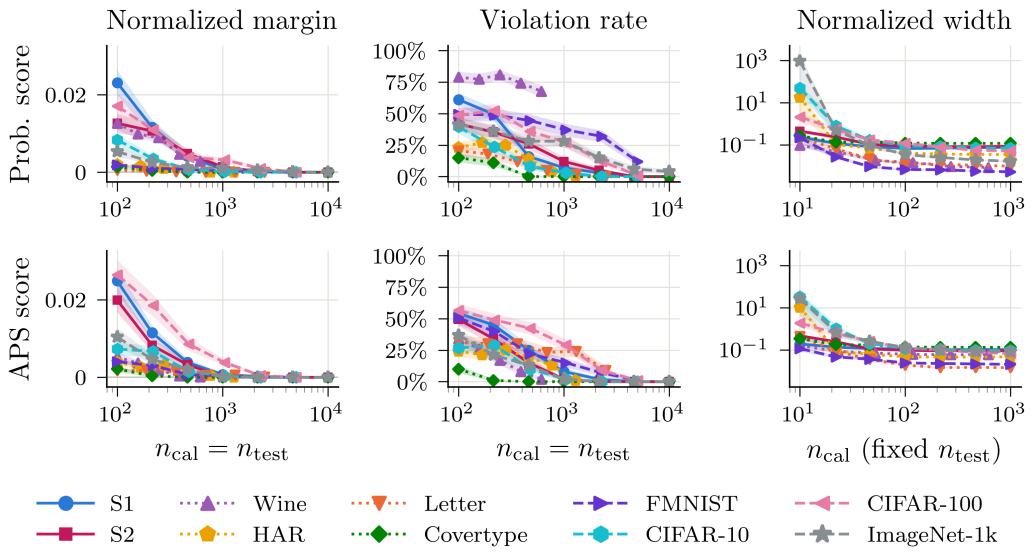

Figure 6: We show the normalized margin and violation rate as we jointly vary n_{cal} and n_{test} (left and middle) and the normalized width as we vary n_{cal} while keeping n_{test} fixed (right). All panels report the mean $\pm$ standard error based on 100 random draws of the calibration and test sets from the fixed evaluation pool.

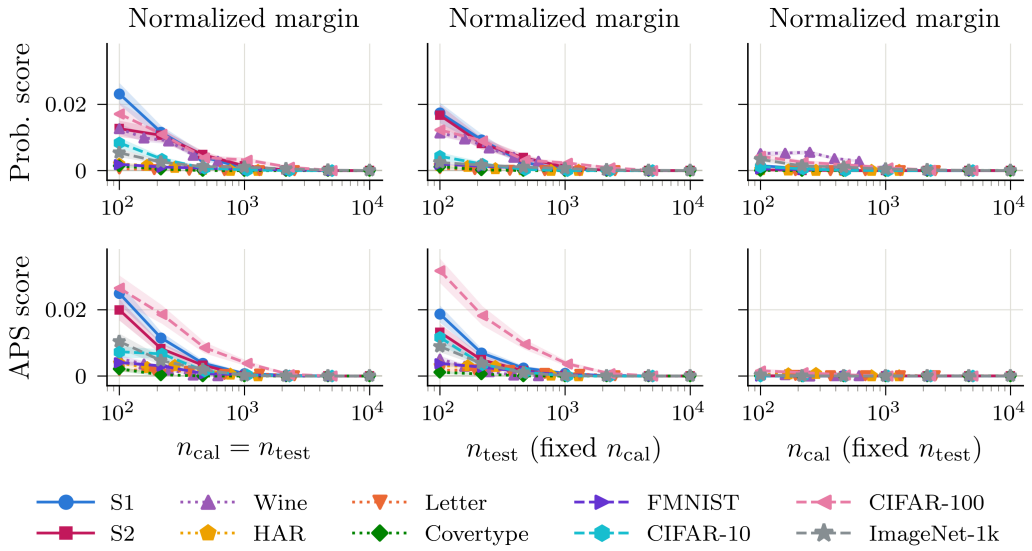

Figure 7: We show the normalized margin as we vary n_{cal} and n_{test} jointly (left) or independently (middle and right). All panels report the mean $\pm$ standard error based on 100 random draws of the calibration and test sets from the fixed evaluation pool.

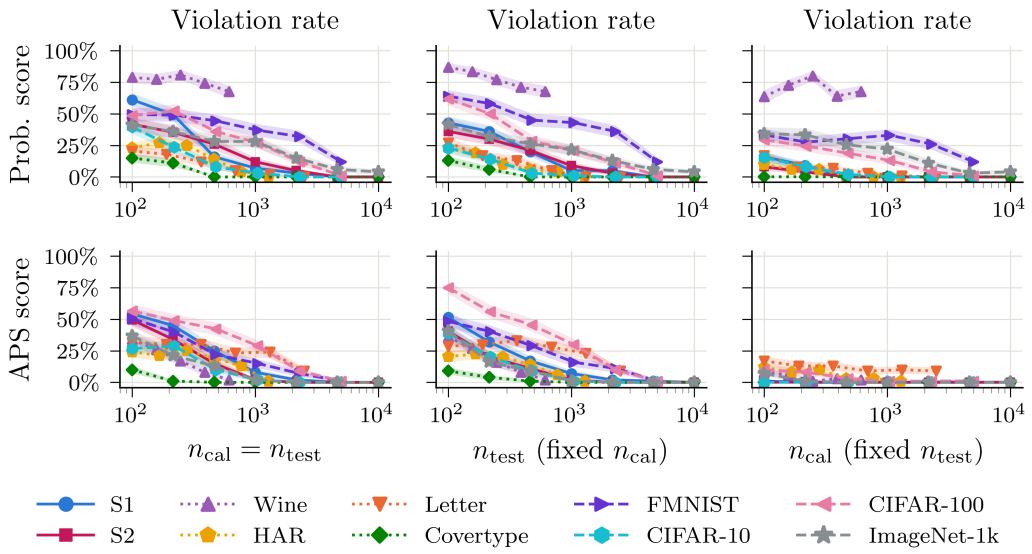

Figure 8: We show the violation rate as we vary n_{cal} and n_{test} jointly (left) or independently (middle and right). All panels report the mean $\pm$ standard error based on 100 random draws of the calibration and test sets from the fixed evaluation pool.

Violation rate (Figure 8). The violation rate exhibits a similar dependence on n_{test}. When n_{test} is fixed, the violation rate for APS is low and largely insensitive to n_{cal}, whereas for the probability score it decreases toward zero as n_{cal} increases. The Wine dataset is the sole exception, where the violation rate remains between roughly 60% and 80%.

Normalized width (Figure 9). The normalized width mostly depends on the size of the calibration set and decreases quickly with n_{cal}. This behavior agrees with our theoretical analysis of the finite-sample calibration terms. As the calibration size increases, the training-conditional miscoverage rates concentrate around the nominal coverage level, causing the finite-sample calibration terms to shrink and the bounds to tighten.

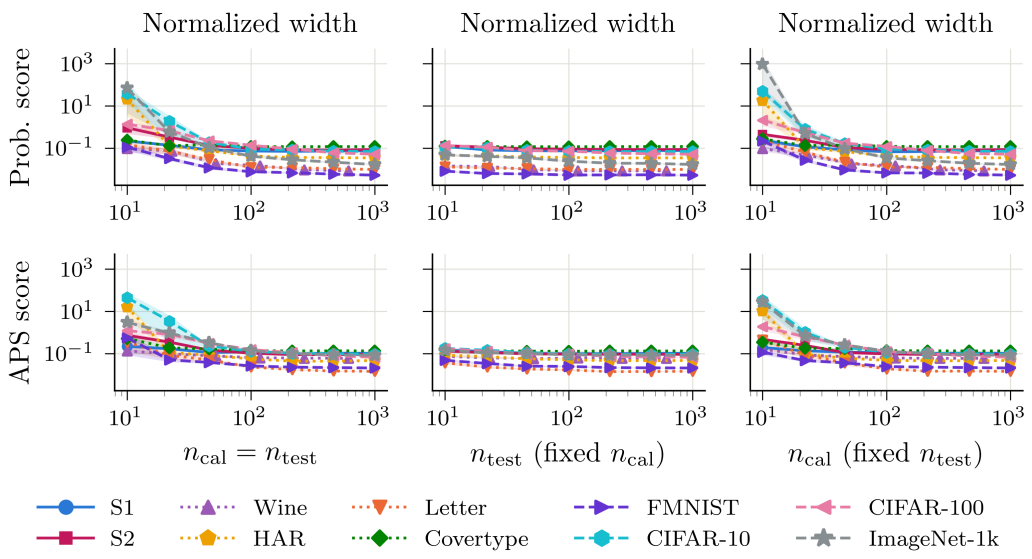

Figure 9: We show the normalized width as we vary n_{cal} and n_{test} jointly (left) or independently (middle and right). All panels report the mean $\pm$ standard error based on 100 random draws of the calibration and test sets from the fixed evaluation pool.

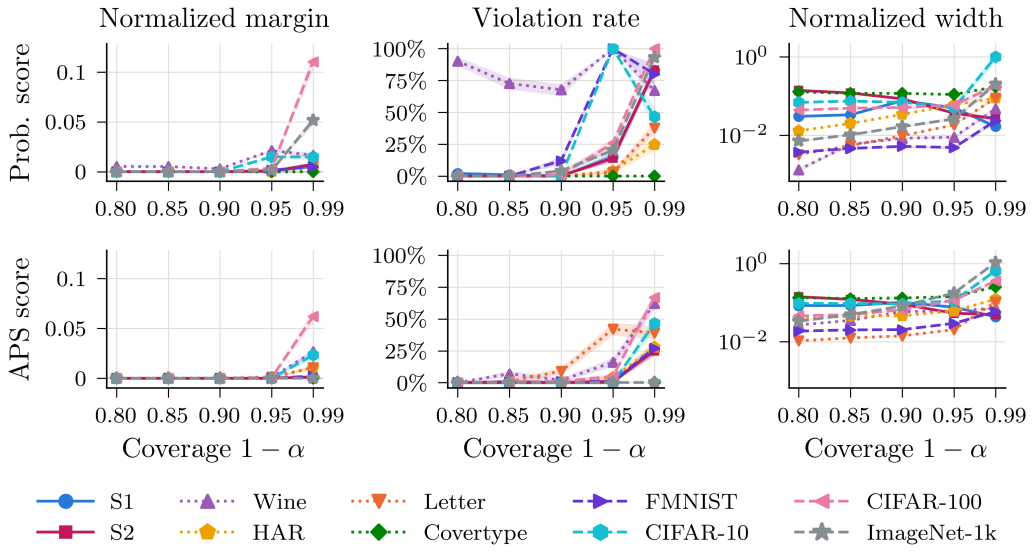

Figure 10: We show the normalized margin, violation rate, and normalized width as we vary the nominal coverage level $1-\alpha \in \{0.8, 0.85, 0.9, 0.95, 0.99\}$, with n_{cal} and n_{test} fixed at their largest values for each dataset. All panels report the mean $\pm$ standard error based on 100 random draws of the calibration and test sets from the fixed evaluation pool.

Nominal coverage levels (Figure 10). The bounds remain well-behaved across nominal coverage levels $1-\alpha \leq 0.95$ for most datasets (with exceptions for Wine, FMNIST, and CIFAR-10). However, at $1-\alpha = 0.99$, all three metrics increase sharply for most datasets. We attribute this behavior to sensitivity to the calibrated thresholds $\hat{\lambda}_\alpha^{(i)}$ and $\hat{\Lambda}_\alpha^{(i)}$. As the nominal coverage level increases, these thresholds become smaller, which amplifies the finite-sample calibration terms and consequently increases the width of the bounds. The smaller thresholds also increase the variance of the corresponding empirical estimates, leading to larger normalized margins and violation rates at fixed calibration and test sample sizes.

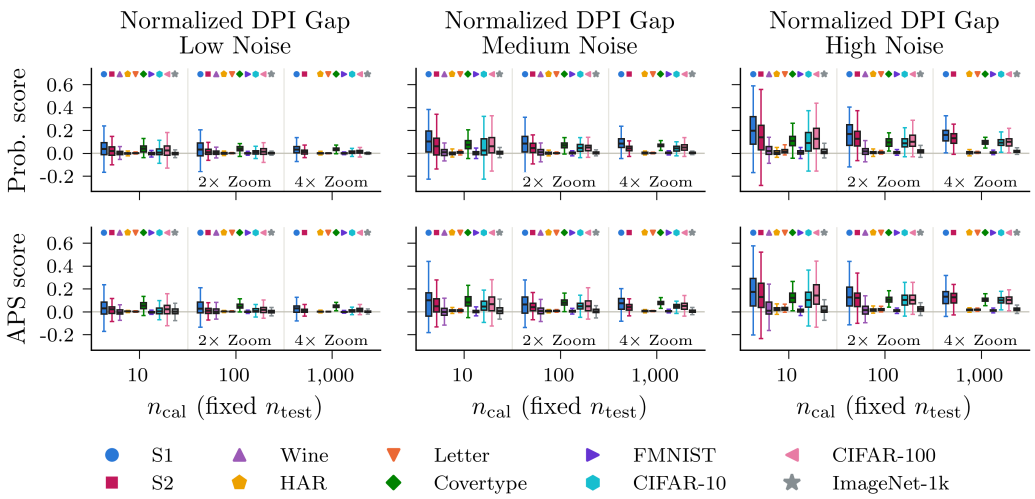

Figure 11: We show the normalized DPI gap across calibration sample sizes and degradation levels, using boxplots to summarize 100 random draws of the calibration and test sets. Each column corresponds to a degradation level, and the marker for each dataset is indicated above the corresponding boxplots. For visual clarity, we magnify the y-axis by $2\times$ and $4\times$ for $n_{\text{cal}} = 100$ and $n_{\text{cal}} = 1{,}000$, respectively. Datasets whose evaluation pool is too small for a given calibration size are omitted.

E.2 DPI VALIDATION RESULTS

We show the distribution of the normalized DPI gap across calibration sample sizes and degradation levels in Figure 11. The results for the medium degradation level ($\gamma = 0.5$) correspond to those reported in Section 5.

Across all degradation levels, the realized gap can be negative for particular draws of the calibration and test data. As the calibration sample size increases, however, the variability of the gap decreases and its distribution generally concentrates around a non-negative mean. Increasing the degradation level also generally increases the average gap as expected, since stronger noise removes more of the information present in X. These trends hold for both probability score and APS score.

These results further support the interpretation of conformal set-size reduction as a measure of information gain. Theorems 3 and 5 establish an approximate data processing inequality, with finite-sample calibration and model-error terms that allow the DPI gap to be negative. Empirically, the gap is generally non-negative on average, providing evidence of this data processing behavior in practice. At the same time, the negative gaps observed for individual replicates demonstrate that the deviations permitted by our theoretical results do indeed occur in practice.

E.3 FEATURE SELECTION RESULTS

We report feature selection results for all datasets in Table 5, excluding synthetic setting S1 since it has too few candidate features to meaningfully distinguish the criteria.

Across most settings, selection based on set-size reduction yields the smallest prediction sets while typically achieving the best top-5 accuracy. The main exceptions occur for the APS score on Wine and FMNIST. The two Δ_α criteria perform similarly to each other but do not always yield the smallest sets under their own score. Since these differences are often within one standard error, they likely reflect estimation noise in the selection set.

We also evaluate selection using I_λ, which performs similarly to Δ_α under the probability score in the synthetic settings but often yields larger sets on the real-world datasets. We attribute this gap to model imperfection. Theorem 2 brackets Δ_α by conformal mutual information only up to the model error terms $D^{(i)}$, so the plug-in I_λ at the geometric mean of the calibrated thresholds becomes a less precise proxy for Δ_α when these terms are nonzero.

Table 5: We evaluate feature selection performance using conformal set-size reduction. Each row represents one selection criterion and values are reported as mean ± standard error over 10 random splits of the evaluation pool. Arrows indicate the direction of improvement, and the best-performing method for each metric is bolded. Empirical coverage is reported for reference without bolding to indicate that differences in set size are not due to undercoverage.

Dataset	Criterion	Prob.		APS		Acc.	
		$\mathbb{E}[\lvert\mathcal{C}\rvert]$ (↓)	Cov.	$\mathbb{E}[\lvert\mathcal{C}\rvert]$ (↓)	Cov.	Top-1 (↑)	Top-5 (↑)
S2	Δ_α (Prob.)	11.464 ± 0.032	0.900 ± 0.001	11.952 ± 0.029	0.900 ± 0.001	0.366 ± 0.001	$\mathbf{0.722 \pm 0.000}$
	Δ_α (APS)	11.459 ± 0.033	0.900 ± 0.001	11.950 ± 0.029	0.900 ± 0.001	0.366 ± 0.001	$\mathbf{0.722 \pm 0.001}$
	I_λ (Prob.)	$\mathbf{11.456 \pm 0.032}$	0.900 ± 0.001	$\mathbf{11.948 \pm 0.028}$	0.900 ± 0.001	0.366 ± 0.001	$\mathbf{0.722 \pm 0.001}$
	Shannon MI	11.473 ± 0.031	0.900 ± 0.001	11.959 ± 0.027	0.900 ± 0.001	$\mathbf{0.367 \pm 0.001}$	$\mathbf{0.722 \pm 0.001}$
	Accuracy	11.503 ± 0.029	0.900 ± 0.001	11.986 ± 0.026	0.900 ± 0.001	$\mathbf{0.367 \pm 0.001}$	$\mathbf{0.722 \pm 0.001}$
	Random	12.207 ± 0.085	0.900 ± 0.001	12.670 ± 0.085	0.900 ± 0.001	0.354 ± 0.002	0.707 ± 0.002
S3	Δ_α (Prob.)	2.239 ± 0.002	0.900 ± 0.000	2.629 ± 0.002	0.900 ± 0.000	0.750 ± 0.001	0.952 ± 0.000
	Δ_α (APS)	2.243 ± 0.002	0.900 ± 0.000	$\mathbf{2.611 \pm 0.002}$	0.900 ± 0.000	0.752 ± 0.000	0.952 ± 0.000
	I_λ (Prob.)	$\mathbf{2.237 \pm 0.003}$	0.900 ± 0.000	2.626 ± 0.002	0.900 ± 0.000	0.753 ± 0.000	$\mathbf{0.953 \pm 0.000}$
	Shannon MI	2.447 ± 0.005	0.900 ± 0.000	2.876 ± 0.006	0.900 ± 0.000	0.767 ± 0.000	0.952 ± 0.000
	Accuracy	3.613 ± 0.020	0.900 ± 0.000	4.487 ± 0.029	0.900 ± 0.000	$\mathbf{0.768 \pm 0.001}$	0.914 ± 0.001
	Random	3.046 ± 0.100	0.900 ± 0.000	3.591 ± 0.116	0.900 ± 0.000	0.672 ± 0.009	0.931 ± 0.004
Wine	Δ_α (Prob.)	$\mathbf{2.265 \pm 0.034}$	0.904 ± 0.008	2.387 ± 0.029	0.905 ± 0.006	$\mathbf{0.553 \pm 0.006}$	$\mathbf{0.996 \pm 0.001}$
	Δ_α (APS)	2.279 ± 0.042	0.902 ± 0.008	2.378 ± 0.036	0.904 ± 0.007	0.548 ± 0.005	0.995 ± 0.001
	I_λ (Prob.)	2.334 ± 0.050	0.903 ± 0.010	2.417 ± 0.040	0.904 ± 0.009	0.529 ± 0.005	0.995 ± 0.001
	Shannon MI	2.288 ± 0.042	0.900 ± 0.009	2.383 ± 0.036	0.903 ± 0.008	0.550 ± 0.005	0.995 ± 0.001
	Accuracy	2.284 ± 0.042	0.904 ± 0.009	$\mathbf{2.371 \pm 0.032}$	0.904 ± 0.007	0.552 ± 0.004	$\mathbf{0.996 \pm 0.001}$
	Random	2.420 ± 0.046	0.906 ± 0.007	2.506 ± 0.046	0.902 ± 0.007	0.507 ± 0.005	0.995 ± 0.001
HAR	Δ_α (Prob.)	0.912 ± 0.005	0.899 ± 0.005	0.963 ± 0.002	0.900 ± 0.002	$\mathbf{0.971 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Δ_α (APS)	$\mathbf{0.909 \pm 0.006}$	0.896 ± 0.006	$\mathbf{0.960 \pm 0.002}$	0.899 ± 0.001	0.970 ± 0.001	$\mathbf{1.000 \pm 0.000}$
	I_λ (Prob.)	0.918 ± 0.005	0.897 ± 0.005	0.961 ± 0.001	0.898 ± 0.002	0.968 ± 0.002	$\mathbf{1.000 \pm 0.000}$
	Shannon MI	0.916 ± 0.005	0.898 ± 0.005	$\mathbf{0.960 \pm 0.001}$	0.899 ± 0.002	0.968 ± 0.002	$\mathbf{1.000 \pm 0.000}$
	Accuracy	0.910 ± 0.005	0.898 ± 0.004	0.961 ± 0.003	0.900 ± 0.001	$\mathbf{0.971 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Random	0.999 ± 0.015	0.900 ± 0.004	1.064 ± 0.017	0.899 ± 0.002	0.934 ± 0.005	$\mathbf{1.000 \pm 0.000}$
Letter	Δ_α (Prob.)	$\mathbf{2.753 \pm 0.016}$	0.898 ± 0.003	$\mathbf{3.358 \pm 0.019}$	0.900 ± 0.001	0.756 ± 0.001	$\mathbf{0.926 \pm 0.000}$
	Δ_α (APS)	2.835 ± 0.038	0.898 ± 0.002	3.398 ± 0.032	0.900 ± 0.001	0.758 ± 0.002	0.924 ± 0.001
	I_λ (Prob.)	2.824 ± 0.022	0.898 ± 0.003	3.385 ± 0.028	0.901 ± 0.002	0.760 ± 0.001	0.924 ± 0.001
	Shannon MI	2.801 ± 0.019	0.899 ± 0.003	3.399 ± 0.026	0.901 ± 0.001	0.758 ± 0.001	0.924 ± 0.001
	Accuracy	2.879 ± 0.020	0.899 ± 0.002	3.415 ± 0.022	0.901 ± 0.001	$\mathbf{0.763 \pm 0.001}$	0.922 ± 0.001
	Random	4.459 ± 0.184	0.897 ± 0.002	5.054 ± 0.174	0.900 ± 0.001	0.642 ± 0.011	0.871 ± 0.007
Covertype	Δ_α (Prob.)	$\mathbf{1.010 \pm 0.001}$	0.900 ± 0.001	1.139 ± 0.001	0.900 ± 0.000	$\mathbf{0.906 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Δ_α (APS)	1.011 ± 0.001	0.900 ± 0.001	1.140 ± 0.001	0.900 ± 0.000	$\mathbf{0.906 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	I_λ (Prob.)	$\mathbf{1.010 \pm 0.001}$	0.900 ± 0.001	$\mathbf{1.138 \pm 0.001}$	0.900 ± 0.000	$\mathbf{0.906 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Shannon MI	$\mathbf{1.010 \pm 0.001}$	0.900 ± 0.001	$\mathbf{1.138 \pm 0.001}$	0.900 ± 0.000	$\mathbf{0.906 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Accuracy	$\mathbf{1.010 \pm 0.001}$	0.900 ± 0.001	1.139 ± 0.001	0.900 ± 0.000	$\mathbf{0.906 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$
	Random	1.593 ± 0.035	0.901 ± 0.001	1.731 ± 0.036	0.900 ± 0.000	0.717 ± 0.011	0.996 ± 0.000
FMNIST	Δ_α (Prob.)	$\mathbf{1.020 \pm 0.002}$	0.900 ± 0.002	1.220 ± 0.004	0.898 ± 0.001	0.903 ± 0.001	$\mathbf{0.998 \pm 0.000}$
	Δ_α (APS)	1.028 ± 0.002	0.901 ± 0.002	1.167 ± 0.002	0.899 ± 0.001	0.901 ± 0.001	$\mathbf{0.998 \pm 0.000}$
	I_λ (Prob.)	1.040 ± 0.002	0.901 ± 0.002	1.166 ± 0.002	0.899 ± 0.001	0.899 ± 0.001	$\mathbf{0.998 \pm 0.000}$
	Shannon MI	1.037 ± 0.002	0.901 ± 0.002	$\mathbf{1.163 \pm 0.002}$	0.899 ± 0.001	0.901 ± 0.001	$\mathbf{0.998 \pm 0.000}$
	Accuracy	1.023 ± 0.002	0.902 ± 0.002	1.220 ± 0.003	0.898 ± 0.001	$\mathbf{0.904 \pm 0.001}$	0.997 ± 0.000
	Random	1.064 ± 0.004	0.902 ± 0.002	1.259 ± 0.004	0.899 ± 0.001	0.896 ± 0.001	0.997 ± 0.000
CIFAR-10	Δ_α (Prob.)	$\mathbf{1.489 \pm 0.007}$	0.899 ± 0.003	1.774 ± 0.007	0.899 ± 0.001	0.795 ± 0.002	$\mathbf{0.986 \pm 0.000}$
	Δ_α (APS)	1.507 ± 0.006	0.900 ± 0.002	$\mathbf{1.742 \pm 0.004}$	0.900 ± 0.001	0.793 ± 0.002	$\mathbf{0.986 \pm 0.000}$
	I_λ (Prob.)	1.592 ± 0.010	0.900 ± 0.002	1.774 ± 0.008	0.899 ± 0.001	0.781 ± 0.001	0.984 ± 0.000
	Shannon MI	1.561 ± 0.007	0.899 ± 0.002	1.756 ± 0.005	0.899 ± 0.001	0.785 ± 0.001	0.984 ± 0.000
	Accuracy	1.502 ± 0.008	0.901 ± 0.002	1.780 ± 0.004	0.900 ± 0.001	$\mathbf{0.800 \pm 0.002}$	0.985 ± 0.000
	Random	1.711 ± 0.012	0.900 ± 0.002	1.996 ± 0.015	0.900 ± 0.002	0.772 ± 0.003	0.978 ± 0.001
CIFAR-100	Δ_α (Prob.)	12.683 ± 0.100	0.904 ± 0.002	15.661 ± 0.127	0.902 ± 0.002	0.466 ± 0.002	$\mathbf{0.753 \pm 0.002}$
	Δ_α (APS)	$\mathbf{12.594 \pm 0.088}$	0.902 ± 0.002	$\mathbf{15.472 \pm 0.129}$	0.902 ± 0.002	$\mathbf{0.467 \pm 0.002}$	$\mathbf{0.753 \pm 0.002}$
	I_λ (Prob.)	13.700 ± 0.115	0.901 ± 0.002	16.438 ± 0.110	0.901 ± 0.002	0.456 ± 0.003	0.740 ± 0.002
	Shannon MI	13.763 ± 0.115	0.902 ± 0.002	16.597 ± 0.145	0.902 ± 0.002	0.458 ± 0.003	0.741 ± 0.002
	Accuracy	12.872 ± 0.130	0.902 ± 0.002	15.877 ± 0.158	0.901 ± 0.002	$\mathbf{0.467 \pm 0.003}$	0.752 ± 0.002
	Random	14.243 ± 0.253	0.902 ± 0.003	17.189 ± 0.224	0.902 ± 0.003	0.454 ± 0.002	0.737 ± 0.002
ImageNet-1k	Δ_α (Prob.)	$\mathbf{47.159 \pm 0.297}$	0.901 ± 0.001	58.993 ± 0.352	0.900 ± 0.001	0.476 ± 0.001	0.711 ± 0.001
	Δ_α (APS)	47.262 ± 0.322	0.901 ± 0.001	$\mathbf{58.155 \pm 0.337}$	0.901 ± 0.001	0.475 ± 0.001	0.711 ± 0.001
	I_λ (Prob.)	51.958 ± 0.443	0.901 ± 0.001	63.887 ± 0.359	0.900 ± 0.001	0.472 ± 0.001	0.707 ± 0.001
	Shannon MI	55.959 ± 0.358	0.900 ± 0.001	68.675 ± 0.320	0.900 ± 0.001	0.471 ± 0.001	0.705 ± 0.001
	Accuracy	51.504 ± 0.494	0.900 ± 0.001	64.796 ± 0.581	0.900 ± 0.001	$\mathbf{0.480 \pm 0.001}$	$\mathbf{0.713 \pm 0.001}$
	Random	58.187 ± 0.522	0.901 ± 0.001	72.623 ± 0.616	0.900 ± 0.001	0.472 ± 0.001	0.704 ± 0.001

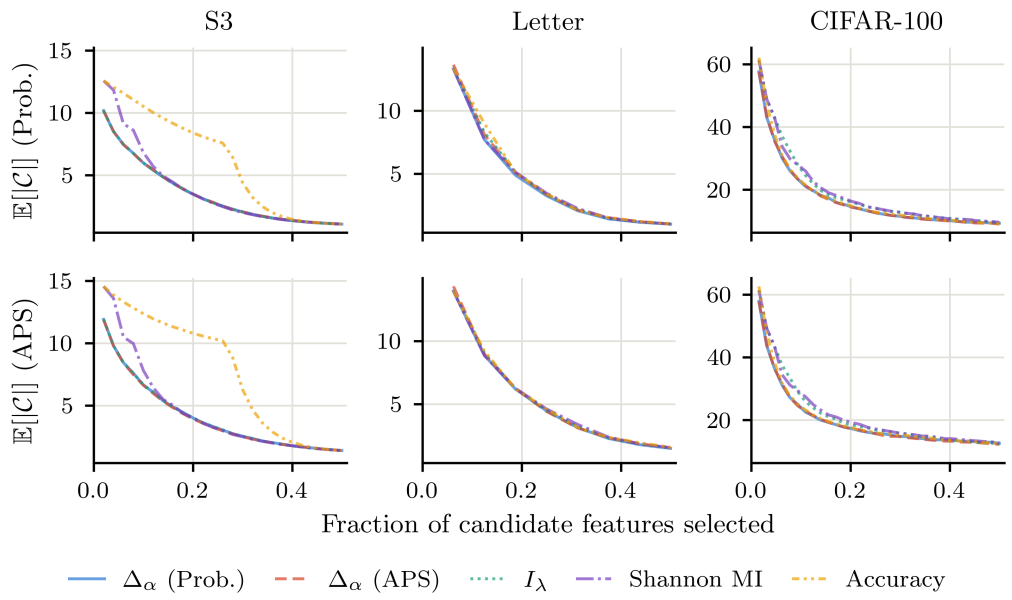

Figure 12: We show the expected set size using the probability score and APS score against the fraction of candidate features selected. Each curve is the mean over 10 random splits of the evaluation pool. We show only the first half of the selection trajectory, after which the curves largely coincide. On S3 and CIFAR-100, the gap between Δ_α and Shannon MI is concentrated in the early steps, which becomes diluted when averaging over the full trajectory.

The metrics in Table 5 are averaged over all intermediate steps of the feature ordering. Since every criterion eventually selects all candidate features, the set sizes converge as more features are acquired, and averaging over the full trajectory can mask differences that occur early in selection. To examine this, we show the expected set size at each step for the three datasets from Table 1 in Figure 12. The Δ_α criteria yield noticeably smaller sets than Shannon MI when only a small fraction of the candidates has been selected for S3 and CIFAR-100. On Letter, Δ_α using the probability score provides only a small improvement throughout the trajectory.

E.4 Subset-specific model ablation

In the feature selection experiments, the selection criteria and the evaluation both rely on a single mask-conditioned model. A difference between two criteria could therefore reflect how well this model fits particular subsets of features, rather than the information those features carry about Y. To test for this effect, we retrain a subset-specific model for each selected subset on CIFAR-100.

For each criterion and random split, we take the feature ordering from the main experiment and consider its first $k \in \{1, 2, 4, 8\}$ candidates, since the difference between the criteria is concentrated in the early steps of selection (Figure 12). We train a separate classification head on each resulting subset of the frozen backbone representation, using the same architecture and hyperparameters as the mask-conditioned head but without masking. Each subset-specific head is then calibrated and evaluated on the same data as the mask-conditioned model.

Table 6 reports the difference in expected set size between Shannon MI and Δ_α under both models. For simplicity, each Δ_α criterion is evaluated under its own score. The subset-specific models yield smaller sets than the mask-conditioned model for every criterion, and the advantage of Δ_α decreases when averaged over k from 4.07 to 1.89 under the probability score and from 3.84 to 2.11 under the APS score. However, Δ_α continues to yield smaller sets at every k, with the largest differences in the first two steps.

Table 6: Difference in expected set size between Shannon MI and Δ_α on CIFAR-100 when each feature ordering is evaluated with the mask-conditioned model or with subset-specific models. Positive values indicate that Δ_α yields smaller sets. Values are reported as mean $\pm$ standard error over 10 random splits of the evaluation pool.

	Prob.		APS	
k	Mask-cond.	Subset-specific	Mask-cond.	Subset-specific
1	3.75 ± 0.36	2.49 ± 0.32	3.59 ± 0.40	2.33 ± 0.35
2	5.81 ± 0.46	4.06 ± 0.35	5.24 ± 0.54	3.87 ± 0.46
4	3.71 ± 0.33	0.42 ± 0.24	2.98 ± 0.67	1.32 ± 0.31
8	3.00 ± 0.46	0.58 ± 0.32	3.56 ± 0.64	0.91 ± 0.28
Avg.	4.07 ± 0.27	1.89 ± 0.14	3.84 ± 0.41	2.11 ± 0.14

F COMPUTE RESOURCES

All training and analysis were performed on a single machine running Ubuntu 24.04 with 32 GB of RAM and an NVIDIA GeForce RTX 5070 Ti GPU. The hyperparameter grid search required approximately 10 hours and subsequent model training required approximately 12 hours. The validation experiments were run on a single CPU thread without GPU acceleration, whereas the feature selection experiments evaluated the mask-conditioned models on the GPU. All downstream analyses required approximately 4 hours in total.

G EXPERIMENTAL ASSETS

We obtain the Human Activity Recognition dataset (Reyes-Ortiz et al., 2013) directly from the UCI Machine Learning Repository, available under the CC BY 4.0 license. For ImageNet-1k, we use the 64×64 downsampled version of Chrabaszcz et al. (2017), which is made available through the `ChocolateDave/imagenet-64` mirror on Hugging Face. The pretrained ResNet-18 weights (He et al., 2016) are the `IMAGENET1K_V1` weights distributed with torchvision. All remaining datasets are obtained through OpenML or torchvision, with their original sources cited in Section 5 and Appendix B. Code for reproducing our analyses will be released upon publication.