CONFORMAL ELICITATION: LOSS MINIMIZATION CONTROLS SET SIZE REGRET

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction turns a probabilistic classifier into a prediction set with a prescribed coverage level, but coverage alone does not determine how useful that set is. We view expected set size at fixed marginal coverage as a downstream decision objective and ask how well ordinary probability estimation, via loss minimization, serves this objective. For conformal prediction sets induced by *probability conformity scores*, we prove that excess expected size over the Bayes-optimal set is at most $K^{3/2}\alpha^{-1}\sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)}$, where K is the number of classes, α is the target miscoverage, and $\mathsf{Reg}_{\mathsf{Brier}}$ is excess Brier loss. The guarantee requires no margin or density assumption and has sharp dependence on K and α up to a universal constant. We also prove a more general result that bounds the gap to the minimum set size within a model class $\mathcal{F}$ using the loss improvement achievable via a family of post-processings corresponding to $\mathcal{F}$. These results suggest that standard loss minimization using a proper loss (e.g. Brier loss or log loss) can simultaneously achieve the goal of minimizing conformal set size. We run experiments to compare this method with ConfTr (Stutz et al., 2022), a previous method specifically designed for set size minimization. These experiments show, somewhat surprisingly, that loss minimization consistently achieves smaller set sizes compared to ConfTr at miscoverage level $\alpha = 0.05$. These results demonstrate the power of standard proper loss minimization in providing high-quality uncertainty quantification.

1 INTRODUCTION

As predictions from AI systems increasingly inform decision-making, quantifying their uncertainty is important for assessing the associated risks. Conformal prediction is a popular method for this purpose (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Given any fitted predictor, it defines a conformity score function and constructs a prediction set for a new unlabeled sample by including all labels whose conformity scores are at or above a threshold calibrated on held-out labeled data. It has attracted significant interest in part due to its model-agnostic nature: the prediction sets satisfy marginal coverage, i.e., they contain the true label with a prescribed probability. This holds as long as the labeled data used to calibrate the threshold is exchangeable with the new test point, without requiring access to the predictor's internal implementation or any additional fine-tuning.

Coverage alone, however, does not ensure the conformal sets are practically useful. A prediction set that includes all labels achieves coverage but is useless. Expected set size at a fixed coverage level is therefore a well-studied measure of conformal efficiency (Vovk et al., 2016; Sadinle et al., 2019; Stutz et al., 2022). Intuitively, a smaller set size predicts fewer labels and is more informative while retaining the prescribed coverage probability. Prior work has characterized *exact optimality* for set size: (Sadinle et al., 2019) shows that the set size minimizer is achieved by thresholding the true conditional probability $p^\star$, as recalled in Proposition 2. However, $p^\star$ is often impossible to learn exactly from finite samples. This raises the question of *approximate optimality*: how can we learn predictors and construct conformity scores that yield small conformal prediction sets?

A line of work addresses this question by optimizing set size directly (Stutz et al., 2022; Shi et al., 2025). This optimization problem is not straightforward due to the discrete nature of set size as the optimization objective, which is commonly addressed by minimizing certain differentiable

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.

approximations of set size, or even more complicated objectives. As such, these approaches typically rely on specially designed optimization programs, and can suffer from the intrinsic training error.

In parallel, an arguably more heuristic approach is to first characterize a desired form of prediction sets (alternatively, the conformity score yielding the sets) which often involves unknown "ideal" predictors (e.g., the true conditional probability $p^\star$). Then, one can learn a predictor that approximates the "ideal" predictor by minimizing a standard prediction loss, such as the log loss (a.k.a. cross-entropy) or Brier loss (a.k.a. squared loss), and plug this trained predictor into the conformity score (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021). Thus far, there have been few theoretical guarantees for this strategy, motivating the following question.

When does loss minimization directly yield regret bounds for set size in conformal prediction?

Our work formulates conformal set size optimization as a downstream task of standard prediction-loss minimization. We characterize when training a predictor q to achieve low prediction loss, and using it to construct conformal prediction sets, achieves near-optimal set size. We connect these objectives by showing that near-optimal Brier or log loss guarantees near-optimal set size for threshold-based conformal classification. This guarantee shows the predicted probabilities can be used directly as conformity scores, without additional transformations for conformal efficiency.

Loss minimization controls set size regret. Theorem 1 quantifies how closely an estimated predictor approaches optimal conformal prediction. Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ be a probabilistic predictor which outputs a distribution over K labels $\mathcal{Y}$, and let $p^\star : \mathcal{X} \to \Delta(\mathcal{Y})$ denote the true conditional distribution of the label given the features. Write $\mathsf{size}_\alpha(q)$ for the expected size of the population-calibrated set $S_{q,\alpha}$ at miscoverage α, and $\mathsf{SSR}_\alpha(q) = \mathsf{size}_\alpha(q) - \mathsf{size}^\star_\alpha$ for its set size regret, where $\mathsf{size}^\star_\alpha := \mathsf{size}_\alpha(p^\star)$ is the (optimal) expected set size achieved by $p^\star$. We prove that

$$0 \le \mathsf{SSR}_\alpha(q) = \mathsf{size}_\alpha(q) - \mathsf{size}_\alpha(p^\star) \le \frac{K}{\alpha} \left\| q - p^\star \right\|_1 \le \frac{K^{3/2}}{\alpha} \sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)}, \tag{1}$$

where $\|q - p^\star\|_1 := \mathbb{E}\|q(X) - p^\star(X)\|_1$, $\mathsf{Reg}_{\mathsf{Brier}}(q) := \mathbb{E}\|q(X) - p^\star(X)\|_2^2$ is excess multiclass Brier loss, and $1 - \alpha$ is the prescribed coverage. This bound shows that the set size regret is linearly controlled by the ℓ_1 distance between the predictor q and $p^\star$, and polynomially bounded by q's Brier regret. The guarantee holds for every distribution D and predictor q, without margin or density assumptions. Thus vanishing Brier regret implies optimal conformal efficiency. The same argument also bounds excess set size in terms of log-loss regret (Corollary 1), and more generally, regret in any *strongly convex proper loss* (cf. Definition 1 and Definition 2). We complement (1) with a lower bound in Appendix B.3, which shows that its dependences on K and α are sharp up to a universal constant. In Appendix B.4, we give a finite-sample extension of the population-level bound (1), as well as extensions to the *split conformal prediction* setup widely used in practice.

Relation to prior work. To our knowledge, (1) is the first quantitative regret transfer showing that excess risk under standard proper prediction losses directly controls set size regret of conformal prediction sets, without using any margin or distributional restrictions. Prior work has established exact optimality guarantees and has given plug-in guarantees for set size regret under additional regularity conditions Sadinle et al. (2019), or develops objectives tailored specifically to conformal efficiency Wang & Qiao (2023); Jia et al. (2026). In contrast, our result complements this literature by showing that ordinary loss minimization already controls set size regret under minimal assumptions. For completeness, we give a thorough discussion of related work in Appendix A.

Extension under misspecification. The bound (1) is meaningful when our learned q is close to $p^\star$. However, in practice q is often selected among a class $\mathcal{F}$ of predictors meant to model $p^\star$ (e.g., neural networks). In misspecified settings where $\mathcal{F}$ poorly approximates $p^\star$, the best in-class model may have a large Brier regret. Indeed, in Appendix D we demonstrate strong separations between prediction loss minimizers and conformal set size minimizers under misspecification.

To circumvent the hardness of applying (1) directly in the misspecified setting, in Section 4, we prove Theorem 2, a more general result that uses a weaker assumption about loss minimization and applies more broadly. Instead of requiring that our predictor q closely approximates $p^\star$ in ℓ_1 distance or with respect to losses, Theorem 2 only requires a *local* optimality condition introduced in the work of Błasiok et al. (2023; 2024). This condition requires the predictor to achieve competitive loss against a family of *post-processings* of the predictor, and as argued by Błasiok et al. (2023; 2024), is expected to hold in real-world, large-scale neural networks optimized for loss minimization. We

show that this local optimality condition suffices for competitive set size against a benchmark class $\mathcal{F}$ of predictors, as long as the family of post-processings is rich enough relative to $\mathcal{F}$ (Section 4.2). As an intermediate result, we show that *indistinguishability* notions such as *multiaccuracy* and *weight-restricted calibration* jointly imply competitive set sizes (Section 4.1).

Empirical validation. Our experiments demonstrate the efficiency of loss minimization for producing small conformal prediction sets and clarify the boundary of our theory (Section 5). We compare loss minimization and the set size-training framework of ConfTr (Stutz et al., 2022) on both synthetic data and image classification datasets (Krizhevsky, 2009; LeCun et al., 2010). In the synthetic experiments following Romano et al. (2020), standard loss minimization consistently outperforms ConfTr across model classes, including cases where the model class covers the ground truth and when it does not. On image classification datasets, loss minimization outperforms ConfTr once the model class is reasonably expressive, with a few exceptions when the model class is too restricted (such as linear models, and 1-layer MLPs with 8 hidden nodes for CIFAR-100). These findings are consistent across cross-entropy and Brier losses. One possible explanation is that proper loss training may offer a more favorable optimization landscape and avoid the additional approximation introduced by smoothing the discrete prediction-set construction in ConfTr. Our experiments do not isolate these effects, so we view them as possible contributing factors to the observed performance gap. These results illustrate the practical value of proper loss training yet also clarify the particular benefit of the direct size optimization approach within highly restricted model classes.

Our results suggest loss minimization as a simple solution to optimal conformal efficiency: predictors with small excess Brier or log loss, or even appropriate local optimality guarantees, yield small excess set size. For practitioners, our work provides theoretical justification for training probabilistic classifiers with standard losses, and supply predictions directly as conformal scores.

2 PRELIMINARIES

Notation. Let $[K] \coloneqq \{i \in \mathbb{N} : i \leq K\}$ for $K \in \mathbb{N}$. We use $\mathcal{P}(S)$ to denote the power set (set of subsets) of a set S. For $v \in \mathbb{R}^K$ and $k \in [K]$, we let $v_{(k)} \in \mathbb{R}$ denote its kth largest element.

We use $\Delta(\Omega)$ to denote all distributions over a sample space Ω, so $\Delta([K])$ is the K-dimensional probability simplex. For $P \in \Delta(\Omega)$, we use $\omega \sim P$ to mean ω is a random element of Ω with law P. We use $\mathbb{I}(\mathcal{E})$ to denote the 0-1 indicator of an event $\mathcal{E}$. We use e_i to denote the ith standard basis vector when the ambient dimension is clear, e.g., if $y \in [K]$ then $e_y \in \mathbb{R}^K$.

Our study focuses on *batch learning* classification problems. To this end, we let $\mathcal{X}$ denote an abstract feature space and $\mathcal{Y} = [K]$ denote a label space, where K is the number of classes. We also let D be a fixed distribution over $\mathcal{X} \times \mathcal{Y}$, and denote its $\mathcal{X}$-marginal by $D_{\mathcal{X}}$.

We define the ground truth predictor $p^\star : \mathcal{X} \to \Delta(\mathcal{Y})$ by

$$p^\star_y(x) \coloneqq \mathbb{P}\left[Y = y \mid X = x\right], \text{ for all } y \in \mathcal{Y}. \tag{2}$$

Thus, an equivalent way of modeling $(X, Y) \sim D$ is to draw $X \sim D_{\mathcal{X}}$ and then $Y \sim p^\star(X)$. This notation (i.e., $\mathcal{X}, \mathcal{Y}, K, D, D_{\mathcal{X}}, p^\star$) is explicitly reserved throughout the paper.

Finally, we state one convenient piece of notation used throughout: for $q : \mathcal{X} \to \Delta(\mathcal{Y})$, we let

$$\|q - p^\star\|_1 \coloneqq \mathbb{E}_{X \sim D_{\mathcal{X}}}\left[\|q(X) - p^\star(X)\|_1\right]. \tag{3}$$

Loss minimization. In classification problems, the goal is to learn a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$ from samples drawn from D. Typically, the goal is for $q \approx p^\star$, the ground truth.

A particularly convenient family of losses in realizing this goal is *proper losses*.

Definition 1 (Proper loss)**.** Let $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ be a loss. We say that ℓ is *proper* if

$$q \in \arg\min_{p \in \Delta(\mathcal{Y})} \mathbb{E}_{Y \sim q}\left[\ell(p, Y)\right] \text{ for all } q \in \Delta(\mathcal{Y}).$$

Under a mild curvature condition (e.g., strong convexity, Definition 2), proper losses incentivize driving each prediction $q(x)$ to the ground truth $p^\star(x)$. Several standard losses in machine learning

are proper, e.g., the (multiclass) Brier and log losses,

$$\ell_{\mathsf{Brier}}(q, y) \coloneqq \|q - e_y\|_2^2 = \sum_{k \in [K]} (q_k - \mathbb{I}(k = y))^2, \quad \ell_{\mathsf{log}}(q, y) \coloneqq -\log q_y. \tag{4}$$

Treating the first input as a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$, (2) implies $\mathbb{E}_{(X,Y)\sim D}[\ell(q(X), Y)]$ is minimized pointwise by $q = p^\star$. We thus define for any $q : \mathcal{X} \to \Delta(\mathcal{Y})$, the *regret*

$$\mathsf{Reg}_\ell(q) \coloneqq \mathbb{E}_{(X,Y)\sim D}\left[\ell(q(X), Y) - \ell(p^\star(X), Y)\right]. \tag{5}$$

We respectively use $\mathsf{Reg}_{\mathsf{Brier}}, \mathsf{Reg}_{\mathsf{log}}$ for $\ell = \ell_{\mathsf{Brier}}, \ell_{\mathsf{log}}$, which have the convenient representations

$$\mathsf{Reg}_{\mathsf{Brier}}(q) = \mathbb{E}_{X\sim D_{\mathcal{X}}}\|q(X) - p^\star(X)\|_2^2, \quad \mathsf{Reg}_{\mathsf{log}}(q) = \mathbb{E}_{X\sim D_{\mathcal{X}}}\mathrm{KL}(p^\star(X)\|q(X)). \tag{6}$$

In the *batch learning* setting, we are given n i.i.d. samples (X_i, Y_i) drawn from D, and the goal is to learn a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$ minimizing Reg_ℓ for some loss $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$.

Conformal prediction. We provide a brief overview on conformal prediction, with a detailed exposition deferred to Vovk et al. (2022); Angelopoulos & Bates (2023).

In conformal prediction, the goal is to learn a *set-valued* predictor, $S : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. Intuitively, $S(x) \subseteq \mathcal{Y}$ represents a predicted set of labels for given features $x \in \mathcal{X}$. For a misclassification threshold $\alpha \in (0, 1)$, we say that $S : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is α-*valid* if

$$\mathbb{P}\left[Y \in S(X)\right] \geq 1 - \alpha. \tag{7}$$

Conformal prediction is a standard strategy for guaranteeing validity in batch learning settings. Let $B \coloneqq \{(X_i, Y_i)\}_{i\in[n]}$ be a calibration dataset, and let (X, Y) be a new sample we wish to classify, where we know $X \in \mathcal{X}$ but not its label Y. Then conformal prediction takes

$$S(X) \coloneqq \{y \in \mathcal{Y} : s(X, y) \geq \tau\}, \tag{8}$$

where $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a *conformity score*, and

$$\hat{s}_i \coloneqq s(X_i, Y_i) \text{ for all } i \in [n], \quad \tau \coloneqq \hat{s}_{(\lceil (1-\alpha)(n+1) \rceil)}. \tag{9}$$

We use the convention $\hat{s}_{(n+1)} \coloneqq -\infty$, so $\alpha < 1/(n+1)$ gives $\tau = -\infty$ and $S(X) = \mathcal{Y}$. In other words, conformal prediction sets $S(X)$ to include all labels $y \in \mathcal{Y}$ whose conformity score is above a threshold τ, roughly the αth quantile of the dataset's scores. The threshold $\tau = \hat{s}_{(\lceil (1-\alpha)(n+1) \rceil)}$ is standard in *split conformal prediction*, but slightly different rules, e.g., random inclusion at the threshold, are also sometimes used: we give a brief discussion in Appendix B.4.

Remarkably, the coverage condition (7) holds for any s, under fairly minimal assumptions.

Proposition 1 (Theorem 1, Angelopoulos & Bates (2023))**.** *If* $\{(X_i, Y_i)\}_{i\in[n]} \cup \{(X, Y)\}$ *is exchangeable, then the predictor* (9) *satisfies* (7) *for any choice of the conformity score* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Finally, our paper also considers the *idealized setting*, a population level analog of (8), (9). In the idealized setting, the test point (X, Y) is drawn from D, and τ in (8) is the population-level α^{th} quantile of s according to D. Intuitively, this corresponds to batch learning applications where $\{(X_i, Y_i)\}_{i\in[n]} \cup \{(X, Y)\}$ are i.i.d. draws from D, and $n \to \infty$. Our main results are proven under both the idealized setting, and the finite-sample ($n < \infty$) setting of (9).

Set size regret. Proposition 1 guarantees coverage for any choice of the conformity score s, but says nothing about the efficiency of the resulting sets. We use the standard notion of *expected set size* as our measure of efficiency. For a set-valued predictor $S : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, we define

$$\mathsf{size}(S) \coloneqq \mathbb{E}\left[|S(X)|\right], \quad \mathsf{mis}(S) \coloneqq \mathbb{P}\left[Y \notin S(X)\right],$$

where all expectations are taken over the randomness of both the test point (X, Y), and the randomness of the predictor S. For example, in conformal prediction (8), the set S is a function of the random calibration set $\{(X_i, Y_i)\}_{i\in[n]}$, so (7) reads equivalently as $\mathsf{mis}(S) \leq \alpha$.

For a target miscoverage rate α, we define the optimal expected size by

$$\mathsf{size}^\star_\alpha \coloneqq \inf\left\{\mathsf{size}(S) : \mathsf{mis}(S) \leq \alpha\right\}.$$

In the idealized setting, there is a natural conformity score that achieves expected size $\mathsf{size}^\star_\alpha$.

Proposition 2 (Theorem 1, Vovk et al. (2016) and Theorem 1, Sadinle et al. (2019)). *Let $(X, Y) \sim D$, and for each $\alpha \in [0, 1]$, let τ be the population-level α^{th} quantile of a conformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and define $S : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ according to* (8)*. Then $\mathsf{size}(S) = \mathsf{size}^\star_\alpha$ for all $\alpha \in [0, 1]$ if and only if*

$$p^\star_y(x) < p^\star_{y'}(x') \implies s(x, y) < s(x', y') \text{ for all } (x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}.$$

In other words, Proposition 2 states that in the idealized setting, the conformal set predictor (8) with $s(x, y) = p^\star_y(x)$ is not only optimal among all conformal set predictors, it is in fact optimal among *all* predictors S satisfying $\mathsf{mis}(S) \leq \alpha$. Further, any conformity score s that refines the ordering of the ground truth also induces S that achieves $\mathsf{size}^\star_\alpha$ at all α.

Finally, for a set-valued predictor $S : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, we define the *set size regret* at level α by

$$\mathsf{SSR}_\alpha(S) \coloneqq \mathsf{size}(S) - \mathsf{size}^\star_\alpha.$$

Probability conformity scores. Motivated by Proposition 2, in this paper, we focus on a specific set of conformity scores $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, called *probability conformity scores*:

$$s_q(x, y) \coloneqq q_y(x), \text{ for } q : \mathcal{X} \to \Delta(\mathcal{Y}). \tag{10}$$

In other words, to score (x, y), we output the mass that the prediction $q(x)$ places on the label y. Under this interpretation, (8) includes all labels with at least τ mass according to $q(x)$.

Restricting to probability conformity scores imposes that the vector of $\{s(x, y)\}_{y \in \mathcal{Y}}$ lives in $\Delta(\mathcal{Y})$ for all $x \in \mathcal{X}$. While this restriction is not without loss of generality,[1] probability conformity scores are a standard and theoretically well-motivated specialization of conformal prediction Vovk et al. (2016); Sadinle et al. (2019). Indeed, under this restriction, exact set size optimality is still achievable (Proposition 2). More generally, this restriction is sensible in the context of our focus: relating *proper loss minimization*, which naturally yields a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$, to set size regret.

3 Loss regret bounds set size regret

In this section, we give our first main result, which shows that learning a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$ that approximately minimizes an appropriate proper loss ℓ also approximately minimizes the expected set size. We first state a basic version of this result in Theorem 1, expressed in terms of the ℓ_1 distance to $p^\star$ and under the idealized (population-level) setting. We lift this basic version to handle regret under common proper losses in Corollary 1.

We first work in the idealized setting, where we wish to label $(X, Y) \sim D$ and have access to a learned $q : \mathcal{X} \to \Delta(\mathcal{Y})$. For each $\alpha \in [0, 1]$, we let $S_{q,\alpha} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ be the set-valued predictor (8) where $s(x, y) = q_y(x)$ and τ is the population-level α^{th} quantile of s. We also define

$$\mathsf{size}_\alpha(q) \coloneqq \mathsf{size}\left(S_{q,\alpha}\right), \quad \mathsf{SSR}_\alpha(q) \coloneqq \mathsf{SSR}_\alpha\left(S_{q,\alpha}\right) = \mathsf{size}_\alpha(q) - \mathsf{size}_\alpha(p^\star). \tag{11}$$

We next state our basic result, which relates SSR_α and the ℓ_1 distance to $p^\star$.

Theorem 1 (Core reduction from set size to ℓ_1 distance). *Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$. Then following the notation* (2)*,* (3)*,* (11)*,*

$$\mathsf{SSR}_\alpha(q) \leq \frac{K}{\alpha} \left\| q - p^\star \right\|_1. \tag{12}$$

We next state some simple corollaries of Theorem 1 in terms of minimizing standard losses. The losses considered in this section will satisfy the following definition.

Definition 2 (Strong convexity). We say that $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ is *μ-strongly convex in ℓ_1* if

$$\mathbb{E}_{Y \sim p}[\ell(q, Y) - \ell(p, Y)] \geq \frac{\mu}{2} \left\| p - q \right\|_1^2, \text{ for all } p, q \in \Delta(\mathcal{Y}).$$

[1] Applying a monotone transform ϕ to all scores $s(x, y)$ does not affect the sets (8), which depend only on relative order. However, this precludes applying a transformation (e.g., softmax) to force $\{s(x, y)\}_{y \in \mathcal{Y}}$ to lie in $\Delta(\mathcal{Y})$ for all $x \in \mathcal{X}$, because the normalization depends on x; there exist s for which no $\phi \circ s$ has this property.

Lemma 1. *Let $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ be proper and μ-strongly convex in ℓ_1. Then following the notation* (5)*,* (11)*, for all predictors $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$, we have*

$$\mathsf{SSR}_\alpha(q) \le \frac{K}{\alpha}\sqrt{\frac{2}{\mu}\mathsf{Reg}_\ell(q)}.$$

For the Brier and log losses, Lemma 1 gives the following bounds.

Corollary 1. *Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$. Then following the notation* (2)*,* (6)*,*

$$\mathsf{SSR}_\alpha(q) \le \frac{K^{3/2}}{\alpha}\sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)}, \quad \mathsf{SSR}_\alpha(q) \le \frac{K}{\alpha}\sqrt{2\mathsf{Reg}_{\mathsf{log}}(q)}.$$

We defer the proofs of Theorem 1, Lemma 1, and Corollary 1 to Appendix B.

The $K^{3/2}/\alpha$ dependence of the Brier bound is tight up to a constant factor, as shown in Appendix B.3. We also give a finite-sample analog that accounts for sampling error, together with its extension to standard split conformal prediction, in Appendix B.4.

4 Local optimality bounds set size regret

Our main results in Section 3 show that when q closely approximates the ground truth predictor $p^\star$ in average ℓ_1 distance or with respect to appropriate losses, it also has a near-optimal set size. In practice, such a *global* optimality condition may be hard to achieve. For example, in a typical batch learning setting, we choose a class $\mathcal{F}$ of benchmark predictors $f : \mathcal{X} \to \Delta(\mathcal{Y})$, and use empirical risk minimization to compete with the best predictor in $\mathcal{F}$. This strategy may not achieve small population regret if the ground truth $p^\star$ is poorly approximated by $\mathcal{F}$. We call this the *misspecified setting*, a.k.a. the *agnostic learning setting*: the benchmark class $\mathcal{F}$ is not assumed to contain $p^\star$.

In this section, we complement Section 3 by showing that a *local* optimality condition for loss minimization also implies competitive set size against a general class $\mathcal{F}$ of benchmark predictors (Section 4.2). Towards this goal, we first show that appropriate *indistinguishability* conditions from complexity-theoretic regularity (Trevisan et al., 2009) and the *multicalibration/multiaccuracy* literature (Hebert-Johnson et al., 2018; Kim et al., 2019; Dwork et al., 2021; Hu & Vadhan, 2025) imply competitive set size (Section 4.1). This complements the counterexamples shown in Appendix D, where loss optimality within a restricted, misspecified family of models, e.g. multi-class logistic regression, does not guarantee competitive set sizes. Instead, our notion of local optimality follows the work of Błasiok et al. (2023; 2024), requiring a predictor q to achieve competitive loss against a class of post-processings of q. As argued in (Błasiok et al., 2023; 2024), optimality against post-processings is often expected for real-world, large-scale neural networks optimized for loss minimization.

4.1 Set size regret bounds via indistinguishability

Fix a target miscoverage $\alpha \in (0, 1)$. Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ be arbitrary, and fix a nonempty benchmark class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \Delta(\mathcal{Y})\}$; we do not require $q \in \mathcal{F}$. Define the optimal set size $\mathsf{size}_{\mathcal{F},\alpha}$ within class $\mathcal{F}$ and the set size gap $\mathsf{SSR}_{\mathcal{F},\alpha}(q)$ of a general predictor q as follows:

$$\mathsf{size}_{\mathcal{F},\alpha} := \inf_{f\in\mathcal{F}} \mathsf{size}_\alpha(f), \qquad \mathsf{SSR}_{\mathcal{F},\alpha}(q) := \mathsf{size}_\alpha(q) - \mathsf{size}_{\mathcal{F},\alpha}.$$

Here $\mathsf{size}_\alpha(f)$ is the expected set size defined in (11). The signed difference $\mathsf{SSR}_{\mathcal{F},\alpha}(q)$ can be negative because q may outperform all benchmarks. We seek an upper bound on this signed gap.

Following the notion of *indistinguishability* from complexity theory Trevisan et al. (2009); Hu & Vadhan (2025), we consider distinguishers of the form $h : \mathcal{X} \to [0, 1]^{\mathcal{Y}}$. We write $h(x, y) := h(x)_y$ for the coordinate corresponding to label y. The *distinguishing advantage* of h on a predictor q is

$$e_q(h) := \mathbb{E}_{X\sim D_{\mathcal{X}}}\langle q(X) - p^\star(X), h(X)\rangle.$$

We say q is indistinguishable from the ground-truth $p^\star$ w.r.t. distinguisher h if $e_q(h) = 0$.

Given a predictor f, define a distinguisher $h_f = h_{f,\alpha}$ by its inclusion probabilities at miscoverage α:

$$h_f(x, y) := \mathbb{P}\left[y \in S_{f,\alpha}(x)\right] = \mathbb{I}\left(f_y(x) > \tau_f\right) + \gamma_f \mathbb{I}\left(f_y(x) = \tau_f\right).$$

Here the probability is over the tie randomization in the population set $S_{f,\alpha}$ from Section 3; τ_f and γ_f are chosen to give miscoverage exactly α. We use $h_f(x) \in [0,1]^{\mathcal{Y}}$ for the vector of these inclusion probabilities, and write $h_\star = h_{p^\star}$ for the Bayes inclusion vector.

For every predictor q, Theorem 2 shows that indistinguishability w.r.t. distinguishers h_f for $f \in \mathcal{F} \cup \{q\}$ guarantees competitive set size against $\mathcal{F}$. In fact, one-sided bounds suffice: we only need an upper bound on $e_q(h_q)$ and a lower bound on $e_q(h_f)$ for each benchmark $f \in \mathcal{F}$.

Theorem 2. *For every distribution D, nonempty benchmark class $\mathcal{F}$, arbitrary predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$, and $\alpha \in (0,1)$, let $\varepsilon_1, \varepsilon_2 \in \mathbb{R}$ satisfy $e_q(h_q) \le \varepsilon_1$, and $\sup_{f\in\mathcal{F}}\big(-e_q(h_f)\big) \le \varepsilon_2$. Defining $(t)_+ := \max\{t, 0\}$, the set size regret $\mathsf{SSR}_{\mathcal{F},\alpha}(q)$ satisfies*

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \le (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{(\varepsilon_1 + \varepsilon_2)_+}{\alpha + \varepsilon_2}\right\}. \tag{13}$$

When $\alpha + \varepsilon_2 = 0$, the fraction is interpreted as 1, giving the trivial bound $\mathsf{SSR}_{\mathcal{F},\alpha}(q) \le K - \mathsf{size}_{\mathcal{F},\alpha}$.

We defer the proof to Appendix C.1.

At this point, it is appropriate to relate the conditions in Theorem 2 to existing notions in the indistinguishability literature. Having small ε_1 and ε_2 can be viewed as a relaxation of *multicalibration* (Hebert-Johnson et al., 2018). Specifically, ε_2 is a one-sided relaxation of *multiaccuracy* Kim et al. (2019); Hu & Vadhan (2025), whereas ε_1 can be viewed as *weight-restricted calibration* Gopalan et al. (2022b); Hu & Vadhan (2025). Most thematically related to Theorem 2, the design of predictors that simultaneously satisfy calibration and multiaccuracy has been an influential framework for *omniprediction* (Gopalan et al., 2022a) in binary (Gopalan et al., 2023a) and multiclass (Hu et al., 2026) settings; Theorem 2 shows that a similar decomposition suffices for set size regret.

Theorem 2 recovers Theorem 1 by using the benchmark class $\mathcal{F} = \{p^\star\}$; the derivation appears in Appendix C.2. Theorem 2 also extends to the finite-sample setting of Appendix B.4, with additional terms accounting for calibration error for q and the benchmarks in $\mathcal{F}$: see Appendix C.3.

4.2 Consequences for Brier loss and log loss

Previous work (Błasiok et al., 2023; 2024) shows that local optimality conditions for loss minimization yield calibration or multicalibration guarantees. We use this connection to bound ε_1 and ε_2 in Theorem 2, obtaining competitive set size from local optimality of loss minimization.

Let $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ be a loss. For a predictor q with $\mathbb{E}_{(X,Y)\sim D}\ell(q(X), Y) < \infty$ and a family $\mathcal{C}$ of *post-processings* containing q, define the largest possible loss improvement

$$\mathsf{Reg}_\ell(q; \mathcal{C}) := \mathbb{E}_{(X,Y)\sim D}\ell(q(X), Y) - \inf_{r\in\mathcal{C}} \mathbb{E}_{(X,Y)\sim D}\ell(r(X), Y). \tag{14}$$

The condition $\mathsf{Reg}_\ell(q; \mathcal{C}) \le \varepsilon$ means that no post-processing in $\mathcal{C}$ reduces loss by more than ε, and we use this as our local optimality condition for loss minimization, following (Błasiok et al., 2023; 2024). If q is ε-optimal over any enlarged class containing these post-processings, this condition follows. The post-processings below may use the inclusion vectors $h_f(x)$ from the benchmark family $\mathcal{F}$, which accounts for the dependence on $\mathcal{F}$. The following corollary shows that this local optimality condition with small ε for Brier loss implies a competitive set size.

Corollary 2. *Let $\Pi_{\Delta(\mathcal{Y})}$ denote Euclidean projection onto the simplex, and define*

$$T^{\mathsf{Brier}}_{\lambda,h}(q)(x) := \Pi_{\Delta(\mathcal{Y})}(q(x) + \lambda h(x)).$$

For an arbitrary $q : \mathcal{X} \to \Delta(\mathcal{Y})$, use the post-processing family

$$\mathcal{C}_{\mathsf{Brier}}(q, \mathcal{F}, \alpha) := \{T^{\mathsf{Brier}}_{-s,h_q}(q) : s \ge 0\} \cup \{T^{\mathsf{Brier}}_{s,h_f}(q) : f \in \mathcal{F},\ s \ge 0\}.$$

If $\mathsf{Reg}_{\mathsf{Brier}}(q; \mathcal{C}_{\mathsf{Brier}}(q, \mathcal{F}, \alpha)) \le \varepsilon$ for $\varepsilon \ge 0$, then

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \le (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{\sqrt{K\varepsilon}}{\alpha + \frac{1}{2}\sqrt{K\varepsilon}}\right\}.$$

The analogous guarantee for log loss uses exponential post-processings and is stated and proved in Appendix C.4 (Corollary 3). The proof of Corollary 2 also appears there.

5 EXPERIMENTS

Our theoretical results from Sections 3 and 4 suggest that standard loss minimization can be directly used for set size minimization in conformal prediction. In this section, we present experiments that test the performance of this simple approach relative to existing approaches designed specifically for set size minimization (Stutz et al., 2022). We present our results for synthetic and real-world data in the following two subsections, respectively. In both settings, standard loss minimization achieves improved set sizes than previous methods across a range of learning tasks and model architectures.

Specifically, we first consider a synthetic task with known class probabilities and optimal set size. We vary model expressiveness and task complexity controlling the numbers of active features in the prediction model and the data-generating process. Our results show that cross-entropy training approaches optimal efficiency in realizable classes and yields smaller mean sets than direct size training even in the misspecified classes we test. Second, we conduct real-data experiments across model architectures and coarse- and fine-label classification tasks. Proper-loss training consistently produces smaller sets in expressive models, while direct size minimization wins only when the model class is too restricted for the task.

5.1 SYNTHETIC EXPERIMENTS

Setup. Following Romano et al. (2020), we consider a classification task with $K = 10$ classes and features $x \in \mathbb{R}^{10}$. Labels follow $Y \mid X = x \sim p^\star(x)$, where $p^\star(x) = \mathrm{softmax}((W_s^\star)^\top x)$ and only the first s rows of $W_s^\star$ are nonzero. Here $p^\star$ denotes the ground-truth predictor for the chosen task. We fit linear softmax models using the first r input coordinates, varying $s, r \in \{2, 5, 10\}$. The model class contains the true conditional probabilities when $r \geq s$ and is generally misspecified when $r < s$. We compare cross-entropy (CE), multiclass Brier loss, and conformal training (ConfTr; Stutz et al., 2022), which minimizes a differentiable surrogate for prediction-set size. For each method, we select the checkpoint on a held-out validation set: CE and Brier-loss training minimize their respective validation losses, while ConfTr minimizes the validation THR prediction-set size. We evaluate prediction-set sizes against the Bayes THR oracle at nominal 90% coverage. We write n_{train} for the training-sample size, reserving n for the calibration-sample size as in Appendix B.4. Full experimental details and results for both training sizes are provided in Appendix E.

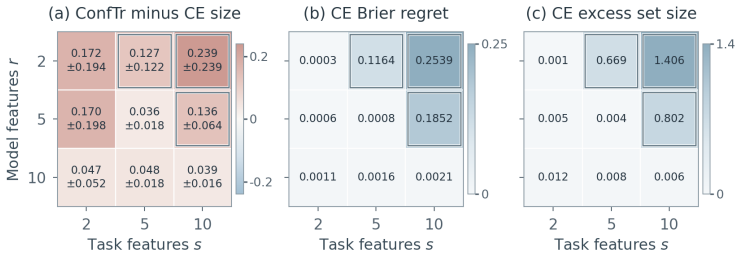

Figure 1: Synthetic active-feature study at nominal 90% coverage and $n_{\mathrm{train}} = 10{,}000$. Columns vary task features s and rows vary model features r. Outlined cells are misspecified ($r < s$). (a) Paired ConfTr-minus-CE set size, with 95% Student-t half-widths; positive values favor CE. (b) Brier regret of CE. (c) Excess set size of CE over the Bayes THR oracle. CE selects checkpoints by validation loss and ConfTr by validation set size. The corresponding Brier-loss results are in Figure 2.

Results. Figure 1(a) presents results on ConfTr and CE. CE yields smaller mean prediction sets than ConfTr in all nine configurations. This holds even though CE selects checkpoints by validation loss, whereas ConfTr selects them by validation set size. Figure 1(b) and (c) are consistent with Corollary 1: in the realizable cases ($r \geq s$), CE attains near-zero Brier regret and nearly matches the oracle set size. In the misspecified cases ($r < s$), CE still outperforms ConfTr. Brier loss training behaves almost identically to CE and also outperforms ConfTr in all nine configurations (See Section E.3).

One possible explanation for this performance gap is that proper loss training has a more favorable optimization landscape than ConfTr's differentiable set-size objective. Moreover, cross-entropy and Brier loss can be optimized directly, without smoothing the discrete prediction-set construction. ConfTr's surrogate introduces an additional potential mismatch between the training objective and the evaluated set size. These optimization and approximation effects may help explain why proper loss training achieves smaller sets, although our comparisons do not isolate their individual contributions.

5.2 REAL-DATA EXPERIMENTS

Setup. We compare CE and ConfTr (Stutz et al., 2022) on CIFAR-100 with its 100 fine or 20 coarse labels (Krizhevsky, 2009). All CIFAR experiments use 45,000 training examples and 5,000 calibration examples, with evaluation on the official 10,000-example test set. The calibration examples are used for determining the final threshold τ in the downstream conformal set generation procedure. Models include raw-pixel affine classifiers, MLPs with one 8-unit or two 64-unit hidden layers, and ResNet-50 with first-stage width of 256 channels. We first train each model with CE for 150 epochs, followed by 50 epochs with either CE or ConfTr. Affine models continue from the same per-seed checkpoint, while MLP and ResNet models freeze the CE-trained features and train new classification heads. Plain ConfTr uses a smooth set-size objective without an auxiliary classification loss. We consider nominal coverage levels of 95% and 99%. Architectures, optimization, the ConfTr objectives, and the calibration rule are detailed in Appendix F.

Evaluation. We report mean THR size and realized coverage, with accuracy and empirical Brier loss as predictive-quality measures. Since the true probabilities are unknown, we report the Brier loss w.r.t. the true labels instead of the Brier regret w.r.t. the ground-truth probabilities. Let Δ_{size} denote the set size difference between ConfTr and CE, where positive values favor CE.

Table 1: Real-data THR comparisons at nominal 95% coverage. Boldface marks the smaller mean size and Brier loss, and the higher accuracy within each row. Δ_{size} is the paired ConfTr-minus-CE difference with a 95% confidence-interval half-width. All rows use 45k training examples, a 5k calibration set, and the official 10k test set. Each row uses ten model/head seeds.

		Mean size			Coverage (%)		Brier loss		Accuracy (%)	
Task	Model	CE	ConfTr	Δ_{size}	CE	ConfTr	CE	ConfTr	CE	ConfTr
CIFAR-20	Affine	**14.8277**	15.0271	$+0.1994 \pm 0.0655$	95.345	95.144	**0.86195**	1.41626	**25.152**	17.355
CIFAR-20	MLP-1 $\times$ 8	**14.1820**	14.7721	$+0.5901 \pm 0.1820$	95.110	94.891	**0.86576**	0.92889	**23.260**	15.163
CIFAR-20	MLP-2 $\times$ 64	**11.8819**	12.7505	$+0.8686 \pm 0.0662$	94.769	95.037	**0.80858**	0.85348	**31.149**	25.181
CIFAR-20	ResNet-50 c256	**1.8775**	1.8892	$+0.0117 \pm 0.0046$	95.413	95.345	**0.25052**	0.26508	**82.350**	82.268
CIFAR-100	Affine	**65.7320**	69.3064	$+3.5745 \pm 0.3797$	94.824	95.272	**0.93046**	1.38400	**16.749**	8.017
CIFAR-100	MLP-1 $\times$ 8	**62.4687**	62.6357	$+0.1670 \pm 0.5792$	95.022	95.045	**0.94996**	0.97077	**12.171**	7.587
CIFAR-100	MLP-2 $\times$ 64	**47.3222**	49.2935	$+1.9713 \pm 0.1791$	94.964	94.997	**0.89385**	0.93472	**21.586**	15.767
CIFAR-100	ResNet-50 c256	**3.8187**	3.9032	$+0.0845 \pm 0.0628$	95.095	95.009	**0.35630**	0.38308	**73.770**	73.216

Results. At 95% coverage, CE yields smaller mean prediction sets in all eight comparisons (Table 1). The confidence intervals exclude zero in favor of CE in every row except CIFAR-100 with MLP-1×8. CE also achieves lower empirical Brier loss and higher accuracy in every row, indicating that better probability prediction and smaller prediction sets are achieved together. Additionally, CE also outperforms the tested classification-loss ConfTr variant on both ResNet tasks (Table 9).

Table 2: Prediction-set size differences across model classes and label granularities at nominal 99% coverage, using the same training/calibration/test split and exact order statistic as Table 1. Entries are paired ConfTr-minus-CE mean size differences Δ_{size} with 95% normal confidence-interval half-widths over ten model/head seeds. Positive Δ_{size} favors CE. Full metrics appear in Table 7.

Model	CIFAR-20	CIFAR-100
Affine	-0.359 ± 0.033	-3.189 ± 0.391
MLP-1×8	$+0.125 \pm 0.027$	-4.459 ± 0.171
MLP-2×64	$+0.953 \pm 0.037$	$+1.582 \pm 0.452$

We also consider the effect of increasing the coverage level from 95% to 99%, which corresponds to reducing α from 0.05 to 0.01. For this higher coverage level, we no longer observe the universal advantage of CE over ConfTr as in Table 1. This is consistent with our theoretical results in Sections 3 and 4 that scale as $\frac{1}{\alpha}$ and weaken as α decreases. As we report in Table 2, ConfTr yields smaller mean sets for the least expressive affine class, whereas CE yields smaller sets for the most expressive MLP-2×64 architecture. For the intermediate MLP-1×8 architecture, CE is favored on CIFAR-20 and ConfTr on CIFAR-100. This sign flip already appears at 98% coverage (Section F.5). These results show that CE need not yield the smallest sets when a demanding task is paired with a restricted model class. This is again well aligned with our theory in Sections 3 and 4 where the strongest set-size guarantee is proved when the model class is expressive enough to allow small loss w.r.t. the ground-truth probabilities.

AI USE STATEMENT

We used generative AI tools to assist in writing of proofs, design and provide feedback on experiments, and implement methods (i.e., generate code for the experiments). We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, provide critical ingredients for proving mathematical claims, design research methodology or generate synthetic data sets. Translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. Every mathematical proof generated by AI tools is verified by at least one author, and all LLM-generated code has been verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs of the results in Sections 3 and 4 are given in Appendices B and C. The synthetic experiments are specified in Appendix E, including the data-generating process, model selection, and all hyperparameters (Table 3). The CIFAR experiments are specified in Appendix F, including the data split, architectures, training schedules and hyperparameters (Table 6), the ConfTr objectives, and the calibration and evaluation procedures. Code is attached as supplementary material; it contains the training and evaluation code and the configurations for all experiments, together with instructions for running them.

REFERENCES

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 4–22. PMLR, 2015. URL https://proceedings.mlr.press/v40/Agarwal15.html.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

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. URL https://arxiv.org/abs/2009.14193.

Tian Bai and Ying Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024.

Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 525–547. PMLR, 2023. URL https://proceedings.mlr.press/v195/bao23a.html.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. doi: 10.1198/016214505000000907. URL https://doi.org/10.1198/016214505000000907.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72071–72095. Curran Associates, Inc., 2023. doi: 10.52202/075280-3154. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e4165c96702bac5f4962b70f3cf2f136-Paper-Conference.pdf.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss Minimization Yields Multicalibration for Large Neural Networks. In Venkatesan Guruswami (ed.), *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*, volume 287 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 17:1–17:21, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-309-6. doi: 10.4230/LIPIcs.ITCS.2024.17. URL https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2024.17.

Evgenii Chzhen, Christophe Denis, and Mohamed Hebiri. Minimax semi-supervised set-valued approach to multi-class classification. *Bernoulli*, 27(4):2389–2412, 2021. doi: 10.3150/20-BEJ1313.

Nicolo Colombo and Vladimir Vovk. Training conformal predictors. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pp. 55–64. PMLR, 2020. URL https://proceedings.mlr.press/v128/colombo20a.html.

Christophe Denis and Mohamed Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(102):1–28, 2017. URL https://jmlr.org/papers/v18/16-596.html.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1095–1108. Association for Computing Machinery, 2021. doi: 10.1145/3406325.3451064. URL https://arxiv.org/abs/2011.13426.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman (ed.), *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 79:1–79:21, Dagstuhl, Germany, 2022a. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79.

Parikshit Gopalan, Michael P. Kim, Mihir A. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3193–3234. PMLR, 2022b. URL https://proceedings.mlr.press/v178/gopalan22a.html.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss Minimization Through the Lens Of Outcome Indistinguishability. In Yael Tauman Kalai (ed.), *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 60:1–60:20, Dagstuhl, Germany, 2023a. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.60. URL https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.60.

Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39936–39956. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7d693203215325902ff9dbdd067a50ac-Paper-Conference.pdf.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Lunjia Hu and Salil Vadhan. Generalized and unified equivalences between hardness and pseudoentropy. In *Theory of Cryptography: 23rd International Conference, TCC 2025, Proceedings, Part IV*, volume 16271 of *Lecture Notes in Computer Science*, pp. 258–288. Springer, 2025. doi: 10.1007/978-3-032-12290-2_9. URL https://arxiv.org/abs/2507.05972.

Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13497–13527. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hu23b.html.

Lunjia Hu, Kevin Tian, and Chutong Yang. Simultaneous Blackwell approachability and applications to multiclass omniprediction. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3593–3634. PMLR, 2026. URL https://proceedings.mlr.press/v336/hu26c.html.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36:26699–26721, 2023.

Xuesong Jia, Yuanjie Shi, Ziquan Liu, Yi Xu, and Yan Yan. Cost-sensitive conformal training with provably controllable learning bounds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22274–22282, 2026. doi: 10.1609/aaai.v40i27.39384.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 247–254, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314287. URL https://doi.org/10.1145/3306618.3314287.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b41907dd4df5c60f86216b73fe0c7465-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, (just-accepted):1–26, 2026.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.02544.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341. URL https://arxiv.org/abs/1609.00451.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Yuanjie Shi, Hooman Shahrokhi, Xuesong Jia, Xiongzhi Chen, Janardhan Rao Doppa, and Yan Yan. Direct prediction set minimization via bilevel conformal classifier training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55016–55045, 2025. URL https://proceedings.mlr.press/v267/shi25i.html.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity*, pp. 126–136. IEEE Computer Society, 2009. doi: 10.1109/CCC.2009.41. URL https://people.seas.harvard.edu/~salil/research/regularity-abs.html.

Vladimir Vovk, Ilia Nouretdinov, Valentina Fedorova, Ivan Petej, and Alex Gammerman. Criteria of efficiency for conformal prediction. *arXiv preprint arXiv:1603.04416*, 2016. URL https://arxiv.org/abs/1603.04416.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, 2 edition, 2022. doi: 10.1007/978-3-031-06649-8.

Wenbo Wang and Xingye Qiao. Set-valued support vector machine with bounded error rates. *Journal of the American Statistical Association*, 118(544):2847–2859, 2023. doi: 10.1080/01621459.2022.2089573.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

A Related work

Optimal prediction sets and plug-in guarantees. This work is related to the literature on characterizing the optimal form of prediction sets subject to coverage constraints. Sadinle et al. (2019) characterize minimum-expected-size sets under coverage constraints through thresholding conditional label probabilities and establish oracle convergence for plug-in and conformal methods. Vovk et al. (2016) study idealized conformal efficiency criteria and their relationship to probabilistic conformity measures. A complementary formulation fixes the expected set size and minimizes miscoverage, leading to probability-threshold oracles, surrogate-learning guarantees, and minimax convergence rates (Denis & Hebiri, 2017; Chzhen et al., 2021). Kiyani et al. (2024) develop Conformal Prediction with Length-Optimization (CPL), a minimax framework for minimizing expected prediction-set length subject to coverage constraints, including marginal and group-conditional coverage. Their population analysis establishes length optimality through strong duality under regularity assumptions. Building on these oracle and optimization perspectives, we complement these results with an explicit quantitative transfer from Brier regret to the excess size when thresholding a probabilistic predictor to a fixed marginal coverage.

Learning for conformal efficiency. How to leverage data to optimize efficiency is an important thread in conformal prediction. An early approach uses conformal efficiency criteria, such as observed fuzziness, as training objectives (Colombo & Vovk, 2020). A line of work explicitly sets the efficiency (size) of prediction sets as an empirical objective in learning the conformity score, including ConfTr which simulates conformal calibration during training using smooth quantiles and membership functions (Stutz et al., 2022), DPSM which formulates training as a bilevel problem to optimize the upper-level size objective (Shi et al., 2025), and CF-GNN which trains additional graph layers to improve prediction sets for graph neural networks (Huang et al., 2023). Jia et al. (2026) propose rank-weighted cross-entropy (RWCE), motivated by an upper bound on expected set size in terms of expected true-label rank and an additive term depending on coverage and calibration size. A complementary line of work studies model selection and optimization without sample splitting while preserving finite-sample coverage guarantees (Liang et al., 2026; Bai & Jin, 2024). Our focus is the transfer from ordinary proper-loss regret to excess calibrated set size; our experiments compare this approach with ConfTr.

Design of conformal scores. A complementary line of work designs conformity scores from a fitted probabilistic classifier and then applies conformal calibration. Adaptive prediction sets (APS) construct scores from cumulative predicted probabilities in descending label order, aiming to provide stronger adaptivity (Romano et al., 2020). Subsequent ideas include regularized adaptive prediction sets (RAPS) which adds a rank penalty to avoid exceedingly large sets due to tiny tail probabilities (Angelopoulos et al., 2021). These methods use plug-in probability estimates within different score constructions, with conformal calibration providing marginal coverage. Our work complements this perspective by fixing the score wrapper and quantifying how the quality of the underlying probability estimates affects the efficiency of the resulting calibrated sets.

Omniprediction. Omniprediction studies how a single predictor can support multiple downstream objectives through loss-specific post-processing. Gopalan et al. (2022a) show that multicalibrated predictors can compete with a hypothesis class simultaneously for a broad family of losses. Subsequent work develops this connection through loss outcome indistinguishability (Gopalan et al., 2023a), characterizes swap omniprediction through multicalibration (Gopalan et al., 2023b), and extends omniprediction to constrained optimization (Hu et al., 2023). Particularly relevant to our motivation, Błasiok et al. (2024) show that squared-loss minimization over neural networks implies multicalibration against smaller neural networks for all but a bounded number of exceptional network sizes. These results connect learning under one criterion to guarantees for other decision objectives. Our work establishes such a connection for conformal efficiency: small Brier or log-loss regret implies small excess expected set size after probability thresholding and calibration to a prescribed marginal coverage. The same probability predictor can serve different coverage targets by changing its calibration threshold. Our guarantee in Section 3 is relative to the Bayes optimum and depends on probability-estimation error; omniprediction seeks guarantees relative to a comparator class without requiring accurate estimation of the full conditional distribution. Our convex loss-minimization example in Appendix D clarifies why ordinary proper-loss minimization within a restricted class need not provide such a guarantee for conformal set size. Section 4 gives competitive set size for an arbitrary predictor, whether or not it belongs to the benchmark class, under multiaccuracy on

benchmark rules and a restricted calibration condition on its own rule. Local loss optimality supplies these conditions through specified post-processings of the learned predictor.

Surrogate losses. Surrogate risk minimization replaces a difficult target objective with a tractable training loss and asks whether small excess surrogate risk ensures small excess target risk. Classical results establish such transfers from convex surrogate losses to classification error (Bartlett et al., 2006). More directly, Wang & Qiao (2023) study set-valued classification under class-specific noncoverage constraints and bound excess ambiguity by excess risk under a truncated-hinge surrogate tailored to that problem. Proper losses elicit conditional label probabilities (Gneiting & Raftery, 2007), supporting downstream decisions through property elicitation (Agarwal & Agarwal, 2015) and quantitative control of probability-estimation error (Bao, 2023). Related work on predict-then-optimize constructs surrogates for decision error, including the convex SPO+ loss of Elmachtoub & Grigas (2022), while decision-focused learning differentiates through continuous relaxations of discrete optimization problems Wilder et al. (2019); Mandi et al. (2024) survey this broader approach. Our result identifies a standard probability-estimation loss, the multiclass Brier loss, as a consistent surrogate for expected THR set size subject to a marginal coverage constraint: its regret controls set-size regret at rate $O(K^{3/2}\alpha^{-1}\sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)})$. The target objective involves discrete membership decisions and a calibration threshold that changes with the predictor. Our analysis controls both effects without margin or density assumptions, with randomized ties at exact population coverage. This complements approaches that design a surrogate around the downstream optimization problem by establishing when ordinary probability estimation already suffices for conformal efficiency.

B Deferred results and proofs from Section 3

B.1 Idealized reduction

Theorem 1 (Core reduction from set size to ℓ_1 distance)**.** *Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$. Then following the notation* (2)*,* (3)*,* (11)*,*

$$\mathsf{SSR}_\alpha(q) \le \frac{K}{\alpha} \left\| q - p^\star \right\|_1 . \tag{12}$$

Proof. Throughout this proof, for each fixed pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let

$$h_q(x, y) \coloneqq \mathbb{P}[y \in S_{q,\alpha}(x)], \quad h_\star(x, y) \coloneqq \mathbb{P}[y \in S_{p^\star,\alpha}(x)],$$

where the probabilities are over the randomness used by $S_{q,\alpha}$ and $S_{p^\star,\alpha}$ (i.e., randomized thresholding to break ties at the boundary). For convenience we define

$$\Delta \coloneqq \mathbb{E}_{X\sim D_\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} (h_q(X, y) - h_\star(X, y)) \right] = \mathsf{SSR}_\alpha(q), \quad R \coloneqq \left\| q - p^\star \right\|_1 ,$$

so our goal is to bound $\Delta \le \frac{KR}{\alpha}$. Also let τ_q be the population quantile defining $S_{q,\alpha}$, so that $h_q(x, y) = 1$ if $q_y(x) > \tau_q$, and $h_q(x, y) = 0$ if $q_y(x) < \tau_q$. Because $h_\star(x, y) \in [0, 1]$,

$$\tau_q(h_q(x, y) - h_\star(x, y)) \le q_y(x)(h_q(x, y) - h_\star(x, y))$$

for every pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Summing this inequality over $\mathcal{Y}$ and taking expectation over $D_\mathcal{X}$,

$$\tau_q\Delta \le \mathbb{E}_{X\sim D_\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} q_y(X)(h_q(X, y) - h_\star(X, y)) \right] . \tag{15}$$

Further, recall that $\mathsf{mis}(S_{q,\alpha}) = \mathsf{mis}(S_{p^\star,\alpha}) = \alpha$ by construction, implying

$$\mathsf{mis}(S_{p^\star,\alpha}) - \mathsf{mis}(S_{q,\alpha}) = \mathbb{E}_{X\sim D_\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} p^\star_y(X)(h_q(X, y) - h_\star(X, y)) \right] = 0.$$

Combining the above two displays and using that $h_q(x, y) - h_\star(x, y) \in [-1, 1]$ pointwise,

$$\tau_q\Delta \le \mathbb{E}_{X\sim D_\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} (q_y(X) - p^\star_y(X))(h_q(X, y) - h_\star(X, y)) \right] \le R. \tag{16}$$

On the other hand, since $q_y(x) \leq \tau_q$ whenever $h_q(x, y) < 1$,

$$\begin{aligned} \alpha = \mathsf{mis}(S_{q,\alpha}) &= \mathbb{E}_{X \sim D_{\mathcal{X}}} \left[\sum_{y \in \mathcal{Y}} p^\star_y(X)(1 - h_q(X, y)) \right] \\ &\leq \mathbb{E}_{X \sim D_{\mathcal{X}}} \left[\sum_{y \in \mathcal{Y}} q_y(X)(1 - h_q(X, y)) \right] + R \\ &\leq \tau_q(K - \mathsf{size}_\alpha(q)) + R. \end{aligned} \tag{17}$$

Multiplying (17) by $\Delta \geq 0$ (Proposition 2), and substituting (16) then gives the conclusion:

$$\alpha\Delta \leq \tau_q \Delta(K - \mathsf{size}_\alpha(q)) + \Delta R \leq (K - \mathsf{size}_\alpha(q) + \Delta)R = (K - \mathsf{size}_\alpha(p^\star))R \leq KR.$$

□

We briefly comment on the structure of Theorem 1's proof. The first half establishes (16), which relates the set size regret to the ℓ_1 distance R, up to a factor of τ_q^{-1}. This can be interpreted as every extra included label according to q costing at least τ_q in probability mass.

At this point it is tempting to directly lower bound τ_q; unfortunately, it is simple to construct examples where τ_q is 0 (e.g., q predicts point masses everywhere). In (17), we are instead able to directly relate τ_q and R in a way that introduces a cancellation with our earlier bound (16), by leveraging the fact that $\mathsf{size}_\alpha(q) - \Delta = \mathsf{size}_\alpha(p^\star)$ is nonnegative.

B.2 Strongly convex losses

Lemma 1. *Let $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ be proper and μ-strongly convex in ℓ_1. Then following the notation* (5)*,* (11)*, for all predictors $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$, we have*

$$\mathsf{SSR}_\alpha(q) \leq \frac{K}{\alpha} \sqrt{\frac{2}{\mu} \mathsf{Reg}_\ell(q)}.$$

Proof. In light of Theorem 1 and $\mathbb{E}[Z] \leq \sqrt{\mathbb{E}Z^2}$ for random variables Z, it suffices to show

$$\sqrt{\mathbb{E}_{X \sim D_{\mathcal{X}}} \left[\|q(X) - p^\star(X)\|_1^2 \right]} \leq \sqrt{\frac{2}{\mu} \mathsf{Reg}_\ell(q)}.$$

To see this, applying Definitions 1 and 2 pointwise over $x \in \mathcal{X}$ gives

$$\frac{\mu}{2} \|q(x) - p^\star(x)\|_1^2 \leq \mathbb{E}_{Y \sim p^\star(x)} \left[\ell(q(x), Y) - \ell(p^\star(x), Y) \right].$$

Rearranging and taking expectations over $X \sim D_{\mathcal{X}}$ gives the claim. □

Corollary 1. *Let $q : \mathcal{X} \to \Delta(\mathcal{Y})$ and $\alpha \in (0, 1]$. Then following the notation* (2)*,* (6)*,*

$$\mathsf{SSR}_\alpha(q) \leq \frac{K^{3/2}}{\alpha} \sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)}, \quad \mathsf{SSR}_\alpha(q) \leq \frac{K}{\alpha} \sqrt{2\mathsf{Reg}_{\mathsf{log}}(q)}.$$

Proof. Recall the standard facts that the losses ℓ_{Brier} and ℓ_{log} in (4) are respectively $\mu = \frac{2}{K}$ and $\mu = 1$-strongly convex in ℓ_1. The former is due to an ℓ_1-to-ℓ_2 norm conversion, and the latter is Pinsker's inequality. Combining these facts with Lemma 1 gives Corollary 1. □

B.3 Tightness

We give a hard instance showing that the $\frac{K^{3/2}}{\alpha}$ blowup in Corollary 1 is tight up to a constant factor. Notably, this also shows that the $\frac{K}{\alpha}$ dependence of Theorem 1 is tight.

Proposition 3. *For all $\alpha \in (0, \frac{1}{4})$ and sufficiently large K, there is a distribution D over $\mathcal{X}, \mathcal{Y}$ and a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$ such that*

$$\mathsf{SSR}_\alpha(q) = \Omega\left(\frac{K^{3/2}}{\alpha} \sqrt{\mathsf{Reg}_{\mathsf{Brier}}(q)} \right).$$

Proof. Let $\mathcal{X}$ be a singleton, so we identify $p^\star$ and q with elements of $\Delta(\mathcal{Y})$. Set

$$n \coloneqq \left\lfloor \frac{K}{2} \right\rfloor, \quad m \coloneqq \left\lfloor \frac{K}{4} \right\rfloor,$$

partition the labels $\mathcal{Y}$ into $A \cup B \cup C \cup \{K\}$ where $|A| = n$ and $|B| = m$, and let

$$a \coloneqq \frac{\alpha}{n}, \quad b \coloneqq \frac{\alpha}{m}, \quad r \coloneqq 1 - 2\alpha, \quad t \coloneqq \frac{2\alpha}{m+n}.$$

Note that $a < t < b < r$. We now define our predictors:

$$p^\star_y = \begin{cases} a & y \in A \\ b & y \in B \\ 0 & y \in C \\ r & y = K \end{cases}, \quad q_y = \begin{cases} t & y \in A \cup B \\ 0 & y \in C \\ r & y = K \end{cases}.$$

The optimal set-valued predictor $S_{p^\star,\alpha}$ therefore accepts $B \cup \{K\}$, so $\mathsf{size}_\alpha(p^\star) = m + 1$.

On the other hand, all labels in $A \cup B$ are tied according to q, with a total mass of 2α. Thus, randomizing the acceptance to guarantee an α miscoverage level gives a $\frac{1}{2}$ chance of acceptance for each element, so $\mathsf{size}_\alpha(q) = 1 + \frac{m+n}{2}$ and $\mathsf{SSR}_\alpha(q) = \Omega(K)$.

Finally, since n, m, and $n - m$ are all $\Theta(K)$,

$$\mathsf{Reg}_{\mathsf{Brier}}(q) = \|q - p^\star\|_2^2 = n(t-a)^2 + m(t-b)^2 = \frac{\alpha^2(n-m)^2}{mn(m+n)} = \Theta\left(\frac{\alpha^2}{K}\right).$$

Together with $\mathsf{SSR}_\alpha(q) = \Omega(K)$, this proves the claimed lower bound. □

B.4 Finite-sample reduction

We give Theorem 3, a finite-sample analog of Theorem 1. Theorem 3 focuses on a slight variant of the split conformal prediction rule (9) to ease our analysis, but below we discuss how it extends to the more standard thresholding rule (9).

Fix a predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$ trained independently of a calibration dataset $B = \{(X_i, Y_i)\}_{i \in [n]}$ drawn i.i.d. from D. We denote the empirical CDF of the conformity score by

$$\widehat{F}_n(t) \coloneqq \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left(q_{Y_i}(X_i) \le t\right),$$

and the threshold to be

$$\widehat{\tau}_n \coloneqq \inf\left\{t : \widehat{F}_n(t) \ge \alpha\right\}.$$

Given features X, the set (8) includes all labels $y \in \mathcal{Y}$ with $q_y(X) > \widehat{\tau}_n$. To ensure that the empirical miscoverage is exactly α, we also accept y with $q_y(X) = \widehat{\tau}_n$ with probability $\widehat{\gamma}_n$. Here,

$$\widehat{\gamma}_n \coloneqq \frac{\widehat{F}_n(\widehat{\tau}_n) - \alpha}{\widehat{F}_n(\widehat{\tau}_n) - \widehat{F}_n(\widehat{\tau}_n^-)} \implies \widehat{\gamma}_n(\widehat{F}_n(\widehat{\tau}_n) - \widehat{F}_n(\widehat{\tau}_n^-)) + 1 - \widehat{F}_n(\widehat{\tau}_n) = 1 - \alpha,$$

where $\widehat{F}_n(\widehat{\tau}_n^-)$ is the empirical fraction of scores $q_{Y_i}(X_i) < \widehat{\tau}_n$. Thus, letting $\widehat{h}_n(x, y)$ denote the probability that (8) accepts label $y \in \mathcal{Y}$ for features $x \in \mathcal{X}$,

$$\widehat{h}_n(x, y) = \mathbb{I}\left(q_y(x) > \widehat{\tau}_n\right) + \widehat{\gamma}_n \mathbb{I}\left(q_y(x) = \widehat{\tau}_n\right). \tag{18}$$

Analogously to the notation in Appendix B.1, in the finite-sample setting we define $\widehat{S}_{q,\alpha,n}$ to be the set-valued predictor (8) using the acceptance rule (18). We also define

$$\widehat{\mathsf{size}}_{\alpha,n}(q; B) \coloneqq \mathsf{size}\left(\widehat{S}_{q,\alpha,n} \mid B\right),$$
$$\widehat{\mathsf{SSR}}_{\alpha,n}(q; B) \coloneqq \mathsf{SSR}_\alpha(\widehat{S}_{q,\alpha,n} \mid B),$$
$$\widehat{\mathsf{mis}}_{\alpha,n}(q; B) \coloneqq \mathsf{mis}\left(\widehat{S}_{q,\alpha,n} \mid B\right),$$

to specify the dependence on the calibration dataset B. All of these quantities take expectations over only the sample (X, Y) and threshold randomization, holding B fixed.

Theorem 3. *Let $\alpha, \delta, \varepsilon \in (0,1)$, and let $q : \mathcal{X} \to \Delta(\mathcal{Y})$. Then if $n \geq \frac{1}{2\varepsilon^2}\log(\frac{2}{\delta})$, with probability $\geq 1-\delta$ over $B = \{(X_i, Y_i)\}_{i\in[n]}$ drawn i.i.d. from D,*

$$\widehat{\mathsf{mis}}_{\alpha,n}(q; B) \leq \alpha + \varepsilon, \tag{19}$$

and

$$\widehat{\mathsf{SSR}}_{\alpha,n}(q; B) \leq \frac{2K}{\alpha}\left(\|q - p^\star\|_1 + \varepsilon\right) \leq \frac{2K}{\alpha}\left(\sqrt{\min\left\{K\mathsf{Reg}_{\mathsf{Brier}}(q), 2\mathsf{Reg}_{\mathsf{log}}(q)\right\}} + \varepsilon\right). \tag{20}$$

Proof. We first prove (19). Let $F(t) \coloneqq \mathbb{P}_{(X,Y)\sim D}[q_Y(X) \leq t]$ denote the population CDF of the conformity score. The variant of the Dvoretzky-Kiefer-Wolfowitz inequality shown in Corollary 1, Massart (1990) implies that with probability $\geq 1-\delta$ over the calibration dataset B,

$$\sup_t \left|\widehat{F}_n(t) - F(t)\right| \leq \varepsilon. \tag{21}$$

Condition on this event henceforth. By the acceptance rule (18),

$$\beta \coloneqq \widehat{\mathsf{mis}}_{\alpha,n}(q; B) = F(\widehat{\tau}_n^-) + (1-\widehat{\gamma}_n)(F(\widehat{\tau}_n) - F(\widehat{\tau}_n^-)).$$

On the other hand, $\widehat{\gamma}_n$ was chosen to ensure empirical miscoverage exactly α, i.e.,

$$\alpha = \widehat{F}_n(\widehat{\tau}_n^-) + (1-\widehat{\gamma}_n)(\widehat{F}_n(\widehat{\tau}_n) - \widehat{F}_n(\widehat{\tau}_n^-)).$$

Since $\widehat{\gamma}_n \in [0,1]$, subtracting the above two displays and using (21) implies (19):

$$|\beta - \alpha| = \left|(1-\widehat{\gamma}_n)(F(\widehat{\tau}_n) - \widehat{F}_n(\widehat{\tau}_n)) + \widehat{\gamma}_n(F(\widehat{\tau}_n^-) - \widehat{F}_n(\widehat{\tau}_n^-))\right| \leq \varepsilon. \tag{22}$$

We move on to proving (20). The second inequality follows immediately from strong convexity of ℓ_{Brier} and ℓ_{log}, in exactly the same way as shown in Lemma 1, so we focus on the first.

To prove the first inequality in (20), we give a robust version of Theorem 1. We follow its notation, and let $R \coloneqq \|q - p^\star\|_1$, $h_\star(x,y) \coloneqq \mathbb{P}[y \in S_{p^\star,\alpha}(x)]$, and

$$\Delta \coloneqq \mathbb{E}_{X\sim D_{\mathcal{X}}}\left[\sum_{y\in\mathcal{Y}}\left(\widehat{h}_n(X,y) - h_\star(X,y)\right)\right] = \widehat{\mathsf{SSR}}_{\alpha,n}(q; B).$$

Then, exactly following the derivation of (16) implies

$$\widehat{\tau}_n\Delta \leq \mathbb{E}_{X\sim D_{\mathcal{X}}}\left[\sum_{y\in\mathcal{Y}} q_y(X)\left(\widehat{h}_n(X,y) - h_\star(X,y)\right)\right].$$

On the other hand, expanding the difference of the population miscoverages of q and $p^\star$ gives

$$\alpha - \beta = \mathbb{E}_{X\sim D_{\mathcal{X}}}\left[\sum_{y\in\mathcal{Y}} p^\star_y(X)\left(\widehat{h}_n(X,y) - h_\star(X,y)\right)\right].$$

Summing these displays and using $|\beta - \alpha| \leq \varepsilon$ gives our robust version of (16),

$$\widehat{\tau}_n\Delta \leq R + \varepsilon. \tag{23}$$

Next, exactly following the derivation of (17) implies

$$\beta \leq \widehat{\tau}_n K + R \implies \widehat{\tau}_n \geq \frac{\alpha - \varepsilon - R}{K}.$$

Now if $R + \varepsilon \leq \frac{\alpha}{2}$, we have $\widehat{\tau}_n \geq \frac{\alpha}{2K}$ and plugging this into (23) gives the first inequality in (20). Otherwise, the trivial bound $\Delta \leq K$ again implies this inequality, because

$$\Delta \leq K < \frac{2K}{\alpha}(R + \varepsilon).$$

□

Standard split conformal threshold. Theorem 3 applies to the finite-sample thresholding rule (18), which is theoretically convenient due to its empirical miscoverage level of exactly α. However, it also extends straightforwardly to the more standard split conformal prediction setting (9), up to a constant factor, under the relatively mild assumption that (over $(X, Y) \sim D$) $q_Y(X)$ has a continuous CDF.

Let $j = \lfloor (n+1)\alpha \rfloor$, and let β denote the population miscoverage under (9). If $j = 0$, the sentinel convention gives $\tau = -\infty$, so all labels are included and $\beta = 0$. In this case $\alpha < 1/(n+1) < 1/n$, hence $|\beta - \alpha| < 1/n$. The size bound with $\varepsilon + 1/n$ in place of ε also holds directly: the excess size is at most K, whereas the right-hand side of the first inequality in (20) is at least $2K/(\alpha n) > K$.

Now suppose $j \geq 1$. The threshold τ is the jth smallest calibration score. Since the score CDF is continuous, we have $\beta = F(\tau)$ and $\widehat{F}_n(\tau) = j/n$ almost surely. Thus, under (21),

$$|\beta - \alpha| \leq |F(\tau) - \widehat{F}_n(\tau)| + \left| \frac{j}{n} - \alpha \right| \leq \varepsilon + \frac{1}{n}.$$

Leaving the rest of the proof unchanged, the conclusion of Theorem 3 then holds with $\varepsilon + \frac{1}{n} = O(\varepsilon)$ in place of ε, i.e., the same up to a low-order term.

C Deferred proofs from Section 4

C.1 Set size regret via indistinguishability

Theorem 2. *For every distribution D, nonempty benchmark class $\mathcal{F}$, arbitrary predictor $q : \mathcal{X} \to \Delta(\mathcal{Y})$, and $\alpha \in (0, 1)$, let $\varepsilon_1, \varepsilon_2 \in \mathbb{R}$ satisfy $e_q(h_q) \leq \varepsilon_1$, and $\sup_{f \in \mathcal{F}} \big(-e_q(h_f)\big) \leq \varepsilon_2$. Defining $(t)_+ := \max\{t, 0\}$, the set size regret $\mathsf{SSR}_{\mathcal{F},\alpha}(q)$ satisfies*

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{(\varepsilon_1 + \varepsilon_2)_+}{\alpha + \varepsilon_2}\right\}. \tag{13}$$

When $\alpha + \varepsilon_2 = 0$, the fraction is interpreted as 1, giving the trivial bound $\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq K - \mathsf{size}_{\mathcal{F},\alpha}$.

Proof. The proof generalizes Theorem 1, and follows the same two-step argument. First, exactly the same derivation as used in (15), with a predictor $f \in \mathcal{F}$ in place of $p^\star$, shows that

$$\tau_q \left(\mathsf{size}_\alpha(q) - \mathsf{size}_\alpha(f)\right) \leq \mathbb{E}_{X \sim D_\mathcal{X}} \langle q(X), h_q(X) - h_f(X) \rangle = e_q(h_q) - e_q(h_f) \leq \varepsilon_1 + \varepsilon_2.$$

Here we used $\mathbb{E}_{X \sim D_\mathcal{X}} \langle p^\star(X), h_q(X) \rangle = \mathbb{E}_{X \sim D_\mathcal{X}} \langle p^\star(X), h_f(X) \rangle = 1 - \alpha$. Supremizing over $\mathcal{F}$ gives

$$\tau_q \mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq \varepsilon_1 + \varepsilon_2. \tag{24}$$

Next, letting $\mathbf{1}_\mathcal{Y}$ denote the all-ones vector in $\mathbb{R}^\mathcal{Y}$,

$$\begin{aligned} \alpha &= \mathbb{E}_{X \sim D_\mathcal{X}} \langle p^\star(X), \mathbf{1}_\mathcal{Y} - h_q(X) \rangle \\ &= \mathbb{E}_{X \sim D_\mathcal{X}} \langle q(X), \mathbf{1}_\mathcal{Y} - h_q(X) \rangle + e_q(h_q) \\ &\leq \tau_q (K - \mathsf{size}_\alpha(q)) + \varepsilon_1, \end{aligned} \tag{25}$$

where the first line used that prediction with probability conformity scores given by q has miscoverage α, the second used the definition of e_q and that $e_q(\mathbf{1}_\mathcal{Y}) = 0$, and the third line used the same argument as in (17), namely, that $q_y(x)(1 - h_q(x, y)) \leq \tau_q(1 - h_q(x, y))$ pointwise, so

$$\begin{aligned} \mathbb{E}_{X \sim D_\mathcal{X}} \langle q(X), \mathbf{1}_\mathcal{Y} - h_q(X) \rangle &\leq \tau_q \mathbb{E}_{X \sim D_\mathcal{X}} \langle \mathbf{1}_\mathcal{Y}, \mathbf{1}_\mathcal{Y} - h_q(X) \rangle \\ &= \tau_q (K - \mathsf{size}_\alpha(q)). \end{aligned}$$

We now restrict to the setting where $\mathsf{SSR}_{\mathcal{F},\alpha}(q) \geq 0$, $\varepsilon_1 + \varepsilon_2 \geq 0$. Indeed, if $\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq 0$, then (13) is immediate. If $\varepsilon_1 + \varepsilon_2 < 0$, then (24) implies $\mathsf{SSR}_{\mathcal{F},\alpha}(q) < 0$, and thus (13) again follows.

Multiplying (25) by $\mathsf{SSR}_{\mathcal{F},\alpha}(q)$ and substituting (24) then gives

$$\begin{aligned} \alpha \mathsf{SSR}_{\mathcal{F},\alpha}(q) &\leq (\varepsilon_1 + \varepsilon_2)(K - \mathsf{size}_\alpha(q)) + \varepsilon_1 \mathsf{SSR}_{\mathcal{F},\alpha}(q) \\ &= (\varepsilon_1 + \varepsilon_2)(K - \mathsf{size}_{\mathcal{F},\alpha}) - \varepsilon_2 \mathsf{SSR}_{\mathcal{F},\alpha}(q). \end{aligned} \tag{26}$$

Moreover, for any $f \in \mathcal{F}$, because h_f achieves population coverage $1 - \alpha$,

$$-e_q(h_f) = \mathbb{E}_{X \sim D_\mathcal{X}} \langle p^\star(X) - q(X), h_f(X) \rangle = (1 - \alpha) - \mathbb{E} \langle q(X), h_f(X) \rangle \geq -\alpha,$$

so $\alpha + \varepsilon_2 \geq 0$. If $\alpha + \varepsilon_2 = 0$, then we use the trivial bound $\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq K - \mathsf{size}_{\mathcal{F},\alpha}$ to show (13), because $\mathsf{size}_\alpha(q) \leq K$. Otherwise, rearranging (26) and dividing by $\alpha + \varepsilon_2 > 0$ gives (13). □

C.2 Recovering the global guarantee

We provide a simple derivation that shows how Theorem 2 strictly generalizes Theorem 1. Take $\mathcal{F} = \{p^\star\}$ so that Proposition 2 implies

$$\mathsf{size}_{\mathcal{F},\alpha} = \mathsf{size}^\star_\alpha, \quad \mathsf{SSR}_{\mathcal{F},\alpha}(q) = \mathsf{SSR}_\alpha(q).$$

Moreover, let $\varepsilon_1 := e_q(h_q)$ and $\varepsilon_2 := \max\{0, -e_q(h_\star)\}$. Since every coordinate of h_q and $h_q - h_\star$ lies in $[-1, 1]$, Hölder's inequality implies

$$\varepsilon_1 + \varepsilon_2 = \max\left\{e_q(h_q), e_q(h_q) - e_q(h_\star)\right\} \le \|q - p^\star\|_1 .$$

Substituting these bounds into Theorem 2 gives $\mathsf{SSR}_\alpha(q) \le \frac{K}{\alpha}\|q - p^\star\|_1$, i.e., Theorem 1.

C.3 Finite-sample extension

We extend Theorem 2 to the finite-sample setting of Appendix B.4. For the calibration dataset $B = \{(X_i, Y_i)\}_{i\in[n]}$, let $\widehat{h}_{f,n}(x, y) = \mathbb{P}[y \in \widehat{S}_{f,\alpha,n}(x) \mid B]$ denote the empirical inclusion rule for predictor f. Thus $\widehat{h}_{q,n}$ is the rule $\widehat{h}_n$ in (18); write $\widehat{\tau}_n$ for its empirical threshold. Define

$$\widehat{\mathsf{size}}_{\mathcal{F},\alpha,n}(B) := \inf_{f\in\mathcal{F}} \widehat{\mathsf{size}}_{\alpha,n}(f; B), \qquad \widehat{\mathsf{SSR}}_{\mathcal{F},\alpha,n}(q; B) := \widehat{\mathsf{size}}_{\alpha,n}(q; B) - \widehat{\mathsf{size}}_{\mathcal{F},\alpha,n}(B).$$

The conclusion $\widehat{\mathsf{mis}}_{\alpha,n}(q; B) \le \alpha + \varepsilon$ in (19) is unchanged. For the class-relative analog of (20), suppose that

$$\left|\widehat{\mathsf{mis}}_{\alpha,n}(f; B) - \alpha\right| \le \varepsilon_{\mathcal{F}} \quad \text{for all } f \in \mathcal{F}.$$

For finite $\mathcal{F}$, we can obtain a uniform bound $\varepsilon_{\mathcal{F}}$ by taking a union bound over (21); otherwise, a structural assumption (e.g., appropriately bounded VC dimension) is typically required. Choose $\varepsilon_1 \ge e_q(\widehat{h}_{q,n})$ and $\varepsilon_2 \ge \sup_{f\in\mathcal{F}}(-e_q(\widehat{h}_{f,n}))$. Accounting for deviations in miscoverage, (24) becomes

$$\widehat{\tau}_n \widehat{\mathsf{SSR}}_{\mathcal{F},\alpha,n}(q; B) \le \varepsilon_1 + \varepsilon_2 + \varepsilon + \varepsilon_{\mathcal{F}},$$

and (25) becomes

$$\alpha \le \widehat{\tau}_n \left(K - \widehat{\mathsf{size}}_{\alpha,n}(q; B)\right) + \varepsilon_1 + \varepsilon.$$

Combining these inequalities as in Theorem 2 gives

$$\widehat{\mathsf{SSR}}_{\mathcal{F},\alpha,n}(q; B) \le \left(K - \widehat{\mathsf{size}}_{\mathcal{F},\alpha,n}(B)\right) \min\left\{1, \frac{(\varepsilon_1 + \varepsilon_2 + \varepsilon + \varepsilon_{\mathcal{F}})_+}{\alpha + \varepsilon_2 + \varepsilon_{\mathcal{F}}}\right\},$$

with the same zero-denominator convention as in Theorem 2. This argument also extends to the split conformal rule (9) up to an additional low-order $1/n$ term, as in Appendix B.4.

C.4 Consequences for Brier loss and log loss

In this subsection, when ℓ is a loss and $q : \mathcal{X} \to \Delta(\mathcal{Y})$ is a predictor, we use the following shorthand notation for the population risk:

$$R_\ell(q) := \mathbb{E}_{(X,Y)\sim D}\left[\ell(q(X), Y)\right].$$

As with the regret notation in Section 2, we abbreviate $R_{\ell_{\mathsf{Brier}}}$ and $R_{\ell_{\mathsf{log}}}$ by R_{Brier} and R_{log}, respectively.

Corollary 2. *Let $\Pi_{\Delta(\mathcal{Y})}$ denote Euclidean projection onto the simplex, and define*

$$T^{\mathsf{Brier}}_{\lambda,h}(q)(x) := \Pi_{\Delta(\mathcal{Y})}(q(x) + \lambda h(x)).$$

For an arbitrary $q : \mathcal{X} \to \Delta(\mathcal{Y})$, use the post-processing family

$$\mathcal{C}_{\mathsf{Brier}}(q, \mathcal{F}, \alpha) := \{T^{\mathsf{Brier}}_{-s,h_q}(q) : s \ge 0\} \cup \{T^{\mathsf{Brier}}_{s,h_f}(q) : f \in \mathcal{F},\ s \ge 0\}.$$

If $\mathsf{Reg}_{\mathsf{Brier}}(q; \mathcal{C}_{\mathsf{Brier}}(q, \mathcal{F}, \alpha)) \le \varepsilon$ for $\varepsilon \ge 0$, then

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \le (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{\sqrt{K\varepsilon}}{\alpha + \frac{1}{2}\sqrt{K\varepsilon}}\right\}.$$

Proof. Write $\mathcal{C} = \mathcal{C}_{\mathsf{Brier}}(q, \mathcal{F}, \alpha)$. The local optimality assumption and (14) imply

$$R_{\mathsf{Brier}}(r) - R_{\mathsf{Brier}}(q) \geq -\mathsf{Reg}_{\mathsf{Brier}}(q; \mathcal{C}) \geq -\varepsilon \qquad (r \in \mathcal{C}).$$

Fix any $f \in \mathcal{F} \cup \{q\}$, write $h = h_f$, and center the coordinates of this distinguisher at each input:

$$\overline{h}(x) := K^{-1} \sum_y h(x)_y, \qquad h^\circ(x) := h(x) - \overline{h}(x)\mathbf{1}_{\mathcal{Y}}.$$

Because $h(x)_y \in [0,1]$, its coordinate variance is at most $1/4$, so $\|h^\circ(x)\|_2^2 \leq K/4$. Projection onto the simplex is unchanged by adding a multiple of $\mathbf{1}_{\mathcal{Y}}$ to its argument. Also, for every $v \in \mathbb{R}^K$ and $p^\star(x) \in \Delta(\mathcal{Y})$, $\|\Pi_{\Delta(\mathcal{Y})}(v) - p^\star(x)\|_2 \leq \|v - p^\star(x)\|_2$. For any $\lambda \in \mathbb{R}$, the Brier decomposition gives

$$\begin{aligned} R_{\mathsf{Brier}}(T^{\mathsf{Brier}}_{\lambda,h}(q)) - R_{\mathsf{Brier}}(q) &\leq 2\lambda \mathbb{E}_{X \sim D_{\mathcal{X}}} \langle q(X) - p^\star(X), h^\circ(X) \rangle + \lambda^2 \mathbb{E}_{X \sim D_{\mathcal{X}}} \|h^\circ(X)\|_2^2 \\ &\leq 2\lambda e_q(h) + \frac{K}{4}\lambda^2. \end{aligned}$$

The equality of the residual correlations uses $\langle q(x) - p^\star(x), \mathbf{1}_{\mathcal{Y}} \rangle = 0$. The left-hand side is at least $-\varepsilon$ whenever $T^{\mathsf{Brier}}_{\lambda,h}(q) \in \mathcal{C}$. If $e_q(h_q) > 0$, take $h = h_q$ and $\lambda = -4e_q(h_q)/K < 0$. This post-processing belongs to $\mathcal{C}$ and gives $4e_q(h_q)^2/K \leq \varepsilon$. If $e_q(h_q) \leq 0$, an upper bound is automatic. Similarly, for each $f \in \mathcal{F}$ with $e_q(h_f) < 0$, take $h = h_f$ and $\lambda = -4e_q(h_f)/K > 0$, so that the post-processing again belongs to $\mathcal{C}$. It follows that

$$e_q(h_q) \leq \beta := \tfrac{1}{2}\sqrt{K\varepsilon}, \qquad e_q(h_f) \geq -\beta \quad (f \in \mathcal{F}).$$

Applying (13) with $\varepsilon_1 = \varepsilon_2 = \beta$ gives the claimed

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{2\beta}{\alpha + \beta}\right\}.$$

□

Corollary 3. *For any $q : \mathcal{X} \to \Delta(\mathcal{Y})$, inclusion vector h, and $\lambda \in \mathbb{R}$, define*

$$T^{\mathsf{log}}_{\lambda,h}(q)(x)_y := \frac{q_y(x)\exp(\lambda h(x)_y)}{\sum_j q_j(x)\exp(\lambda h(x)_j)}. \tag{27}$$

Use the post-processing family

$$\mathcal{C}_{\mathsf{log}}(q, \mathcal{F}, \alpha) := \{T^{\mathsf{log}}_{-s,h_q}(q) : s \geq 0\} \cup \{T^{\mathsf{log}}_{s,h_f}(q) : f \in \mathcal{F},\ s \geq 0\}.$$

If $\mathbb{E}_{(X,Y)\sim D}\ell_{\mathsf{log}}(q(X), Y) < \infty$ and $\mathsf{Reg}_{\mathsf{log}}(q; \mathcal{C}_{\mathsf{log}}(q, \mathcal{F}, \alpha)) \leq \varepsilon$ for $\varepsilon \geq 0$, then

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{\sqrt{2\varepsilon}}{\alpha + \sqrt{\varepsilon/2}}\right\}.$$

Proof. Write $\mathcal{C} = \mathcal{C}_{\mathsf{log}}(q, \mathcal{F}, \alpha)$. As in the Brier case, the local optimality assumption gives

$$R_{\mathsf{log}}(r) - R_{\mathsf{log}}(q) \geq -\mathsf{Reg}_{\mathsf{log}}(q; \mathcal{C}) \geq -\varepsilon \qquad (r \in \mathcal{C}).$$

Fix any $f \in \mathcal{F} \cup \{q\}$, write $h = h_f$, and let $\lambda \in \mathbb{R}$. The normalizing denominator in (27) is positive. Finite expected log loss ensures $q_Y(X) > 0$ almost surely, so

$$\begin{aligned} R_{\mathsf{log}}(T^{\mathsf{log}}_{\lambda,h}(q)) - R_{\mathsf{log}}(q) &= \mathbb{E}_{X \sim D_{\mathcal{X}}} \log\left(\sum_y q_y(X) e^{\lambda h(X,y)}\right) - \lambda \mathbb{E}_{(X,Y)\sim D} h(X,Y) \\ &\leq \lambda e_q(h) + \frac{\lambda^2}{8}. \end{aligned}$$

The inequality is Hoeffding's lemma applied, conditionally on X, to the random value $h(X, J) \in [0,1]$ with $J \sim q(X)$. For the post-processings in $\mathcal{C}$, this loss difference is at least $-\varepsilon$. If $e_q(h_q) > 0$, choose $h = h_q$ and $\lambda = -4e_q(h_q) < 0$ to obtain $2e_q(h_q)^2 \leq \varepsilon$. If $e_q(h_f) < 0$ for a benchmark $f \in \mathcal{F}$, choose $h = h_f$ and $\lambda = -4e_q(h_f) > 0$ to obtain $2e_q(h_f)^2 \leq \varepsilon$. For the other signs, the corresponding one-sided bounds hold automatically. Thus

$$e_q(h_q) \leq \beta := \sqrt{\varepsilon/2}, \qquad e_q(h_f) \geq -\beta \quad (f \in \mathcal{F}).$$

Applying (13) with $\varepsilon_1 = \varepsilon_2 = \beta$ yields the claimed

$$\mathsf{SSR}_{\mathcal{F},\alpha}(q) \leq (K - \mathsf{size}_{\mathcal{F},\alpha}) \min\left\{1, \frac{2\beta}{\alpha + \beta}\right\}.$$

□

D FAILURE OF CONVEX LOSS MINIMIZATION UNDER MISSPECIFICATION

In contrast to our results in Section 3 where *global* loss minimization implies small set size, we present examples where minimizing a convex loss among a *restricted* family of models may result in a highly sub-optimal set size. These examples hold for affine score models with *admissible* convex losses and corresponding link functions. We verify these conditions for standard linear regression with simplex projection, multinomial logistic regression, and coordinatewise Bernoulli and Poisson generalized linear models (GLMs) with normalized predicted means.

We distinguish the scores used for training from the probabilities used for calibration. Let

$$u = \frac{1}{K}\mathbf{1}_{\mathcal{Y}}, \qquad H = \{x \in \mathbb{R}^K : \textstyle\sum_j x_j = 0\}.$$

An orthonormal coordinate system identifies H with the ordinary feature space $\mathbb{R}^{K-1}$. For an affine score function $z_{W,b}(x) = b + Wx \in \mathbb{R}^K$, train by minimizing

$$R(W, b) := \mathbb{E}_{(X,Y)\sim D}\phi_Y(b + WX),$$

where ϕ_y is the loss when the observed label is y. We use *link function* for a map $\Psi : \mathbb{R}^K \to \Delta(\mathcal{Y})$ from scores to probabilities. Write

$$q_{W,b}(x) = \Psi(b + Wx), \qquad \mathcal{Q}_\Psi = \{q_{W,b} : W : H \to \mathbb{R}^K \text{ linear}, \ b \in \mathbb{R}^K\}.$$

The convexity assumption concerns the training loss ϕ_y as a function of the scores. The definition below collects the conditions used in our construction.

Definition 3 (Admissible convex loss and link function)**.** A family of finite score losses $\phi = (\phi_y : \mathbb{R}^K \to \mathbb{R})_{y\in\mathcal{Y}}$ is *admissible* if each ϕ_y is convex and the balanced risk

$$\overline{\phi}(z) := \frac{1}{K}\sum_{y\in\mathcal{Y}} \phi_y(z)$$

attains its minimum at $z = \beta\mathbf{1}_{\mathcal{Y}}$ for some $\beta \in \mathbb{R}$. For this choice of β, a link function $\Psi : \mathbb{R}^K \to \Delta(\mathcal{Y})$ is *admissible* if it satisfies:

1. **Symmetry and preservation of order.** For every permutation matrix P and score vector z,
$$\Psi(Pz) = P\Psi(z), \qquad z_i \le z_j \implies \Psi_i(z) \le \Psi_j(z).$$
2. **Confident predictions along class directions.** For every $y \in [K]$,
$$\lim_{\lambda\to\infty} \Psi_y\big(\beta\mathbf{1}_{\mathcal{Y}} + \lambda(e_y - u)\big) = 1.$$

Symmetry implies $\Psi(\beta\mathbf{1}_{\mathcal{Y}}) = u$. The link may return boundary points of the simplex; this allows Euclidean projection as well as softmax. As verified below, admissible choices include squared loss with simplex projection, multinomial log loss with softmax, and Bernoulli and Poisson GLM losses with normalized mean predictions.

Theorem 4. *Fix $K \ge 3$, $\alpha \in (0, 1)$, and $0 < p < \alpha$. Let ϕ be an admissible convex loss with balanced minimizer $\beta\mathbf{1}_{\mathcal{Y}}$, and let Ψ be an admissible link for this choice of β, as in Definition 3. There is a distribution D, depending only on K and p, with $2K$ feature values and deterministic labels such that $(W, b) = (0, \beta\mathbf{1}_{\mathcal{Y}})$ globally minimizes $R(W, b)$ and induces the uniform predictor $q_{\mathrm{cvx}}(x) = u$. Nevertheless, there is $q_{\mathrm{alt}} \in \mathcal{Q}_\Psi$ with*

$$\mathsf{size}_\alpha(q_{\mathrm{cvx}}) = K(1 - \alpha), \qquad \mathsf{size}_\alpha(q_{\mathrm{alt}}) = \mathsf{size}^\star_\alpha = 1 - \alpha. \tag{28}$$

Consequently,

$$\mathsf{size}_\alpha(q_{\mathrm{cvx}}) - \inf_{q\in\mathcal{Q}_\Psi} \mathsf{size}_\alpha(q) = (K - 1)(1 - \alpha). \tag{29}$$

If each ϕ_y is strictly convex between any two score vectors whose difference is not a multiple of $\mathbf{1}_{\mathcal{Y}}$, then every global training minimizer induces q_{cvx}.

These conclusions persist after adding any finite nonnegative parameter penalty that vanishes at $(0, \beta\mathbf{1}_{\mathcal{Y}})$, including an L^2 penalty on W. For $p = 1/m < \alpha$, $m \ge 2$, they also hold for empirical risk minimization on an explicit dataset of Km examples.

To prove Theorem 4, the key observation is that if $\mathbb{E}[X \mid Y] = 0$, conditional Jensen's inequality implies that a constant predictor attains minimum expected loss over the affine class. An admissible link allows another non-constant affine model to produce much smaller calibrated sets. Throughout this appendix, set sizes use exact population calibration at target miscoverage α, with a common inclusion probability for labels tied at the threshold.

Proof. We first construct the distribution and use convexity to identify a training minimizer. We then show how the link allows a second model to attain the optimal calibrated size.

Step 1: Make every class have mean-zero features. For each label y, let

$$v_y = e_y - u, \qquad c = \frac{1-p}{p} > 0.$$

Draw Y uniformly from $[K]$, and conditionally on $Y = y$ set

$$X = \begin{cases} v_y, & \text{with probability } 1-p, \\ -cv_y, & \text{with probability } p. \end{cases}$$

Since $K \geq 3$, no vertex v_y is a negative multiple of another. These $2K$ feature values are therefore distinct, making the label a deterministic function of the feature. The two feature values for a fixed label lie on opposite sides of the origin, with their magnitudes chosen so that

$$\mathbb{E}[X \mid Y = y] = (1-p)v_y - pcv_y = 0 \quad \text{for every } y.$$

For small p, the rare examples have large feature magnitude and cancel the conditional means of the more frequent examples.

Step 2: Apply conditional Jensen's inequality. Fix any affine score function $b + Wx$. Conditional on $Y = y$, its mean score vector is b. Convexity therefore gives

$$\mathbb{E}[\phi_y(b + WX) \mid Y = y] \geq \phi_y\big(b + W\mathbb{E}[X \mid Y = y]\big) = \phi_y(b).$$

Averaging over the balanced labels yields

$$R(W, b) \geq \overline{\phi}(b) \geq \overline{\phi}(\beta \mathbf{1}_\mathcal{Y}) = R(0, \beta \mathbf{1}_\mathcal{Y}). \tag{30}$$

Thus the constant scores $\beta \mathbf{1}_\mathcal{Y}$ are globally optimal among all affine score functions. Symmetry of Ψ implies $\Psi(\beta \mathbf{1}_\mathcal{Y}) = u$, so this minimizer predicts uniformly. More generally, the same argument gives $R(W, b) \geq \mathbb{E}_Y \phi_Y(b)$ for any distribution with $\mathbb{E}[X \mid Y] = 0$, independently of the link.

Step 3: When strict convexity holds, every minimizer is uniform. Suppose the additional strict-convexity condition holds and (W, b) is a global minimizer. Both inequalities in (30) must be equalities. For a fixed label y, the two score vectors in Step 2 differ by $(1 + c)Wv_y$. If Wv_y were not a multiple of $\mathbf{1}_\mathcal{Y}$, strict convexity would make the conditional Jensen inequality strict. Hence $Wv_y \in \operatorname{span}\{\mathbf{1}_\mathcal{Y}\}$ for every y. Since the v_y span H, the same holds for Wx at every $x \in H$.

Also, $\overline{\phi}(b) = \overline{\phi}(\beta \mathbf{1}_\mathcal{Y})$ forces $b \in \operatorname{span}\{\mathbf{1}_\mathcal{Y}\}$. Otherwise strict convexity at their midpoint would give a value below the minimum of $\overline{\phi}$. Thus every score vector $b + Wx$ has equal coordinates. Symmetry of Ψ maps each such vector to u. This proves uniqueness of the minimizing probability predictor, even when the score parameters have a common-coordinate redundancy.

Step 4: Use another affine model to attain the optimal size. Consider the affine scores

$$z_{\text{alt}}(x) = \beta \mathbf{1}_\mathcal{Y} + \lambda x.$$

At $x = v_y$, symmetry makes the true-label probability

$$T_\lambda := \Psi_y(\beta \mathbf{1}_\mathcal{Y} + \lambda v_y)$$

the same for every y, and makes all incorrect-label probabilities equal to $(1 - T_\lambda)/(K-1)$. The confidence assumption gives $T_\lambda \to 1$. Since $K \geq 3$, choose a finite $\lambda > 0$ with

$$T_\lambda > \frac{1}{K-1} > \frac{1}{K}.$$

Every incorrect-label probability at these vertices is then below T_λ.

At an opposite point $x = -cv_y$, the true-label score is smaller than all incorrect-label scores, which are equal to one another. Admissibility of Ψ therefore gives a true-label probability $r_\lambda \le 1/K$ and equal incorrect-label probabilities

$$\frac{1 - r_\lambda}{K - 1} \le \frac{1}{K - 1} < T_\lambda.$$

Again these values are the same for every y. Consequently, $q_{\mathrm{alt}} = \Psi \circ z_{\mathrm{alt}}$ has true-label score $r_\lambda < T_\lambda$ on mass $p < \alpha$ and score T_λ on the remaining mass. The calibration threshold and its tie-inclusion probability are therefore

$$\tau_{q_{\mathrm{alt}}} = T_\lambda, \qquad \gamma_{q_{\mathrm{alt}}} = \frac{1 - \alpha}{1 - p}.$$

All incorrect labels are excluded. The expected size is exactly the true-label inclusion probability $1 - \alpha$, which is the smallest possible size at that coverage.

For the uniform predictor, calibration includes every label with probability $1-\alpha$, giving size $K(1-\alpha)$. This proves (28) and (29).

Step 5: Include regularization and empirical risk minimization. A nonnegative penalty vanishing at $(0, \beta \mathbf{1}_\mathcal{Y})$ preserves that minimizer. Moreover, any penalized global minimizer must also minimize the unpenalized risk, so the uniqueness conclusion continues to hold whenever the strict-convexity condition holds.

For $p = 1/m < \alpha$, form a dataset with $m - 1$ copies of (v_y, y) and one copy of $(-(m-1)v_y, y)$ for each class y. Its Km examples have exactly the distribution used above. The empirical objective therefore equals $R(W, b)$, and all conclusions follow. □

Linear regression with squared loss and simplex projection. Take squared loss on one-hot labels and the Euclidean projection link:

$$\phi_y(z) = \|z - e_y\|_2^2, \qquad \Psi(z) = \Pi_{\Delta(\mathcal{Y})}(z) := \arg\min_{r \in \Delta(\mathcal{Y})} \|r - z\|_2^2.$$

Here $R(W, b)$ is the ordinary least-squares objective for an affine model. The training loss is evaluated on the raw affine outputs; projection is used afterward to obtain probability conformity scores. Each ϕ_y is strictly convex, and

$$\overline{\phi}(z) = 1 - \frac{1}{K} + \|z - u\|_2^2,$$

so the loss is admissible with $\beta = 1/K$.

The projection is symmetric because permuting coordinates preserves both the simplex and Euclidean distance. Its coordinates have the form

$$\Psi_j(z) = \max\{z_j - \theta, 0\}, \qquad \sum_j \max\{z_j - \theta, 0\} = 1,$$

which also shows that it preserves coordinate order. Finally, for $\lambda \ge 1$,

$$\Psi\big(u + \lambda(e_y - u)\big) = e_y.$$

Indeed, in the projection formula one may take $\theta = (\lambda - 1)(K - 1)/K$; the projected y-coordinate is one and all other coordinates are zero. Thus the link is admissible.

Theorem 4 therefore covers ordinary linear regression on one-hot labels followed by simplex projection. Its training minimizer predicts u, while the alternative $q_{\mathrm{alt}}(x) = \Pi_{\Delta(\mathcal{Y})}(u + x)$ attains the optimal calibrated size $1 - \alpha$. This uses $\lambda = 1$ in the proof, for which $T_\lambda = 1 > 1/(K - 1)$.

Logistic regression with log loss and the softmax link. Take the softmax link and its log-loss training objective,

$$\Psi_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}, \qquad \phi_y(z) = \log \sum_j e^{z_j} - z_y.$$

This is the full multinomial logistic class with affine logits and an intercept. For $r = \Psi(z)$, the second derivative of ϕ_y in any score direction d is

$$\sum_j r_j d_j^2 - \left(\sum_j r_j d_j \right)^2 .$$

This variance is positive unless all coordinates of d are equal, verifying the convexity and strict-convexity conditions. Also,

$$\overline{\phi}(z) = \log \sum_j e^{z_j} - \frac{1}{K} \sum_j z_j \geq \log K,$$

with equality when the coordinates of z are equal. We may therefore take $\beta = 0$. Softmax is symmetric and preserves coordinate order, and

$$\Psi_y(\lambda v_y) = \frac{e^\lambda}{e^\lambda + K - 1} \longrightarrow 1.$$

Thus the loss and link are admissible, and the strict-convexity condition in Theorem 4 holds. In particular, $\lambda = \log(2(K-1))$ makes $T_\lambda = 2/3$, giving an explicit alternative. Every log-loss minimizer predicts uniformly, while another logistic model achieves Bayes-optimal calibrated size. Because $\beta = 0$, an L^2 penalty on both weights and intercepts also preserves this conclusion.

For example, with $K = 3$ and 99% target coverage, the logistic training minimizer has expected size 2.97, while its alternative has size 0.99.

Generalized linear models: Bernoulli and Poisson regression. The theorem applies to GLMs whose score losses and probability maps satisfy Definition 3. Two examples are coordinatewise Bernoulli and Poisson regression of the one-hot labels. With affine natural parameters $z = b + Wx$, their negative log-likelihood losses are, respectively,

$$\phi_y^{\mathrm{Bern}}(z) = \sum_j \log(1 + e^{z_j}) - z_y, \qquad \phi_y^{\mathrm{Pois}}(z) = \sum_j e^{z_j} - z_y.$$

Both are strictly convex. Their balanced risks are minimized at $\beta \mathbf{1}_\mathcal{Y}$ with $\beta = -\log(K-1)$ and $\beta = -\log K$, respectively. To obtain class probabilities, normalize the predicted means of the K coordinatewise regressions:

$$\Psi_j(z) = \frac{g(z_j)}{\sum_k g(z_k)}, \qquad g(t) = \frac{1}{1+e^{-t}} \text{ or } g(t) = e^t.$$

Here $g(z_j)$ is the mean response of the jth Bernoulli or Poisson model. The normalization is used to construct probability conformity scores after fitting the GLMs with the losses above. Both links satisfy the symmetry and order assumptions. Along $\beta\mathbf{1} + \lambda v_y$, the incorrect-coordinate means tend to zero while the true-coordinate mean stays bounded away from zero, so the confidence assumption holds as well. Thus both GLM examples satisfy Theorem 4's assumptions. They also satisfy the strict convexity assumption, so every training minimizer predicts uniformly and has a highly sub-optimal set size.

E Synthetic experiments

This appendix specifies the synthetic study in Section 5 and reports results for both training-set sizes.

E.1 Setup

Tasks and models. We adapt the multinomial generator of Romano et al. (2020, Section 3.2) with ten features and ten classes. The first feature takes values 1 and -8 with probabilities 0.2 and 0.8; the remaining features are independent standard Gaussians. For each replicate, we draw $W^\star \in \mathbb{R}^{10\times 10}$ with independent standard Gaussian entries. A task with s active features uses $p^\star(x) = \mathrm{softmax}((W_s^\star)^\top x)$, where $W_s^\star$ keeps the first s rows of $W^\star$ and sets the rest to zero. Thus $p^\star$ is the ground-truth predictor for the selected task. The model is $q_{W,b}(x) = \mathrm{softmax}(W^\top x + b)$, with a free intercept and $W_{j,:} = 0$ for $j > r$. Both s and r range over $\{2, 5, 10\}$; the model class is realizable when $r \geq s$ and misspecified when $r < s$.

Data and training. We use ten independent task/data replicates with training sizes $n_{\text{train}} \in \{1{,}000, 10{,}000\}$, the smaller sample being a subset of the larger. Validation, reference, and test samples are drawn independently (Table 3). Within each replicate, all methods share the training data, initialization, minibatch order, and evaluation samples. Each objective is trained with Adam for 200 epochs under six hyperparameter configurations. CE and Brier training use an optional penalty $\lambda\|W\|_F^2/2$. ConfTr follows Stutz et al. (2022), splitting each minibatch into calibration and prediction halves, and minimizes the log soft set size without a size hinge, classification term, or weight penalty.

Model selection. Every ten epochs, we score the current checkpoint on the validation sample and select the configuration and checkpoint with the best score: validation loss for CE and Brier, and hard THR set size for ConfTr, with half of the validation sample used for calibration and half for measuring size. Selection uses sampled labels only.

Table 3: Synthetic experiment settings. Each objective uses six hyperparameter configurations and the same training budget.

Setting	Value
Training / validation	$n_{\text{train}} \in \{1{,}000, 10{,}000\}$ / 4,000 examples
Reference / test	50,000 independent examples each
Epochs / batch size	200 / 256
Learning rates	$\{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$
CE and Brier grid	Learning rate $\times$ penalty $\lambda \in \{0, 10^{-4}\}$
ConfTr grid	Learning rate $\times$ temperature $T_{\text{soft}} \in \{0.1, 1\}$
ConfTr dispersion / size weight	0.1 / 0.005

E.2 Evaluation

For each selected predictor q, we fit a THR threshold on the reference sample using the known probabilities $p^\star$, so that the sets cover 90% of the reference distribution, with randomized inclusion at the threshold when needed. Applying the same procedure to $p^\star$ gives the Bayes THR oracle. Both rules are then fixed and evaluated on the independent test sample, where we report the excess set size over the oracle and the test estimate of Brier regret $\mathsf{Reg}_{\mathsf{Brier}}(q) = \mathbb{E}_{X\sim D_X}\|q(X) - p^\star(X)\|_2^2$. Test coverage lies between 90.00% and 90.08% for every method and cell, so the excess size closely approximates, but is not exactly, the set-size regret at matched coverage. Intervals are cellwise 95% Student-t intervals across the ten replicates; method comparisons use within-replicate differences.

E.3 Results

Figure 2 compares CE and Brier training at $n_{\text{train}} = 10{,}000$ in the format of Figure 1. Table 4 reports excess set size and Brier regret for both training sizes, and Table 5 gives the paired set-size differences. CE yields smaller mean sets than ConfTr in all nine cells at both training sizes. At $n_{\text{train}} = 10{,}000$, CE and Brier training differ in mean set size by at most 0.0052, and Brier training also outperforms ConfTr in every cell. At $n_{\text{train}} = 1{,}000$, the two proper losses agree less closely, and ConfTr slightly outperforms Brier training at $(s, r) = (2, 10)$. Some paired intervals include zero.



Figure 2: CE and Brier training on the synthetic tasks at nominal 90% coverage with $n_{\text{train}} = 10{,}000$, averaged over ten replicates. Each cell shows CE (top) and Brier (bottom); outlined cells are misspecified ($r < s$). (a) Paired ConfTr-minus-CE and ConfTr-minus-Brier set size; positive values favor the proper loss. (b) Brier regret. (c) Excess set size over the Bayes THR oracle. Color scales match Figure 1; intervals for the paired differences are in Table 5.

Table 4: Synthetic results at nominal 90% coverage. Excess size is the mean test set size minus that of the Bayes THR oracle; the Brier-regret columns estimate $\mathsf{Reg}_{\mathsf{Brier}}(q) = \mathbb{E}_{X \sim D_{\mathcal{X}}} \|q(X) - p^\star(X)\|_2^2$. Entries are means over ten replicates; paired intervals for the set-size comparisons are in Table 5. Stars mark misspecified cells ($r < s$).

			Excess set size			Brier regret		
n_{train}	s	r	CE	Brier	ConfTr	CE	Brier	ConfTr
1,000	2	2	0.021	0.024	0.330	0.0020	0.0022	0.4252
		5	0.061	0.078	0.242	0.0063	0.0073	0.2991
		10	0.113	0.199	0.190	0.0118	0.0155	0.1731
	5	2*	0.708	0.706	0.891	0.1184	0.1189	0.3815
		5	0.036	0.048	0.152	0.0074	0.0097	0.2078
		10	0.075	0.110	0.231	0.0145	0.0200	0.1065
	10	2*	1.436	1.441	1.687	0.2556	0.2560	0.5707
		5*	0.858	0.894	1.192	0.1910	0.1929	0.5164
		10	0.057	0.074	0.165	0.0176	0.0220	0.0640
10,000	2	2	0.001	0.003	0.174	0.0003	0.0003	0.3240
		5	0.005	0.008	0.175	0.0006	0.0008	0.3261
		10	0.012	0.016	0.058	0.0011	0.0014	0.1641
	5	2*	0.669	0.670	0.795	0.1164	0.1164	0.3251
		5	0.004	0.006	0.040	0.0008	0.0012	0.0613
		10	0.008	0.011	0.056	0.0016	0.0022	0.0670
	10	2*	1.406	1.408	1.645	0.2539	0.2538	0.6045
		5*	0.802	0.807	0.938	0.1852	0.1852	0.3178
		10	0.006	0.008	0.045	0.0021	0.0029	0.0197

Table 5: Paired differences in synthetic test set size. Positive ConfTr-minus-CE and ConfTr-minus-Brier values favor the proper loss. Entries are means ± 95% Student-t half-widths over ten replicates; stars mark misspecified cells ($r < s$).

n_{train}	s	r	Brier − CE	ConfTr − CE	ConfTr − Brier
1,000	2	2	$+0.0034 \pm 0.0060$	$+0.3089 \pm 0.2634$	$+0.3055 \pm 0.2610$
		5	$+0.0172 \pm 0.0137$	$+0.1814 \pm 0.1721$	$+0.1641 \pm 0.1685$
		10	$+0.0860 \pm 0.0604$	$+0.0769 \pm 0.0985$	-0.0091 ± 0.0690
	5	2^*	-0.0015 ± 0.0085	$+0.1828 \pm 0.1760$	$+0.1843 \pm 0.1786$
		5	$+0.0119 \pm 0.0073$	$+0.1154 \pm 0.0999$	$+0.1035 \pm 0.0960$
		10	$+0.0350 \pm 0.0175$	$+0.1563 \pm 0.1395$	$+0.1213 \pm 0.1291$
	10	2^*	$+0.0043 \pm 0.0085$	$+0.2508 \pm 0.2224$	$+0.2465 \pm 0.2239$
		5^*	$+0.0367 \pm 0.0144$	$+0.3349 \pm 0.1584$	$+0.2981 \pm 0.1633$
		10	$+0.0177 \pm 0.0099$	$+0.1079 \pm 0.0480$	$+0.0902 \pm 0.0412$
10,000	2	2	$+0.0016 \pm 0.0022$	$+0.1725 \pm 0.1938$	$+0.1709 \pm 0.1939$
		5	$+0.0028 \pm 0.0020$	$+0.1702 \pm 0.1975$	$+0.1673 \pm 0.1970$
		10	$+0.0043 \pm 0.0030$	$+0.0468 \pm 0.0517$	$+0.0425 \pm 0.0508$
	5	2^*	$+0.0010 \pm 0.0030$	$+0.1266 \pm 0.1221$	$+0.1256 \pm 0.1221$
		5	$+0.0020 \pm 0.0017$	$+0.0360 \pm 0.0180$	$+0.0341 \pm 0.0172$
		10	$+0.0034 \pm 0.0020$	$+0.0483 \pm 0.0177$	$+0.0448 \pm 0.0166$
	10	2^*	$+0.0025 \pm 0.0046$	$+0.2394 \pm 0.2386$	$+0.2369 \pm 0.2391$
		5^*	$+0.0052 \pm 0.0057$	$+0.1359 \pm 0.0639$	$+0.1307 \pm 0.0624$
		10	$+0.0023 \pm 0.0014$	$+0.0391 \pm 0.0160$	$+0.0368 \pm 0.0153$

F Real-data experiments

This appendix specifies the CIFAR experiments in Section 5 and reports additional results.

F.1 Data, models, and training

Data. CIFAR-20 uses the canonical coarse-label annotation of the CIFAR-100 images. We train on the first 45,000 training examples, calibrate on the remaining 5,000, and test on the official 10,000 test examples. All main comparisons use this fixed split.

Architectures. Affine models flatten the standardized image and apply an affine map to K logits. MLP-1×8 has one eight-unit affine layer followed by BatchNorm (decay 0.9) and ReLU; MLP-2×64 repeats this block twice with width 64. The ResNet-50 is a preactivation ResNet v2 with stage widths $(256, 512, 1024, 2048)$.

Training stages. Every model is first trained with CE for 150 epochs. Affine models use one pretraining run per seed (0–9); CE and ConfTr then continue from the same checkpoint for 50 epochs, updating the full affine map. For MLPs and ResNets, we train one CE representation per configuration, freeze its parameters and BatchNorm statistics, and train ten paired heads (seeds 0–9) for 50 epochs, with the final layer identically initialized within each pair. During head training, the frozen ResNet backbone runs BatchNorm in training mode. The CIFAR-100 ResNet comparison reuses a previously trained backbone and its CE heads; only the ConfTr heads are newly trained. We evaluate the final epoch, with no checkpoint selection or temperature scaling. Because the CE objective does not depend on α, CE models and representations are shared across the two coverage levels, while ConfTr is trained separately at $\alpha = 0.05$ and $\alpha = 0.01$ with all other settings unchanged.

Optimization. All stages use SGD with Nesterov momentum 0.9 and the learning rates and batch sizes in Table 6. A step schedule keeps the base learning rate for the first two fifths of training and multiplies it by 0.1, 0.01, and 0.001 at the starts of the last three fifths. CE minimizes the average negative log-likelihood plus $5 \times 10^{-4}\|\theta\|_2^2$. For the ResNet heads, CE and ConfTr share the epoch budget but not the batch size or learning rate, so their numbers of parameter updates differ. In the 50-epoch affine and MLP stages, the training set for each seed is formed by concatenating five copies of the 45,000 examples, shuffling with the seed, and truncating to 45,000 examples. Inputs are scaled to $[0, 1]$ and standardized channelwise with means $(0.4914, 0.4822, 0.4465)$ and standard

deviations $(0.2470, 0.2435, 0.2616)$. All models use random horizontal flips and four-pixel padded crops; ResNets additionally use the CIFAR AutoAugment policy and cutout with padding eight.

Table 6: Base learning rate / batch size by training stage.

Model	Initial CE	Final CE	Final ConfTr
Affine, MLPs	0.05/100	0.005/100	0.005/100
ResNet-50	0.05/100	0.01/500	0.005/100

F.2 ConfTr objectives

Let $\mathcal{B}$ index a training minibatch of b examples, split into calibration and prediction halves $\mathcal{B}_{\text{cal}}$ and $\mathcal{B}_{\text{pred}}$ of size $b_{\text{cal}} = b/2$. For logits $z_\theta(x)$, write $q_\theta(x) = \text{softmax}(z_\theta(x))$. During training, ConfTr uses log-probability conformity scores $s_\theta(x, y) = \log q_{\theta,y}(x)$, which have the same ordering as the predicted probabilities. Its differentiable threshold is

$$\widetilde{\tau}_\theta = \text{SmoothQuantile}_{(1+1/b_{\text{cal}})\alpha,\, d}\big(\{s_\theta(X_i, Y_i) : i \in \mathcal{B}_{\text{cal}}\}\big),$$

computed with a variational sorting network with dispersion $d = 0.1$. The soft inclusion probabilities and size surrogate are

$$\widetilde{h}_{\theta,i}(y) = \sigma\left(\frac{s_\theta(X_i, y) - \widetilde{\tau}_\theta}{T_{\text{soft}}}\right), \qquad \widetilde{\mathsf{size}}_\theta(\mathcal{B}) = \frac{1}{b_{\text{cal}}} \sum_{i \in \mathcal{B}_{\text{pred}}} \sum_{y \in \mathcal{Y}} \widetilde{h}_{\theta,i}(y),$$

where σ is the logistic function. Plain ConfTr minimizes

$$\mathcal{L}_{\text{ConfTr}}(\theta; \mathcal{B}) = \log\big(\lambda_{\text{size}} \widetilde{\mathsf{size}}_\theta(\mathcal{B}) + 10^{-8}\big) + 5 \times 10^{-4} \|\theta\|_2^2, \tag{31}$$

with $b = 100$, $T_{\text{soft}} = 1$, and $\lambda_{\text{size}} = 0.005$, and has no CE or coverage term.

ConfTr+$\mathcal{L}_{\text{class}}$, used only for the ResNet heads, follows the `lclass` configuration of the ConfTr code (Stutz et al., 2022): it adds a classification coverage loss with loss matrix I_K and uses the valid-size loss with size weight 0.01 and a logarithmic transform. It has no auxiliary CE term. Both variants use $T_{\text{soft}} = 1$, $d = 0.1$, and half-batch calibration.

F.3 Calibration and evaluation

Use the calibration notation from Section 2: let $B = \{(X_i, Y_i)\}_{i \in [n]}$, where $n = 5{,}000$, and let $\hat{s}_i = s_q(X_i, Y_i) = q_{Y_i}(X_i)$ be the true-label conformity score. Writing $\hat{s}_{(k)}$ for the kth largest calibration score, the standard split conformal rule (9) is

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad \tau = \hat{s}_{(k)}, \qquad S(x) = \{y \in \mathcal{Y} : q_y(x) \geq \tau\}. \tag{32}$$

Equivalently, the threshold has rank $j = n + 1 - k$ in ascending order: $j = 250$ at $\alpha = 0.05$ and $j = 50$ at $\alpha = 0.01$. Coverage, accuracy, Brier loss, and NLL are measured on the test set. Intervals are cellwise 95% normal intervals $\overline{z} \pm 1.96 s_z/\sqrt{10}$ over the ten model or head seeds, and paired differences are computed within seed. For MLPs and ResNets, the intervals condition on one frozen representation and do not reflect variability from retraining it.

F.4 Full results at 99% coverage

Table 7 gives the full metrics behind Table 2, using the same data, CE models, and seeds as Table 1. In every configuration where ConfTr has smaller mean sets, it also has higher Brier loss and lower accuracy than CE. For the narrow CIFAR-100 MLP, the paired ConfTr-minus-CE realized coverage difference is -0.106 ± 0.034 percentage points, so this comparison is at a common nominal target rather than at identical realized coverage.

Table 7: Affine/MLP results at nominal 99% coverage. Size intervals are 95% normal intervals over ten seeds. Coverage and accuracy are percentages; Brier is the test loss.

Model	Method	Size	Coverage	Accuracy	Brier
CIFAR-20					
Affine	CE	18.282 ± 0.014	98.968	25.152	0.86195
	ConfTr	17.923 ± 0.036	98.913	16.531	1.38420
MLP-1 $\times$ 8	CE	17.127 ± 0.006	99.120	23.260	0.86576
	ConfTr	17.252 ± 0.028	99.183	20.926	0.94466
MLP-2 $\times$ 64	CE	14.925 ± 0.019	98.979	31.149	0.80858
	ConfTr	15.879 ± 0.044	99.021	20.182	0.93907
CIFAR-100					
Affine	CE	89.188 ± 0.159	98.837	16.749	0.93046
	ConfTr	85.999 ± 0.294	99.021	7.090	1.34339
MLP-1 $\times$ 8	CE	86.465 ± 0.083	99.099	12.171	0.94996
	ConfTr	82.006 ± 0.191	98.993	5.688	0.98124
MLP-2 $\times$ 64	CE	62.717 ± 0.149	98.746	21.586	0.89385
	ConfTr	64.299 ± 0.398	98.859	14.064	0.93987

F.5 Results at 98% coverage

To check whether the sign flips in Table 2 are specific to 99% coverage, we repeat the affine and MLP comparisons at 98% nominal coverage ($\alpha = 0.02$). The protocol is identical to Table 1: the same fixed 45k/5k/10k training/calibration/test split, exact order statistic, and ten seeds. ConfTr is trained at $\alpha = 0.02$ from the same pretraining checkpoints, with all other settings unchanged. Because the CE objective does not depend on α, the CE models are those of Table 1.

Table 8 shows that the paired set-size difference again changes sign across model classes. On CIFAR-20, ConfTr yields smaller mean sets for the affine model, whereas CE yields smaller sets for MLP-2 $\times$ 64. On CIFAR-100, ConfTr yields smaller mean sets for the affine and MLP-1 $\times$ 8 models, whereas the MLP-2 $\times$ 64 difference favors CE in the mean but its interval includes zero. The CIFAR-20 MLP-1 $\times$ 8 difference ($+0.059 \pm 0.044$) is small relative to the paired realized-coverage difference ($+0.21 \pm 0.06$ percentage points, ConfTr minus CE). Thus the sign flip of the set-size difference is not specific to 99% coverage: it already appears at 98%.

Table 8: Prediction-set size differences across model classes and label granularities at nominal 98% coverage, using the same fixed 45k/5k/10k training/calibration/test split and exact order statistic as Table 1. Entries are paired ConfTr-minus-CE mean size differences Δ_{size} with 95% normal confidence-interval half-widths over ten model/head seeds. Positive Δ_{size} favors CE.

Model	CIFAR-20	CIFAR-100
Affine	-0.391 ± 0.054	-0.764 ± 0.457
MLP-1 $\times$ 8	$+0.059 \pm 0.044$	-1.704 ± 0.487
MLP-2 $\times$ 64	$+1.158 \pm 0.067$	$+0.110 \pm 0.517$

F.6 Standard-width ResNet

Table 9 expands the two ResNet rows of Table 1 and adds ConfTr+$\mathcal{L}_{\text{class}}$ (Section F.2). At 95% coverage, both direct-training variants produce larger mean sets than CE on both labelings; the paired ConfTr+$\mathcal{L}_{\text{class}}$-minus-CE differences are 0.911 ± 0.078 on CIFAR-20 and 0.051 ± 0.030 on CIFAR-100. The plain-ConfTr ordering is unchanged under ten paired random resplits of the pooled calibration and test examples, and under a midpoint empirical quantile in place of the exact order statistic: across all four combinations of split and threshold, the paired ConfTr-minus-CE difference ranges from 0.012 to 0.021 on CIFAR-20 and from 0.084 to 0.139 on CIFAR-100, with every interval excluding zero.

99% coverage. Table 10 reports the CIFAR-100 ResNet at 99% coverage. Paired differences relative to CE are -0.154 ± 0.312 for ConfTr and -0.535 ± 0.215 for ConfTr+$\mathcal{L}_{\text{class}}$, with paired

Table 9: Standard-width ResNet-50 at nominal 95% coverage. Size intervals are 95% normal intervals over ten head seeds sharing one backbone per task. Coverage and accuracy are percentages; Brier and NLL are test losses.

Method	Size	Coverage	Accuracy	Brier	NLL
CIFAR-20					
CE	1.87746 ± 0.00427	95.413	82.350	0.25052	0.56368
ConfTr	1.88918 ± 0.00521	95.345	82.268	0.26508	0.64931
ConfTr+$\mathcal{L}_{\text{class}}$	2.78888 ± 0.07726	95.027	77.348	0.36176	1.09692
CIFAR-100					
CE	3.81871 ± 0.02090	95.095	73.770	0.35630	0.91681
ConfTr	3.90319 ± 0.05316	95.009	73.216	0.38308	1.04887
ConfTr+$\mathcal{L}_{\text{class}}$	3.87020 ± 0.03050	95.132	73.621	0.37181	0.99619

realized-coverage differences of $+0.053 \pm 0.055$ and -0.011 ± 0.056 percentage points. Both variants thus have smaller mean sets, although the plain-ConfTr interval includes zero. The advantage of CE for this expressive model is therefore specific to the 95% target and the training protocols evaluated here.

Table 10: Standard-width CIFAR-100 ResNet-50 at nominal 99% coverage. Size intervals are 95% normal intervals over ten head seeds sharing one backbone; ConfTr heads are trained at $\alpha = 0.01$. Coverage and accuracy are percentages; Brier and NLL are test losses.

Method	Size	Coverage	Accuracy	Brier	NLL
CE	10.9915 ± 0.1199	98.946	73.770	0.35630	0.91681
ConfTr	10.8371 ± 0.2976	98.999	72.709	0.38007	0.99947
ConfTr+$\mathcal{L}_{\text{class}}$	10.4568 ± 0.1432	98.935	73.653	0.37137	0.99155