QUANTILE MERGING FOR BOOTSTRAPPED TRANSDUCTIVE CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transductive conformal prediction constructs joint prediction sets for groups of test samples, but practical implementations face a fundamental challenge: calibration data are typically available as i.i.d. individual samples, whereas inference is performed on groups whose size may vary at deployment time. We study how to construct transductive calibration sets from a fixed pool of calibration samples while preserving coverage. We consider bootstrap sampling from the empirical calibration distribution and show that valid marginal coverage requires an adjustment of the calibration quantile. To address the variance increase from bootstrap sampling, we introduce a K-fold transductive conformal framework that repeats bootstrap calibration and aggregates the resulting predictors using p-value and e-value merging techniques. We show that all considered aggregation schemes admit an equivalent quantile-level implementation, reducing inference to standard split conformal prediction with a merged calibration quantile. We further theoretically characterize the effects of calibration size, group size, and the number of folds on coverage and efficiency. Empirical results validate the theory and demonstrate substantial improvements in stability over single fold bootstrap calibration.

1 INTRODUCTION

Conformal prediction (CP) constructs prediction sets using training and calibration data while providing finite-sample coverage guarantees. Whereas full CP repeatedly retrains the predictive model and is therefore computationally demanding, split CP, (Papadopoulos et al., 2002; Lei et al., 2018), requires only a single training step (or a pre-trained model) and uses a dedicated calibration set to determine prediction set thresholds. Transductive CP (Vovk, 2013; Gazin et al., 2024; Fermanian et al., 2025) generalizes this paradigm from individual test instances to groups of test samples, yielding joint prediction sets defined over the group. This is motivated by various applications where a batch of tests is regularly encountered: testing molecules and solutions in drug discovery, fraud detection in finance, and medical screening. Split transductive CP methods as in (Behboodi et al., 2025) are the preferred approaches in practice given their efficiency.

However, calibration data are typically observed as i.i.d. individual samples, whereas transductive inference operates on groups whose size may vary at deployment time. This raises a fundamental question: given a collection of calibration samples, how should calibration groups be generated for transductive calibration? Answering this question requires understanding how the grouping and sampling strategy affects both the validity and efficiency of the resulting conformal predictor. This paper aims to characterize these effects and identify principled procedures for constructing transductive calibration sets. As a natural starting point, we consider bootstrap sampling from the empirical distribution of the calibration data to construct transductive calibration groups. We show, however, that standard bootstrap calibration does not achieve the desired marginal coverage level and therefore requires an adjustment of the target quantile. Furthermore, bootstrap sampling introduces increased variance in conditional coverage and conditional efficiency.

To mitigate this effect, we propose a K-fold transductive split conformal framework. The procedure generates K independent bootstrap-based calibration folds and performs conformal calibration on each fold separately. The resulting prediction sets are then combined using a family of aggregation methods based on p-value and e-value merging. We show that, for all considered aggregation schemes, instead of explicitly merging prediction sets, one can merge the fold-specific calibration

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.

quantiles through the corresponding merging operator and then perform transductive split conformal inference using the resulting merged quantile. This yields a simple and computationally efficient implementation while preserving the theoretical guarantees of the underlying merging procedure. We further characterize the impact of the calibration sample size and the choice of K on coverage and efficiency. Our theoretical analysis reveals the trade-offs introduced by repeated sampling and quantile aggregation, and our experiments support the theory. **Notation.** Table 3 in Appendix contains the notations used in this paper.

Preliminaries. Standard conformal prediction (CP) constructs a prediction set for each test input that contains the true label with probability at least $1-\alpha$. Let $Z_1^m = \big((X_i, Y_i) : i \in [m]\big)$ be a sequence of labeled examples, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. Given a test input X_{m+1}, CP produces a prediction set $\Gamma^\alpha(X_{m+1})$ satisfying the marginal coverage guarantee

$$\mathbb{P}\left(Y_{m+1} \notin \Gamma^\alpha(X_{m+1})\right) \leq \alpha.$$

In split conformal prediction (SCP) (Papadopoulos et al., 2002; Lei et al., 2018), a pre-trained predictor and an independent calibration set are used to estimate a threshold for a nonconformity score $S(x, y)$. Let $S_i = S(X_i, Y_i), i \in [m]$, denote the calibration scores, and define $k_m = \lceil (m+1)(1-\alpha) \rceil$. Writing $S_{(1)} \leq \cdots \leq S_{(m)}$ for the order statistics of the calibration scores, the conformal threshold is $\widehat{q}_m = S_{(k_m)}$ for $k_m \leq m$, and $+\infty$ for $k_m = m+1$. The resulting prediction set is

$$\Gamma^\alpha(x) = \{y \in \mathcal{Y} : S(x, y) \leq \widehat{q}_m\}.$$

This is equivalent to p-value based conformal prediction using the term $p(s) = \frac{1}{m+1}\sum_{i=1}^{m} \mathbf{1}(S_i \geq s)$ as a p-variable (see App. D and B for more details). Many applications require guarantees that hold jointly across multiple predictions. In autonomous driving, ranking, and decision-making systems, a single erroneous prediction may invalidate the entire outcome (Fermanian et al., 2025). Transductive conformal prediction (TCP) addresses such settings by constructing a joint prediction set for a batch of test samples (Vovk, 2013; Vovk et al., 2022a). Given calibration data Z_1^m and test inputs $X_{m+1}^{m+n} = (X_{m+1}, \ldots, X_{m+n})$, TCP outputs a set of candidate label vectors satisfying the joint coverage guarantee

$$\mathbb{P}\left(Y_{m+1}^{m+n} \notin \Gamma^\alpha\big(Z_1^m, X_{m+1}^{m+n}\big)\right) \leq \alpha. \tag{1}$$

By efficiency of CP, we mean the size of the prediction set. For transductive cases, we also use the efficiency rate as proposed in Behboodi et al. (2025), defined as $\gamma_{n,m} = \log|\Gamma^\alpha|/n$. A simple baseline applies a Bonferroni correction by constructing $\Gamma^{\alpha/n}(X_{m+i})$ for each test input and taking the Cartesian product $\prod_{i=1}^{n}\Gamma^{\alpha/n}(X_{m+i})$. Although this construction satisfies the coverage in eq. 1, it becomes increasingly inefficient as the batch size n grows. Recently, Behboodi et al. (2025) proposed transductive split conformal prediction (TSCP), which replaces per-sample scores with transductive conformity scores defined over candidate label sequences using the approximate conditional probability of the underlying model. In particular, TSCP evaluates a transductive score for $\big(X_{m+1}^{m+n}, \widetilde{Y}_{m+1}^{m+n}\big), \widetilde{Y}_{m+1}^{m+n} \in \mathcal{Y}^n$, and calibrates it using groups of n labeled samples from the calibration set. The calibration outcome is the quantile τ. The resulting joint prediction set can be computed efficiently using dynamic programming. However, TSCP assumes that the calibration data are already given as groups of size n. We have included the details of transductive methods in Appendix B. Under a fixed calibration budget of m samples, partitioning the data into such groups reduces the effective calibration sample size to approximately m/n, potentially degrading efficiency as n increases. At the other extreme, constructing all possible ordered groups would yield m^n calibration sequences and is therefore computationally prohibitive. This paper investigates whether resampling can bridge this gap to achieve coverage and efficiency. Given a fixed pool of calibration samples, can one construct transductive calibration groups that preserve joint coverage while recovering the efficiency lost through partitioning? More broadly, how do different resampling strategies affect marginal coverage, conditional coverage variability, and prediction-set efficiency?

2 Related Works

We focus on the most related works here and present the extended version in Appendix C. Transductive conformal prediction (TCP) was introduced to provide joint guarantees over multiple test samples (Vovk, 2013; Vovk et al., 2022a). Bonferroni correction to individual conformal predictors was proposed (Vovk, 2013). Transductive Split Conformal Prediction (TSCP) (Behboodi et al., 2025)

introduced transductive conformity scores and efficient dynamic-programming inference for large output spaces. However, existing methods largely assume that calibration data are already available in groups matching the transductive batch size. How to construct such calibration groups from a fixed pool of calibration samples is the primary focus of this work.

Cross-conformal prediction and its variants aim to improve the utilization of calibration data (Vovk et al., 2022a) by partitioning the data into multiple folds, constructing conformal predictors on each fold, and aggregating the resulting evidence. Recent developments have connected such aggregation procedures to the broader literature on combining conformal p-values and e-values (Vovk et al., 2022b; Vovk & Wang, 2024; Gasparin et al., 2025; Vovk & Wang, 2020; Fisher, 1948; Brown, 1975). Our proposed method builds on this line of work and shows that, for split conformal prediction, the aggregation can be performed entirely at the quantile level.

The efficiency of conformal prediction and optimal conformal predictors have been studied extensively (Vovk et al., 2022a; Sadinle et al., 2019; Correia et al., 2024; Vovk et al., 2016; Gasparin & Ramdas, 2025; Behboodi et al., 2025; Lei & Wasserman, 2014; Lei, 2014). Recent work has analyzed finite-sample expected prediction-set size directly and derived performance bounds for conformal predictors (Dhillon et al., 2024). For split conformal prediction, the probabilistic analysis of efficiency and coverage has been discussed in Tibshirani (2023); Vovk et al. (2022a); Vovk (2012); Hulsman (2022); Marques F. (2025); Bian & Barber (2023). The fact that sample conditional coverage has a Beta distribution is indicated in Vovk (2012) and expanded in some of the above papers. Our work provides various new asymptotics for sample-conditional efficiency and coverage of SCP.

3 Bootstrapping for Transductive Conformal Prediction

The calibration samples are typically m i.i.d. samples. For classical split conformal prediction ($n = 1$), the common practice is to use all the samples, and finding the threshold costs at worst $O(m \log m + mC_S)$ where C_S is the cost of computing nonconformity scores. For the case of transductive conformal prediction with $n > 1$, one approach is to convert m calibration samples to m/n groups of n samples (as in Behboodi et al. (2025)), which leads to a small calibration set size. However, instead of one fixed partition of the calibration set into groups of size n, we can resample K folds, each built from $n \cdot N_s$ samples drawn from the m available calibration points (N_s groups of size n per fold). Each fold produces its own transductive calibration quantile and its own inclusion/exclusion decision for a candidate test label sequence y_{m+1}^{m+n}. In what follows, we discuss how to aggregate these K per-fold outputs into a single joint decision that still satisfies eq. 1. We call these methods transductive cross conformal prediction (TCCP).

The single-fold bootstrapping. For the case $K = 1$, no aggregation is needed. We sample N_s groups using the empirical distribution obtained from the calibration dataset. When the number of test points in transductive prediction is one, resampling adds Monte Carlo sampling noise and has higher variance than using all samples. However, for the general transductive size n, using all samples amounts to using m^n candidates, which is generally cumbersome. Resampling creates a tractable way of estimating the quantile with the cost of higher variance due to sampling (see Section 4). The K-fold provides a way of reducing the variance with the aggregation methods that we describe below.

p-value based aggregation. For each fold $k \in [K]$ and each candidate label sequence y_{m+1}^{m+n}, a transductive conformal prediction method on fold k yields a transductive p-value $p_k(y_{m+1}^{m+n}) \in [0, 1]$. We can aggregate these K (generally dependent) p-values into a single combined p-value $\bar{p}(y_{m+1}^{m+n})$ and use that for building the prediction set. The methods for merging p-values with arbitrary dependence have been discussed in prior works Vovk et al. (2022b); Vovk & Wang (2020); Gasparin et al. (2025).

In fold k, there are N_s groups. Denote the nonconformity scores for this group as $t_k^{(1)}, \ldots, t_k^{(N_s)}$, and the nonconformity score for the candidate labels y_{m+1}^{m+n} as $t_k(y_{m+1}^{m+n})$. First, define the p-variable $p_k(t)$ as follows:

$$p_k(t) := \frac{1 + \#\{i : t_k^{(i)} \geq t\}}{N_s + 1}. \tag{2}$$

The transductive p-value for the candidate label is given by $p_k(y_{m+1}^{m+n})$ A function that combines multiple p-values into a single p-value is called *a merging function*. Suppose that M is a merging function, and the final p-value is $\bar{p}(y_{m+1}^{m+n}) = M(p_1(y_{m+1}^{m+n}), \ldots, p_K(y_{m+1}^{m+n}))$. The prediction set is constructed as follows:

$$\Gamma^\alpha = \{y_{m+1}^{m+n} \in \mathcal{Y}^n : \bar{p}(y_{m+1}^{m+n}) > \alpha\}. \tag{3}$$

Since $\bar{p}$ is a p-value, $\mathbb{P}(Y_{m+1}^{m+n} \notin \Gamma^\alpha) \leq \alpha$. Naively, we would need to compute these p-values $p_i(y_{m+1}^{m+n})$ for $i \in [K]$, and for each candidate labels y_{m+1}^{m+n}, namely $k|\mathcal{Y}|^n$ times, which is cumbersome. However, for split-conformal values, the nonconformity scores $t_k^{(i)}$ are independent of y_{m+1}^{m+n}. Also suppose that we use the same nonconformity function across folds, and therefore, the corresponding score $t_k(y_{m+1}^{m+n})$ is independent of k (for instance by using the logits of a model). In this case, the prediction set can be built similarly to the usual split-conformal prediction by thresholding the nonconformity score y_{m+1}^{m+n}. For a merging function M, define $\bar{p}(t) := M(p_1(t), \ldots, p_K(t))$

Proposition 3.1. Consider K calibration sets with arbitrary dependence and a non-increasing $\bar{p}(t)$ defined from the merging function M as above. Consider a split conformal prediction setting, and suppose that the nonconformity score does not depend on the fold index. Then, the prediction set obtained by merging p-values in eq. 3 using $M(\cdot)$ is equivalent to the following prediction set:

$$\{y_{m+1}^{m+n} \in \mathcal{Y}^n : t(y_{m+1}^{m+n}) < t_\alpha\},$$

where the threshold $t_\alpha := \bar{p}^{-1}(\alpha)$ and $\bar{p}^{-1}(u) := \inf\{t : \bar{p}(t) \leq u\}$.

The proof is given in Appendix D.1. The proposition shows that p-value merging can be done by pre-computing t_α using the folds and the merging function.

A general form of the merging function is given by $c_{\phi,K} M_{\phi,K}$ where $M_{\phi,K} := \psi\left(\frac{\phi(p_1)+\cdots+\phi(p_K)}{K}\right)$ is the averaging function with ϕ as a continuous strictly monotonic function, and ψ as its inverse. $c_{\phi,K}$ is the calibration constant needed to obtain a p-value. For this choice, the final p-value is computed as $\bar{p}(t) := c_{\phi,K} M_{\phi,K}(p_1(t), \ldots, p_K(t))$. We can compute the constant from Theorem D.2, although the main challenge is to obtain a constant with mild dependence on K and independence from ϵ. We have included the technical details in Appendix D. In this paper, we consider the two merging functions. First, we consider **arithmetic averaging** defined as $\bar{p} = \frac{2}{K}\sum_{k=1}^K p_k$, and then **harmonic averaging** defined as $\bar{p} = e \ln K (K / \sum_{k=1}^K p_k^{-1})$. As can be seen, the constants 2 and $e \ln K$ are chosen to yield a theoretically valid p-value (Vovk & Wang, 2020). Other merging functions require their own calibration constants given in (Vovk & Wang, 2020).

To summarize, p-value merging can be done by pre-computing and combining the thresholds (quantiles) into a combined quantile t_α, then using it in a classical prediction method. The fixed cost for computing t_α is $O(KN_s n)$. *TCCP with p-value merging* is explained in Algorithm 1.

Practical nonconformity score. Building the prediction set still requires a search over $|\mathcal{Y}|^n$ pair of candidate labels y_{m+1}^{m+n}. However, we can still improve the procedure with dynamic programming-based approaches as presented in the transductive split conformal prediction (TSCP) of Behboodi et al. (2025). TSCP uses the joint log-probability of the group's labels $t(y_{m+1}^{m+n}) = \sum_j \log p(y_j|x_j)$ as the *conformity score*, and compute the log-quantile $\log \tau$. Note that this is a conformity, and not a nonconformity score, we need to update the algorithm by inverting the inequalities as needed, for example using $t_k^{(i)} \leq t$ instead in p-value computation. Next, we reuse the pruning strategy and dynamic programming method previously proposed in TSCP method of Behboodi et al. (2025) out of the box for building the prediction sets. The adapted algorithm for the TSCP variant is given in the Appendix, Algorithm 2. We monitor $\log \tau$, which is the calibrated log-quantile, during the experiments.

e-value based aggregation. Alternatively, for each fold k, we can compute a transductive prediction set $\Gamma_k^\alpha(x)$ (the TSCP set calibrated on fold k at level α) and then combine them. Combining confidence sets is related to the idea of combining e-variables discussed in (Ramdas & Wang, 2025, Sec. 13.3). Each per-fold set induces an e-variable for the inclusion test of a candidate label sequence y_{m+1}^{m+n}, $e_k(y_{m+1}^{m+n}) = \alpha^{-1}\mathbf{1}\{y_{m+1}^{m+n} \notin \Gamma_k^\alpha(x)\}$ (refer to Appendix E for more details on e-variables and their combination). Averaging e-values is valid under arbitrary dependence (the average of e-values is itself an e-value). Such averaging is equivalent to combining the underlying prediction sets by a vote

Algorithm 1 TCCP with p-value merging via threshold scan

Require: calibration pool of N_{cal} labeled samples; number of folds K; transductive (group) size n; groups per fold N_s; confidence level α; merging function $c_{\phi,K} M_{\phi,K}(p_1, \ldots, p_K)$

Ensure: threshold t_α such that $\Gamma^\alpha(x_{m+1}^{m+n}) = \{y_{m+1}^{m+n} \in \mathcal{Y}^n : t(y_{m+1}^{m+n}) \le t_\alpha\}$

for $k = 1, \ldots, K$ **do**
  draw N_s groups of size n, with replacement, from the calibration pool ▷ fold k
  for $i = 1, \ldots, N_s$ **do**
    compute the nonconformity score $t_k^{(i)}$ ▷ group score
  end for
end for
$B \leftarrow$ sorted union of $\{t_k^{(i)} : k \in [K],\ i \in [N_s]\}$
for $t \in B$, in increasing order **do**
  $p_k(t) \leftarrow \dfrac{1 + \#\{i : t_k^{(i)} \ge t\}}{N_s + 1}$ for each $k = 1, \ldots, K$ ▷ eq. 2
  $\bar{p}(t) \leftarrow c_{\phi,K} M_{\phi,K}(p_1(t), \ldots, p_K(t))$
  if $\bar{p}(t) \le \alpha$ **then**
    return $t_\alpha \leftarrow t$ ▷ $\bar{p}(\cdot)$ is non-decreasing in t: first crossing is the threshold
  end if
end for

threshold $k^\star \in \{1, \ldots, K\}$: include y_{m+1}^{m+n} in the aggregate set if and only if it appears in *at least* $k^\star$ of $\Gamma_1^{\alpha_1}, \ldots, \Gamma_K^{\alpha_K}$, where each fold's own level α_k may be adjusted to keep the aggregate valid at α.

We fix the confidence levels to α. Then given Γ_i^α, $i \in [K]$, all satisfying the confidence level α, we consider the following majority voting scheme:

$$\Gamma^\tau := \{y \in \mathcal{Y} : \sum_{i=1}^{K} \mathbf{1}\{y \in \Gamma_i^\alpha\} > K\tau\}. \tag{4}$$

Theorem 13.10 of Ramdas & Wang (2025) states that the aggregated set achieves the marginal coverage of $1 - \frac{\alpha}{1-\tau}$. We have also provided the bound on the confidence when each set has its own error level α_i in Appendix, Corollary 1. In particular, *the standard majority voting* ($k^\star = \lceil K/2 \rceil + 1$, $\tau = 1/2$) with $\sum_{i=1}^K \alpha_i = K\alpha$ satisfies the error level 2α. To guarantee the overall confidence level $1 - \alpha$, each fold should satisfy the confidence $1 - \alpha/2$. The penalty factor of 2 is known to be tight Ramdas & Wang (2025). In general, for a vote threshold τ, calibrating every fold at $\alpha_{\text{fold}} = \alpha(1 - \tau)$ guarantees $\mathbb{P}(Y_{m+1}^{m+n} \notin \Gamma^\tau) \le \alpha$ under arbitrary dependence. This is the tightest α_{fold} obtainable from a Markov argument.

However, the general e-value based aggregation in eq. 4 can also be reduced to a simple threshold similar to the union rule and the p-value based merging in the previous section. The following proposition formalizes this idea.

Proposition 3.2. Suppose that the threshold for building the prediction set at fold k is given by $t_{k,\alpha}$ for $k \in [K]$. Let $t_{(1)} \ge \cdots \ge t_{(K)}$ be these K numbers sorted descending. Then, for any $\tau \in [0, 1)$, the prediction set Γ^τ in eq. 4 is equal to: $\{y \in \mathcal{Y} : t(y) \le t_\alpha\}$, where $t_\alpha = t_{(k^\star)}$ for $k^\star = \lfloor K\tau \rfloor + 1$.

The proof is given in Appendix E.1. Proposition 3.2 states that Γ^τ can also be implemented by computing the K per-fold thresholds once and reading off their m-th largest. Then, only t_α is needed for inference. The union rule is the special case $\tau = 0$ ($m = 1$, $t_{(1)} = \max_k t_{k,\alpha}$). Note that our propositions are all for nonconformity scores. In our implementation, we use conformity scores instead, for which the signs should be flipped in the theorem.

4 Calibration Set Size's Impact on Coverage and Efficiency

We studied different methods for bootstrapping calibration samples and aggregating them to build the final prediction set. The main question of this section is to characterize how calibration set size affects efficiency and coverage. We start with classical prediction without resampling, restate some known results, and derive new ones. All analyses are for split conformal prediction and for $n = 1$.

Classical Conformal Prediction. To analyze the impact on coverage and efficiency, we compute the distribution of calibration-conditional coverage, which we then use to compute efficiency statistics. The starting point is that the conditional coverage is a Beta random variable (eq. 6). See Appendix F for more details. The Appendix summarizes the impacts of calibration set size in Table 5.

Impact on coverage. The marginal coverage of classical conformal prediction is known to be $\frac{k_m}{m+1}$ under exchangeable and almost surely distinct $S_1, \ldots, S_m, S_{m+1}$, where $k_m \leq m$ is the finite-sample corrected quantile. The marginal coverage approaches $1-\alpha$ as m increases but can change non-monotonically. For small m, the gap relative to $1-\alpha$ can be substantial. For example, $1-\alpha = 0.9$ with only 8 calibration samples will achieve the marginal coverage of 1.

Consider the conditional coverage now. Suppose that the CDF and density of the score S is respectively F_S and f_S. The sample-conditional coverage, or calibration-conditional coverage, is given by

$$U_m := \mathbb{P}\left(S_{m+1} \leq \widehat{q}_m \mid S_1, \ldots, S_m\right) = F_S(\widehat{q}_m). \tag{5}$$

For continuous F_S[1], the random variable U_m is distributed according to Beta distribution (see Tibshirani (2023), and Appendix F for details and discussions):

$$\boxed{U_m = F_S(\widehat{q}_m) \sim \text{Beta}(k_m, m+1-k_m).} \tag{6}$$

The expected value of U_m is the finite-sample coverage. The calibration size m is inversely related to the variance: $\text{Var}(U_m) = \alpha(1-\alpha)/m + o(m^{-1})$. As $m \to \infty$, the distribution of conditional coverage can be approximated as follows:

$$\boxed{\sqrt{m}(U_m - (1-\alpha)) \xrightarrow{d} \mathcal{N}(0, \alpha(1-\alpha)).} \tag{7}$$

Equation 7 is fundamental for the rest of our derivation.

Impact on the approximate quantile. The approximate quantile $\widehat{q}_m$ converges almost surely to $q_\alpha := F_S^{-1}(1-\alpha)$ as $m \to \infty$. Since $U_m = F_S(\widehat{q}_m)$, we can find similar Gaussian asymptotics for $\widehat{q}_m$ using the Delta method and Gaussian asymptotics for U_m. When $f_S(q_\alpha) > 0$, we have

$$\boxed{\sqrt{m}(\widehat{q}_m - q_\alpha) \Rightarrow \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f_S(q_\alpha)^2}\right).}$$

The quantile variance decreases as well with m, however, if the CDF F_S is flat around q_α (small f_S), the variance will be higher.

Impact on efficiency. Define the expected prediction-set size at deterministic threshold q:

$$H(q) := \mathbb{E}_X\left[|\{y \in \mathcal{Y} : S(X,y) \leq q\}|\right]. \tag{8}$$

For finite-label classification, $H(q) = \sum_{y \in \mathcal{Y}} \mathbb{P}_X\{S(X,y) \leq q\}$. Conditionally on the calibration sample, $H(\widehat{q}_m)$ is the conditional efficiency $\mathbb{E}\left[|\Gamma^\alpha(X_{m+1})| \mid S_1, \ldots, S_m\right]$, and when H is continuous, $H(\widehat{q}_m) \longrightarrow H(q_\alpha)$ as $m \to \infty$. Using Gaussian asymptotics for the quantile $\widehat{q}_m$ above, the conditional efficiency asymptotically tends to a Gaussian distribution.

Proposition 4.1. Assume that F_S is differentiable at q_α, with $0 < f_S(q_\alpha) < \infty$. Then, the conditional efficiency $H(\widehat{q}_m)$ satisfies

$$\boxed{\sqrt{m}\{H(\widehat{q}_m) - H(q_\alpha)\} \xrightarrow{d} \mathcal{N}\left(0, \frac{H'(q_\alpha)^2 \alpha(1-\alpha)}{f_S(q_\alpha)^2}\right).} \tag{9}$$

The proof is presented in Appendix F.3. From this result, it can be seen that the calibration-induced variance is approximately given by $\frac{\alpha(1-\alpha)}{m}\left(\frac{H'(q_\alpha)}{f_S(q_\alpha)}\right)^2$. Here again, the calibration set size m appears to be inversely related to the variance; however, the variance itself depends on $H'(q_\alpha)$ and $f_S(q_\alpha)$. Therefore, when $H(q)$ varies widely around q_α or $f_S(q_\alpha)$ is small, the variance is higher.

[1]This is an important assumption for the result. However, when F_S is not continuous, we can use randomized smoothing scores to get similar results. See Appendix F.2 for further discussions.

The marginal efficiency is given by:

$$\mathbb{E}\left[|\Gamma^{\alpha}(X_{m+1})|\right] = \mathbb{E}[H(\widehat{q}_m)]. \tag{10}$$

when H is continuous and suitable uniform-integrability conditions hold, we have $\mathbb{E}[H(\widehat{q}_m)] \longrightarrow H(q_\alpha)$. From this equation, given the score distribution F_S, we can compute the expected efficiency. For continuous F_S, define $Q(u) = F_S^{-1}(u)$ and let $g(u) := H(Q(u))$. We call $g(u)$ the **coverage-efficiency function**. Now, the efficiency can be computed using U_m:

$$\boxed{\mathbb{E}\left[|\Gamma^{\alpha}(X_{m+1})|\right] = \mathbb{E}[g(U_m)], \qquad U_m \sim \mathrm{Beta}(k_m, m+1-k_m).} \tag{11}$$

The calibration set size affects the marginal efficiency in three ways: effective marginal coverage $k_m/(m+1)$, the score distribution F_S and the conditional efficiency function $H(q)$. The latter two combine into the function $g(u) = H(Q(u))$. These factors can change the marginal efficiency with m in a generally non-monotonic way.

Analysis for K folds. As we saw in the previous section, once the asymptotics of $U_m = F_S(\widehat{q}_m)$ are found, the rest of the asymptotics follow mechanically using the Delta rule. Following Propositions 3.1 and 3.2, the merging schemes reduce to exactly a single scalar threshold t_α, built from K (generally dependent) per-fold quantiles. In this section, we derive the conditional coverage statistics and limiting distribution of t_α for majority voting with τ as the threshold.

Single fold $K = 1$. Marginal coverage. Our first step was to sample a subset of calibration data points. Consider the ordered non-conformity scores of the calibration set given as $t_{(1)} \leq \cdots \leq t_{(m)}$. Sampling N_s samples from the set amounts to sampling N_s of $t_{(i)}$ with replacement. The p-quantile for the fold is approximated by the ordered element $k_{p,N_S} = \lceil p(N_s+1) \rceil$, which corresponds to one of the nonconformity scores denoted by $t_{(J)}$. J is a random variable for the selected index.

Bootstrapped quantile estimation affects marginal coverage. To obtain the correct marginal guarantee, we should adjust the per-fold confidence level, as indicated by the following proposition.

Proposition 4.2. Consider a single-fold bootstrap sample of N_s samples from the m calibration data points. If the p-quantile is approximated using the $\lceil (N_s+1)p \rceil$-order sample of the fold, the marginal coverage probability is given as:

$$\frac{m}{m+1}p \leq \mathbb{P}(Y^{\text{test}} \in \Gamma(X^{\text{test}})) \leq \frac{m}{m+1}p + \frac{m+N_s+1}{(m+1)(N_s+1)}.$$

The choice of $p = \frac{m+1}{m}(1-\alpha)$ provides the desired marginal coverage guarantee.

The proof is given in Appendix G.1. Quantile estimation using a single fold yields a biased estimator. However, as the number of samples N_s in the fold tends to infinity, the quantile estimate converges to the p-quantile of the calibration dataset. Appendix G.1 provides details of this claim, including comments on the variance of the marginal coverage.

Asymptotics of conditional coverage for K-fold. Next, we compute the asymptotic statistics for K-fold, which, as a special case, includes the case $K = 1$. The key insight used for the derivation is that the random variable J, the index picked by each fold, has a neat Binomial CDF that only depends on N_s, p, m (see Appendix G.2 for more details). The following proposition gives the main result.

Proposition 4.3. Consider K-fold bootstrapping of the calibration dataset with m samples, and N_s samples for fold. Suppose that $t_{(J_{K,\tau})}$ is the threshold aggregated using τ-quantiles of the estimated p-quantile threshold $t_{(J_k)}$ for each fold. Suppose that $m, N_s, K \to \infty$ such that the term $\frac{\tau(1-\tau)}{f(\tau)^2}\frac{m}{N_s.K}$ converges to a constant γ_τ with $f(\tau) := \phi(\Phi^{-1}(\tau))$ as Gaussian density ϕ evaluated at τ-quantile $\Phi^{-1}(\tau)$ with Φ^{-1} as probit function. Let $p_\tau^{(N_s)} := p + \Phi^{-1}(\tau)\sqrt{p(1-p)/N_s}$. We have:

$$\sqrt{m}(F_S(t_{(J_{K,\tau})}) - p_\tau^{(N_s)}) \xrightarrow{d} \mathcal{N}(0, p(1-p)(1+\gamma_\tau)).$$

The proof is given in Appendix G.2. When the median of quantile is used with $\tau = 1/2$, we have $p_\tau^{(N_s)} = p$ and $\frac{\tau(1-\tau)}{f(\tau)^2} = \pi/2$. From the above proposition, we can see that the variance of the conditional coverage scales as:

$$\mathrm{Var}(F_S(t_{(J_{K,\tau})})) \to p(1-p)\left(\frac{1}{m} + \frac{\tau(1-\tau)}{f(\tau)^2}\frac{1}{K.N_s}\right).$$

In other words, K-fold approximation reduces the Monte Carlo sampling variance as $1/(KN_s)$, although it does not affect the term $\frac{1}{m}$ due to finite calibration samples. Single fold variance is $O(\frac{1}{m} + \frac{1}{N_s})$. Therefore, the main impact of K-fold estimation is reduced sampling variance.

5 Experimental Results

For the experiments, we follow the experimental setup in Behboodi et al. (2025), using the dataset-model pairs of MNIST-LeNet, and (FashionMNIST, CIFAR10 and CIFAR100)-ResNet20 and flipping the labels randomly with a prescribed correct probability. Different noise regimes can help to differentiate between various methods. We sample multiple pairs of calibration-test data points. For each pair, we compute the threshold and build the prediction set, which gives a sample of the prediction set size and the inclusion of correct labels. We use the joint log-probability of the group's labels as conformity scores, introduced in (Behboodi et al., 2025). For the bootstrapping and merging experiments, we take samples of m calibration data (default $m = 200$). Then, we sample $K = 1, 5, 10$ folds, each with $N_s = 200$ samples, using 250 fresh seeds, which gives us a population for sample conditional statistics. Throughout the paper and in our experiments, we run experiments for $n = 1, \ldots, 20$ and show the results for different choices of n.

TSCP of Behboodi et al. (2025) is our baseline. We denote the bootstrapping method with $K = 1$ (without folds) as **TCCP-K=1** in our tables. For K-fold methods, we implemented three merging techniques: two p-value merging methods for arithmetic and harmonic averaging, respectively, denoted by **TCCP-PA** and **TCCP-PH**, and one e-value based method using majority voting, denoted by **TCCP-E**. Our implementation uses the doubling constant for arithmetic averaging. For harmonic averaging, the constant $e \ln K$ is used. Refer to Algorithm 2 for more details.

We include only a few selected results in the main paper and discuss more ablations in Appendix H, with key takeaways highlighted per section. We encourage readers to consider all results in the main paper and appendix.

Marginal efficiency and coverage. As expected from theory, there is no general rule for how calibration set size affects average efficiency. In general, the main impact of calibration set size is to bring the marginal coverage closer to the target confidence. Our numerical results confirm this hypothesis. We include these results in App. H.2 and H.3, where we ran extensive experiments for various transductive numbers n and calibration sizes while monitoring coverage and efficiency statistics.

Bootstrapping impact. The impact of resampling using the empirical distribution ("TCCP-K=1" with $N_s = 200$ groups) is given in Tables 1 and 14 (including Bonferroni methods as baseline). The reported coverage differences in the tables are all within the standard error, so they are comparable. We see a clear trend that resampling improves the marginal efficiency for larger transductive sizes. For smaller transductive sizes ($n = 3, 6$), the efficiency changes are within the standard error and no gain can be claimed marginally.

To sanity-check the result, we can consider $n = 1$, where using all calibration samples directly should be best compared to bootstrapping and aggregation. Table 11 shows this. Bootstrapping ($K = 1$) is comparable to using all samples in marginal efficiency and coverage metrics, while K-fold merging methods for larger K underperform at $n = 1$. The main gain shows itself for larger n where using all samples is prohibitive.

Sample-conditional variance reduction. In this part, we verify the impact of folding on reducing the variance of coverage, efficiency, and the quantile conditioned on the calibration samples. The overall expected trend of variance reduction is $O(\sqrt{1/m + 1/K.N_s})$. We start by monitoring the quantile estimation variance. In our experiments, we compute the standard deviation of $\log \tau$ with different total calibration budgets m and fold numbers $K \in \{1, 5, 10\}$ for all aggregators. We have plotted it for (LeNet5, MNIST), with a confidence level $\alpha = 0.1$, the correct label with probability $= 0.9$. We have averaged the standard deviation over $n \in \{10, 13, 16, 20\}$. The fold sampling uses 250 seeds. All three plots in Figure 1 show absolute variance decreasing with m, and increasing folds reduces the variance. Table 2 reports, for each aggregator and noise level, the trend of the standard deviation of $\log \tau$ at $K = 1, 5, 10$. When m is large, the dominant term in the standard deviation is $1/\sqrt{K.N_s}$. For fixed N_s, we expect $1/\sqrt{5} = 0.447$ and $1/\sqrt{10} = 0.316$ ratio in the standard

Table 1: Efficiency and coverage comparison of TSCP and TCCP-K=1 for (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20) for transductive sizes $n = 12, 15, 18$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level. TCCP-K=1 resamples each fold from $N_s = 200$ groups.

Method	$n = 12$			$n = 15$			$n = 18$		
	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.
TSCP (MNIST)	1.3471	0.0095	0.9000	0.7704	0.0314	0.9400	0.9850	0.0116	0.9000
TCCP-K=1 (MNIST)	0.6355	0.0460	0.9400	0.6319	0.0587	0.9000	0.6018	0.0426	0.9000
TSCP (FMNIST)	1.4496	0.0478	0.9800	2.9457	0.0000	0.9000	3.1520	0.0574	0.9000
TCCP-K=1 (FMNIST)	0.9960	0.0189	0.9000	0.9421	0.0428	0.9200	0.9880	0.0167	0.8600
TSCP (CIFAR10)	2.9837	0.0948	1.0000	3.0790	0.1034	0.9200	3.2254	0.0230	0.9200
TCCP-K=1 (CIFAR10)	1.1915	0.0476	0.9000	1.1896	0.0374	0.9000	3.1195	0.0862	0.9000
TSCP (CIFAR100)	6.6439	0.0000	0.9000	6.6439	0.0000	0.9000	6.6439	0.0000	0.9200
TCCP-K=1 (CIFAR100)	6.6439	0.0000	0.8400	6.6439	0.0000	0.8400	6.6439	0.0000	0.8200

Table 2: Standard deviation of $\log \tau$, (LeNet5, MNIST), $\alpha = 0.1$, compared with prediction $1/\sqrt{5} = 0.447$ and $1/\sqrt{10} = 0.316$.

Aggregator		correct_prob= 0.9			correct_prob= 0.5		
		K=1	K=5	K=10	K=1	K=5	K=10
TCCP-E	std($\log \tau$)	0.849	0.596	0.436	0.520	0.345	0.245
	ratio vs. K=1	1.00	0.70	0.51	1.00	0.66	0.47
TCCP-PA	std($\log \tau$)	1.106	0.502	0.355	0.636	0.286	0.200
	ratio vs. K=1	1.00	0.45	0.32	1.00	0.45	0.31
TCCP-PH	std($\log \tau$)	0.849	0.742	0.640	0.520	0.431	0.371
	ratio vs. K=1	1.00	0.88	0.75	1.00	0.83	0.71

deviation reduction. TCCP-PA reduction tracks the prediction almost exactly at both noise levels. TCCP-E shrinks more weakly, and TCCP-PH shrinks the weakest of all. The impact of more folds on the variance reduction is also demonstrated separately for all merging variants and dataset-model pairs in Tables 15, 16, and 17. This consistently confirms our framework.

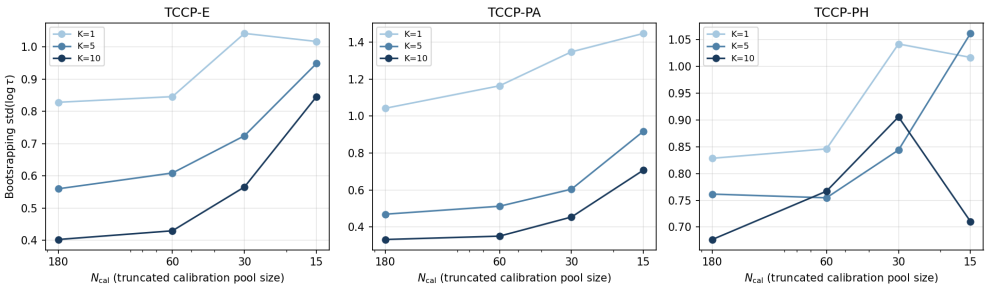

Figure 1: Standard deviation of $\log \tau$ vs. calibration budget (log-scaled, decreasing left to right), $K \in \{1, 5, 10\}$, per aggregator.

Tables 12 and 13 show the marginal coverage and efficiency for all merging variants across datasets. Although the merging methods do not show a clear winner across datasets, e-value merging methods seem to have more consistent coverage and better efficiency, particularly for larger n. We discussed that the impact of calibration sample size on final efficiency is unclear, and that bootstrapping and merging aim to reduce the variance of sample-conditional coverage.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for editorial purposes and to automate experiments. We also used generative AI to search for proof techniques in Section 4, refine our suggestions and provide alternatives, and verify the results. The main benefit of AI for the theory of the paper was pointing us to the work Hulsman (2022). We did NOT use generative AI tools for the main idea of bootstrapping and quantile merging, p-value and e-value merging methods, or the ultimate proofs for the asymptotics and statistics of conditional coverage and efficiency. We found the immediate AI proposed derivations convoluted, at times in our assessment incorrect, and not immediately useful. The overarching proof strategy of starting from the conditional coverage distribution, and using Delta rule for deriving other asymptotics are by the authors. We have reviewed all AI-assisted work. We manually checked the references from LLM search to validate their authenticity. We independently searched for the proposed techniques to credit prior work. Given the theoretical support for the paper's results, we checked the LLM-generated code by monitoring result consistency under controlled experiments and also reviewed the full code at selected points. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work is about uncertainty estimation, and does not raise ethical concern to the best of our knowledge.

REPRODUCIBILITY STATEMENT

We have provided details of the experiments we ran, the models used, and, in particular, the main algorithms in standard Algorithm format. You can use the format to generate code to reproduce the results. In particular, Algorithm 2 includes all the main steps. We added details on the experiments and how various quantities are numerically approximated in Appendix H. Similarly, for our theoretical results, we include all proofs, key lemmas and results from past work, and all assumptions in the Appendix and in various sections.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, 27 March 2023.

R R Bahadur. A note on quantiles in large samples. *Ann. Math. Stat.*, 37(3):577–580, June 1966.

Arash Behboodi, Alvaro H. C. Correia, Fabio Valerio Massoli, and Christos Louizos. Fundamental bounds on the efficiency–confidence trade-off for transductive conformal prediction. *arXiv preprint arXiv:2509.04631*, 2025.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electron. J. Stat.*, 17(2), 1 January 2023.

Morton B Brown. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987, December 1975.

Alvaro Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 12 2024.

Herbert A David and Haikady N Nagaraja. *Order Statistics: David/Order Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN, 3 edition, 21 July 2003.

Luc Devroye. Laws of the iterated logarithm for order statistics of uniform spacings. *Ann. Probab.*, 9 (5):860–867, 1 October 1981.

Guneet S Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1549–1557. PMLR, 18 April 2024.

John Duchi. Sample-conditional coverage in split-conformal prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 29 October 2025.

Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for ranking. *arXiv [cs.LG]*, 1 2025.

R. A. Fisher. Combining independent tests of significance. *American Statistician*, 2(5):30, 1948.

Matteo Gasparin and Aaditya Ramdas. Improving the statistical efficiency of cross-conformal prediction. In *Forty-second International Conference on Machine Learning*, 18 June 2025.

Matteo Gasparin, Ruodu Wang, and Aaditya Ramdas. Combining exchangeable P-values. *Proc. Natl. Acad. Sci. U. S. A.*, 122(11):e2410849122, 18 March 2025.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 4 2024.

Ulysse Gazin, Ruth Heller, Etienne Roquain, and Aldo Solari. Powerful batch conformal prediction for classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 3250–3258. PMLR, 23 April 2025.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *J. R. Stat. Soc. Series B Stat. Methodol.*, 87(4):1100–1126, 15 September 2025.

Roel Hulsman. Distribution-free finite-sample guarantees and split conformal prediction. *arXiv [stat.ML]*, 26 October 2022.

Yonghoon Lee, Eric Tchetgen Tchetgen, and Edgar Dobriban. Batch predictive inference. *arXiv [stat.ME]*, 21 September 2024.

Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 1 December 2014.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Series B Stat. Methodol.*, 76(1):71–96, 1 January 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Paulo C Marques F. Universal distribution of the empirical coverage in split conformal prediction. *Stat. Probab. Lett.*, 219(110350):110350, 4 2025.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pp. 345–356. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

R Pyke. Spacings. *J. R. Stat. Soc. Series B Stat. Methodol.*, 27(3):395–436, 1 September 1965.

Aaditya Ramdas and Ruodu Wang. *Hypothesis testing with E-values*, volume 1. Foundations and Trends® in Statistics, Emerald, 31 July 2025.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

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.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Stat. Probab. Lett.*, 184(109395): 109395, 1 May 2022.

Ryan J. Tibshirani. Conformal prediction. Advanced Topics in Statistical Learning lecture notes, University of California, Berkeley, 2023. URL https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf. Accessed 2026-08-20.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pp. 475–490. PMLR, 11 2012.

Vladimir Vovk. Transductive conformal predictors. In *IFIP Advances in Information and Communication Technology*, IFIP advances in information and communication technology, pp. 348–360. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 1 December 2020.

Vladimir Vovk and Ruodu Wang. Merging sequential e-values via martingales. *Electron. J. Stat.*, 18 (1), 1 2024.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pp. 23–39. Springer, 2016.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, 2022a.

Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *Annals of Statistics*, 50(1), 1 February 2022b.

A M Walker. A note on the asymptotic distribution of sample quantiles. *J. R. Stat. Soc. Series B Stat. Methodol.*, 30(3):570–575, 1 September 1968.

Appendix Contents

A Notation **14**

B Transductive Conformal Prediction: Background **15**

C Related works: Extended Version **15**

D Combining p-values **17**

D.1 Proof of Proposition 3.1 17

E Combining e-values **18**

E.1 Proof of Proposition 3.2 19

F Impact of Calibration Set Size on Efficiency and Coverage - Proofs and Details **20**

F.1 The conditional coverage is Beta distributed 20

F.2 Discontinuous scores and ties 22

F.3 Gaussian asymptotics for the efficiency 22

F.4 The delta method and efficiency variance 23

G Bootstrapping and K-fold analysis **23**

G.1 Single fold analysis 23

G.2 Proof of Proposition 4.3: multi-fold analysis 25

H Experimental results - follow-up **27**

H.1 Computing statistics 27

H.2 How does the calibration set size impact the marginal coverage of TSCP? 27

H.3 How does the calibration set size impact the marginal coverage for cross-conformal prediction? 28

H.4 How does the single-fold bootstrapping impact the marginal coverage and efficiency? 29

H.5 How do bootstrapping and merging impact the sample-conditional coverage and efficiency? 30

I Estimates of the score density, conditional coverage function and the coverage-efficiency function **37**

A NOTATION

Table 3: Main notation used throughout the paper.

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}$	Input space and output (label) space.
Z_1^m	Training/calibration sequence $(Z_1, \ldots, Z_m)$.
X_{m+1}^{m+n}	Batch of n test inputs $(X_{m+1}, \ldots, X_{m+n})$.
Y_{m+1}^{m+n}	Corresponding unknown test labels.
α	Target error level (confidence level $1-\alpha$).
$\Gamma^\alpha(\cdot)$	Prediction set at level α.
$\gamma_{n,m}$	Efficiency rate defined as $\log \mathbb{E}\lvert\Gamma^\alpha\rvert/n$.
P_e	Transductive joint error probability $\mathbb{P}(Y_{m+1}^{m+n} \notin \Gamma^\alpha)$.
n	Transduction size (number of test points in a batch).
m	Calibration-set size, often written as N_{cal}.
Split conformal prediction	
$S(x, y)$	Nonconformity score.
S_i	Calibration score $S(X_i, Y_i)$.
F_S	CDF of calibration scores.
f_S	Density of calibration scores.
q_α	Population $(1-\alpha)$-quantile, $q_\alpha = F_S^{-1}(1-\alpha)$.
k_m	Conformal rank index, $k_m = \lceil (m+1)(1-\alpha) \rceil$.
$S_{(k)}$	k-th order statistic of calibration scores.
$\widehat{q}_m$	Empirical conformal threshold, $\widehat{q}_m = S_{(k_m)}$.
$\Gamma^\alpha(x)$	Split conformal prediction set $\{y : S(x, y) \le \widehat{q}_m\}$.
U_m	Conditional coverage random variable, $U_m = F_S(\widehat{q}_m)$.
Efficiency analysis	
$H(q)$	Expected prediction-set size at threshold q: $H(q) = \mathbb{E}_X[\lvert\{y : S(X, y) \le q\}\rvert]$.
$g(u)$	Coverage-to-efficiency mapping $g(u) = H(F_S^{-1}(u))$.
$\mathbb{E}\lvert\Gamma^\alpha(X_{m+1})\rvert$	Marginal expected prediction-set size.
Transductive conformal prediction (TCP)	
$t(y_{m+1}^{m+n})$	Transductive score assigned to a candidate label sequence.
$t_k^{(i)}$	Calibration group score for group i in fold k.
N_s	Number of calibration groups per fold.
K	Number of resampling folds.
p-value aggregation	
$p_k(y)$	Fold-k transductive conformal p-value.
$p_k(t)$	Fold-k p-value viewed as a function of score threshold t.
$\bar{p}$	Merged p-value.
$M(\cdot)$	p-value merging function.
$M_{\phi,K}$	Generalized averaging operator used for p-value merging.
t_α	Merged score threshold satisfying $\bar{p}(t_\alpha) = \alpha$.
e-value aggregation	
$E(y)$	e-variable associated with candidate label y.
$e_k(y)$	Fold-k e-variable.
Γ_k^α	Prediction set produced by fold k.
Γ^τ	Aggregated prediction set based on voting threshold τ.
τ	Vote fraction required for inclusion.
$k^\star$	Equivalent vote count threshold.
$t_{k,\alpha}$	Individual fold threshold.
$t_{(1)} \ge \cdots \ge t_{(K)}$	Ordered fold thresholds.

B TRANSDUCTIVE CONFORMAL PREDICTION: BACKGROUND

In this section, we review the key results from transductive prediction literature.

Transductive conformal predictor was introduced in (Vovk, 2013). The proposed procedure was similar to full conformal prediction. We are given a transductive nonconformity score $S(\cdot)$, labeled data $Z_1^m = ((X_i, Y_i), i \in \{1, \ldots, m\})$, and the test sequence $X_{m+1}^{m+n} = (X_{m+1}, \ldots, X_{m+n})$.

For each candidate labels $\tilde{Y}_{m+1}^{m+n} = (\tilde{Y}_{m+1}, \ldots, \tilde{Y}_{m+n})$, we form the labeled test sequence $\tilde{Z}_{m+1}^{m+n}$. The nonconformity of the labeling $\tilde{Y}_{m+1}^{m+n}$ is measured by $S(Z_1^m, \tilde{Z}_{m+1}^{m+n})$. Note that the nonconformity score is invariant to the order of data points in the first argument, namely Z_1^m. Denote the value of this score by S_o. Leveraging the exchangeability assumption, the method considers all permutations of the pairs of test and training data in which the first m samples are unordered and the last n samples are ordered. For each permutation π, we compute the nonconformity score in a similar fashion. Denote this by S_π. We have overall $(m+n)!/m!$ of such permutations. Then, for each candidate label, the p-variable is constructed as follows:

$$p(\tilde{Y}_{m+1}^{m+n}) = \frac{|\pi : S_\pi \geq S_o|}{(m+n)!/m!}.$$

And the final prediction set is constructed as follows:

$$\Gamma^\alpha(Z_1^m, X_{m+1}^{m+n}) = \{\tilde{Y}_{m+1}^{m+n} \in \mathcal{Y}^n : p(\tilde{Y}_{m+1}^{m+n}) \geq \alpha\}.$$

Computing the score $(m+n)!/m!$ times and repeating it for each candidate label has the complexity $O(|\mathcal{Y}|^n(m+n)!/m!)$, which is clearly prohibitive for larger n.

Vovk (2013) proposed building the transductive score from classical nonconformity scores. In particular, they proposed the Bonferroni method. Bonferroni predictors have a simple expression in terms of the product of individual prediction sets:

$$\Gamma^\alpha(Z_1^m, X_{m+1}^{m+n}) = \prod_{i=1}^{n} \Gamma^{\alpha/n}(Z_1^m, X_{m+i}).$$

The Bonferroni method was shown to generate highly inefficient sets of large size Behboodi et al. (2025). Transductive split conformal prediction (TSCP) was proposed as a proper transductive prediction method with efficient implementation. The proposed score is based on using the likelihood of the candidate label computed using a model that approximates the conditional distribution: $p_\theta(\tilde{Y}_{m+1}^{m+n}|X_{m+1}^{m+n}) = \prod_{i=1}^{n} p_\theta(\tilde{Y}_{m+i}|X_{m+i})$. For conformal calibration of the score, we organize the calibration dataset into groups of n samples and compute the $(1-\alpha)$-quantile of the transductive nonconformity scores S_i.

Using the quantile, we can use the classical split conformal prediction method to build the final prediction set. Denote the quantile by τ. For each candidate label, we need to compute $p_\theta(\tilde{Y}_{m+1}^{m+n}|X_{m+1}^{m+n})$ and compare with τ to decide on inclusion. This is still computationally expensive (for CIFAR100 and $n = 10$, we should repeat this 10^{20} times). Therefore, the authors proposed a dynamic programming scheme to prune some candidates. This is done by tree search over $\log p_\theta(\tilde{Y}_{m+i}|X_{m+i})$ and comparing with $\log \tau$. In our paper, we have implemented the algorithm based on the details provided in Behboodi et al. (2025).

As a reminder, the thresholding based on empirical $(1-\alpha)$-quantile q_α is equivalent to computing the p-value as follows:

$$s \leq q_\alpha \iff \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(S_i \geq s) > \alpha.$$

The term $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(S_i \geq s)$ constructed as a p-value is called conformal transducer in Vovk et al. (2022a).

C RELATED WORKS: EXTENDED VERSION

Split Conformal Prediction. A central theme in conformal prediction research is improving the efficiency of prediction sets while preserving finite-sample validity. Split conformal prediction

(SCP) separates model fitting from calibration by training a predictor once, estimating a calibration quantile on a held-out calibration set, and constructing prediction sets using this quantile. Under exchangeability, SCP provides exact finite-sample *marginal* coverage guarantees in a distribution-free manner (Papadopoulos et al., 2002; Lei et al., 2018; Vovk et al., 2022a; Angelopoulos & Bates, 2023). In the terminology of Vovk et al. (2022a), these methods are also known as *inductive conformal predictors*. Numerous variants have been developed to improve statistical efficiency through alternative conformity scores, including adaptive prediction sets (APS) and related methods for classification (Romano et al., 2020; Angelopoulos & Bates, 2023).

Cross-conformal prediction and p-value merging. Cross-conformal prediction and its variants aim to improve the utilization of calibration data while avoiding the computational expense of full conformal prediction (Vovk et al., 2022a). These methods partition the data into multiple folds, construct conformal predictors on each fold, and aggregate the resulting evidence. Recent developments have connected such aggregation procedures to the broader literature on combining conformal p-values and e-values (Vovk et al., 2022b; Vovk & Wang, 2024; Gasparin et al., 2025; Vovk & Wang, 2020; Fisher, 1948; Brown, 1975). Our proposed method builds on this line of work. However, unlike existing approaches that merge prediction sets, p-values, or e-values directly, we show that for split conformal prediction the aggregation can be performed entirely at the quantile level, reducing the final procedure to a standard split conformal predictor with a merged calibration quantile.

Transductive conformal prediction. Transductive conformal prediction (TCP) was introduced to provide joint guarantees over multiple test samples (Vovk, 2013; Vovk et al., 2022a). A straightforward approach applies a Bonferroni correction to individual conformal predictors, but the resulting prediction sets become increasingly conservative as the transductive batch size grows. More efficient transductive procedures have been proposed in both general and structured settings (Vovk, 2013; Gazin et al., 2024; Fermanian et al., 2025; Behboodi et al., 2025). In particular, Transductive Split Conformal Prediction (TSCP) (Behboodi et al., 2025) introduced transductive conformity scores and efficient dynamic-programming inference for large output spaces. The main assumption is that the calibration data are already available in groups matching the transductive batch size.

Conformal Efficiency and Conditional Analysis. The efficiency of conformal prediction has been studied extensively, both from the perspective of efficiency criteria and the characterization of optimal conformal predictors (Vovk et al., 2022a; Sadinle et al., 2019; Correia et al., 2024; Vovk et al., 2016; Gasparin & Ramdas, 2025; Behboodi et al., 2025; Lei & Wasserman, 2014; Lei, 2014). Common measures of efficiency include prediction-set cardinality in classification and interval length in regression, although several alternative criteria have also been proposed (Vovk et al., 2022a; 2016). Recent work has analyzed finite-sample expected prediction-set size directly and established performance bounds for conformal predictors (Dhillon et al., 2024).

Beyond marginal validity, a growing body of work studies stronger notions of conditional, sample-conditional, and related validity guarantees, together with their implications for efficiency (Vovk, 2012; Duchi, 2025; Gibbs et al., 2025). For split conformal prediction in particular, the probabilistic behavior of coverage and efficiency has been investigated in (Tibshirani, 2023; Vovk et al., 2022a; Vovk, 2012; Hulsman, 2022; Marques F., 2025; Bian & Barber, 2023). In this literature, the sample-conditional coverage of split conformal prediction has been shown to follow a Beta distribution, a result first noted in (Vovk, 2012) and further developed in subsequent works.

Simultaneous batch inference from unpartitioned calibration pools. A parallel line of work constructs joint prediction sets for a batch of test points directly from a pool of *individual* calibration observations, thereby avoiding the m/n partitioning penalty by a route other than resampling. Gazin et al. (2025) work with the n per-test-point conformal p-values and combine them through the Simes inequality, obtaining a batch prediction set that is uniformly more powerful than the Bonferroni product of Vovk (2013); they further give a general recipe for arbitrary p-value combining functions and a shortcut for computing the resulting bounds. Lee et al. (2024) instead target prediction sets for a *function* of the batch (mean, median, quantile), with simultaneous coverage as one special case; validity rests on quantiles of a rank-ordering function over subsets of ranks, which is NP-hard in general but tractable by dynamic programming for compositionally structured batch scores. Both are valid on unpartitioned calibration data and are complementary rather than competing with our contribution, for three reasons. First, they fix the calibration-efficiency problem by changing the

inference rule, whereas we keep the TSCP inference rule, which is a single joint likelihood score over label sequences, thresholded by one scalar quantile. We change only how calibration *groups* are formed. Our merging results therefore apply to any transductive split predictor whose acceptance region is monotone in a group score, including, in principle, a Simes-combined one. Second, their constructions are per-coordinate, a vector of individual p-values, or a compositional batch score. TSCP's joint log-likelihood score is not decomposable into n exchangeable univariate p-values, so Simes-type combination does not directly apply to it. Third, and most importantly for the present paper, neither addresses the *sample-conditional* variability of the calibrated threshold, which is the quantity our K-fold aggregation is designed to stabilize. On the choice of resampling scheme, subsampling without replacement (Solari & Djordjilović, 2022) and the recent efficiency-improved cross-conformal variants of Gasparin & Ramdas (2025) preserve exchangeability and so avoid the quantile bias corrected in Proposition 4.2. We adopt sampling with replacement because, at a fixed pool of m points, it is the only scheme that permits N_s to be chosen *independently of* m and n, and does not restrict the number of disjoint groups at $\lfloor m/n \rfloor$. Benchmarking against subsampling without replacement at matched N_s would isolate how much of the observed gain is attributable to data reuse rather than to replacement, and we leave this to future work.

D COMBINING P-VALUES

A p-variable is a random variable P that satisfies $\mathbb{P}(P \leq \epsilon) \leq \epsilon$. It is a fundamental tool in statistical hypothesis testing and is intuitively understood as evidence for the null hypothesis, i.e., the small p-value amounts to rejecting the null hypothesis.

In conformal prediction, the null hypothesis is the exchangeability of the pair of test sample and its candidate label with the dataset samples. Therefore, a label is included in the prediction set if the null hypothesis is not rejected, which means the corresponding p-value is large. (Vovk & Wang, 2020) discussed a class of functions that can combine multiple p-values with arbitrary dependence into a valid p-value. These functions are called merging functions.

Definition D.1 (From Vovk & Wang (2020))**.** A merging function is an increasing Borel function $M : [0,1]^K \to [0,\infty)$ such that for any marginally $[0,1]$-uniformly distributed random variables $U_1, \ldots, U_K$ with arbitrary joint distribution, $M(U_1, \ldots, U_K)$ is a p-variable.

They used the general notion of averaging for merging:

$$M_{\phi,K} := \psi\left(\frac{\phi(p_1) + \cdots + \phi(p_K)}{K}\right), \tag{12}$$

where $\phi : [0,1] \to [-\infty, \infty]$ is a continuous, strictly monotonic function, and ψ is its inverse.

The averaging function $M_{\phi,K}$ needs a scaling factor to yield a p-variable. The scaling factor follows from the following theorem.

Theorem D.2 (Theorem 2 Vovk & Wang (2020))**.** *Consider a continuous strictly monotonic function $\phi : [0,1] \to [-\infty, \infty]$, and suppose that ϕ is integrable ($\int_0^1 |\phi(t)|\mathrm{d}t < \infty$). Then, for any $K \in \{2, 3, \ldots\}$ and any $\epsilon > 0$, we have:*

$$\mathbb{P}\left(M_{\phi,K}(p_1, \ldots, p_K) \leq \psi\left(\frac{1}{\epsilon}\int_0^{\epsilon} \phi(u)\mathrm{d}u\right)\right) \leq \epsilon \tag{13}$$

$$M_{r,K}(p_1, \ldots, p_K) := \left(\frac{p_1^r + \cdots + p_K^r}{K}\right)^{1/r} \tag{14}$$

D.1 PROOF OF PROPOSITION 3.1

Proof. Note that the p-value $\bar{p}(y_{m+1}^{m+n})$ depends on y_{m+1}^{m+n} only through the nonconformity score $t(y_{m+1}^{m+n})$, which is the same across folds. Let t be any threshold. Using any merging function, we

Algorithm 2 TCCP with p-value merging via threshold scan and log-probability scores

Require: calibration pool of N_{cal} labeled samples; classifier $p(\cdot \mid \cdot)$; number of folds K; group size n; groups per fold N_s; level α; merging rule $\in \{\text{ARITH}, \text{HARM}\}$ (harmonic calibrator $c \in \{e \ln K, \ln K\}$, 'formal' vs. 'practical')

Ensure: threshold $\log \hat{\tau}$ such that $\Gamma^\alpha(x_{m+1}^{m+n}) = \{y_{m+1}^{m+n} \in \mathcal{Y}^n : t(y_{m+1}^{m+n}) > \log \hat{\tau}\}$

for $k = 1, \ldots, K$ **do**
  draw N_s groups of size n, with replacement, from the calibration pool ▷ fold k
  for $i = 1, \ldots, N_s$ **do**
    $t_k^{(i)} \leftarrow \sum_{j \in \text{group } i} \log p(y_j \mid x_j)$ ▷ group score
  end for
end for
$B \leftarrow$ sorted union of $\{t_k^{(i)} : k \in [K],\ i \in [N_s]\}$
for $t \in B$, in increasing order **do**
  $p_k(t) \leftarrow \dfrac{1 + \#\{i : t_k^{(i)} \le t\}}{N_s + 1}$ for each $k = 1, \ldots, K$ ▷ Eq. eq. 2
  if merging rule $=$ ARITH **then**
    $\bar{p}(t) \leftarrow \dfrac{2}{K} \sum_{k=1}^{K} p_k(t)$ ▷ doubling constant (Vovk & Wang, 2020)
  else
    $\bar{p}(t) \leftarrow \frac{cK}{\sum_{k=1}^{K} p_k(t)^{-1}}$ ▷ harmonic mean, calibrator c (Vovk & Wang, 2020)
  end if
  if $\bar{p}(t) > \alpha$ **then**
    return $\log \hat{\tau} \leftarrow t$ ▷ $\bar{p}(\cdot)$ is non-decreasing in t: first crossing is the threshold
  end if
end for

Table 4: Examples of merging functions

range of r	**merging function**	**inculdes:**
$r = \infty$	$M_{\infty,K} := \max(p_1, \ldots, p_K)$	maximum
$r = -\infty$	$K\, M_{-\infty,K} := K \min(p_1, \ldots, p_K)$	a.k.a. Bonferroni
$r = 0$	$e\, M_{0,K} = e \left(\prod_{i=1}^{K} p_i\right)^{1/K}$	geometric
$r = -1$	$e(\ln K) M_{-1,K} = e(\ln K) \left(K \big/ \sum_{k=1}^{K} p_k^{-1}\right)$	harmonic
$r \in [K-1, \infty)$	$K^{1/r} M_{r,K}$	
$r \in \left[\frac{1}{K-1},\ K-1\right]$	$(r+1)^{1/r} M_{r,K}$	arithmetic
$r \in (-1, \infty]$	$(r+1)^{1/r} M_{r,K}$	arithmetic
$r \in (-\infty, -1)$	$\dfrac{r}{r+1} K^{1+1/r} M_{r,K}$	

obtain the final p-value $\bar{p}(t)$ for a given t. Note that $p_k(t)$ is non-increasing in t, therefore, we can find a threshold t_α below which $\bar{p}(t)$ is bigger than or equal to the confidence level α:

$$\bar{p}(t) > \alpha \iff t < t_\alpha,$$

where $t_\alpha = \bar{p}^{-1}(\alpha)$. The final joint prediction set in eq. 3 includes y_{m+1}^{m+n} if and only if its calibrated combined p-value $\bar{p}(t(y_{m+1}^{m+n}))$ exceeds α. However, this means that the nonconformity score $t(y_{m+1}^{m+n})$ is less than t_α. □

E COMBINING E-VALUES

In what follows, we assume the output space $\mathcal{Y}$ is a general abstract probability space, which includes both classical and transductive cases as special cases. We denote the true output by the random variable Γ. The first starting point is the notion of e-value. We refer the reader to the monograph Ramdas & Wang (2025) for a mathematical and detailed treatment of e-values.

Definition E.1 (e-variable and e-value)**.** An e-variable E for a class of probability distributions $\mathcal{P}$ is a $[0, \infty]$-valued random variable satisfying $\mathbb{E}_{\mathbb{P}}[E] \leq 1$ for all $\mathbb{P} \in \mathcal{P}$. The value of the e-variable is called an e-value.

For the space $\mathcal{Y}$, denote the e-variable by $E(y)$ for $y \in \mathcal{Y}$. Then, it can be used to construct the confidence set as follows:

$$\Gamma^{\alpha} = \{y \in \mathcal{Y} : E(y) < \frac{1}{\alpha}\}.$$

Then, using Markov's inequality, we can see that:

$$\mathbb{P}(Y \notin \Gamma^{\alpha}) = \mathbb{P}(E(Y) \geq \frac{1}{\alpha}) \leq \alpha \mathbb{E}(E(Y)) \leq \alpha.$$

So, the corresponding set satisfies the confidence level α, namely:

$$\mathbb{P}(Y \in \Gamma^{\alpha}) \geq 1 - \alpha.$$

On the other hand, given any confidence set Γ^{α} satisfying the above condition, define $E(y) := \frac{1}{\alpha}\mathbf{1}(y \notin \Gamma^{\alpha})$. We have:

$$\mathbb{E}(E(Y)) = \frac{1}{\alpha}\mathbb{P}(Y \notin \Gamma^{\alpha}) \leq \frac{\alpha}{\alpha} = 1.$$

Therefore, we can build an e-variable from a confidence set. Figure 2 gives the details.

Now suppose that we have constructed K confidence interval Γ_i^{α}, $i \in [K]$, all satisfying the confidence level α. Consider the following majority voting scheme:

$$\Gamma^{\tau} := \left\{ y \in \mathcal{Y} : \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}\{y \in \Gamma_i^{\alpha}\} > \tau \right\}.$$

The following theorem bounds the confidence level of majority voting.

Theorem E.2 (Theorem 13.10 Ramdas & Wang (2025))**.** *Let Γ_i^{α}, $i \in [K]$, be prediction sets with confidence level α. We have:*

$$\mathbb{P}(Y \in \Gamma^{\tau}) \geq 1 - \frac{\alpha}{1 - \tau}.$$

In particular, for $\tau = 1/2$, the prediction set Γ^{τ} satisfies the confidence level 2α.

We can extend the theorem to the case where each set is constructed with a different confidence level:

$$\Gamma_{\boldsymbol{\alpha}}^{\tau} := \left\{ y \in \mathcal{Y} : \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}\{y \in \Gamma_i^{\alpha_i}\} > \tau \right\}.$$

We have the following corollary:

Corollary 1. Consider the prediction set $\Gamma_{\boldsymbol{\alpha}}^{\tau}$. Then,

$$\mathbb{P}(Y \in \Gamma_{\boldsymbol{\alpha}}^{\tau}) \geq 1 - \frac{\sum_{i=1}^{K} \alpha_i}{K(1 - \tau)}.$$

Proof. The proof follows the simple application of Markov's inequality:

$$\mathbb{P}(Y \notin \Gamma_{\boldsymbol{\alpha}}^{\tau}) = \mathbb{P}\left(\frac{1}{K}\sum_{i=1}^{K} \mathbf{1}\{Y \notin \Gamma_i^{\alpha_i}\} \geq 1 - \tau \right) \leq \frac{\sum_{i=1}^{K} \alpha_i}{K(1 - \tau)}.$$

□

E.1 Proof of Proposition 3.2

Proof. Fold k's inclusion indicator is monotone in $t(y)$: $\mathbf{1}\{y \in \Gamma_k^{\alpha}\} = \mathbf{1}\{t(y) < t_{k,\alpha}\}$. Hence the vote count is given by $\sum_k \mathbf{1}\{y \in \Gamma_k^{\alpha}\} = \#\{k : t(y) < t_{k,\alpha}\}$. Then:

$$\sum_k \mathbf{1}\{y \in \Gamma_k^{\alpha}\} > \tau K \iff \#\{k : t(y) < t_{k,\alpha}\} > \tau K.$$

But this means that $t(y)$ is smaller than or equal to top $\lfloor \tau K \rfloor + 1$ thresholds, which proves the results. □

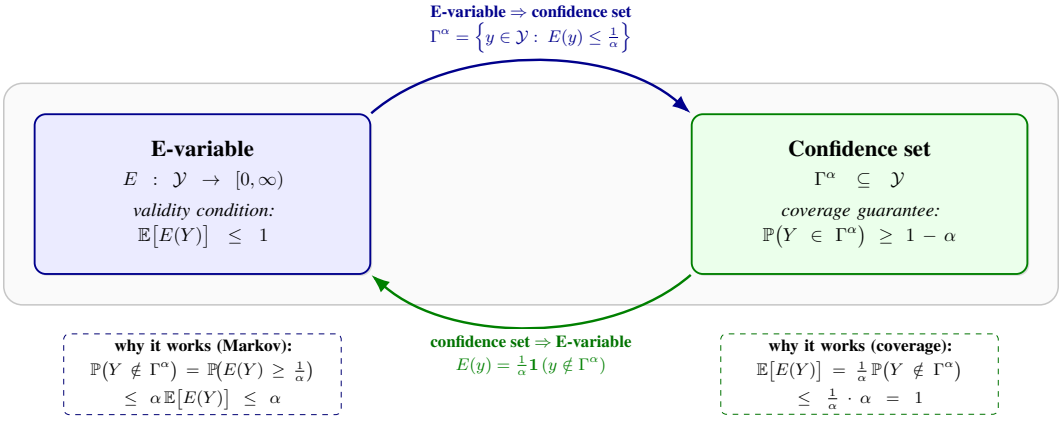

Figure 2: Duality between e-variables and confidence sets: an e-variable induces a valid confidence set via $\Gamma^\alpha = \{y : E(y) \le 1/\alpha\}$, and conversely any valid set yields an e-variable $E(y) = \frac{1}{\alpha}\mathbf{1}(y \notin \Gamma^\alpha)$.

F Impact of Calibration Set Size on Efficiency and Coverage - Proofs and Details

Table 5 summarizes the results. We provide the technical details and proofs in this section.

F.1 The conditional coverage is Beta distributed

Assume F_S is continuous. Define

$$V_i = F_S(S_i), \qquad i = 1, \dots, m.$$

By the probability integral transform, $V_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. Since a monotone transformation preserves order,

$$U_m = F_S(S_{(k_m)}) = V_{(k_m)}.$$

For $0 \le u \le 1$, the event $V_{(k_m)} \le u$ is equivalent to at least k_m of the m uniforms lying in $[0, u]$. Hence

$$\mathbb{P}(U_m \le u) = \mathbb{P}\{\text{Binomial}(m, u) \ge k_m\} \tag{15}$$

$$= \sum_{j=k_m}^{m} \binom{m}{j} u^j (1-u)^{m-j}. \tag{16}$$

Differentiation yields

$$f_{U_m}(u) = \frac{m!}{(k_m - 1)!(m - k_m)!} u^{k_m - 1}(1-u)^{m-k_m},$$

so that

$$\boxed{U_m \sim \text{Beta}(k_m, m + 1 - k_m).} \tag{17}$$

This result can be found in Standard order-statistic literature for the uniform distribution (David & Nagaraja, 2003; Walker, 1968). The mean and variance are

$$\mathbb{E}[U_m] = \frac{k_m}{m+1} = p_m, \qquad \text{Var}(U_m) = \frac{k_m(m + 1 - k_m)}{(m+1)^2(m+2)}. \tag{18}$$

Therefore $\text{Var}(U_m) = \alpha(1-\alpha)/m + o(m^{-1})$.

F.1.1 Gaussian asymptotics for Beta distribution

Next, we consider the asymptotics of the Beta distribution. We first state a useful lemma for the rest of the paper.

Table 5: Impact of calibration size m on coverage and efficiency.

Case	Impact of calibration size m
Marginal coverage	$$\Pr\{Y_{m+1} \in \Gamma^\alpha(X_{m+1})\} = \frac{k_m}{m+1}, \qquad k_m = \lceil (m+1)(1-\alpha) \rceil.$$ Depends only on (m, α) and is distribution-free. The excess above $1-\alpha$ is smaller than $1/(m+1)$ and vanishes as $m \to \infty$.
Conditional coverage	$$U_m = F_S(\widehat{q}_m) \sim \mathrm{Beta}(k_m, m+1-k_m).$$ $$\mathbb{E}[U_m] = \frac{k_m}{m+1}, \qquad \mathrm{Var}(U_m) \approx \frac{\alpha(1-\alpha)}{m}.$$ Depends on (m, α) but not on the score density f_S. Larger m reduces calibration-to-calibration variability.
Approximate quantile	$$\widehat{q}_m \to q_\alpha = F_S^{-1}(1-\alpha), \qquad \mathrm{Var}(\widehat{q}_m) \approx \frac{\alpha(1-\alpha)}{m\, f_S(q_\alpha)^2}.$$ The $1/m$ decay is scaled by the score density $f_S(q_\alpha)$; small $f_S(q_\alpha)$ implies larger quantile uncertainty.
Conditional efficiency	$$H(\widehat{q}_m) \to H(q_\alpha),$$ with $$\mathrm{Var}\big(H(\widehat{q}_m)\big) \approx \frac{\alpha(1-\alpha)}{m} \left(\frac{H'(q_\alpha)}{f_S(q_\alpha)} \right)^2.$$ Calibration-induced variability decreases as $1/m$, but is amplified when $H'(q_\alpha)$ is large or $f_S(q_\alpha)$ is small.
Marginal (expected) efficiency	$$\mathbb{E}\lvert\Gamma^\alpha(X_{m+1})\rvert = \mathbb{E}[g(U_m)], \qquad g(u) = H(F_S^{-1}(u)).$$ Unlike coverage, there is generally no monotone dependence on m. The effect of m is governed by: (i) the effective coverage level $k_m/(m+1)$, (ii) the Beta variability $\mathrm{Var}(U_m) \approx \alpha(1-\alpha)/m$, and (iii) the curvature of g: $$\mathbb{E}[g(U_m)] = g(1-\alpha) + \frac{1}{2} g''(1-\alpha) \frac{\alpha(1-\alpha)}{m} + o(m^{-1}).$$ Hence finite-m corrections depend on both F_S and H.

Lemma F.1 (Central-order-statistic Gaussian limit)**.** *Let* $k_m/(m+1) \to p \in (0,1)$ *and* $U_m \sim \mathrm{Beta}(k_m, m+1-k_m)$. *If*

$$\frac{k_m - (m+1)p}{\sqrt{m}} \to c,$$

then

$$\sqrt{m}(U_m - p) \xrightarrow{d} \mathcal{N}\big(c, p(1-p)\big).$$

In particular, for conformal prediction, we have $k_m = \lceil (m+1)p \rceil$, $p = 1-\alpha$, $c = 0$ *and*

$$\sqrt{m}(U_m - p) \xrightarrow{d} \mathcal{N}(0, \alpha(1-\alpha)).$$

Proof. First, we can write the Beta random variable as the ratio of Gamma random variables:

$$U_m \overset{d}{=} \frac{A_m}{A_m + B_m}, \qquad A_m \sim \mathrm{Gamma}(k_m, 1), \quad B_m \sim \mathrm{Gamma}(m+1-k_m, 1), \tag{19}$$

with A_m and B_m independent. Write $l_m = m + 1 - k_m$. Using this representation, we first find the asymptotics of the Gamma variables and combine them to obtain the final result. Gamma variables with integer shape are sums of independent unit-rate exponentials (Pyke, 1965; Devroye, 1981),

$$A_m = \sum_{i=1}^{k_m} E_i, \qquad B_m = \sum_{j=1}^{l_m} E'_j,$$

where all exponentials are independent, have mean one, and variance one. The joint CLT gives

$$\left(\frac{A_m - k_m}{\sqrt{m}}, \frac{B_m - l_m}{\sqrt{m}} \right) \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} p & 0 \\ 0 & 1-p \end{bmatrix}\right).$$

Apply the multivariate delta method to $h(a, b) = a/(a + b)$ around (k_m, l_m). Its gradient at the normalized limit $(p, 1 - p)$ is $(1 - p, -p)$. The resulting variance is

$$(1-p)^2 p + p^2(1-p) = p(1-p).$$

The deterministic centering contributes c. For conformal ranks, $k_m - (m+1)p \in [0, 1)$, so $c = 0$. □

Classical proofs of sample-quantile normality can be found in (Walker, 1968) using the binomial CLT.

The exact moments in eq. 18 already suggest a similar scaling:

$$\mathbb{E}[U_m] = \alpha + O(m^{-1}), \qquad m\mathrm{Var}(U_m) \to \alpha(1-\alpha).$$

F.2 Discontinuous scores and ties

The exact identity $F_S(S_i) \sim \mathrm{Unif}(0, 1)$ requires continuity of F_S. If F_S has atoms, as can occur for discrete or rounded scores, then $F_S(S_i)$ is not exactly uniform, order-statistic ties have positive probability, and the simple Beta distribution for $F_S(\widehat{q}_m)$ generally fails. Conservative marginal coverage can still be obtained through the standard rank construction with appropriate weak inequalities, but equality and the continuous-score upper bound require care (Tibshirani, 2023). A randomized probability integral transform restores exact uniformity, similar to smoothed conformal predictors Vovk et al. (2022a). Let $W_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0, 1)$ be independent and define

$$V_i = F_S(S_i^-) + W_i\big(F_S(S_i) - F_S(S_i^-)\big), \tag{20}$$

where $F_S(s^-) = \mathbb{P}(S < s)$. Then $V_i \sim \mathrm{Unif}(0, 1)$ even when F_S has jumps. A conformal procedure with compatible randomized tie-breaking recovers exact rank uniformity and the associated Beta/order-statistic description on the randomized probability scale. Tibshirani (2023) describe randomized empirical-CDF conformalization.

F.3 Gaussian asymptotics for the efficiency

Assume that F_S is differentiable at q_α, with

$$0 < f_S(q_\alpha) < \infty.$$

Since

$$\widehat{q}_m = Q(U_m), \qquad Q = F_S^{-1},$$

and the inverse-function theorem gives $Q'(1-\alpha) = \frac{1}{f_S(q_\alpha)}$, the delta method and Lemma F.1 imply

$$\boxed{\sqrt{m}(\widehat{q}_m - q_\alpha) \xrightarrow{d} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f_S(q_\alpha)^2}\right).} \tag{21}$$

This is the classical Gaussian asymptotic for a sample quantile (Bahadur, 1966; Walker, 1968).

The same limit follows from the Bahadur representation

$$\widehat{q}_m - q_\alpha = \frac{(1-\alpha) - \widehat{F}_m(q_\alpha)}{f_S(q_\alpha)} + o_{\mathbb{P}}(m^{-1/2}), \tag{22}$$

where

$$\widehat{F}_m(q_\alpha) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{S_i \le q_\alpha\}.$$

The summands are Bernoulli$(1-\alpha)$, so

$$\sqrt{m}\{\widehat{F}_m(q_\alpha) - (1-\alpha)\} \xrightarrow{d} \mathcal{N}(0, \alpha(1-\alpha)),$$

and eq. 21 follows by Slutsky's theorem. The factor $1/f_S(q_\alpha)$ is the correction to the Gaussian asymptotics for conditional coverage. A small density means a locally flat CDF and hence large quantile uncertainty.

F.4 The delta method and efficiency variance

The one-dimensional delta method states that if

$$\sqrt{m}(T_m - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

and g is differentiable at θ, then

$$\sqrt{m}\{g(T_m) - g(\theta)\} \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

It follows from the Taylor expansion

$$g(T_m) = g(\theta) + g'(\theta)(T_m - \theta) + o(m^{-1/2}).$$

Apply this first to $Q = F_S^{-1}$ to obtain eq. 21. Apply it again to H:

$$H(\widehat{q}_m) = H(q_\alpha) + H'(q_\alpha)(\widehat{q}_m - q_\alpha) + o_{\mathbb{P}}(m^{-1/2}).$$

Therefore,

$$\sqrt{m}\{H(\widehat{q}_m) - H(q_\alpha)\} \xrightarrow{d} \mathcal{N}\left(0, \frac{H'(q_\alpha)^2\alpha(1-\alpha)}{f_S(q_\alpha)^2}\right), \tag{23}$$

and the calibration-induced variance is approximately

$$\boxed{\mathrm{Var}_{\mathcal{D}_{\mathrm{cal}}}\{H(\widehat{q}_m)\} \approx \frac{\alpha(1-\alpha)}{m}\left(\frac{H'(q_\alpha)}{f_S(q_\alpha)}\right)^2.} \tag{24}$$

Equivalently, since $g(u) = H(Q(u))$,

$$g'(1-\alpha) = \frac{H'(q_\alpha)}{f_S(q_\alpha)}, \qquad \mathrm{Var}\{g(U_m)\} \approx g'(1-\alpha)^2\frac{\alpha(1-\alpha)}{m}.$$

G Bootstrapping and K-fold analysis

Consider the calibration dataset with $N_{\mathrm{cal}} = m$ samples with nonconformity scores $t_1, \ldots, t_m$. The ordered scores are denoted as:

$$t_{(1)} \le \cdots \le t_{(m)}.$$

G.1 Single fold analysis

On the bias of quantile estimation. Suppose that we sample a fold with N_s samples with replacement and compute the p-quantile of the fold, denoted by $\hat{t}_{p,N_s}$. Suppose that the approximated quantile corresponds to the index $t_{(J)}$ from the ordered quantiles:

$$\hat{t}_{p,N_s} = t_{(J)}.$$

Conditioned on the calibration set, the event that J is smaller than or equal to i is the event that at least $k_{p,N_s} = \lceil p(N_s+1)\rceil$ of the samples in the fold are from $t_{(1)}, \ldots, t_{(i)}$. This is a binomial

random variable with success probability i/m when the samples are uniformly distributed. Denote this random variable by $\text{Bin}_{N_s,i/m}$. Therefore:

$$\mathbb{P}(J \leq i) = \mathbb{P}(\text{Bin}_{N_s,i/m} \geq k_{p,N_s}) = \sum_{\ell=k_{p,N_s}}^{N_s} \binom{N_s}{\ell} \left(\frac{i}{m}\right)^{\ell} \left(1 - \frac{i}{m}\right)^{N_s-\ell}.$$

The CDF above yields the probability mass function for J. Note that the probability that N_s Bernoulli trials of probability p have at least k_{p,N_S} successes is identical to the probability that a Beta random variable $\text{Beta}(k_{p,N_s}, N_s + 1 - k_{p,N_s})$ is smaller than or equal to p. Therefore, we have:

$$J = \lceil mB \rceil$$

where B is Beta random variable distributed according to $\text{Beta}(k_{p,N_s}, N_s + 1 - k_{p,N_s})$.

Define $w_i := \mathbb{P}(J = i)$. We have:

$$\mathbb{E}[\hat{t}_{p,N_s}|t_1, \ldots, t_m] = \sum_{i=1}^{m} w_i t_{(i)}, \qquad \mathbb{E}[\hat{t}^2_{p,N_s}|t_1, \ldots, t_m] = \sum_{i=1}^{m} w_i t^2_{(i)}.$$

This is the conditional expectation and moment; however, the marginal expectation of the quantile estimate is:

$$\mathbb{E}[\hat{t}_{p,N_s}] = \sum_{i=1}^{m} w_i \mathbb{E}[t_{(i)}].$$

Therefore, **the quantile estimate is not, in general, unbiased.**

As $N_s \to \infty$, the binomial random variable $\text{Bin}_{N_s,i/m}$ will converge to its mean $N_s(i/m)$. The CDF of J will be step-function around i such that $N_s(i/m) = k_{p,N_s}$, which means that J as well converges to $k_{p,N_s}m/N_s \to pm$. To make the argument more precise, note that as $N_s \to \infty$ for $B \sim \text{Beta}(k_{p,N_s}, N_s + 1 - k_{p,N_s})$, we have:

$$\sqrt{N_s}(B - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Therefore:

$$\sqrt{N_s}(mB - mp) \xrightarrow{d} \mathcal{N}(0, m^2 p(1-p)). \tag{25}$$

So $J = \lceil mB \rceil$ will converge to $\lceil mp \rceil$. Below, we summarize this finding informally.

A single-fold quantile estimate is in general biased; however, as the number of samples N_s in the fold tends to infinity, the quantile estimate converges to the p-quantile of the calibration dataset.

Proof of Proposition 4.2 - Impact of the marginal coverage. We characterize the impact on the marginal coverage. Consider the following:

$$\mathbb{P}(t_{m+1} \leq t_{(j)}) = F_S(t_{(j)}),$$

where t_{m+1} is the score of the new test sample. We know from the analysis of Section F that $F_S(t_{(j)})$ is distributed according to $\text{Beta}(j, m + 1 - j)$. The conditional coverage in our case is $F_S(t_{(J)})$, therefore the marginal coverage is given as:

$$\mathbb{E}[F_S(t_{(J)})] = \sum_{j=1}^{m} \mathbb{P}(J = j)\mathbb{E}[F_S(t_{(j)})] = \sum_{j=1}^{m} \mathbb{P}(J = j)\frac{j}{m+1} = \frac{\mathbb{E}[J]}{m+1}.$$

The CDF $\mathbb{P}(J \leq i)$ has the form of Beta distribution $\text{Beta}(k_{p,N_s}, N_s + 1 - k_{p,N_s})$ evaluated as i/m. In other words, we have $J = \lceil mB \rceil$, where B is a Beta random variable. We have:

$$\frac{mk_{p,N_s}}{N_s + 1} \leq \mathbb{E}[J] \leq \frac{mk_{p,N_s}}{N_s + 1} + 1$$

Using this inequality, and that $k_{p,N_s} = \lceil (N_s + 1)p \rceil$, we can see that the marginal coverage is given as:

$$\frac{m}{m+1}p \leq \mathbb{P}(Y^{\text{test}} \in \Gamma(X^{\text{test}})) \leq \frac{m}{m+1}p + \frac{m + N_s + 1}{(m+1)(N_s+1)}.$$

Choosing $p = \frac{m+1}{m}(1 - \alpha)$ gives the desired coverage and completes the proof of Proposition 4.2.

Impact on the variance. We know $J = \lceil mB \rceil$ where B is another Beta random variable distributed according to $\text{Beta}(k_{p,N_s}, N_s + 1 - k_{p,N_s})$. We know that $F_S(t_{(i)})$ is distributed according to $\text{Beta}(i, m + 1 - i)$.

We used these properties to compute the marginal coverage. We can use it again for the conditional coverage variance. Conditioned on $J = i$, the variance of $F_S(t_{(i)})$ is equal to:

$$\text{Var}(F_S(t_{(J)})|J = i) = \frac{i(m+1-i)}{(m+1)^2(m+2)},$$

Therefore, using the law of total variance:

$$\text{Var}(F_S(t_{(J)})) = \mathbb{E}[\text{Var}(F_S(t_{(J)})|J)] + \text{Var}(\mathbb{E}[F_S(t_{(J)})|J]).$$

We have:

$$\text{Var}(F_S(t_{(J)})) = \mathbb{E}\left(\frac{J(m+1-J)}{(m+1)^2(m+2)}\right) + \frac{\text{Var}(J)}{(m+1)^2}.$$

As $m, N_s \to \infty$, we can informally analyze the expression. The random variable J/m can be approximated by the random variable B and asymptotically tends to a normal distribution of variance $p(1-p)/N_s$. For these values, $J \approx mp$, and the $F_S(t_{(J)})$ is also normal with the variance $\approx p(1-p)/m$. Therefore, the variance of the conditional coverage has two components: the first one relates to the total calibration size, and the second one relates to the bootstrap sampling. The total variance of the conditional coverage is $O(p(1-p)(\frac{1}{m} + \frac{1}{N_s}))$.

G.2 Proof of Proposition 4.3: multi-fold analysis

We start by restating the general form of the proposition 4.3.

Proposition G.1. Consider K-fold bootstrapping of a calibration dataset with m samples and N_s groups per fold. Let $t_{(J_{K,\tau})}$ be the threshold obtained by aggregating the per-fold p-quantile thresholds $t_{(J_k)}$ via their τ-quantile. Define:

$$p_\tau^{(N_s)} := p + \Phi^{-1}(\tau)\sqrt{\frac{p(1-p)}{N_s}}.$$

Suppose $m, N_s, K \to \infty$ such that $\gamma_\tau := \frac{\tau(1-\tau)}{f(\tau)^2}\frac{m}{N_s K}$ converges to a constant, where $f(\tau) := \phi(\Phi^{-1}(\tau))$. Then

$$\sqrt{m}\Big(F_S\big(t_{(J_{K,\tau})}\big) - p_\tau^{(N_s)}\Big) \xrightarrow{d} \mathcal{N}\big(0,\ p(1-p)(1+\gamma_\tau)\big).$$

Equivalently, centering at p, we have:

$$\sqrt{m}\Big(F_S\big(t_{(J_{K,\tau})}\big) - p\Big) - \delta_{m,N_s,\tau} \xrightarrow{d} \mathcal{N}\big(0,\ p(1-p)(1+\gamma_\tau)\big), \quad \delta_{m,N_s,\tau} = \Phi^{-1}(\tau)\sqrt{p(1-p)}\,\sqrt{m/N_s}.$$

The bias term $\delta_{m,N_s,\tau}$ vanishes if and only if $\tau = 1/2$ or $m/N_s \to 0$. In particular, for median aggregation the limit is centered at p for every joint scaling of (m, N_s, K).

Proof. We start with the case of $K = 1$, which makes the essence of the proof clearer.

Normal asymptotics for $K = 1$. We analyze the normal asymptotics of the coverage as the total number of calibrations, m, tends to infinity, with $m/N_s = \gamma$. Consider the following decomposition:

$$\begin{aligned} F_S(t_{(J)}) - p &= (F_S(t_{(J)}) - \mathbb{E}(F_S(t_{(J)})|J)) + (\mathbb{E}(F_S(t_{(J)})|J) - p) \\ &= (F_S(t_{(J)}) - \frac{J}{m+1}) + (\frac{J}{m+1} - p). \end{aligned}$$

For the term $\frac{J}{m+1} - p$, we can use eq. 25 and $m = \gamma N_s$ to show that

$$\sqrt{N_s}(\frac{J}{m+1} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)) \implies \sqrt{m}(\frac{J}{m+1} - p) \xrightarrow{d} \mathcal{N}(0, \gamma p(1-p)) \tag{26}$$

For the first term $(F_S(t_{(J)}) - \frac{J}{m+1})$, we know that given J, $F_S(t_{(J)})$ is also Beta distributed as $\text{Beta}(J, m+1-J)$. However, using $\frac{J}{m+1} \to p$, we can use Lemma F.1 to show that

$$\sqrt{m}(F_S(t_{(J)}) - \frac{J}{m+1}) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Therefore, $\sqrt{m}(F_S(t_{(J)}) - p)$ is the sum of two Gaussians, and we asymptotically get:

$$\sqrt{m}(F_S(t_{(J)}) - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)(1+\gamma)).$$

Once the Gaussian asymptotics of $F_S(t_{(J)})$ are given, we can obtain the asymptotics of $t_{(J)}$ and the conditional efficiency similar to the method described earlier.

Normal asymptotics for general K. Suppose that we have K folds, and each fold $k \in [K]$ produces N_s group scores $t_k^{(1)}, \ldots, t_k^{(N_s)}$ and estimate p-quantile with $k_{p,N_s} = \lceil (N_s+1)p \rceil$. The estimated quantile per fold is denoted by $\hat{t}_{k,\alpha}$, equally with $t_{(J_k)}$. We know that asymptotically $\sqrt{N_s}(J_k - mp) \sim \mathcal{N}(0, m^2 p(1-p))$. Each J_k is distributed according to these asymptotics. We have discussed the aggregation of the quantiles using the τ-quantile of the samples $\hat{t}_{k,\alpha} = t_{(J_k)}, k \in [K]$. Denote the estimated τ-quantile by $J_{K,\tau}$. Since $J_k \sim \mathcal{N}(mp, m^2 p(1-p)/N_s)$, we can have similar asymptotics for the quantile estimation (assuming the τ-quantile of $\mathcal{N}(mp, m^2 p(1-p)/N_s)$ is given by $m.p_\tau$):

$$\sqrt{K}(J_{K,\tau} - mp_\tau) \sim \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f(\tau)^2} . \frac{m^2 p(1-p)}{N_s}\right), \tag{27}$$

where $f(\tau) := \phi(\Phi^{-1}(\tau))$ is the Gaussian density ϕ, evaluated at τ-quantile $\Phi^{-1}(\tau)$ with Φ as probit function. For the median $\tau = 1/2$, the constant term in the variance is $\pi/2$.

We can use the same derivation as the single-fold analysis:

$$\begin{aligned} F_S(t_{J_{K,\tau}}) - p_\tau &= (F_S(t_{J_{K,\tau}}) - \mathbb{E}(F_S(t_{K,\tau})|J_{K,\tau})) + (\mathbb{E}(F_S(t_{J_{K,\tau}})|J_{K,\tau}) - p) \\ &= (F_S(t_{J_{K,\tau}}) - \frac{J_{K,\tau}}{m+1}) + (\frac{J_{K,\tau}}{m+1} - p_\tau). \end{aligned}$$

We can use Lemma F.1 for the first term to get the Gaussian asymptotics. For the second term, the equation 27 yields:

$$\sqrt{m}(F_S(t_{(J_{K,\tau})}) - p_\tau) \xrightarrow{d} \mathcal{N}(0, p(1-p)(1+\gamma_\tau)),$$

where $\gamma_\tau := \frac{\tau(1-\tau)}{f(\tau)^2} \frac{m}{N_s.K}$. However, we need to know that p_τ is centered around $p + \Phi^{-1}(\tau)\sqrt{\frac{p(1-p)}{N_s}}$, which finishes the proof.

□

Remark G.2. As $K \to \infty$ with N_s held fixed, $J_{K,\tau}/m$ concentrates at:

$$p_\tau^{(N_s)} := F^{-1}_{\text{Beta}(k_{p,N_s},\, N_s+1-k_{p,N_s})}(\tau),$$

which is the τ-quantile of the exact Beta law of J/m established in Appendix G.1. In the degenerate case $N_s = 1$ the fold threshold is a single uniformly drawn calibration score, $B \sim \text{Beta}(1,1) = \text{Unif}(0,1)$, and aggregation returns exactly the τ-quantile of the calibration set, $t_{(\lceil \tau m \rceil)}$, irrespective of the nominal level p. As $N_s \to \infty$, by contrast, $p_\tau^{(N_s)} \to p$ with $p_\tau^{(N_s)} = p + \Phi^{-1}(\tau)\sqrt{p(1-p)/N_s} + O(N_s^{-1})$. Here, the aggregation is asymptotically unbiased. In general, τ (used for aggregation across folds) and p (the target quantile) are distinct, and $p_\tau^{(N_s)}$ interpolates between them as N_s grows.

Table 6: 95% Monte Carlo half-width $1.96\sqrt{\widehat{c}(1-\widehat{c})/R}$ of the coverage estimate $\widehat{c}$, for $R = 50$ test batches, swept over $\widehat{c} \in [0.70, 1.00]$.

$\widehat{c}$	95% half-width
0.70	0.1270
0.75	0.1200
0.80	0.1109
0.85	0.0990
0.90	0.0832
0.95	0.0604
1.00	0.0000

H EXPERIMENTAL RESULTS - FOLLOW-UP

H.1 COMPUTING STATISTICS

We use Monte Carlo evaluation to approximate marginal coverage and efficiency. Consider R repetitions:

$$I_r = \mathbf{1}\{Y_r^{\text{test}} \in \Gamma^{(\alpha,r)}(X_r^{\text{test}})\}, \qquad L_r = \left|\Gamma^{(\alpha,r)}(X_r^{\text{test}})\right| .r \in [R]$$

Averaging I_r estimates joint or marginal coverage over both calibration and test randomness. Averaging L_r estimates marginal expected set size.

As standard error for the coverage, we can report 95% Monte Carlo interval using $\widehat{c} \pm 1.96\sqrt{\frac{\widehat{c}(1-\widehat{c})}{R}}$. This can be done mechanically, since coverage is the average of Bernoulli random variables. The number of samples is 50 ($R = 50$) for all four test sets (MNIST, FashionMNIST, CIFAR10, CIFAR100). They have exactly 10,000 images; we have used a batch size of 200 for sampling calibration and tests. We have precomputed the standard error for a range of $\widehat{c}$ in Table 6.

For efficiency, we recorded the median $m_{0.5}$ and 95% quantile $q_{0.95}$ to capture the distribution concisely. We can obtain the standard error using the Gaussian assumption. First, see $q_{0.95} = m_{0.5} + 1.645\sigma$, therefore:

$$SE \approx \frac{q_{0.95} - m_{0.5}}{1.645\sqrt{R}},$$

and with $R = 50$,

$$SE(\hat{\mu}) \approx 0.086\,(q_{0.95} - m_{0.5})$$

which is what we can use.

Using multiple test points for each calibration draw permits separation of calibration-induced variability from test-input variability, which we use for studying the statistics of conditional coverage

Finally, in various experiments, the median is 0.0000 for the reported methods (for instance in Table 10). That's a real artifact of the underlying data (median joint set size at that operating point rounds to a set size of 1, i.e., $\log_2(1)/3 = 0$).

H.2 HOW DOES THE CALIBRATION SET SIZE IMPACT THE MARGINAL COVERAGE OF TSCP?

Key Takeaway. Increasing the calibration set size does not monotonically change the efficiency. It can improve coverage stability.

The impact of calibration-set size on TSCP's marginal efficiency (average, median, 95th-percentile quantile, and Standard Error (SE)) and coverage (Cov. and SE) at transductive size $n = 1, 3, 6, 9$ is computed for MNIST (LeNet5), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as the noise level. Each row is drawn from an independent $R = 50$ test batches, where in each batch we have a total calibration set size N_t and n test samples. SE is expected to follow $1/\sqrt{R}$ scaling. For n test samples, the effective calibration set size is $N_s = \lfloor \text{calib_size}/n \rfloor$, which is the number of full calibration groups of size n available to the TSCP algorithm.

Table 7: Impact of calibration-set size on TSCP's marginal efficiency (average, median, 95th-percentile quantile, and SE) and coverage (Cov. and SE) at transductive size $n = 1$, for MNIST (LeNet5), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Calib. size	N_s	Avg	Median	$q_{0.95}$	SE(Eff.)	Cov.	SE(Cov.)
9	9	1.3561	0.0000	3.3219	0.4362	0.9000	0.0832
20	20	0.2630	0.0000	1.4060	0.1706	0.9400	0.0658
30	30	-0.0291	0.0000	0.0000	0.0000	0.9800	0.0388
60	60	-0.0893	0.0000	0.0000	0.0000	0.9400	0.0658
91	91	-0.0893	0.0000	0.0000	0.0000	0.9400	0.0658
120	120	-0.0893	0.0000	0.0000	0.0000	0.9400	0.0658
150	150	-0.2176	0.0000	0.0000	0.0000	0.8400	0.1016
180	180	-0.1844	0.0000	0.0000	0.0000	0.8400	0.1016

Table 8: Impact of calibration-set size on TSCP's marginal efficiency (average, median, 95th-percentile quantile, and SE) and coverage (Cov. and SE) at transductive size $n = 3$, for MNIST (LeNet5), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Calib. size	N_s	Avg	Median	$q_{0.95}$	SE(Eff.)	Cov.	SE(Cov.)
9†	3	0.7984	0.0000	1.5480	0.1887	0.8200	0.1065
20	6	0.9162	0.0000	1.6135	0.1702	0.8600	0.0962
30	10	0.9147	0.0000	1.5584	0.1515	0.9200	0.0752
60	20	0.6545	0.0000	1.3926	0.1813	0.9000	0.0832
91	30	0.5771	0.0000	1.3196	0.1812	0.8600	0.0962
120	40	0.4595	0.0000	1.1990	0.1766	0.9200	0.0752
150	50	0.3475	0.0000	0.9722	0.1315	0.9400	0.0658
180	60	0.2014	0.0000	0.6785	0.0843	0.9000	0.0832

Notes. †Rows: $N_s \leq 3$ calibration groups.

We report the efficiency values using $\gamma_{n,m} = \log_2(\cdot)/n$ of the mean joint prediction-set size. We have computed SE for efficiency using $SE(\hat{\mu}) \approx 0.086\,(q_{0.95} - m)$ (computed on the joint prediction-set size and not the efficiency rate, with m its median and $q_{0.95}$ its 95th percentile). Coverage (Cov.) is the empirical coverage over the row's $R = 50$ test batches. SE(Cov.) is the 95% Monte Carlo half-width $1.96\sqrt{\widehat{c}(1-\widehat{c})/R}$.

In the tables, the rows with † sign have $N_s \leq 3$. We characterize these cases as degenerate cases (e.g. $N_s = 1$ takes the single available calibration score as the threshold). This degeneracy manifests as coverage below the $1 - \alpha = 0.9$ target.

The results are presented in Tables 7,8, 9, and 10. The main impact of calibration set size on the coverage and efficiency is to move the operating point to a non-degenerate case where the prescribed coverage is met. The coverage SE is independent of the calibration set size. However, the median, 0.95-quantile, and SE for efficiency change with m. Here, the effect is not monotonic, and the dependency is not very clear. This observation is compatible with our theory, where the calibration set size is a function of $g(\cdot)$ function too. Asymptotically, the standard error should diminish as $1/\sqrt{m}$ using Proposition 4.1 but not necessarily monotonically.

H.3 How does the calibration set size impact the marginal coverage for cross-conformal prediction?

Key Takeaway. Using quantile merging for transductive cross-conformal prediction does not have a clear impact on marginal coverage and efficiency. The theoretically grounded merging functions can overestimate the quantile and thereby lead to lower efficiency.

We have studied the change of efficiency and coverage for standard conformal prediction $n = 1$. We compared TSCP with TCCP-E, TCCP-PH and TCCP-PA for the chosen datasets and model pairs: (MNIST, LeNet5), (FashionMNIST, CIFAR10, CIFAR100: ResNet20). In Table 11, we have

Table 9: Impact of calibration-set size on TSCP's marginal efficiency (average, median, 95th-percentile quantile, and SE) and coverage (Cov. and SE) at transductive size $n = 6$, for MNIST (LeNet5), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Calib. size	N_s	Avg	Median	$q_{0.95}$	SE(Eff.)	Cov.	SE(Cov.)
9†	1	0.6781	0.0000	0.8840	0.0474	0.5800	0.1368
20†	3	0.8486	0.0000	0.9105	0.0261	0.8000	0.1109
30	5	1.3454	0.4099	1.3116	0.0175	0.8400	0.1016
60	10	0.9189	0.7880	1.1398	0.0398	0.9000	0.0832
91	15	1.0981	0.8371	1.4730	0.0913	0.9000	0.0832
120	20	0.7441	0.7432	0.9421	0.0265	0.8400	0.1016
150	25	0.7911	0.8012	0.9571	0.0197	0.9000	0.0832
180	30	0.6963	0.6590	0.9211	0.0350	0.9600	0.0543

Notes. †Rows: $N_s \leq 3$ calibration groups.

Table 10: Impact of calibration-set size on TSCP's marginal efficiency (average, median, 95th-percentile quantile, and SE) and coverage (Cov. and SE) at transductive size $n = 9$, for MNIST (LeNet5), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Calib. size	N_s	Avg	Median	$q_{0.95}$	SE(Eff.)	Cov.	SE(Cov.)
9†	1	1.1980	0.0000	0.6394	0.0004	0.5000	0.1386
20†	2	0.3652	0.0000	0.6224	0.0672	0.5800	0.1368
30†	3	0.4087	0.0000	0.7057	0.0869	0.6600	0.1313
60	6	1.0649	0.5953	1.0754	0.0140	0.9200	0.0752
91	10	1.1333	0.5873	1.4409	0.0935	0.8400	0.1016
120	13	0.9046	0.6787	1.1434	0.0578	0.9800	0.0388
150	16	1.1388	0.6764	1.3092	0.0391	0.9400	0.0658
180	20	0.8280	0.5852	1.1280	0.0866	0.8600	0.0962

Notes. †Rows: $N_s \leq 3$ calibration groups.

included TCCP-K=1, where only a single fold is considered, but with resampling. We compute the expectation of the prediction-set size and convert it to the efficiency rate $\gamma_{n,m}$ using $\log_2(\cdot)/n$. The empirical joint coverage is computed over $R = 50$ test batches. We followed a similar procedure as in previous experiments to compute SE for coverage and efficiency. TCCP-E, TCCP-PH, and TCCP-PA are all reported at $K = 5$ folds. The upper bound on the efficiency rate is $\log_2(\text{number-of-classes})$, which is 3.3219 for 10 classes, and 6.6439 for 100 classes.

As we have reported before, resampling in TCCP-K=1 improves the efficiency in a consistent way across models and datasets. However, because we used merging functions with theoretical guarantees for TCCP-PH and TCCP-PA, this choice generally yields higher confidence when the dependence structure is more relaxed, as the results show. The cost of higher confidence, in general, is poorer efficiency.

We also studied changes in efficiency and coverage for larger transductive sizes. We have reported these results for transductive sizes $n = 3, 6, 9$ in Table 12 and for $n = 12, 15, 18$ in Table 13. The conclusion remains similar: higher confidence and lower efficiency for theoretically grounded TCCP variants.

H.4 How does the single-fold bootstrapping impact the marginal coverage and efficiency?

Key Takeaway. Resampling from the empirical distribution improves the marginal efficiency for larger transductive sizes.

Within the same experimental setup, we have considered the impact of resampling using the empirical distribution. The method "TCCP-K=1" resamples each fold from $N_s = 200$ groups. We have

Table 11: Efficiency (average, median, 95th-percentile quantile, and SE) and coverage (Cov. and SE) for TSCP, TCCP-K=1, TCCP-E, TCCP-PA, and TCCP-PH at transductive size $n = 1$, for (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20), $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Dataset	**Method**	Avg	Median	$q_{0.95}$	SE(Eff.)	Cov.	SE(Cov.)
MNIST	TSCP	-0.1844	0.0000	0.0000	0.0000	0.8400	0.1016
	TCCP-K=1	-0.1203	0.0000	0.0000	0.0000	0.8400	0.1016
	TCCP-E	-0.0291	0.0000	0.0000	0.0000	0.9400	0.0658
	TCCP-PA	0.1375	0.0000	0.0000	0.0000	1.0000	0.0000
	TCCP-PH	1.4114	0.0000	3.0959	0.3522	1.0000	0.0000
FMNIST	TSCP	0.0286	0.0000	0.0000	0.0000	0.9400	0.0658
	TCCP-K=1	-0.0589	0.0000	0.0000	0.0000	0.9200	0.0752
	TCCP-E	0.5850	0.0000	2.4725	0.3764	0.9600	0.0543
	TCCP-PA	0.4647	0.0000	1.3505	0.1394	0.9600	0.0543
	TCCP-PH	1.7655	1.0000	3.3219	0.2919	0.9400	0.0658
CIFAR10	TSCP	0.3103	0.0000	1.0000	0.1001	0.9000	0.0832
	TCCP-K=1	0.1635	0.0000	1.0000	0.1108	0.9400	0.0658
	TCCP-E	0.6781	0.0000	2.0000	0.2326	0.9800	0.0388
	TCCP-PA	1.1110	0.0000	2.7115	0.3188	0.9400	0.0658
	TCCP-PH	1.7485	1.3219	3.0179	0.2068	0.9800	0.0388
CIFAR100	TSCP	2.7698	1.8074	4.5423	0.3602	0.8600	0.0962
	TCCP-K=1	2.3334	1.5850	3.9588	0.3089	0.9400	0.0658
	TCCP-E	4.1342	3.3923	5.8099	0.3222	0.9000	0.0832
	TCCP-PA	3.9505	3.1699	5.3611	0.2576	0.9600	0.0543
	TCCP-PH	4.5922	4.0000	6.2288	0.3035	1.0000	0.0000

summarized these results in Table 14 and Table 1. Note that the reported coverage differences in the tables are all within the standard error, so they are comparable.

We see a clear trend: resampling improves marginal efficiency for larger transductive sizes. For smaller transductive sizes ($n = 3, 6$), the efficiency changes are within the standard error, so we cannot claim a marginal gain.

H.5 How do bootstrapping and merging impact the sample-conditional coverage and efficiency?

Key Takeaway. The main impact of sampling K folds and merging is to decrease the variance of bootstrapping sampling.

We compare the standard deviation of coverage, efficiency, and the estimated quantile when we compute them conditionally on the given calibration data.

Similar to the other experiments, we take samples of m calibration data (default $m = 200$). Then, we sample $K = 1, 5, 10$ folds, each with $N_s = 200$ samples, using 250 fresh seeds, which gives us a population for sample conditional statistics. In particular, to compute coverage, we need to sample more test groups for each calibration data and fold to compute coverage reliably. For a 0.98 confidence level, we need enough samples to make the estimate valid.

Note that TSCP has no fold parameter K. TCCP-K=1, similarly, is valid only for $K = 1$ by definition. The rows of TCCP-E and TCCP-PH for $K = 1$ are identical to that of TCCP-K=1. TCCP-E uses majority voting at $K = 1/5/10$, meaning it uses the threshold $1, 3, 5$ for inclusion. Harmonic merging is defined for $K > 1$; for $K = 1$, we use the classical prediction. Therefore, TCCP-PH collapses to TCCP-K=1 for $K = 1$, leading to a similar value in the table. TCCP-PA uses calibration constant $c = 2.0$, which is conservative for $K = 1$. Therefore, TCCP-PA for $K = 1$ is different for other TCCP methods. Coverage is evaluated against 200 fixed test groups per (dataset, batch, n), shared identically across every seed and method so the coverage rate itself is a low-noise estimate.

Table 12: Efficiency and coverage comparison of TSCP, TCCP-E, TCCP-PH and TCCP-PA for (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20) for transductive sizes $n = 3, 6, 9$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Method	$n = 3$			$n = 6$			$n = 9$		
	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.
TSCP (MNIST)	0.2014	0.0843	0.9000	0.6963	0.0350	0.9600	0.8280	0.0866	0.8600
TCCP-E (MNIST)	0.9940	0.1050	0.9400	0.8183	0.0194	0.9200	0.7674	0.0354	0.9200
TCCP-PH (MNIST)	1.4021	0.0179	0.9400	1.1574	0.0837	0.9600	1.0541	0.0384	0.9400
TCCP-PA (MNIST)	0.9903	0.0950	0.8400	0.8454	0.0129	1.0000	0.7759	0.0646	0.9000
TSCP (FMNIST)	0.7873	0.1126	0.9000	1.0648	0.0579	0.9000	1.1134	0.0794	0.9200
TCCP-E (FMNIST)	1.3481	0.0790	0.9200	1.2070	0.0570	0.9000	1.1932	0.0460	0.9200
TCCP-PH (FMNIST)	1.8784	0.1054	0.9400	1.4352	0.0498	0.9200	1.3313	0.0428	0.9800
TCCP-PA (FMNIST)	1.3023	0.0652	0.9200	1.2961	0.0542	1.0000	1.1564	0.0412	0.9200
TSCP (CIFAR10)	0.9952	0.0760	0.8600	1.3562	0.0906	0.9200	1.2414	0.0355	0.9400
TCCP-E (CIFAR10)	1.3357	0.0558	0.9200	1.3954	0.0808	0.9600	1.4354	0.0399	0.9400
TCCP-PH (CIFAR10)	1.7307	0.1089	0.9600	1.5672	0.0601	1.0000	1.6112	0.0711	1.0000
TCCP-PA (CIFAR10)	1.5319	0.0815	0.9800	1.5544	0.0774	0.9600	1.6075	0.0224	0.9600
TSCP (CIFAR100)	3.4176	0.0392	0.8800	5.8699	0.0000	0.9200	6.6081	0.0000	0.8800
TCCP-E (CIFAR100)	3.8363	0.2543	0.9400	6.0366	0.2585	0.9200	6.6270	0.0000	0.9600
TCCP-PH (CIFAR100)	4.8050	0.2271	0.9800	6.4112	0.0544	1.0000	6.6406	0.0000	0.9600
TCCP-PA (CIFAR100)	4.1428	0.1249	0.9600	5.8699	0.0000	0.8800	6.6339	0.0000	0.9800

We have three main tables. In each table, we have included TCSP, TCCP-$K = 1$, TCCP-E, TCCP-PH, TCCP-PA for all the dataset and model pairs. We have three main columns corresponding to transductive sizes $n = 10, 13, 16$. Each one has three sub-columns corresponding to the standard deviations of the desired quantity for $K = 1, 5, 10$. Therefore, we have three separate tables: one for coverage, one for efficiency and one for the quantile.

Tables 15, 16, and 17 present the results for all standard deviation variations. The story is quite consistent in all tables. Increasing K reduces the variance.

Table 13: Efficiency and coverage comparison of TSCP, TCCP-E, TCCP-PH and TCCP-PA for (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20) for transductive sizes $n = 12, 15, 18$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level.

Method	$n = 12$			$n = 15$			$n = 18$		
	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.
TSCP (MNIST)	1.3471	0.0095	0.9000	0.7704	0.0314	0.9400	0.9850	0.0116	0.9000
TCCP-E (MNIST)	0.7902	0.0447	0.9200	0.7675	0.0347	1.0000	0.6766	0.0170	0.9000
TCCP-PH (MNIST)	0.9727	0.0318	0.9400	0.9393	0.0260	0.9800	3.0084	0.0000	0.8600
TCCP-PA (MNIST)	0.7103	0.0448	0.8400	0.7086	0.0287	0.9800	0.7531	0.0132	0.8800
TSCP (FMNIST)	1.4496	0.0478	0.9800	2.9457	0.0000	0.9000	3.1520	0.0574	0.9000
TCCP-E (FMNIST)	1.1425	0.0515	0.9400	1.1263	0.0407	0.9200	3.0084	0.0000	0.8400
TCCP-PH (FMNIST)	1.4110	0.0493	1.0000	3.0790	0.1034	0.9800	3.1750	0.0431	0.9200
TCCP-PA (FMNIST)	1.1493	0.0412	0.9800	1.1503	0.0356	0.9800	1.1127	0.0354	0.9400
TSCP (CIFAR10)	2.9837	0.0948	1.0000	3.0790	0.1034	0.9200	3.2254	0.0230	0.9200
TCCP-E (CIFAR10)	1.4212	0.0645	0.9800	1.2909	0.0408	0.9200	3.1643	0.0492	0.9000
TCCP-PH (CIFAR10)	1.4630	0.0358	0.9800	3.1005	0.0827	0.9800	3.2524	0.0164	0.9800
TCCP-PA (CIFAR10)	1.4837	0.0415	0.9200	3.0790	0.1034	0.9200	3.2355	0.0203	0.9200
TSCP (CIFAR100)	6.6439	0.0000	0.9000	6.6439	0.0000	0.9000	6.6439	0.0000	0.9200
TCCP-E (CIFAR100)	6.6439	0.0000	0.9000	6.6439	0.0000	0.9400	6.6439	0.0000	0.9400
TCCP-PH (CIFAR100)	6.6439	0.0000	0.9600	6.6439	0.0000	0.9400	6.6439	0.0000	1.0000
TCCP-PA (CIFAR100)	6.6439	0.0000	0.9200	6.6439	0.0000	0.9200	6.6439	0.0000	0.9200

Table 14: Efficiency and coverage comparison of SCP, APS (both Bonferroni corrected at α/n for joint prediction over n transductive samples), TSCP and TCCP-K=1 for (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20) for transductive sizes $n = 3, 6, 9$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level. TCCP-K=1 resamples each fold from $N_{groups} = 200$ groups.

Method	$n = 3$			$n = 6$			$n = 9$		
	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.	$\gamma_{n,m}$	SE(γ)	Cov.
SCP (MNIST)	0.8584	0.2033	0.8800	2.7143	0.1009	0.9600	2.8548	0.0660	0.9200
APS (MNIST)	0.8145	0.1011	0.9600	2.5360	0.0984	0.9400	2.8524	0.0761	0.9600
TSCP (MNIST)	0.2014	0.0843	0.9000	0.6963	0.0350	0.9600	0.8280	0.0866	0.8600
TCCP-K=1 (MNIST)	0.3137	0.1842	0.9400	0.6193	0.0356	0.9000	0.6046	0.0264	0.9000
SCP (FMNIST)	1.4042	0.0705	0.8800	2.7644	0.2099	0.8200	3.1409	0.0425	0.9600
APS (FMNIST)	1.3742	0.1591	0.9200	2.6491	0.1238	0.9000	3.0022	0.0678	0.8800
TSCP (FMNIST)	0.7873	0.1126	0.9000	1.0648	0.0579	0.9000	1.1134	0.0794	0.9200
TCCP-K=1 (FMNIST)	0.6970	0.1184	0.8200	0.9005	0.0691	0.8800	0.9301	0.0605	0.9000
SCP (CIFAR10)	1.6788	0.1196	0.9000	2.7067	0.1393	0.9200	2.9777	0.0882	0.9200
APS (CIFAR10)	1.5871	0.1485	0.9000	2.5133	0.0994	0.8600	2.9327	0.0741	0.9000
TSCP (CIFAR10)	0.9952	0.0760	0.8600	1.3562	0.0906	0.9200	1.2414	0.0355	0.9400
TCCP-K=1 (CIFAR10)	1.0491	0.1039	0.9600	1.1219	0.0330	0.9200	1.1646	0.0550	0.9000
SCP (CIFAR100)	4.8998	0.2354	0.9600	6.2604	0.1018	0.9200	6.5391	0.0000	0.9200
APS (CIFAR100)	4.9237	0.2177	0.9000	6.0093	0.1213	0.9400	6.2943	0.0831	0.9800
TSCP (CIFAR100)	3.4176	0.0392	0.8800	5.8699	0.0000	0.9200	6.6081	0.0000	0.8800
TCCP-K=1 (CIFAR100)	3.2321	0.0595	0.9200	3.1494	0.1171	0.9800	6.5327	0.0138	0.9000

Table 15: Standard deviation of the estimated log-quantile $\log \tau$, conditional on a fixed calibration set, for TSCP, TCCP-K=1, TCCP-E, TCCP-PH and TCCP-PA on (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20), for transductive sizes $n = 10, 13, 16$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level, with the real calibration data held fixed and only the resampling seed varied.

Method	$n = 10$			$n = 13$			$n = 16$		
	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$
TSCP (MNIST)	0.9647	–	–	2.8736	–	–	3.1781	–	–
TCCP-K=1 (MNIST)	0.5673	–	–	0.4918	–	–	0.7534	–	–
TCCP-E (MNIST)	0.5673	0.3976	0.2908	0.4918	0.5702	0.3976	0.7534	0.5512	0.3812
TCCP-PH (MNIST)	0.5673	0.7879	0.9692	0.4918	1.0586	0.7881	0.7534	0.9181	0.7605
TCCP-PA (MNIST)	0.7495	0.3547	0.2342	1.0741	0.4523	0.3335	1.1607	0.5019	0.3316
TSCP (FMNIST)	4.2861	–	–	7.4788	–	–	8.3106	–	–
TCCP-K=1 (FMNIST)	2.0345	–	–	1.9341	–	–	1.9067	–	–
TCCP-E (FMNIST)	2.0345	1.1891	0.8894	1.9341	1.2704	0.9142	1.9067	1.6189	1.1602
TCCP-PH (FMNIST)	2.0345	1.7338	1.7077	1.9341	2.0522	1.6582	1.9067	1.8775	1.6350
TCCP-PA (FMNIST)	2.1605	1.0309	0.7368	2.4122	1.0666	0.7199	2.9963	1.4026	0.9766
TSCP (CIFAR10)	2.0653	–	–	3.8643	–	–	4.1261	–	–
TCCP-K=1 (CIFAR10)	0.9039	–	–	1.1323	–	–	1.2285	–	–
TCCP-E (CIFAR10)	0.9039	0.9028	0.6431	1.1323	0.7566	0.5550	1.2285	0.7868	0.5952
TCCP-PH (CIFAR10)	0.9039	0.8470	0.7945	1.1323	0.9782	0.8221	1.2285	1.0105	0.9135
TCCP-PA (CIFAR10)	1.5489	0.7373	0.4996	1.4883	0.6499	0.4664	1.5117	0.6786	0.4824
TSCP (CIFAR100)	4.5529	–	–	8.3062	–	–	8.6832	–	–
TCCP-K=1 (CIFAR100)	1.9332	–	–	2.0734	–	–	2.2729	–	–
TCCP-E (CIFAR100)	1.9332	1.3582	0.9424	2.0734	1.4687	1.0196	2.2729	1.5310	1.1372
TCCP-PH (CIFAR100)	1.9332	1.7794	1.5564	2.0734	1.9635	1.8055	2.2729	2.0510	1.8251
TCCP-PA (CIFAR100)	2.4928	1.1236	0.7929	2.7477	1.2613	0.8893	3.0350	1.3536	0.9269

Table 16: Standard deviation of the joint coverage rate, conditional on a fixed calibration set, for TSCP, TCCP-K=1, TCCP-E, TCCP-PH and TCCP-PA on (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20), for transductive sizes $n = 10, 13, 16$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level, with the real calibration data held fixed and only the resampling seed varied.

Method	$n = 10$			$n = 13$			$n = 16$		
	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$
TSCP (MNIST)	0.0195	–	–	0.0277	–	–	0.0455	–	–
TCCP-K=1 (MNIST)	0.0223	–	–	0.0204	–	–	0.0226	–	–
TCCP-E (MNIST)	0.0223	0.0075	0.0062	0.0204	0.0062	0.0025	0.0226	0.0128	0.0106
TCCP-PH (MNIST)	0.0223	0.0044	0.0055	0.0204	0.0093	0.0035	0.0226	0.0085	0.0057
TCCP-PA (MNIST)	0.0113	0.0076	0.0061	0.0149	0.0030	0.0018	0.0208	0.0126	0.0099
TSCP (FMNIST)	0.0314	–	–	0.0326	–	–	0.0328	–	–
TCCP-K=1 (FMNIST)	0.0216	–	–	0.0175	–	–	0.0123	–	–
TCCP-E (FMNIST)	0.0216	0.0030	0.0022	0.0175	0.0059	0.0036	0.0123	0.0089	0.0063
TCCP-PH (FMNIST)	0.0216	0.0064	0.0043	0.0175	0.0085	0.0036	0.0123	0.0032	0.0033
TCCP-PA (FMNIST)	0.0098	0.0024	0.0018	0.0129	0.0046	0.0030	0.0147	0.0077	0.0054
TSCP (CIFAR10)	0.0363	–	–	0.0371	–	–	0.0477	–	–
TCCP-K=1 (CIFAR10)	0.0215	–	–	0.0238	–	–	0.0286	–	–
TCCP-E (CIFAR10)	0.0215	0.0124	0.0082	0.0238	0.0077	0.0057	0.0286	0.0090	0.0072
TCCP-PH (CIFAR10)	0.0215	0.0055	0.0055	0.0238	0.0078	0.0052	0.0286	0.0057	0.0025
TCCP-PA (CIFAR10)	0.0213	0.0097	0.0065	0.0165	0.0066	0.0048	0.0163	0.0078	0.0062
TSCP (CIFAR100)	0.0426	–	–	0.0495	–	–	0.0563	–	–
TCCP-K=1 (CIFAR100)	0.0194	–	–	0.0257	–	–	0.0237	–	–
TCCP-E (CIFAR100)	0.0194	0.0108	0.0077	0.0257	0.0119	0.0078	0.0237	0.0118	0.0084
TCCP-PH (CIFAR100)	0.0194	0.0065	0.0049	0.0257	0.0083	0.0073	0.0237	0.0078	0.0084
TCCP-PA (CIFAR100)	0.0183	0.0087	0.0067	0.0222	0.0096	0.0061	0.0220	0.0100	0.0072

Table 17: Standard deviation of the efficiency rate $\gamma_{n,m} = \log_2(\text{joint prediction-set size})/n$, conditional on a fixed calibration set, for TSCP, TCCP-K=1, TCCP-E, TCCP-PH and TCCP-PA on (MNIST: LeNet5), and (FashionMNIST, CIFAR10, CIFAR100: ResNet20), for transductive sizes $n = 10, 13, 16$, $\alpha = 0.1$, and $1 - \epsilon = 0.98$ as noise level, with the real calibration data held fixed and only the resampling seed varied.

Method	$n = 10$			$n = 13$			$n = 16$		
	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$
TSCP (MNIST)	0.0729	–	–	0.1953	–	–	0.1899	–	–
TCCP-K=1 (MNIST)	0.0597	–	–	0.0308	–	–	0.0453	–	–
TCCP-E (MNIST)	0.0597	0.0300	0.0211	0.0308	0.0410	0.0281	0.0453	0.0276	0.0192
TCCP-PH (MNIST)	0.0597	0.0572	0.0731	0.0308	0.0728	0.0581	0.0453	0.0565	0.0411
TCCP-PA (MNIST)	0.0598	0.0259	0.0149	0.0729	0.0313	0.0228	0.0660	0.0258	0.0165
TSCP (FMNIST)	0.1593	–	–	0.2678	–	–	0.4301	–	–
TCCP-K=1 (FMNIST)	0.0816	–	–	0.0597	–	–	0.0502	–	–
TCCP-E (FMNIST)	0.0816	0.0420	0.0315	0.0597	0.0357	0.0256	0.0502	0.0599	0.0418
TCCP-PH (FMNIST)	0.0816	0.0567	0.0545	0.0597	0.0638	0.0612	0.0502	0.1275	0.1026
TCCP-PA (FMNIST)	0.0767	0.0363	0.0261	0.0680	0.0298	0.0200	0.1296	0.0556	0.0361
TSCP (CIFAR10)	0.1410	–	–	0.3202	–	–	0.5124	–	–
TCCP-K=1 (CIFAR10)	0.0652	–	–	0.0655	–	–	0.0740	–	–
TCCP-E (CIFAR10)	0.0652	0.0580	0.0412	0.0655	0.0410	0.0296	0.0740	0.1053	0.0834
TCCP-PH (CIFAR10)	0.0652	0.0501	0.0460	0.0655	0.0877	0.0892	0.0740	0.1624	0.1310
TCCP-PA (CIFAR10)	0.1003	0.0471	0.0319	0.0911	0.0361	0.0252	0.1840	0.0914	0.0701
TSCP (CIFAR100)	0.6544	–	–	0.0605	–	–	0.0000	–	–
TCCP-K=1 (CIFAR100)	0.3681	–	–	0.0329	–	–	0.0000	–	–
TCCP-E (CIFAR100)	0.3681	0.1034	0.0812	0.0329	0.0000	0.0000	0.0000	0.0000	0.0000
TCCP-PH (CIFAR100)	0.3681	0.0007	0.0083	0.0329	0.0000	0.0000	0.0000	0.0000	0.0000
TCCP-PA (CIFAR100)	0.1744	0.0795	0.0601	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

I ESTIMATES OF THE SCORE DENSITY, CONDITIONAL COVERAGE FUNCTION AND THE COVERAGE-EFFICIENCY FUNCTION

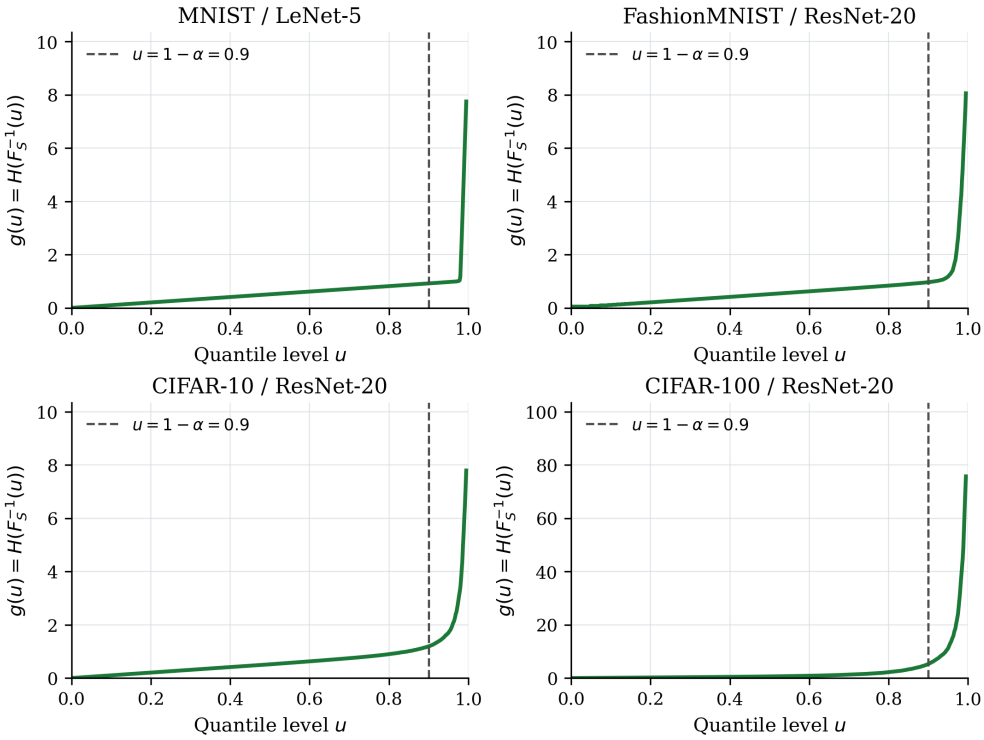

Figure 3: The coverage-efficiency function $\widehat{g}(u) = \widehat{H}(\widehat{F}_S^{-1}(u))$.

We estimate the functions entering the calibration-size asymptotics for the four dataset-model pairs trained at correct-label probability 0.98: MNIST/LeNet-5 and FashionMNIST, CIFAR-10, and CIFAR-100 with ResNet-20. To match the nonconformity convention in defining the prediction set as in $\{y : S(x, y) \leq q\}$, we use the negative model log-probability

$$S(x, y) = -\log p_\theta(y \mid x),$$

obtained by applying log-softmax to the checkpoint logits, which is the reversed sign of the log-probability conformity score. This change does not affect any variance computation.

All expectations over X are Monte Carlo averages over the complete test split ($N = 10000$ for every dataset). We form f_S from the weighted empirical distribution using a Gaussian kernel. For each threshold on a 2,400-point grid, we estimate

$$\widehat{H}(q) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \mathbf{1}\{S(X_i, y) \leq q\}.$$

The displayed H curve is this direct empirical estimate. For the derivative only, we use a centered finite difference on the smoothed version of it with a Gaussian kernel. The variance factor is then $\widehat{H}'(q)^2 / \widehat{f}_S(q)^2$. Finally, F_S^{-1} is a weighted empirical quantile and $\widehat{g}(u) = \widehat{H}(\widehat{F}_S^{-1}(u))$. We added a line to all figures representing the operating point corresponding to $\alpha = 0.1$.

The coverage-efficiency function $\widehat{g}(u) = \widehat{H}(\widehat{F}_S^{-1}(u))$ is shown in Figure 3. The function shows how calibration data affects the marginal efficiency. Simpler datasets with stronger models, like MNIST-LeNet, can yield more efficient sets even at very high confidence, while CIFAR100 has the worst conditional efficiency around the target quantile, as shown in our results.

Figure 4 shows the score density function f_S. As shown, the y-axis is on a logarithmic scale. The MNIST dataset, the simplest and highest-accuracy one, has the most mass concentrated around

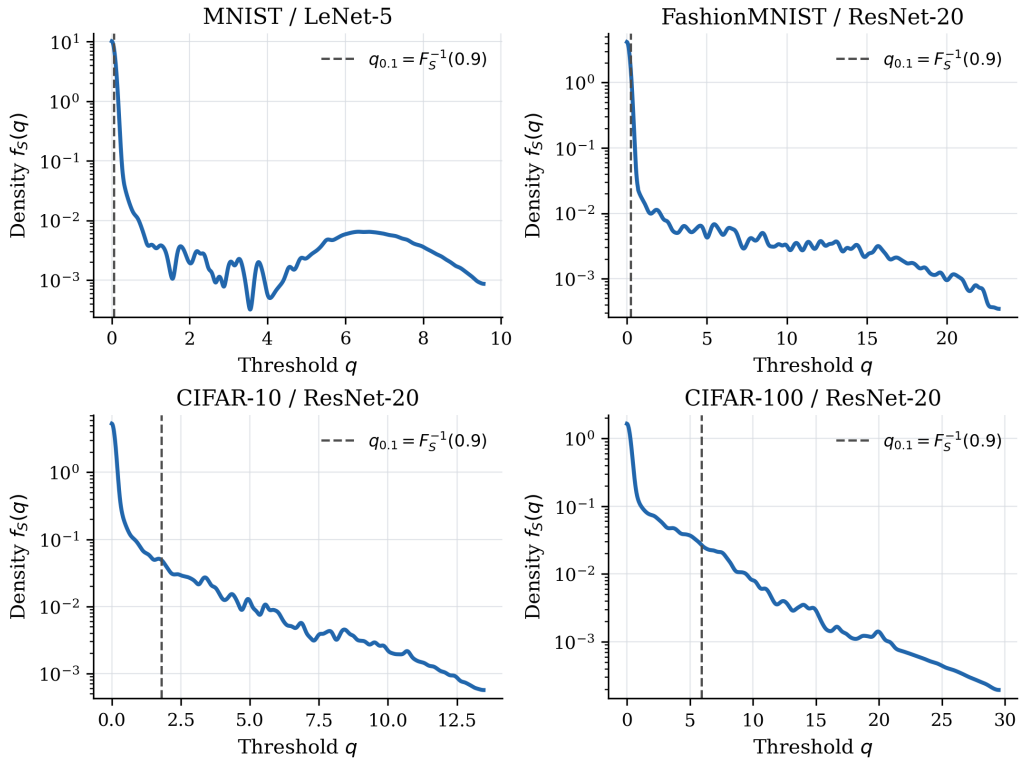

Figure 4: Monte Carlo density estimates $\widehat{f}_S(q)$ for the four checkpoints trained at correct-label probability 0.98.

0, which corresponds to a probability of 1. The main impact of f_S is on the variance of the estimated quantile, where the small density leads to high variance. MNIST should have the lowest variance around the quantile, while CIFAR100 should have the highest. This is compatible with our experimental results, as shown in Table 15.

Figure 5 shows the split-conformal prediction-set size $\widehat{H}(q)$ for a fixed q. This function shows how conditional efficiency converges to marginal efficiency as the number of calibration data samples increases.

Figure 6 plots the term that directly impacts the variance of the conditional efficiency. Again, MNIST and CIFAR100 show, respectively, the smallest and highest values around the quantile, which is reflected in the standard deviation as well. The order-of-magnitude difference in the shown value $\widehat{H}'(q)^2/\widehat{f}_S(q)^2$ for these two datasets also shows itself in the standard deviation differences of the conditional efficiency shown in Table 17.

Finally, note that the variance of conditional coverage is not dependent on these functions. Table 16 also shows that all standard deviations are comparable.

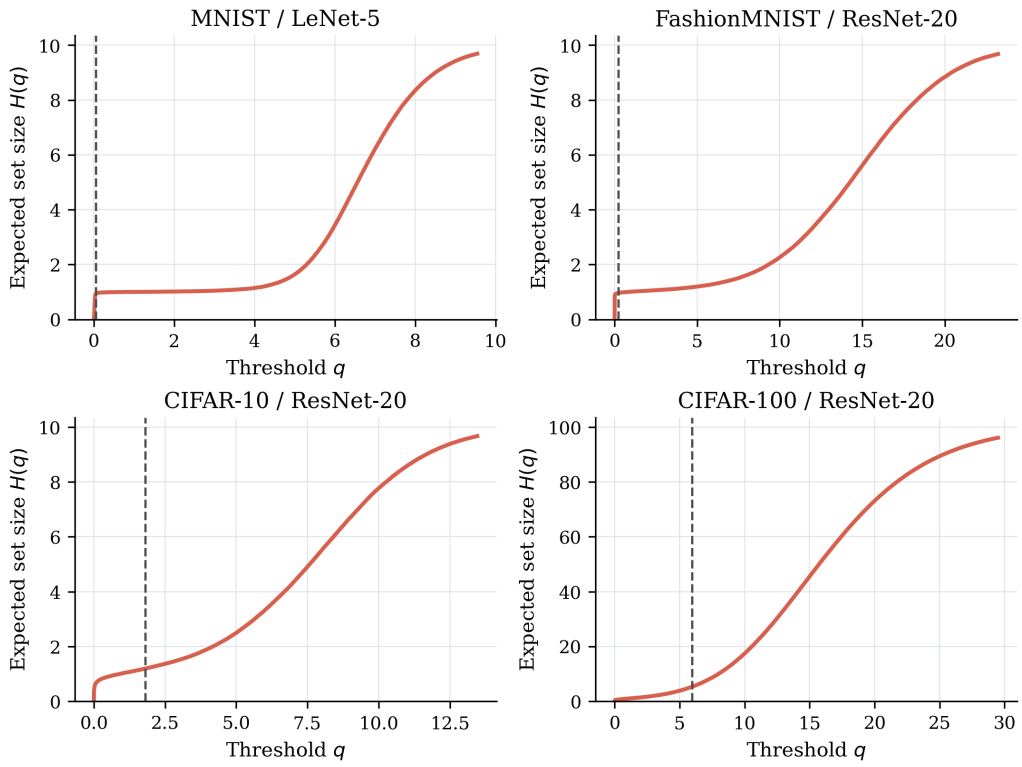

Figure 5: Expected split-conformal prediction-set size $\widehat{H}(q)$ as a function of the deterministic score threshold.

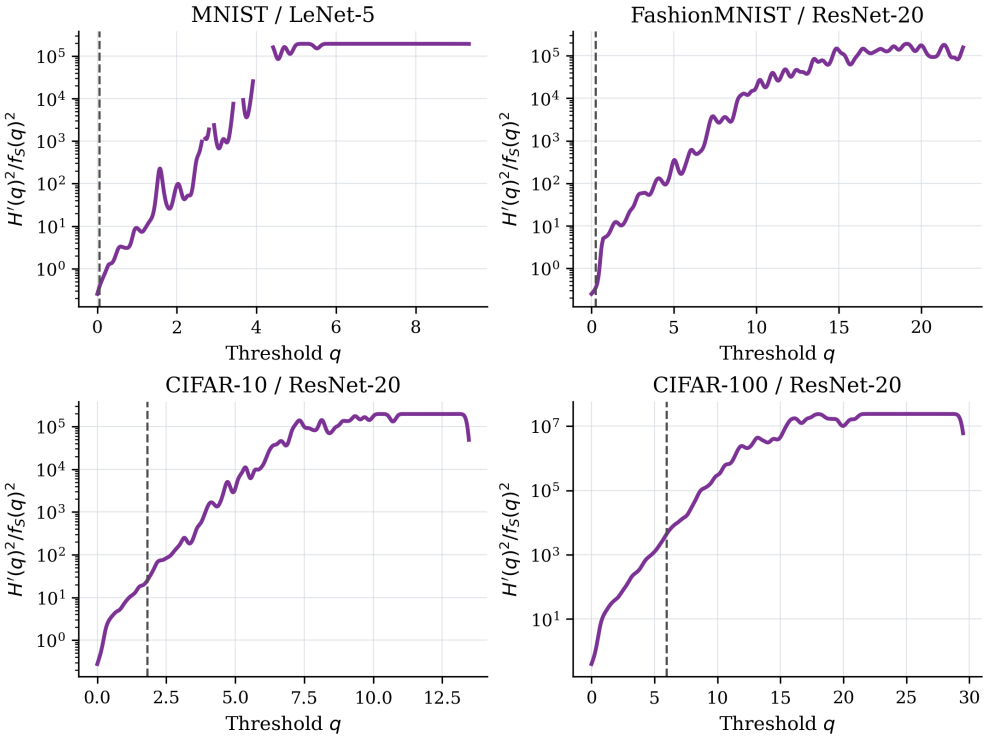

Figure 6: Estimated delta-method efficiency factor $\widehat{H}'(q)^2/\widehat{f}_S(q)^2$. The vertical axis is logarithmic; bins without adequate empirical density support are omitted.