Valid Selection among Conformal Sets

Mahmoud Hegazy[1,2] **Liviu Aolaritei**[3] **Michael I. Jordan**[2,3] **Aymeric Dieuleveut**[1]

[1]CMAP, École polytechnique, Institut Polytechnique de Paris
[2]Inria, École Normale Supérieure, PSL Research University
[3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

mahmoud.hegazy@polytechnique.edu liviu.aolaritei@berkeley.edu
jordan@cs.berkeley.edu aymeric.dieuleveut@polytechnique.edu

Abstract

Conformal prediction offers a distribution-free framework for constructing prediction sets with coverage guarantees. In practice, multiple valid conformal prediction sets may be available, arising from different models or methodologies. However, selecting the most desirable set, such as the smallest, can invalidate the coverage guarantees. To address this challenge, we propose a stability-based approach that ensures coverage for the selected prediction set. We extend our results to the online conformal setting, propose several refinements in settings where additional structure is available, and demonstrate its effectiveness through experiments.

1 Introduction

Conformal Prediction (CP) provides a principled framework for uncertainty quantification by constructing prediction sets with guaranteed marginal coverage [1–3]. For a desired coverage level $1-\alpha$, a conformal predictor outputs a set that on average contains the true label with at least this probability. The appeal of conformal prediction lies in its minimal assumptions about the data distribution and the underlying predictive model. In practice, multiple conformal prediction algorithms may be available for a given task, arising from variations in underlying models or data splits. This multiplicity motivates the choice of the most desirable set, often the smallest. To illustrate, consider a prediction problem with feature X and K sets $\{C_i^\alpha(X)\}_{i=1}^K$, each generated by a different conformal predictor. Suppose each $C_i^\alpha(X)$ satisfies the marginal coverage guarantee $\mathbb{P}\{Y \in C_i^\alpha(X)\} \geq 1-\alpha$, for all $i \in [K]$, where Y denotes the label associated to X. Although each $C_i^\alpha(X)$ individually meets the coverage guarantee, selecting the smallest set generally invalidates the guarantee due to dependencies on the data introduced by the selection.

Contributions and outline. To address this issue, we first introduce a novel perspective on the selection process based on algorithmic stability [4, 5]. The core idea is to employ a *stable* randomized selection mechanism, meaning its output is robust to small input perturbations. Such stability then allows us to transfer the marginal coverage of *individual conformal predictors* to the selected set. We introduce several stable selection rules, in particular MinSE, which we prove to be optimal. We further extend approach, by introducing an adaptive and a derandomized variant. These contributions are given in Section 3, after recalling preliminaries on conformal prediction in Section 2.

Furthermore, we extend in Section 4 our work to the online conformal setting [6–9], where data arrives sequentially, and predictions are made in real-time. We first demonstrate how our stability-based approach integrates seamlessly with existing online conformal methods, particularly with the approach of [10], to enable more adaptable selection among online predictors. Finally, in Section 5, we explore methods to optimize the implementation of our stable selection framework in practice. In particular, we take a closer look at the split conformal setting, where the stability-based bound can be overly conservative, and propose a recalibration mechanism that can achieve better empirical performance. Lastly, we validate our approaches on multiple experimental settings in Section 6.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Related Work. The motivation to select the smallest conformal prediction set has spurred a series of recent works introducing principled selection methods that retain coverage guarantees while favoring smaller sets. For example, Liang et al. [11] proposes to merge multiple predictors by selecting the one with smallest average set size on the calibration data. Moreover, [12] proposes split-conformal methods that either inflate quantiles or use independent splits to preserve coverage after selection. Building on this approach, for the classification setting, Luo and Zhou [13] proposes a method for constructing the best possible conformal score, which may be expressed as a weighted average of different scores. In contrast to such methods, our work enables *pointwise* selection depending on each X while maintaining validity, i.e. our aim is not to select the predictor with best *average performance*, but rather to pointwise select a predictor producing a small set for *each* realization of X.

In the online setting, Gasparin and Ramdas [10] proposed a method for conformal online model aggregation, which adapts model weights over time based on past performance. Their method combines prediction sets using a weighted majority vote with the weights learned in an online fashion. In another recent contribution, Hajihashemi and Shen [14] handles distribution shifts but focuses on temporal adaptation to distribution shifts, our approach addresses the distinct challenge of ensuring valid coverage when selecting across multiple predictors in static or online settings.

Our approach builds on algorithmic stability, a concept with roots in generalization properties of algorithms [15, 16] and differential privacy [17, 18]. Originally introduced to ensure privacy-preserving data analysis, differential privacy has been adapted for other tasks, such as adaptive data analysis [19], where it addresses the challenges of reusing data for multiple adaptive queries. Zrnic and Jordan [4] applied stability-based techniques to establish statistical validity after selection processes. Building on their work, we extend these ideas to the conformal prediction setting, developing stability-based methods for both batch and online conformal frameworks. Within the conformal predictions literature, different notions of algorithmic stability have been leveraged. For example, Barber et al. [20] proved coverage properties of the Jackknife method under a notion of stability albeit very different from the one we use.

While the aforementioned works focus on selecting conformal predictors for a single prediction task, other research has explored related but distinct problems in the context of conformal inference. Conformalized selection methods aim to identify a subset of data points whose unobserved labels exceed a given threshold while controlling the False Discovery Rate (FDR) [21]. In a recent work, Bai and Jin [22] introduced a framework that allows data reuse for both training and selection while maintaining finite-sample FDR control. Although it addresses a different problem than ours, the focus on managing data reuse aligns conceptually with our goal of ensuring valid coverage despite dependencies introduced by the selection.

2 Preliminaries in Conformal Prediction

We consider CP in two key scenarios: the batch setting, which assumes i.i.d. or exchangeable samples, and the online setting, where data arrives sequentially under minimal distributional assumptions.

Batch Setting. We consider a dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, where the points in $\mathcal{D}$, along with any test sample $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, are assumed to be either i.i.d. or exchangeable. In the i.i.d. case, we denote by $\mathcal{P}$ the distribution from which they are drawn, and by $\mathcal{P}_X$ and $\mathcal{P}_Y$ its marginals over X and Y, respectively. Without any further assumptions on the data generating process, conformal prediction allows to construct a (random) prediction set $C^\alpha(X)$ with the guarantee

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

Here, the probability is taken over (X, Y) as well as the randomness employed in the construction of C^α. Arguably, the most common approach for batch conformal prediction is the *split conformal* procedure, which first partitions the dataset $\mathcal{D}$ into two disjoint subsets: $\mathcal{D}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^{n-m}$, used to train the underlying predictor f, and $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{m}$, reserved for calibration. Then, for any nonconformity score function s, which quantifies the error between the predictor f and the true output, it estimates the empirical $\lceil (1-\alpha)(m+1) \rceil / m$-quantile, denoted $\hat{q}_\alpha$, of the set $\{s_i := s(X_i, Y_i, f)\}_{i=1}^{m}$. Finally, for the test point X, the split conformal prediction set is defined as $C^\alpha(X) := \{y \in \mathcal{Y} : \, s(X, y, f) \leq \hat{q}_\alpha\}$, which satisfies (1) through a rank statistic argument.

Online Setting. In this setting, observations (X_t, Y_t) arrive sequentially for $t = 1, 2, \ldots$. At each time step t, we observe X_t and aim to cover Y_t using a prediction set $C^{(t)}(X_t)$, which is constructed

based on a base model trained on all past data $\{(X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})\}$. After making the prediction, the true label Y_t is revealed, and the process continues to the next time step. Unlike the classical conformal prediction setting, where data is assumed to be exchangeable, the online setting allows for the data to be non-stationary or even adversarial. As a result, classical coverage guarantees no longer hold, and alternative notions of asymptotic coverage are required [6]. Specifically, we say that $\{C^{(t)}\}_{t\in\mathbb{N}}$ achieves asymptotic coverage if

$$\liminf_{T\to\infty} \ \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{Y_t \in C^{(t)}\left(X_t\right)\right\} \ \geq \ 1-\alpha. \tag{2}$$

The limit (2) ensures that, in the long run, the fraction of instances where the true label Y_t falls within the prediction set $C^{(t)}(X_t)$ meets or exceeds the desired coverage level, even under non-stationary or adversarial data. Stronger notions of asymptotic coverage can be considered, for example [23].

3 Smallest Confidence Set Selection

This section considers the batch setting without imposing additional restrictions, such as those in split conformal methods. Specifically, we focus on the problem of selecting the smallest among a collection of K conformal prediction sets $\{C_i^{\alpha}(X)\}_{i=1}^{K}$. While such a selection is appealing, it inevitably invalidates the marginal coverage guarantee (1) since the selection process depends on the data. To address this issue, we develop a strategy based on algorithmic stability, ensuring adjusted coverage guarantees even after the data-dependent selection process.

3.1 Valid Selection via Algorithmic Stability

We first recall the notion of algorithmic stability from [4] and extend their framework to tackle the data-dependent selection in conformal prediction. We start by introducing indistinguishability.

Definition 1 (Indistinguishability)**.** *A random variable (r.v.) S is (η,τ)-indistinguishable from a r.v. S_0, denoted $S \approx_{\eta,\tau} S_0$, if for all measurable sets $\mathcal{O}$, it holds that $\mathbb{P}\{S\in\mathcal{O}\} \leq e^{\eta}\mathbb{P}\{S_0\in\mathcal{O}\}+\tau$.*

This definition extends to the conditional case, denoted by $S \approx^{|\xi}_{\eta,\tau} S_0$, if the inequality holds almost surely with respect to the conditioning variable ξ, that is, $\mathbb{P}\{S\in\mathcal{O} \mid \xi\} \leq e^{\eta}\mathbb{P}\{S_0\in\mathcal{O} \mid \xi\}+\tau$. In essence, the parameter η measures the degree of similarity between the distributions of S and S_0, with smaller values of η allowing for greater similarity. Leveraging indistinguishability, we can define a notion of stability for randomized algorithms. For more precision, we define a *randomized algorithm* as a deterministic mapping from $\Xi\times\mathcal{E}$ into $\mathcal{S}$, where Ξ is typically the data space, and $\mathcal{E}$ describes the inner randomness of the algorithm. We also note that the randomness of an algorithm S may be either implicitly or explicitly parameterized by $\mathcal{E}$. Nonetheless, we keep the dependence on $\mathcal{E}$ explicit in order to more precisely separate different sources of randomness in our statements.

Definition 2 (Stability)**.** *A randomized algorithm $\hat{S}:\Xi\times\mathcal{E}\to\mathcal{S}$ is (η,τ,ν)-stable w.r.t. a measure $\mathcal{P}$ on Ξ if there exists a r.v. S_0, possibly dependent on $\mathcal{P}$, such that $\mathbb{P}\{\hat{S}(\xi,\varepsilon) \approx^{|\xi}_{\eta,\tau} S_0\} \geq 1-\nu$.*

In words, a randomized algorithm $\hat{S}$ is stable if there exists a reference r.v. S_0 such that, for almost any inputs $\xi\in\Xi$ (up to a probability ν) sampled from the distribution $\mathcal{P}$, the distribution of $\hat{S}(\xi,\varepsilon)$ resembles that of S_0. Essentially, this means that, for most inputs ξ, the algorithm's output (that randomly depends on ε) behaves as if governed by a fixed distribution, independent of the specific input. We now examine how stability can be leveraged for selection among confidence sets.

Let $\zeta\in\mathcal{Z}$ and $\xi\in\Xi$ be two random variables with arbitrary dependence. Suppose that there exists a set of (possibly random) confidence intervals $\{\mathrm{CI}_s^{\alpha} | s\in\mathcal{S}\}$, each correlated with ξ (for instance, ξ may be a vector of size $|\mathcal{S}|$ containing the size of all sets CI_s^{α}), such that, for all $s\in\mathcal{S}$, we have

$$\mathbb{P}\left\{\zeta \notin \mathrm{CI}_s^{\alpha}\right\} \leq \alpha, \tag{3}$$

for some $\alpha\in(0,1)$. Moreover, let $\hat{S}(\xi,\varepsilon)$ define an arbitrary selection algorithm. For example, $\hat{S}(\xi,\varepsilon)$ might be biased towards selecting smaller size confidence sets. Without further assumptions, individual guarantees (3) do not translate to the selected interval $\mathrm{CI}^{\alpha}_{\hat{S}(\xi,\varepsilon)}$, i.e. $\mathbb{P}\{\zeta\notin\mathrm{CI}^{\alpha}_{\hat{S}(\xi,\varepsilon)}\}\leq\alpha$ is *not* guaranteed to hold. Nonetheless, in the following theorem, we show that if $\hat{S}$ is (η,τ,ν)-stability, then an adjustment of the confidence level in (3) is sufficient to account for the effects of selection.

Theorem 1 (Valid stable selection). *Assume that* $\mathbb{P}\{\zeta \notin \mathrm{CI}_s^\alpha\} \leq \alpha$ *holds for all* $s \in \mathcal{S}$. *If* $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ *is an* (η, τ, ν)*-stable selection algorithm, then,*

$$\mathbb{P}\Big\{\zeta \notin \mathrm{CI}^\alpha_{\hat{S}(\xi,\varepsilon)}\Big\} \leq \alpha e^\eta + \tau + \nu. \tag{4}$$

3.2 Application to Conformal Prediction

In the context of conformal prediction, $\mathcal{S} = \{1, \ldots, K\}$ and $\{\mathrm{CI}_s^\alpha\}_{s\in\mathcal{S}}$ are the conformal prediction sets at X, i.e. $\{C_i^\alpha(X)\}_{i=1}^K$, where each set $C_i^\alpha(X)$ is assumed to satisfy the coverage guarantee (1). The notion of (η, τ, ν)-stability enables one to tackle the challenge of favoring the smallest among the K conformal prediction sets $\{C_i^\alpha(X)\}_{i=1}^K$. The r.v. ζ represents the output Y and we define

$$\xi := [\lambda\left(C_1^\alpha(X)\right), \ldots, \lambda\left(C_K^\alpha(X)\right)], \tag{5}$$

where $\lambda(C_i^\alpha(X))$ represents a "size" (for example, a scaled Lebesgue's measure, counting measure, or more generically, any notion of set desirability) of set $C_i^\alpha(X)$, for all $i \in [K]$. Now note that ν introduced in Definition 2 is a function of the distribution of ξ. In line with conformal prediction methods, which benefit from distribution-free guarantees, we will focus on selection algorithms for which $\nu = 0$, and will actually obtain results that hold almost surely on ξ. With a slight abuse of notation, we will call these algorithms (η, τ)-stable (or simply η-stable if, additionally, $\tau = 0$). We are ready to specialize Theorem 1 to conformal prediction.

Corollary 1 (Smallest conformal set selection). *Let* $\hat{S}$ *be an* (η, τ)*-stable selection algorithm (e.g., for approximating* $\arg\min_{i\in[K]} \lambda(C_i^\alpha(X))$*). Then, we have*

$$\mathbb{P}\left\{Y \in C^\alpha_{\hat{S}(\xi,\varepsilon)}(X)\right\} \geq 1 - \alpha e^\eta - \tau. \tag{6}$$

Note that the standard $1-\alpha$ coverage can be achieved by simply adjusting the confidence level of the K individual sets to $1 - (\alpha - \tau)e^{-\eta}$. In what follows, inspired by the differential privacy literature [18, 24], we provide several examples of easily implementable stable selection algorithms.

Lemma 1 (Stability via Laplace noise). *Assume* $\lambda(C_i^\alpha(X)) \in [0, 1]$, *for all* $i \in [K]$. *If* $\varepsilon \sim (\mathrm{Lap}\,(1/\eta))^{\otimes K}$, $\varepsilon \perp\!\!\!\perp \xi$, *then the selection algorithm* $\hat{S}$ *such that* $\hat{S}(\xi, \varepsilon) := \arg\min_{i\in[K]} \{\lambda\left(\mathrm{C}_i^\alpha(X)\right) + \varepsilon_i\}$ *is* η*-stable.*

Lemma 2 (Stability via exponential mechanism). *Assume* $\lambda(C_i^\alpha(X)) \in [0, 1]$, *for all* $i \in [K]$. *Then, the selection algorithm* $\hat{S}$ *with*

$$\mathbb{P}\left\{\hat{S}(\xi, \varepsilon) = i \middle| \xi\right\} = \frac{\exp(-\eta\lambda(\mathrm{C}_i^\alpha(X)))}{\sum_{j\in[K]} \exp\left(-\eta\lambda\left(\mathrm{C}_j^\alpha(X)\right)\right)}$$

is 2η*-stable. Note that we do not need to make the distribution of* ε *or the mapping* $\hat{S}$ *explicit here.*

We note that the assumption that $\lambda(C_i^\alpha(X)) \in [0, 1]$ for any $i \in [n]$ was used to alleviate notations and may be relaxed up to appropriate scaling of the mechanisms. More importantly, while the two selection algorithms discussed above satisfy the stability requirement, they are adapted from the literature of differential privacy, which is strictly stronger than our required notion of stability. Thus, this additional strength can lead to overly conservative behavior when only stability is required.

To address this, we propose a new Minimum Stable Expectation (MinSE) selection mechanism, designed to achieve stability as tightly as possible. MinSE relies on a *prior* $b \in \Delta^{K-1}$, that encodes prior knowledge on which interval to select, before observing the different predictive intervals at X.

Lemma 3 (Minimum stable expectation). *Let* $\eta, \tau \geq 0$, *and a fixed* $b \in \Delta^{K-1}$, *and consider the following linear program*

$$\begin{aligned} p^\star(b, \xi) = \arg\min_{p} \quad & \sum_{i=1}^K p_i\, \lambda\left(C_i^\alpha\left(X\right)\right) \\ \text{s.t.} \quad & p \in \Delta^{K-1}, \;\; s \in \mathbb{R}_+^K, \;\; p_i \leq e^\eta b_i + s_i, \;\; \sum_{i\in[K]} s_i \leq \tau \end{aligned} \tag{MinSE}$$

Then, the selection algorithm $\hat{S}(\xi, \varepsilon)$ *with* $\mathbb{P}\{\hat{S}(\xi, \varepsilon) = i|\xi\} = p_i^\star(b, \xi)$ *is* (η, τ)*-stable.*

All three lemmas propose randomized selection algorithms which assign the highest probability to the smallest set. However, different from the first two, which assign nonzero probability to all the sets, MinSE assigns zero probability to the largest, whenever feasible.

To develop further intuition of MinSE, consider the case where $b = (1/K)_{i\in[K]}$ represents a uniform prior. Then, if we set $e^\eta = K, \tau = 0$, MinSE reduces to $\arg\min_i \{\lambda(C_i^\alpha(X)), i = 1, \ldots, K\}$, thus selecting deterministically the smallest set. This is reasonable, as then, for $1-\alpha$ after selection, the original confidence sets should have $1 - \alpha/K$ coverage, which corresponds to a Bonferroni correction [25]. Now consider a less extreme example with $e^\eta = 2, \tau = 0$, and b still a uniform prior. By direct checking of the feasibility conditions, one can show that MinSE will never choose any of the $\lfloor K/2 \rfloor$ largest sets. In addition, it is possible to show that MinSE achieves a notion of optimality among all stable selection mechanisms.

Proposition 1 (Optimality of MinSE)**.** *Let $\mathcal{A} : \Xi \times \mathcal{E} \to [K]$ be an (η, τ)-stable algorithm w.r.t. a measure $\mathcal{P}$ on Ξ. Then there exists a prior vector $b \in \Delta^{K-1}$ such that it holds $\mathcal{P}$-almost surely that*

$$\sum_{i=1}^K p_i^\star(b,\xi)\lambda(C_i^\alpha(X)) \le \sum_{i=1}^K p_i^{\mathcal{A}}(\xi)\lambda(C_i^\alpha(X)),$$

where $\xi = [\lambda(C_1^\alpha(X)), \ldots, \lambda(C_K^\alpha(X))]$, the vector $p^\star(b,\xi) = (p_1^\star(b,\xi), \ldots, p_K^\star(b,\xi))$ is the output of MinSE with parameters $\eta, \tau \ge 0$ and prior b, and $p_i^{\mathcal{A}}(\xi) \coloneqq \mathbb{P}_\varepsilon\{\mathcal{A}(\xi,\varepsilon) = i|\xi\}$.

In words, for any (η, τ)-stable algorithm, there exists a prior vector b such that the distribution of the output of MinSE achieves the smallest expected size.

3.3 MinSE examples, tightness, and extensions

Example 1 (Worst-case Oracles)**.** Consider K oracle confidence interval methods such that for any $X \in \mathcal{X}$, an index $j \in [K]$ is chosen uniformly at random, and oracle j outputs $C_j(X) = \emptyset$, while all other oracles $i \neq j$ output $C_i(X) = \mathcal{Y}$. This setup provides a scenario to analyze the tightness of the stability guarantee. For simplicity, let $\lambda(\mathcal{Y}) > 0$ and $\lambda(\emptyset) = 0$. For any datapoint (X, Y), each oracle i individually has marginal miscoverage $\mathbb{P}\{Y \notin C_i(X)\} = 1/K$. Let $\exp(\eta)/K + \tau \le 1$, applying MinSE with parameters (η, τ) and a uniform prior b, we examine the post-selection miscoverage $\mathbb{P}(Y \notin C_{\hat{S}})$. Miscoverage occurs only if the empty oracle is selected. By construction, one and only one set miscovers and has minimal size. MinSE assigns the maximum possible probability $p_j^* = \exp(\eta)/K + \tau$ to the zero-size set j. Thus, the miscoverage probability is $\exp(\eta)/K + \tau$. This result exactly matches the upper bound from Corollary 1, demonstrating that the stability bound is indeed tight in worst-case scenarios. Arguably, this is a pathological setting due to the dependence structure of the confidence sets. Nonetheless, we empirically show in the next example that the stability coverage bound is tight, even with independent confidence sets.

Example 2 (Coin flips)**.** Let $\mathcal{Y} = [0, 1]$ and each of the K sets be constructed using an independent coin flip: for each $i \in [K]$, $C_i^\alpha(X)$ is equal to $\mathcal{Y}$ with probability $1-\alpha$ and to $\emptyset$ with probability α. In this example, we appropriately adjust the coverage of the original coin flips to achieve coverage of $1-\alpha$ after η-stable selection. Contrary to Example 1, here the different oracles are independent. As shown in Figure 1, the η-stable selection using MinSE almost exactly achieves $1-\alpha$ coverage across different stability levels η, particularly as K grows. This suggests that while the stable selection mechanism inflates the probability of yielding the full set $\mathcal{Y}$, selecting the set with the minimum size effectively counterbalances this inflation. Therefore, even with oracle independence, without additional assumptions, the inflation of the original sets is necessary.

Example 3 (Toy regression model)**.** Let $Y = |X| + \mathcal{N}(0, 0.25)$, with $X \sim \text{Uniform}([-1, 1])$, and consider the following two predictors $f_1(X) = X$ and $f_2(X) = -X$. This setup could arise, for example, when the data is split during training, e.g., due to privacy or design constraints, and is meant to be favourable for our method. Indeed, on each half-space ($X \ge 0$ or $X \le 0$), one confidence interval is much smaller than the other one. Numerical results, given to the right of Figure 1, show that despite the inflation in set size, stable selection leads to an improvement in the average set size, in comparison to relying on any individual predictor, while guaranteeing the same coverage. Numerically, the individual sets are about 30% wider on average. This improvement highlights the advantage of using stable selection in settings where different predictors have complementary strengths, i.e. are accurate on different subsets of $\mathcal{X}$.

Remark 1 (Adaptive version—AdaMinSE)**.** A practical consideration when using MinSE is the selection of the stability parameters η and τ. Given K conformal predictors, each achieving at

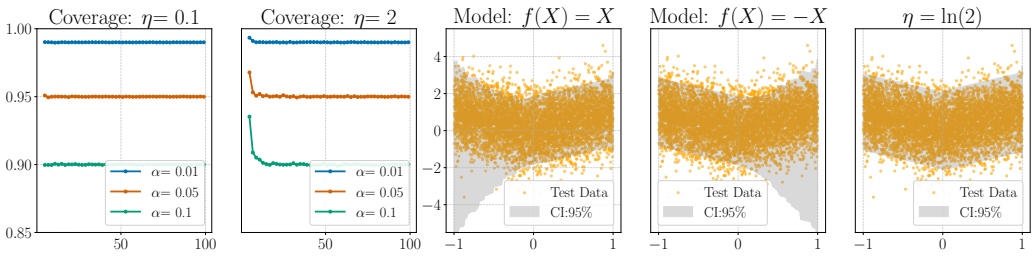


Figure 1: (Left-example 2) Coverage of the coin flips example after selection using MinSE with parameters $\eta \in \{0.1, 2\}$ and $\tau = 0$. Before selection, the oracle returns the full set $[0, 1]$ with probability $1 - \alpha \exp(-\eta)$. (Right-example 3) Second and third figures reflecting adaptive conformal intervals, with miscoverage $\alpha = 0.05$, obtained using the models $f(X) = X$ and $f(X) = -X$. Fifth figure shows the stable selection applied to the conformal sets in the first two figures, adjusted to have miscoverage $\alpha = 0.025$, using MinSE with $\eta = \ln(2)$.

least $1 - \alpha'$ coverage, and a desired post-selection coverage of $1 - \alpha$, any pair (η, τ) satisfying $\alpha' \leq (\alpha - \tau)e^{-\eta}$ is theoretically valid. This presents a choice, as the trade-off between η and τ is not always immediately obvious. To address this, we introduce an adaptive version of MinSE, AdaMinSE, in Appendix A.1. This method automatically optimizes the (η, τ) trade-off to achieve the target $1 - \alpha$ coverage after selection, thereby alleviating the need for manual parameter tuning.

Remark 2 (Derandomization)**.** The randomized output of the stable selection mechanisms, such as MinSE, may be undesirable in certain applications where deterministic prediction sets are required. To address this limitation, it is possible to derandomize the selection process. Building upon techniques from [26], one can construct a single, deterministic prediction set from the output probabilities of the stable selection rule, while still preserving a (correspondingly adjusted) coverage guarantee. The precise results are deferred to Appendix A.2.

Remark 3 (Conditional Coverage)**.** For clarity and ease of exposition, our main analysis has focused on marginal coverage guarantees as defined in (1). However, the stability-based selection framework naturally extends to scenarios where the underlying conformal predictors satisfy stronger guarantees, such as conditional coverage. We provide further details on this extension in Appendix A.3.

4 Extension to Online Conformal Prediction

Next, we show how the framework based on algorithmic stability introduced above extends naturally to the online setting. Consider a collection of K online conformal prediction algorithms that, at each time step $t \in \mathbb{N}$, produce K prediction sets $\{C_i^{(t)}(X_t)\}_{i=1}^K$ for the label Y_t. As noted in Section 2, each prediction set depends on the entire history $\{(X_i, Y_i)\}_{i=1}^{t-1}$. Similarly to Section 3.2, we define $\xi_t := [\lambda(C_1^{(t)}(X_t)), \ldots, \lambda(C_K^{(t)}(X_t))]$, where $\lambda(C_i^{(t)}(X_t))$ denotes the size of the set $C_i^{(t)}(X_t)$, for each $i \in [K]$. The following corollary specializes Theorem 1 to this scenario, showing that an adjusted coverage guarantee can be obtained if a stable selection is applied at each time step t.

Corollary 2 (Smallest online conformal set selection)**.** *At each time $t \in \mathbb{N}$, let $\hat{S}(\xi_t, \varepsilon_t)$ be a (η, τ)-stable selection algorithm (e.g., for approximating $\arg\min_{i \in [K]} \lambda(C_i^{(t)}(X_t))$). Assume each $C_i^{(t)}(\cdot)$ satisfies the guarantee (2). Moreover, let the elements of the sequence $(\varepsilon_t)_{t \in \mathbb{N}}$ be independent. Then,*

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}\left\{ Y_t \in C^{(t)}_{\hat{S}(\xi_t, \varepsilon_t)}(X_t) \right\} \geq 1 - \alpha e^{\eta} - \tau. \tag{7}$$

In essence, given multiple online conformal prediction algorithms, applying a stable selection mechanism provides a practical and systematic way to combine them. It is worth noting that while the coverage guarantee in (7) matches the one established for the batch setting in (6), the guarantee in (7) is achieved only in the long run. This distinction aligns with the nature of online conformal prediction, as discussed in Section 2. Finally, we emphasize that, unlike (2), which ensures coverage for the empirical average $\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left\{ Y_t \in C^{(t)}(X_t) \right\}$, the guarantee in (7) is in probability. However, the only source of randomness in this setting is the selection noise ε_t in the algorithm $\hat{S}(\xi_t, \varepsilon_t)$, as all other quantities can be treated as adversarial or fixed through conditioning.

Algorithm 1 Adaptive COMA (AdaCOMA)

Input: K conformal algorithms $\{C_i^{(t)}\}_{i=1}^K$, stability parameter (η, τ), initial weights $w^{(1)} = (1/k, \ldots, 1/k)$
For: $t = 1, 2, \ldots$
Compute $w^{(t)}$ using COMA.
Compute $p^\star((w^{(t)}), \xi_t) \in \Delta^{K-1}$ using MinSE with $b = w^{(t)}$ and parameters (η, τ)
Output: Any of the following two options:
Option 1: Combined set $C_{\text{comb}}^{(t)}(X_t)$ equal to $\left\{y \in \mathcal{Y} \middle| \sum_{i=1}^K p_i^\star(w^{(t)}, \xi_t) \mathbb{1}\left\{y \in C_i^{(t)}(X_t)\right\} \geq \frac{1}{2}\right\}$
Option 2: Combined predictor leading to $C_{\hat{S}_{(\xi_t, \varepsilon_t)}}^{(t)}(X_t)$, with $\mathbb{P}\left\{\hat{S}_{(\xi_t, \varepsilon_t)} = i | \xi_t\right\} = p_i^\star(w^{(t)}, \xi_t)$

4.1 Adaptive Conformal Online Model Aggregation.

Conformal Online Model Aggregation (COMA) [10] extends online conformal prediction by addressing the challenge of model aggregation. It combines prediction sets from multiple algorithms through a voting mechanism, where weights are dynamically adjusted over time based on past performance. Formally, at each time step t, COMA assigns weights $w^{(t)} = [w_1^{(t)}, \ldots, w_K^{(t)}] \in \Delta^{K-1}$, which reflect the relative importance of each of the K underlying conformal predictors according to the following rule $w_i^{(t)} \propto \exp\left(-\gamma_{t-1} \sum_{j=1}^{t-1} \lambda\big(C_i^{(t)}(X_j)\big)\right)$, where γ_t is the adaptive learning rate from the AdaHedge algorithm [27], an adaptive version of the Hedge algorithm [28]. COMA then outputs the aggregated prediction set $C^{(t)} := \left\{y \in \mathcal{Y} \middle| \sum_{i=1}^K w_i^{(t)} \mathbb{1}\{y \in C_i^{(t)}(X_t)\} \geq \frac{1}{2}\right\}$, which can be interpreted as the aggregated set obtained by selecting the i-th conformal set with probability $w_i^{(t)}$.

Non-adaptiveness of COMA. Crucially, at time t, the COMA framework assigns weights to the K conformal algorithms using only the observations up to time $t-1$, *without* access to X_t or the prediction sets $\{C_i^{(t)}(X_t)\}_{i=1}^K$. Furthermore, due to its AdaHedge formulation, COMA optimizes the weights on average over time, without adapting to each individual X_t.

AdaCOMA. To achieve the best of both worlds, we incorporate COMA into our stable selection algorithm. Specifically, at iteration t, we use COMA's weights $w^{(t)}$ as the prior for a stable selection mechanism, which selects after observing both X_t and the sets $\{C_i^{(t)}(X_t)\}_{i=1}^K$, allowing for pointwise adaptability. The combined procedure, termed AdaCOMA, is detailed in Algorithm 1.

COMA does not have assumption-free coverage guarantees for a fixed target level. However, letting

$$\beta_t := \mathbb{E}\left[\sum_{i=1}^K w_i^{(t)} \mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\}\right], \tag{8}$$

Gasparin and Ramdas [10] show that the following holds

$$\mathbb{P}\left\{Y_t \notin C^{(t)}\right\} \leq 2\beta_t. \tag{9}$$

In the remainder of this section, we analyze the coverage guarantees of AdaCOMA in comparison to the bound in (9) for COMA. Detailed bounds on β_t under additional assumptions, are given in Gasparin and Ramdas [10].

Proposition 2 (Adaptive COMA)**.** *Consider Algorithm 1, and let β_t be defined as in (8). Then,*

- *the set $C_{\text{comb}}^{(t)}(X_t)$ satisfies $\mathbb{P}\left\{Y_t \in C_{\text{comb}}^{(t)}(X_t)\right\} \geq 1 - 2\left(\beta_t e^\eta + \tau\right)$,*
- *the set $C_{\hat{S}_{(\xi_t, \varepsilon_t)}}^{(t)}(X_t)$ satisfies $\mathbb{P}\left\{Y_t \in C_{\hat{S}_{(\xi_t, \varepsilon_t)}}^{(t)}(X_t)\right\} \geq 1 - \beta_t e^\eta - \tau$.*

Proposition 2 demonstrates that AdaCOMA inherits the flexibility of COMA while improving its adaptability to current observations through the stable selection mechanism.

5 Post-selection Calibration in Split Conformal Prediction

The coverage bounds derived from our stability-based approach (e.g., Corollary 1) are distribution-free and hold under minimal assumptions, allowing them to extend to even the adversarial online case. While potentially tight in worst-case scenarios, these bounds can be conservative when additional structure is available, particularly in the batch setting. By leveraging the inherent rank structure of the split conformal method, this section develops a recalibration procedure specifically for split conformal prediction, aiming to achieve tight finite-sample coverage guarantees after selection.

We operate within the standard split conformal setup introduced in Section 2, with a calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^m$ and a test point (X, Y). We assume the sequence of $m+1$ data points $\{(X_1, Y_1), \ldots, (X_m, Y_m), (X, Y)\}$ to be exchangeable. We consider K base predictors $f_1, \ldots, f_K$ and corresponding non-conformity score functions $s_1, \ldots, s_K$. For each base predictor $k \in [K]$ and datapoint $i \in [m]$, denote the non-conformity scores as $s_{k,i} \coloneqq s_k(X_i, Y_i, f_k)$. We use $s_{k,(r)}$ to denote the r-th order statistic of $s_{k,1}, \ldots, s_{k,m}$. Finally, for any k and i, we denote the rank of a score $s_{k,i}$ as $R_{k,i} \coloneqq \sum_{j=1}^m \mathbb{1}\{s_{k,j} \leq s_{k,i}\}$. Using these definitions, we can parameterize the conformal prediction sets using ranks, i.e., for any rank index $R \in [m]$, we define

$$C_k(X, R) \coloneqq \left\{y \in \mathcal{Y} : s_k(X, y, f_k) \leq s_{k,(R)}\right\},$$

which recovers the classical set C^α recalled in Section 2 for $R_\alpha = \lceil (1-\alpha)(m+1) \rceil$, satisfying, for any $k \in [K]$, $\mathbb{P}\left\{Y \in C_k(X, R_\alpha)\right\} \geq 1 - \alpha$.

Calibration After Selection using Effective Ranks. We now introduce an arbitrary (stochastic) selection rule $\hat{S}$. Our goal is to determine an *effective rank* $\hat{R}_\alpha$ such that a similar property holds, but *for the selected interval*, i.e., $\mathbb{P}\{Y \in C_{\hat{S}(X,\varepsilon)}(X, \hat{R}_\alpha)\} \geq 1 - \alpha$. To that end, we use a recalibration process after selection, that uses the effective ranks *as meta-scores*. For each point $i \in [m]$, we apply the selection rule using its feature vector X_i and *independent* randomness $\varepsilon_i \sim \mathcal{P}_\varepsilon$. We now define the effective rank (or the meta-score) for the i-th point as:

$$\hat{R}_i \coloneqq R_{\hat{S}(X_i, \varepsilon_i), i},$$

that is, the rank of the i-th point's score calculated using the selected predictor $\hat{S}(X_i, \varepsilon_i)$. Subsequently, we define the sequence of effective ranks $\mathcal{R} \coloneqq (\hat{R}_1, \ldots, \hat{R}_m)$. Using uniform random tiebreaks between equal ranks in $\mathcal{R}$, for $t \in [m]$, we use $\hat{R}_{(t)}$ to denote the t-th order statistics of $\mathcal{R}$.

Theorem 2. *Assume that $\hat{S}$ is independent of $\mathcal{D}_{\text{cal}}$. Let $\tau_\alpha = \lceil (1-\alpha)(m+1) \rceil \leq m$. Then,*

$$\mathbb{P}\left\{Y \in C_{\hat{S}(X,\varepsilon)}\left(X, \hat{R}_{(\tau_\alpha)}\right)\right\} \geq 1 - \alpha.$$

Theorem 2 provides a method to maintain coverage guarantees after selecting among multiple split conformal predictors. The use of effective ranks acts as a meta-score, allowing for the calibration of predictors even if they utilize different non-conformity score functions $s_k(\cdot)$, offering flexibility in model-specific score design. The theorem states that by selecting the appropriate order statistic $\hat{R}_{(\tau_\alpha)}$ of these effective ranks, derived using an independent selection rule $\hat{S}$, the conformal set $C_{\hat{S}(X,\varepsilon)}(X, \hat{R}_{(\tau_\alpha)})$ for the chosen predictor $\hat{S}(X, \varepsilon)$ achieves the desired coverage.

Constructing an Independent $\hat{S}$. Theorem 2 mandates the selection rule $\hat{S}$ to be independent of the calibration data $\mathcal{D}_{\text{cal}}$. This independence plays a critical role in the proof as it ensures that the effective ranks $\hat{R}_1, \ldots, \hat{R}_m$ are exchangeable with the unobserved effective rank of the test point. Consequently, if $\hat{S}$ aims to select the smallest set, it cannot use set sizes derived from $\mathcal{D}_{\text{cal}}$-based quantiles due to the induced dependency. To ensure independence, we employ an auxiliary dataset $\mathcal{D}_{\text{aux}}$, disjoint from and independent of $\mathcal{D}_{\text{cal}}$. For each predictor $k \in [K]$ and test point X, proxy quantiles $\hat{q}^{\text{aux}}_{\tilde{\alpha},k}$, computed from $\mathcal{D}_{\text{aux}}$ (at a preliminary miscoverage rate $\tilde{\alpha}$), define proxy conformal sets $C^{\text{aux}}_k(X)$ and their corresponding sizes $\lambda(C^{\text{aux}}_k(X))$. The vector of these proxy sizes, $\xi^{\text{aux}}(X) \coloneqq [\lambda(C^{\text{aux}}_1(X)), \ldots, \lambda(C^{\text{aux}}_K(X))]$, is then independent of $\mathcal{D}_{\text{cal}}$. A selection rule $\hat{S}(X, \varepsilon)$ based solely on X and this $\xi^{\text{aux}}(X)$ (e.g., employing stable mechanisms from Section 3.2, such as MinSE) thus satisfies the independence condition. This allows Theorem 2 to be applied directly for

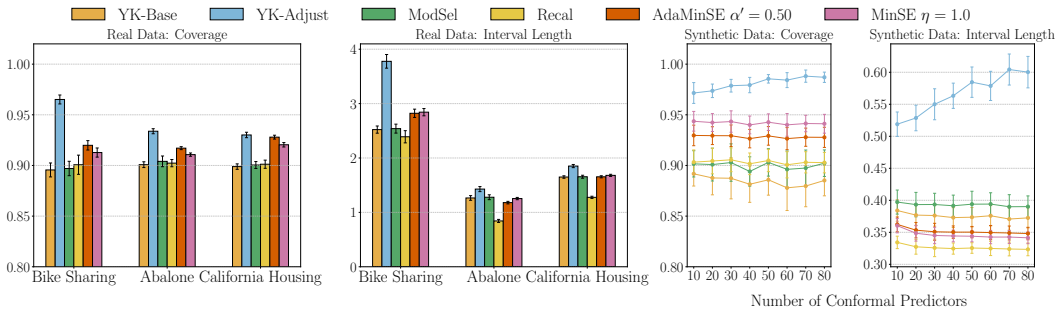


Figure 2: Marginal coverage and average lengths, for real datasets (Bike Sharing, Abalone, California Housing) and synthetic data; Error bars represent twice the standard error of mean estimation using multiple seeds for real datasets and 2 s.d. for the synthetic data.

$1-\alpha$ coverage, avoiding the inflation factors inherent in stability-based bounds that depend on $\mathcal{D}_{\text{cal}}$. The practical effectiveness of such selection hinges on how accurately the proxy information from $\mathcal{D}_{\text{aux}}$ reflects the true characteristics based on $\mathcal{D}_{\text{cal}}$.

6 Experiments

To illustrate our approach in practice, we present here two simple experimental setups, one on synthetic and one on real data. We defer online experiments, additional batch experiments, and further experimental details to Appendix C. We compare our approach to Yang and Kuchibhotla [12] and Liang et al. [11], denoted as YK and ModSel, respectively. As Yang and Kuchibhotla [12] proposed multiple algorithms, adopting the following naming convention from Liang et al. [11], we compare against YK-Adjust and YK-Base. YK-Adjust adjusts the underlying conformal predictors to ensure valid coverage after selecting the best-on-average conformal predictor on the calibration dataset. YK-Base simply selects the best-on-average conformal predictor and does not have coverage guarantees. We report the performance of MinSE with parameters $(\eta = 1, \tau = 0)$ and AdaMinSE with $\alpha' = 0.05$. In addition, we report the performance of Recal, which is based on AdaMinSE with $\alpha' = 0.02$, followed by the recalibration procedure of Section 5. For Recal, the calibration dataset is further split into two blocks to construct $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{aux}}$, so that the selection satisfies the independence requirement of Theorem 2. For the experiments presented here, we target a miscoverage level of $\alpha = 0.1$. Except for YK-Base, all methods guarantee miscoverage $\alpha = 0.1$ after selection.

Throughout the experiments, we aim to design a scenario where the performance of individual predictors varies across the input space. As such, using clustering, we split the feature space into 5 disjoint sets and train each predictor exclusively on a randomly selected subset. We provide additional experiments without such data splitting in Appendix C. Moreover, the code to reproduce the experiments is available in the supplementary material[1].

Synthetic Regression. We generate n data points, $\{(X_i, Y_i)\}_{i=1}^n$, with $X_i \sim \mathcal{N}(0, I_d)$ (before feature space splitting), and the response variable defined as $Y_i = \sin(\langle \beta, X_i \rangle) + 0.1\mathcal{N}(0,1)$, where β is the vector $\{1/d\}_{i \in [d]}$. The feature dimension is set to $d = 10$, and the training data is split into two blocks. In the first block, we train K distinct regression models, $f_1, \ldots, f_K$, using the Kernel Ridge Regression model from scikit-learn [29]. For each model, we randomly sample the kernel function (either linear or radial basis function (RBF)) and the regularization parameter (uniformly chosen between 0.1 and 1). For each model i, we use the second block of training data to train a random forest model g_i that predicts the absolute residuals $|f_i(X) - Y|$, enabling us to use the nonconformity score, defined as $s_i(X, Y) = |f_i(X) - Y|/g_i(X)$. We use 400 datapoints for the calibration dataset.

Real Datasets. In this experiment, we aim to model a more typical data analysis scenario. We conduct experiments on three standard regression datasets: Abalone, California Housing, and Bike Sharing [30, 31]. For each dataset, leveraging `scikit-optimize` for hyperparameter tuning [32], we used the following `scikit-learn` models: AdaBoostRegressor, DecisionTreeRegressor, GradientBoost-

[1]Code also available at Valid-Selection-among-Conformal-Sets.

ingRegressor, ElasticNet, RandomForestRegressor, and LinearRegression. We used 80% of the data for training, 10% for calibration, and 10% for testing. Similar to the synthetic experiments, we used the same adaptive score function, with RandomForestRegressor trained to predict the residuals. Then, up to adjusting coverage to ensure post-selection coverage of $\alpha = 0.1$, we used the same conformal predictors for all selection methods. We normalized the labels across datasets to keep the size of conformal sets comparable.

Both synthetic and real data results are reported in Figure 2. For both experiments, YK-Adjust, MinSE, and AdaMinSE overcover, while Recal and ModSel achieve similar coverage to the target marginal coverage of 0.9. YK-Base undercovers in the synthetic setting. Differences emerge in the resulting average interval lengths. Our proposed Recal consistently performs the best on both the synthetic and real data settings. On real datasets, MinSE and AdaMinSE are competitive with baselines on Abalone and California Housing but produce larger sets on Bike Sharing. Meanwhile, for the synthetic setting, AdaMinSE and MinSE beat the benchmarks.

7 Conclusion

In this paper, we introduced a stability-based framework for selecting among multiple conformal predictors while preserving coverage. By casting selection as an (η, τ)-stable randomized mechanism, we established distribution-free guarantees that transfer the validity of individual predictors to the post-selection set, enabling pointwise (feature-dependent) selection. We instantiated this principle with practical mechanisms and proposed the MinSE mechanism, which is optimal among stable selectors, along with adaptive and derandomized variants. We further extended the framework to the online setting; combined with online aggregation, this yields AdaCOMA, which uses COMA weights as a prior for stable per-time-step selection based on the current features and sets, thereby adapting over time and across each input. Finally, for split conformal prediction, we introduced a post-selection recalibration via effective ranks that mitigates the conservativeness of worst-case stability bounds. Empirically, our methods meet the target coverage and often reduce set sizes relative to existing selection approaches across synthetic and real datasets in heterogeneous settings. Nonetheless, the stability-based guarantees are worst-case and can be conservative in benign regimes; our recalibration reduces this conservativeness but requires an auxiliary dataset, which is independent from calibration, and its effectiveness depends on the quality of proxy quantiles. Moreover, randomized selection and the choice of a stability budget introduce utility/validity trade-offs, and extending the framework by deriving tighter instant-dependent coverage bounds under additional assumptions remains an open direction.

Acknowledgement

The work of Aymeric Dieuleveut and Mahmoud Hegazy is supported by French State aid managed by the Agence Nationale de la Recherche (ANR) under France 2030 program with the reference ANR-23-PEIA-005 (REDEEM project), and ANR-23-IACL-0005, in particular Hi!Paris FLAG chair. Liviu Aolaritei acknowledges support from the Swiss National Science Foundation through the Postdoc.Mobility Fellowship (grant agreement P500PT_222215). Additionally, this project was funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This publication is part of the Chair "Markets and Learning", supported by Air Liquide, BNP PARIBAS ASSET MANAGEMENT Europe, EDF, Orange and SNCF, sponsors of the Inria Foundation.

References

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[2] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[3] Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

[4] Tijana Zrnic and Michael I Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691, 2023.

[5] Raef Bassily and Yoav Freund. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016.

[6] Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[7] Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

[8] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

[9] Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *arXiv preprint arXiv:2205.09095*, 2022.

[10] Matteo Gasparin and Aaditya Ramdas. Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*, 2024.

[11] Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after efficiency-oriented model selection. *arXiv preprint arXiv:2408.07066*, 2024.

[12] Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, pages 1–13, 2024.

[13] Rui Luo and Zhixin Zhou. Conformity score averaging for classification. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Pvfd7NiUS6.

[14] Erfan Hajihashemi and Yanning Shen. Multi-model ensemble conformal prediction in dynamic environments. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[15] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[16] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

[17] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[19] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[20] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife. *The Annals of Statistics*, 49(1):486–507, 2021.

[21] Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

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

[23] Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR, 2023.

[24] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[25] Eric W Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

[26] Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024.

[27] Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

[28] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[30] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[31] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. https://archive.ics.uci.edu, 2023. Accessed: 2023-10-05.

[32] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Zé Vinícius, cmmalone, Christopher Schröder, nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, rene rex, Kejia (KJ) Shi, Justus Schwabedal, carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, Loïc Estève, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch. scikit-optimize/scikit-optimize: v0.5.2, March 2018. URL https://doi.org/10.5281/zenodo.1207017.

[33] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008, 2025.

[34] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

[35] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

[36] Josh Levy-Kramer. k-means-constrained, April 2018. URL https://github.com/joshlk/k-means-constrained.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[38] Ross Wightman. PyTorch Image Models. URL https://github.com/huggingface/pytorch-image-models.

[39] Michael Harries, New South Wales, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

[40] Ruey S Tsay. *Analysis of financial time series*. John wiley & sons, 2005.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the scope and contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While the paper does not summarize all limitations in a single section, it discusses the limitations of the proposed methods at various points. For instance, the potential conservativeness of the stability-based coverage guarantees (e.g., Corollary 1) is acknowledged at the beginning of Section 5, which motivates the development of a specific recalibration approach for split conformal prediction to achieve tighter bounds in that setting. The inherently randomized nature of the primary stable selection mechanisms (like MinSE) is noted as potentially undesirable in some applications (remark after Example 3), leading to the proposal of a derandomization technique in Appendix A.2, though this comes with a degraded coverage guarantee. Furthermore, the recalibration method presented in Section 5 itself has the limitation of requiring the selection rule to be independent of the main calibration data, necessitating an auxiliary dataset, and its effectiveness hinges on the quality of this proxy data, as discussed at the end of that section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Detailed assumptions and complete proofs for all propositions, theorems, and corollaries are provided in Appendix B. All results are properly numbered and referenced, with assumptions clearly stated alongside the formal statements.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information necessary to reproduce the main experimental results—including data generation, evaluation procedures, and implementation details—is provided in Section 6 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Datasets used are either synthetic or open-source under permissive licenses. The code to reproduce all experiments and a guide on how to use it can be found in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the details of the experimental setup in Section 6 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide error bars for all our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational resources, including hardware type (CPU/GPU) and runtime, are described in appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This research adheres to the NeurIPS Code of Ethics. It does not involve human subjects, sensitive data, or real-world deployments with potential societal impact. All results are reproducible, and the code will be released upon acceptance.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is foundational and does not involve specific applications or deployments. While robust prediction can support decision-making in domains such as healthcare or autonomous systems, the paper does not target any particular use case and therefore does not entail direct societal impacts.

    Guidelines: This paper is not directly tied to a particular application.

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not involve the release of pretrained models, large-scale datasets, or tools that pose significant risk of misuse. The research is theoretical and algorithmic in nature, and all released code is for reproducibility of the experiments.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in this work—including code and datasets—are properly cited in the paper, and their licenses have been respected in accordance with the stated terms of use (see Section 6). No proprietary or restricted-access data was used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce new code for our method and experiments, which is documented, included in the supplementary material, and will be made publicly available upon acceptance. The code includes clear instructions for reproducing all results and is structured to facilitate ease of use.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This work does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve human subjects or crowdsourcing and therefore does not require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This research does not involve the use of large language models (LLMs) in any part of the core methodology. Any LLM usage was limited to minor writing assistance and had no impact on the scientific content or originality of the work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Contents

Outline of the Appendix

This appendix provides supplementary material to the main paper. The first part details deferred content from Section 3 of the main paper: Appendix A.1 introduces and proves the Adaptive Minimum Stable Expectation (AdaMinSE) mechanism; Appendix A.2 presents and proves a derandomization technique for prediction sets; and Appendix A.3 discusses and proves an extension to conditional coverage guarantees. Appendix B provides the proofs for the theoretical results presented in Section 3 (on stable selection), Section 4 (on online conformal prediction), and Section 5 (concerning post-selection calibration in split conformal prediction). Finally, Appendix C is dedicated to additional experimental results. These results include further batch experiments under varied settings, online experiments evaluating AdaCOMA, and detailed setup information for these experiments.

A Deferred Content Section 3

A.1 Adaptive Minimum Stable Expectation

One difficulty in using MinSE is tuning the stability parameters η and τ. For instance, assume a user has access to K conformal predictors, each with coverage at least $1-\alpha'$, and wishes to apply MinSE to select among them such that the coverage after selection is at least $1-\alpha$. Then, they may choose any values of τ and η satisfying $\alpha' \leq (\alpha-\tau)e^{-\eta}$. In particular, the utility tradeoff between τ and η is not immediately clear.

To address this, we propose AdaMinSE, an adaptive version of MinSE, which also optimizes over the choice of τ and η. Similarly to MinSE, AdaMinSE also makes use of a prior $b \in \Delta^{K-1}$ (which can be chosen depending on the past as in Algorithm 1). However, instead of requiring the parameters (η, τ) as input, it simply takes the current level of miscoverage α' and the desired miscoverage level α after selection. The following proposition introduces AdaMinSE, together with its coverage guarantee.

Proposition 3 (Adaptive Minimum Stable Expectation)**.** *Let $\alpha', \alpha \in (0,1)$ with $\alpha' \leq \alpha$, and let $b \in \Delta^{K-1}$ be fixed. Consider the following linear program*

$$\begin{aligned} d^\star(b,\xi) = \arg\min_{d} \quad & \sum_{i=1}^{K} d_i \lambda\left(C_i^\alpha(X)\right) \\ \text{s.t.} \quad & d \in \Delta^{K-1}, \quad s \in \mathbb{R}_+^K, \quad \tau, \eta \geq 0, \\ & d_i \leq e^\eta b_i + s_i, \quad \sum_{i\in[K]} s_i \leq \tau, e^\eta \alpha' + \tau \leq \alpha \end{aligned} \tag{AdaMinSE}$$

Let $\mathbb{P}\left\{Y \in C_i^{\alpha'}(X)\right\} \geq 1-\alpha'$, for all $i \in [K]$. Moreover, consider the selection algorithm $\hat{S}(\xi,\varepsilon)$ with $\mathbb{P}\{\hat{S}(\xi,\varepsilon) = i|\xi\} = d^\star(b,\xi)$. Then, $\mathbb{P}\left\{Y \in C^\alpha_{\hat{S}(\xi,\varepsilon)}(X)\right\} \geq 1-\alpha$.

Proof. The result can be recovered as follows

$$\begin{aligned} \mathbb{P}\left\{Y \notin C^\alpha_{\hat{S}(\xi,\varepsilon)}(X)\right\} &= \mathbb{E}\left[\sum_{i=1}^{K} d^\star(b,\xi)\mathbb{1}\left\{Y \notin C_i^{\alpha'}(X)\right\}\right] \\ &\leq \sum_{i=1}^{K} e^\eta b_i \mathbb{P}\left\{Y \notin C_i^{\alpha'}(X)\right\} + \tau \leq e^\eta\alpha' + \tau \leq \alpha, \end{aligned}$$

where the first and last inequalities follow from the two constraint $d_i \leq e^\eta b_i + s_i$ and $e^\eta\alpha' + \tau \leq \alpha$ in AdaMinSE, respectively. □

Proposition 3 ensures that AdaMinSE achieves the desired coverage. By optimizing over (η, τ), it removes the need for manual tuning, making it a practical and reliable approach for selection.

A.2 Derandomizing the Prediction Set

One potential limitation of the stable selection algorithms presented in Section 3.2 is that they produce a random confidence set, which may be undesirable in certain applications. In such cases, the stable selection process can be derandomized using techniques from [26]. This is formalized in the following proposition.

Proposition 4 (Derandomized smallest conformal set)**.** *Let $\hat{S}(\xi, \varepsilon)$ be an (η, τ)-stable selection algorithm, and define $p_i(\xi) \coloneqq \mathbb{P}\left\{\hat{S}(\xi, \varepsilon) = i | \xi\right\}$. Then, consider the derandomized confidence set*

$$C_{\mathrm{dr}}(X, \xi) := \left\{ y \in \mathcal{Y} : \sum_{i=1}^{K} p_i(\xi) \mathbb{1}\left\{y \in C_i^{\alpha}(X)\right\} \geq \frac{1}{2} \right\}.$$

If $C^{\alpha}(\cdot)_1, \ldots, C^{\alpha}(\cdot)_K$ satisfy (1), it holds that $\mathbb{P}\left\{Y \in C_{\mathrm{dr}}(X, \xi)\right\} \geq 1 - 2(\alpha e^{\eta} + \tau)$.

Proof. The proof builds upon the reasoning in [26]. Since $\hat{S}(\xi, \varepsilon)$ is (η, τ)-stability, we know that there exists a fixed point $b \in \Delta^{K-1}$ such that $p_i(\xi) \leq \exp(\eta) b_i + s_i$ and $\sum_{i=1}^{K} s_i \leq \tau$. It follows that

$$\mathbb{E}\left[\sum_{i=1}^{K} p_i(\xi) \mathbb{1}\left\{Y \notin C_i^{\alpha}(X)\right\}\right] \leq \exp(\eta)\, \alpha + \tau.$$

Moreover, using the Markov inequality, we have that

$$\mathbb{P}\left\{Y \notin C_{\mathrm{dr}}(X, \xi)\right\} = \mathbb{P}\left\{\sum_{i=1}^{K} p_i(\xi) \mathbb{1}\left\{Y \notin C_i^{\alpha}(X)\right\} \geq \frac{1}{2}\right\} \leq 2\,\mathbb{E}\left[\sum_{i=1}^{K} p_i(\xi) \mathbb{1}\left\{Y \notin C_i^{\alpha}(X)\right\}\right],$$

from which the result follows immediately. □

We highlight that while the set $C_{\mathrm{dr}}(\cdot, \xi)$ is still random with respect to the randomness of $\{C_i^{\alpha}(\cdot)\}_{i=1}^{K}$, the derandomization here refers to the fact that the stable selection process does not introduce additional randomness due to ε.

A.3 Conditional Coverage

A stronger guarantee than marginal coverage (1) is conditional coverage. Let $G : \mathcal{X} \to \mathcal{G}$ be a function that maps an input X to a group attribute $G(X) \in \mathcal{G}$. A conformal predictor $C(X)$ is said to satisfy $(1 - \alpha)$ conditional coverage with respect to G if,

$$\mathbb{P}\left\{Y \in C(X) | G(X)\right\} \geq 1 - \alpha. \tag{10}$$

This ensures that the coverage guarantee holds not just on average over all X, but also when restricted to specific subpopulations defined by G. An example is the case where $\mathcal{G}$ is finite, which then corresponds to Group-Conditional validity [33]. Such guarantees are crucial for added reliability in many applications.

Our stability-based selection framework can be extended to preserve conditional coverage. Suppose we have K conformal predictors $\{C_i^{\alpha}(X)\}_{i=1}^{K}$, each satisfying $(1 - \alpha)$ conditional coverage with respect to G. That is, for each $i \in [K]$,

$$\mathbb{P}\left\{Y \in C_i^{\alpha}(X) | G(X)\right\} \geq 1 - \alpha. \tag{11}$$

We can now state the conditional coverage guarantee for the selected set.

Proposition 5 (Conditionally valid stable selection)**.** *Assume that each conformal predictor $C_i^{\alpha}(X)$ satisfies $(1 - \alpha)$ conditional coverage with respect to G, for all $i \in [K]$. If $\hat{S} : \Xi \times \mathcal{E} \to [K]$ is an (η, τ)-stable selection algorithm (with $\xi = [\lambda(C_1^{\alpha}(X)), \ldots, \lambda(C_K^{\alpha}(X))]$), then almost surely*

$$\mathbb{P}\left\{Y \in C^{\alpha}_{\hat{S}(\xi, \varepsilon)}(X) \mid G(X)\right\} \geq 1 - (\alpha e^{\eta} + \tau). \tag{12}$$

Proof. We want to bound the conditional miscoverage probability

$$P_{\text{miscover}|G(X)} \coloneqq \mathbb{P}\left\{Y \notin C^{\alpha}_{\hat{S}(\xi(X),\varepsilon)}(X) \mid G(X)\right\}.$$

By the law of total expectation, conditioning on X we have

$$P_{\text{miscover}|G(X)} = \mathbb{E}_{X|G(X)}\left[\mathbb{P}\left\{Y \notin C^{\alpha}_{\hat{S}(\xi(X),\varepsilon)}(X) \mid X, G(X)\right\}\right].$$

Since $G(X)$ is determined by X, the inner probability is $\mathbb{P}\left\{Y \notin C^{\alpha}_{\hat{S}(\xi(X),\varepsilon)}(X) \mid X\right\}$. Since the randomness ε in $\hat{S}(\xi(X),\varepsilon)$ is independent of Y given X, we have

$$\mathbb{P}\left\{Y \notin C^{\alpha}_{\hat{S}(\xi(X),\varepsilon)}(X) \mid X\right\} = \sum_{s=1}^{K} \mathbb{P}\left\{Y \notin C^{\alpha}_{s}(X) \mid X\right\} \mathbb{P}_{\varepsilon}\left\{\hat{S}(\xi(X),\varepsilon) = s \mid X\right\}.$$

Using (η,τ)-stability (conditional on $\xi(X)$) property of $\hat{S}$, we have

$$\begin{aligned}\sum_{s=1}^{K} \mathbb{P}\left\{Y \notin C^{\alpha}_{s}(X) \mid X\right\} \mathbb{P}_{\varepsilon}\left\{\hat{S}(\xi(X),\varepsilon) = s \mid X\right\} &\leq e^{\eta} \sum_{s=1}^{K} \mathbb{P}\left\{Y \notin C^{\alpha}_{s}(X) \mid X\right\} \mathbb{P}\left\{S_0 = s \mid X\right\} + \tau \\ &\leq e^{\eta} \sum_{s=1}^{K} \mathbb{P}\left\{Y \notin C^{\alpha}_{s}(X) \mid X\right\} \mathbb{P}\left\{S_0 = s\right\} + \tau.\end{aligned}$$

for some random S_0 variable with support $[K]$. It follows that

$$\begin{aligned}P_{\text{miscover}|G(X)} &= e^{\eta} \sum_{i=1}^{K} \mathbb{P}\left\{S_0 = i\right\} \mathbb{E}_{X|G(X)}\left[\mathbb{P}\left\{Y \notin C^{\alpha}_{i}(X)\right\} | X\right] + \tau \\ &\leq e^{\eta}\alpha + \tau\,.\end{aligned}$$

□

This result shows that if the original conformal predictors provide conditional coverage, the stable selection mechanism allows for selecting among them while retaining a (correspondingly adjusted) conditional coverage guarantee. The required level for the initial predictors would be $1-(\alpha-\tau)e^{-\eta}$ to achieve $1-\alpha$ conditional coverage post-selection.

B Proofs

B.1 Proofs of Section 3

The proof of Theorem 1 requires the following lemma from Zrnic and Jordan [4], which is inspired by the approach in [34].

Lemma 4. *[4, Lemma 1] Let $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ be an (η,τ,ν)-stable selection algorithm and S_0 be the corresponding random variable. Then, $(\xi, \hat{S}(\xi,\varepsilon)) \approx_{\eta,\tau+\nu} (\xi, S_0)$.*

Proof of Theorem 1. In this proof, similar to analogous results in Zrnic and Jordan [4], a lot of the heavy lifting is done by Lemma 4. Nonetheless, the selection dependence on the data on the confidence sets themselves introduces some subtleties, making direct application of the steps of Zrnic and Jordan [4] non-straightforward. Thus, we take a different starting step by defining a shadow algorithm.

Let us define $\hat{S}' : \Xi \times \mathcal{Z} \times \mathcal{E} \to \mathcal{S}$ such that $\hat{S}'(\xi,\zeta,\varepsilon) = \hat{S}(\xi,\varepsilon)$, for all $(\xi,\zeta) \in \Xi \times \mathcal{Z}$. Then, if $\hat{S}$ is (η,τ,ν)-stable with respect to $\mathcal{P}$ on Ξ, we also have that $\hat{S}'$ is (η,τ,ν)-stable with respect to the product distribution $\mathcal{P}\otimes\mathcal{Q}$, where $\mathcal{Q}$ is the distribution of ζ. Combining this with Lemma 4, we obtain

$(\zeta, \xi, \hat{S}(\xi, \varepsilon)) \approx_{\eta, \tau+\nu} (\zeta, \xi, S_0)$. Now, defining the event $O_\delta \coloneqq \{(\zeta, \xi, \hat{S}(\xi, \varepsilon)) \in \mathcal{Z} \times \Xi \times \mathcal{S} : \zeta \notin \mathrm{CI}^{\delta e^{-\eta}}_{\hat{S}(\xi,\varepsilon)}\}$, we have that

$$\begin{aligned}\mathbb{P}\left\{\zeta \notin \mathrm{CI}^{(\delta e^{-\eta})}_{\hat{S}(\xi,\varepsilon)}\right\} &\leq e^{\eta}\mathbb{P}\left\{(\zeta, \xi, S_0) \in O_\delta\right\} + \tau + \nu \\ &= e^{\eta}\mathbb{P}\left\{\zeta \notin \mathrm{CI}^{\delta e^{-\eta}}_{S_0}\right\} + \tau + \nu \leq e^{\eta}\delta e^{-\eta} + \tau + \nu \leq \delta + \tau + \nu,\end{aligned}$$

where the first inequality follows from the definition of indistinguishability, and the second inequality follows from the assumption that $\mathbb{P}\{\zeta \notin \mathrm{CI}^{\alpha}_{s}\} \leq \alpha$ holds for all $s \in \mathcal{S}$. □

Proof of Corollary 1. The result follows immediately from Theorem 1 with $Y = \zeta$ and $\nu = 0$. □

Proof of Lemma 1. Let $S_0 = \arg\min_{i\in[K]} \varepsilon_i$, with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}\,(1/\eta)$. Moreover, note that $\hat{S}(\xi, \varepsilon) = i$ if and only if $\varepsilon_i \leq \min_{j\neq i}\left\{\varepsilon_j + \lambda\left(\mathrm{C}^{\alpha}_j(X)\right) - \lambda\left(\mathrm{C}^{\alpha}_i(X)\right)\right\} \leq \min_{j\neq i}\varepsilon_j + 1$, where we used the fact that $\lambda(C^{\alpha}_i(X)) \in [0, 1]$. Finally, for all $i \in [K]$

$$\begin{aligned}\mathbb{P}\left\{\hat{S}(\xi, \varepsilon) = i \,\middle|\, \sigma(\xi, \{\varepsilon_j\}_{j\neq i})\right\} &\leq \mathbb{P}\left\{\varepsilon_i \leq \min_{j\neq i}\varepsilon_j + 1 | \sigma(\xi, \{\varepsilon_j\}_{j\neq i})\right\} \\ &\leq e^{\eta}\mathbb{P}\left\{\varepsilon_i \leq \min_{j\neq i}\varepsilon_j | \sigma(\xi, \{\varepsilon_j\}_{j\neq i})\right\} \\ &= e^{\eta}\mathbb{P}\left\{S_0 = i | \sigma(\xi, \{\varepsilon_j\}_{j\neq i})\right\},\end{aligned}$$

where the second inequality uses the fact that the densities ratio $p_{\varepsilon_i - 1}/p_{\varepsilon_i}$ is upper bounded by e^{η}. The result now follows by taking expectation on both sides. □

Proof of Lemma 2. Let S_0 be a uniform r.v. with $\mathbb{P}\{S_0 = i\} = 1/k$. Then,

$$\begin{aligned}\frac{\mathbb{P}\left\{\hat{S}(\xi, \varepsilon) = i \,\middle|\, \xi\right\}}{\mathbb{P}\{S_0 = i\}} &= \frac{k\exp\left(-\eta\lambda\left(\mathrm{C}^{\alpha}_i(X)\right)\right)}{\sum_{i\in[K]}\exp\left(-\eta\lambda\left(\mathrm{C}^{\alpha}_i(X)\right)\right)} \\ &\leq \frac{k\exp(\eta)}{k\exp(-\eta)} = \exp 2\eta,\end{aligned}$$

where we used the fact that $\lambda(C^{\alpha}_i(X)) \in [0, 1]$. □

Proof of Lemma 3. The optimal solution $p^{\star}(b, \xi)$ satisfies $p^{\star}_i(b, \xi) \leq e^{\eta}b_i + s_i$, with $\sum_{i=1}^{K} s_i \leq \tau$, for all $i \in k$. Therefore, letting S_0 be a r.v. with $\mathbb{P}\{S_0 = i\} = b_i$. Then, for all $\mathcal{S} \subseteq [K]$, we have

$$\mathbb{P}\left\{\hat{S}(\xi, \varepsilon) \in \mathcal{S}\right\} \leq e^{\eta}\mathbb{P}\left\{S_0 \in \mathcal{S}\right\} + \tau,$$

which concludes the proof. □

Proof of Proposition 1. By the assumption that $\mathcal{A}$ is (η, τ)-stable w.r.t. $\mathcal{P}$, Definition 2 guarantees the existence of a r.v. S_0 such that $\mathcal{A}(\xi, \varepsilon) \approx^{|\xi}_{\eta,\tau} S_0$ holds $\mathcal{P}$-almost surely. This implies that for any $G \subseteq [K]$, $\mathbb{P}_{\varepsilon}\{\mathcal{A}(\xi, \varepsilon) \in G|\xi\} \leq e^{\eta}\mathbb{P}\{S_0 \in G\} + \tau$ holds $\mathcal{P}$-almost surely.

We set the prior vector $b \in \Delta^{K-1}$ as the distribution of S_0 by choosing $b_i \coloneqq \mathbb{P}\{S_0 = i\}$. Let $p^{\mathcal{A}}_i(\xi) = \mathbb{P}_{\varepsilon}\{\mathcal{A}(\xi, \varepsilon) = i|\xi\}$. Define $s_i(\xi) = \max\left(0, p^{\mathcal{A}}_i(\xi) - e^{\eta}b_i\right)$. Clearly, $s_i(\xi) \geq 0$ and $p^{\mathcal{A}}_i(\xi) \leq e^{\eta}b_i + s_i(\xi)$ for all i. Let $\mathcal{S}^{+}(\xi) = \{i \mid p^{\mathcal{A}}_i(\xi) > e^{\eta}b_i\}$. Then, $\mathcal{P}$-almost surely,

$$\begin{aligned}\sum_{i=1}^{K} s_i(\xi) &= \sum_{i\in\mathcal{S}^{+}(\xi)}\left(p^{\mathcal{A}}_i(\xi) - e^{\eta}b_i\right) = \mathbb{P}_{\varepsilon}\left\{\mathcal{A}(\xi, \varepsilon) \in \mathcal{S}^{+}(\xi)|\xi\right\} - e^{\eta}\mathbb{P}\left\{S_0 \in \mathcal{S}^{+}(\xi)\right\} \\ &\leq \left(e^{\eta}\mathbb{P}\left\{S_0 \in \mathcal{S}^{+}(\xi)\right\} + \tau\right) - e^{\eta}\mathbb{P}\left\{S_0 \in \mathcal{S}^{+}(\xi)\right\} = \tau.\end{aligned}$$

Since $p^{\mathcal{A}}(\xi)$ is a probability distribution, $p^{\mathcal{A}}(\xi) \in \Delta^{K-1}$. Thus, $p^{\mathcal{A}}(\xi)$ satisfies the constraints of the MinSE linear program with prior b for $\mathcal{P}$-almost surely all ξ.

The MinSE algorithm finds the solution $p^\star(b,\xi)$ that minimizes the objective function $\sum_{i=1}^K p_i \lambda\left(C_i^\alpha(X)\right)$ over the set of all feasible distributions satisfying these constraints. Since $p^{\mathcal{A}}(\xi)$ is a feasible solution $\mathcal{P}$-almost surely, its objective value must be greater than or equal to the minimum objective value achieved by the optimal solution $p^\star(b,\xi)$. Therefore, $\mathcal{P}$-almost surely,

$$\sum_{i=1}^K p_i^\star(b,\xi)\lambda\left(C_i^\alpha(X)\right) \leq \sum_{i=1}^K p_i^{\mathcal{A}}(\xi)\lambda\left(C_i^\alpha(X)\right).$$

This completes the proof. □

B.2 Proofs of Section 4

Proof of Corollary 2. Since $\hat{S}(\xi_t,\varepsilon_t)$ is (η,τ)-stability. Using the same starting derivation as the proof of Proposition 1, we know that there exists a fixed point $b \in \Delta^{K-1}$ such that $p_i(\xi_t) \leq \exp(\eta)\, b_i + s_i$ and $\sum_{i=1}^K s_i \leq \tau$. Thus,

$$\begin{aligned}\mathbb{P}\left\{Y_t \notin C^{(t)}_{\hat{S}(\xi_t,\varepsilon_t)}(X_t)\right\} = \mathbb{E}\left[\mathbb{1}\left\{Y_t \notin C^{(t)}_{\hat{S}(\xi_t,\varepsilon_t)}(X_t)\right\}\right] &= \sum_{i=1}^K p_i(\xi_t)\,\mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\} \\ &\leq \sum_{i=1}^K e^\eta b_i \mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\} + \tau,\end{aligned}$$

from which we have that

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathbb{P}\left\{Y_t \notin C^{(t)}_{\hat{S}(\xi_t,\varepsilon_t)}(X_t)\right\} &\leq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\left(\sum_{i=1}^K e^\eta b_i \mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\} + \tau\right) \\ &\leq e^\eta \sum_{i=1}^K b_i\left(\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\}\right) + \tau \\ &\leq e^\eta \alpha + \tau,\end{aligned}$$

where the second inequality follows from Fubini's theorem, which allows us to interchange the two sums, and the subadditivity of $\limsup$. This concludes the proof. □

Proof of Proposition 2. Using the definition of $C^{(t)}_{\text{comb}}$ in Algorithm 1, together with Markov's inequality, we have that

$$\begin{aligned}\mathbb{P}\left\{Y_t \notin C^{(t)}_{\text{comb}}\right\} =& \mathbb{P}\left\{\sum_{i=1}^K p_i^\star(w^{(t)},\xi_t)\mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\} \geq \frac{1}{2}\right\} \\ \leq& 2\,\mathbb{E}\left[\sum_{i=1}^K p_i^\star(w^{(t)},\xi_t)\mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\}\right].\end{aligned}$$

Moreover, from MinSE, we know that $p_i^\star(w^{(t)},\xi_t) \leq e^\eta w_i^{(t)} + s_i$ and $\sum_{i=1}^K s_i \leq \tau$. Therefore,

$$\mathbb{P}\left\{Y_t \notin C^{(t)}_{\text{comb}}\right\} \leq 2\left(e^\eta\,\mathbb{E}\left[\sum_{i=1}^K w^{(t)}\mathbb{1}\left\{Y_t \notin C_i^{(t)}(X_t)\right\}\right] + \tau\right) = 2\left(e^\eta \beta_t + \tau\right),$$

where the equality follows from (8). This concludes the proof.

□

B.3 Proofs of Section 5

Notation reminder. Before providing the proof of Theorem 2, we introduce some additional notation and recall some notation in the main text. We consider a calibration dataset $\mathcal{D}_{\text{cal}} = \left\{(X_i,Y_i)_{i\in[m]}\right\}$ and a test point (X,Y). We interchangeably denote (X,Y) as (X_{m+1},Y_{m+1}) and

define $\mathcal{D}^+_{\text{cal}} = \left\{(X_i, Y_i)_{i\in[m+1]}\right\}$. In addition, for each base predictor $k \in [K]$, we compute the non-conformity scores on the calibration data, for $i \in [m]$,

$$s_{k,i} \coloneqq s_k(X_i, Y_i, f_k)$$

and denote the score of the test point as $s_{k,m+1} = s_k(X_{m+1}, Y_{m+1}, f_k)$. Let $T_k \coloneqq \{s_{k,1}, \ldots, s_{k,m}\}$ denote the set of these calibration scores for model k. We also define $T^+_k \coloneqq T_k \cup \{s_{k,m+1}\}$. We use $s_{k,(r)}$ and $s^+_{k,(r)}$ to denote the r-th order statistic in T_k and T^+_k. Furthermore, we denote the ranks of a score $s_{k,i}$ within the two sets as follows $R_{k,i} \coloneqq \sum_{j=1}^m \mathbb{1}\{s_{k,j} \le s_{k,i}\}$ and $R^+_{k,i} \coloneqq \sum_{j=1}^{m+1} \mathbb{1}\{s_{k,j} \le s_{k,i}\}$. Similarly to Section 5, we parameterize the conformal prediction sets using ranks. For a rank index $R \in [m]$, define, respectively:

$$C_k(X, R) \coloneqq \left\{y \in \mathcal{Y} : s_k(X, y, f_k) \le s_{k,(R)}\right\},$$
$$C^+_k(X, R) \coloneqq \left\{y \in \mathcal{Y} : s_k(X, y, f_k) \le s^+_{k,(R)}\right\}.$$

Both set families are monotonic: $C_k(X, R_1) \subseteq C_k(X, R_2)$ if $R_1 \le R_2$; similarly for C^+_k. The relationship $s_{k,(R)} \ge s^+_{k,(R)}$ holds for $R \in [m]$, implying the set inclusion $C^+_k(X, R) \subseteq C_k(X, R)$. To relate with the standard perspective on split conformal threshold, note that for $i' \in [m]$

$$\mathbb{P}\left\{R^+_{k,m+1} \le i'\right\} = \mathbb{P}\left\{s^+_{k,m+1} \le s^+_{k,(i')}\right\} \le \mathbb{P}\left\{s^+_{k,m+1} \le s_{k,(i')}\right\} = \mathbb{P}\left\{Y_{m+1} \in C_k(X_{m+1}, i')\right\},$$

where $\mathbb{P}\left\{R^+_{k,m+1} \le i'\right\} \ge i'/(m+1)$ by exchangeability. Note that this is equivalent up to a reparameterization to the split conformal method introduced in Section 2, by setting $i' = \lceil (1-\alpha)(m+1)\rceil$.

For each point $i \in [m+1]$, we apply the selection rule using its feature vector X_i and independent randomness $\varepsilon_i \sim \mathcal{P}_\varepsilon$. Let $k_i \coloneqq \hat{S}(X_i, \varepsilon_i)$ be the index of the predictor selected for the i-th data point. By our assumption, k_i is independent of $\mathcal{D}_{\text{cal}}$, conditionally to X_i. We now define the **effective rank** for the i-th point. This is the rank of the i-th point's score calculated using the predictor k_i that was selected specifically for X_i:

$$\hat{R}^+_i \coloneqq R^+_{k_i,i} = R^+_{\hat{S}(X_i,\varepsilon_i),i}.$$

Similarly, for $i \in [m]$, we define

$$\hat{R}_i \coloneqq R_{k_i,i} = R_{\hat{S}(X_i,\varepsilon_i),i}.$$

Finally, we define the following two sequences $\mathcal{R}^+ \coloneqq (\hat{R}^+_1, \ldots, \hat{R}^+_{m+1})$ and $\mathcal{R} \coloneqq (\hat{R}_1, \ldots, \hat{R}_m)$, with $\hat{R}^+_{(i)}$ and $\hat{R}_{(i')}$ denoting their i-th and i'-th order statistics respectively, for $i \in [m+1]$ and $i' \in [m]$.

Proof of Theorem 2. Notice that $\mathcal{R}^+$ forms an exchangeable sequence [3, Lemma 2.2]. This follows from the exchangeability of the original data pairs $\{(X_i, Y_i)\}_{i=1}^{m+1}$ and the fact that the procedure to obtain $\hat{R}^+_i$ is symmetric w.r.t. $\mathcal{D}^+_{\text{cal}}$. In addition, $\mathcal{R} \subset \mathcal{R}^+$ implies that $\hat{R}^+_{(m)} \le \hat{R}_{(m)}$. As a direct consequence, for $i \in [m]$, we get:

$$\begin{aligned}
\frac{i}{m+1} \le \mathbb{P}\left\{\hat{R}^+_{m+1} \le \hat{R}^+_{(i)}\right\} &\le \mathbb{P}\left\{s^+_{k_{m+1},\left(R^+_{k_{m+1},m+1}\right)} \le s_{k_{m+1},\left(\hat{R}^+_{(i)}\right)}\right\} \\
&= \mathbb{P}\left\{s^+_{k_{m+1},m+1} \le s_{k_{m+1},\left(\hat{R}^+_{(i)}\right)}\right\} \\
&\le \mathbb{P}\left\{s^+_{k_{m+1},m+1} \le s_{k_{m+1},\left(\hat{R}_{(i)}\right)}\right\} \\
&= \mathbb{P}\left\{Y_{m+1} \in C_{k_{m+1}}\left(X_{m+1}, \hat{R}_{(i)}\right)\right\}.
\end{aligned}$$

Here, the first line follows from the monotonicity of the order statistics of $\left(s^+_{k_{m+1},1}, \ldots, s^+_{k_{m+1},m+1}\right)$ and the definition of $\hat{R}^+_{m+1}$. The second line is by the fact that for any $k \in [K], i' \in [m+1]$, we have $s^+_{k,(R^+_{k,i'})} = s^+_{k,i'}$. The third by $\hat{R}^+_{(i)} \leq \hat{R}_{(i)}$ for any $i \in [m]$ and the fourth by the monotonicity of the split conformal set w.r.t. increasing score. Finally, choosing $i = \lceil (1-\alpha)(m+1) \rceil$, we get

$$\mathbb{P}\left\{Y_{m+1} \in C_{\hat{S}(X_{m+1})}(X_{m+1}, R_{(i)})\right\} = \mathbb{P}\left\{Y_{m+1} \in C_{k_{m+1}}(X_{m+1}, R_{(i)})\right\} \geq 1-\alpha,$$

and recover the theorem. □

C Additional Experiments

C.1 Batch Experiments

In Section 6, we provided some results in the batch setting. Here, using the same selection algorithms as in Section 6, we extend the experimental setting as detailed in the following subsections. We first provide additional information on the batch experiments.

For the real dataset experiments (Abalone, Bike Sharing [31, CC BY 4.0], California Housing [30, BSD License]), the hyperparameters of several base regression models were optimized prior to their use in the main conformal prediction experiments. This tuning was performed using "BayesSearchCV" from the `scikit-optimize` library [32, BSD-2 license], as mentioned in Section 6. The models subjected to this tuning process included RandomForestRegressor, GradientBoostingRegressor, ElasticNet, DecisionTreeRegressor, AdaBoostRegressor, and ExtraTreesRegressor. All experiments took approximately 200 CPU hours using 16 cores of Intel Xeon CPU Gold 6230 and 32 GB of system memory. For the implementation of ModSel and the structure of our code, we based our implementation on the version open-sourced by Liang et al. [11]. Nonetheless, we significantly deviated from their implementation to allow for more efficient parallel processing.

Both X and Y were standardized. The `BayesSearchCV` process was configured to run for 25 iterations (`n_iter=25`) with 3-fold cross-validation (`cv_folds=3`) for each model and hyperparameter setting. The nonconformity score function also utilized a RandomForestRegressor to predict residuals, and its hyperparameters were the same as the tuned parameters for RandomForestRegressor for prediction.

Finally, for the synthetic experiments to follow, we note that YK-Adjust produced infinitely large sets in some runs, as such it is not plotted.

C.1.1 Homogeneous vs Heterogeneous Data Preprocessing

In Section 6, we preprocessed the dataset, both synthetic and real, by splitting them to 5 disjoint equal subsets using constrained K-means [35, 36, Code under BSD 3-Clause License]. Here, we provide additional results, without this splitting step. We call the setting with no splitting, homogeneous, and the setting with splitting heterogeneous. The homogeneous setting may represent a more challenging environment for our approach and can be more favorable to the methods of Liang et al. [11] and Yang and Kuchibhotla [12]. In particular, in this homogeneous setting, one conformal predictor can typically be superior to all other predictors, as such simply selecting the best-on-average predictor can yield very competitive results, specially that the stability-based coverage guarantees can be conservative.

Using the same synthetic data setting as in Section 6, we report the homogeneous results in Figure 3. In addition, we report the homogeneous results on the real data experiments in Figure 4. For real data experiments, Recal stays competitive with baselines. Nonetheless, for the synthetic experiments, our approach performs worse than baselines.

C.1.2 Varying Data Generation in Synthetic Experiments

To further compare with baselines, we adjust the data generation procedure in the synthetic experiment to match one of the settings in [11]. In particular, we repeat exactly the same steps as the synthetic

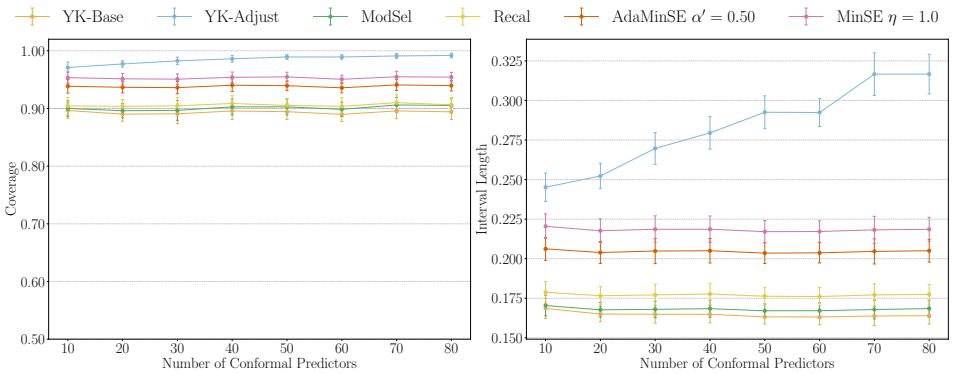


Figure 3: Homogeneous synthetic results. Each plot shows coverage (left) and interval length (right)

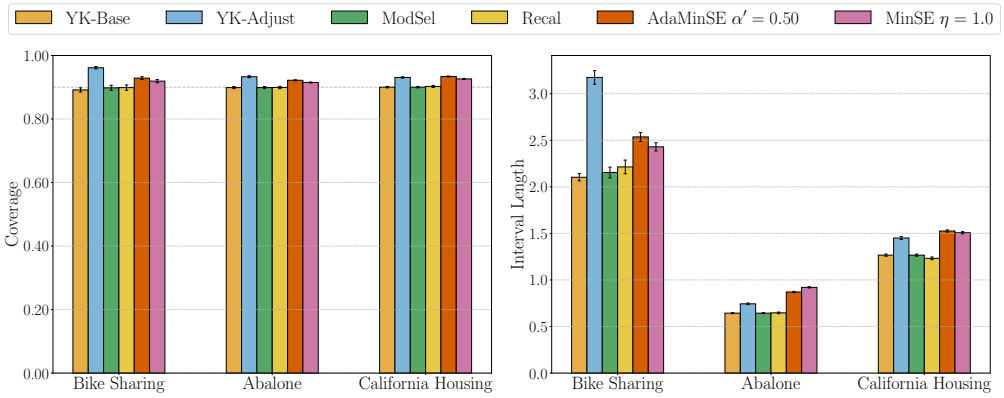


Figure 4: Comparison of UCI dataset results under homogeneous data processing.

experiments of Section 6, but replace the data generation procedure step by

$$X \sim \mathcal{N}(0, I_d), \quad \varepsilon \sim \mathcal{N}(0, 1)$$
$$\theta_i = \mathbb{1}\{i \mod 20 = 0\}$$
$$Y = X^T\theta + \varepsilon,$$

where $d = 300$. This matches the sparse normal setting with Gaussian noise in Liang et al. [11]. Given the linear dependency, this setting behaves more similarly to the homogeneous experiments; thus the globally optimal predictor may be inferred by predictors learning on any subset of the data. We report the results in Figure 5. We observe that Recal achieves very similar results to ModSel. YK-Base produces smaller conformal sets at the cost of higher miscoverage.

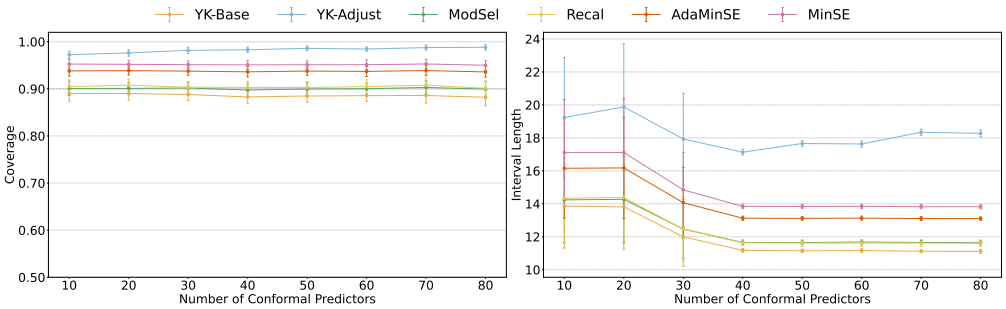


Figure 5: Comparison with baselines in the sparse linear setting with Gaussian noise

C.1.3 Comparison against Average Model Baseline

Furthermore, we repeat the experiments of synthetic experiments of Section 6, but add an additional baseline. In particular, we consider the average model baseline, which averages all predictors trained on the data and conformalizes the average predictor. This baseline may be interpreted as a simple model averaging approach. We report the results in Figure 6, denoting this additional baseline as AvgSplit. We observe that Recal, AdaMinSE, and MinSE perform better than the average model baseline.

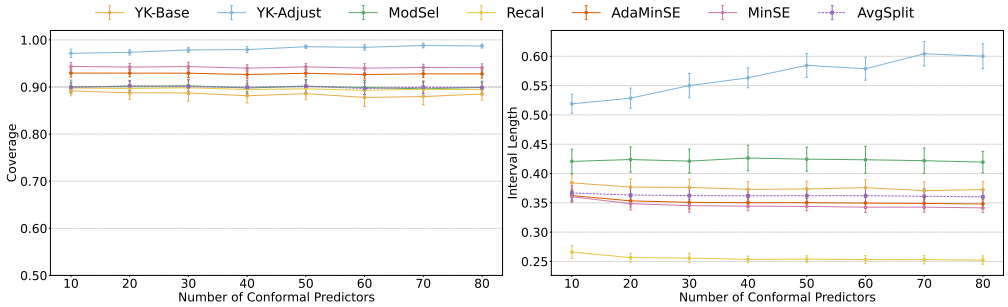


Figure 6: Comparison with average model as a baseline

C.1.4 Effect of the Number of Calibration Points

In addition, we repeat the experiments on synthetic data, while varying the number datapoints in the calibration dataset. This mainly affects the results of YK-Adjust as its performance improves with larger calibration datasets. We report the corresponding results in Figure 7 and Figure 8.

C.1.5 Additional Classification Experiments

We complement our regression studies with a compact ImageNet-1k classification [37, Non-Commercial Use] experiment intended to emulate a heterogeneous setting. Concretely, we construct two ViT-Base models from the same pretrained model: one kept *clean*, and one *degraded* by randomly shuffling the top-k logits on a fraction of examples (we use a moderate corruption fraction of 10% and $k = 20$). Then, each of the two models is used to construct a conformal predictor. For the models, we used the pretrained `timm` checkpoints [38].

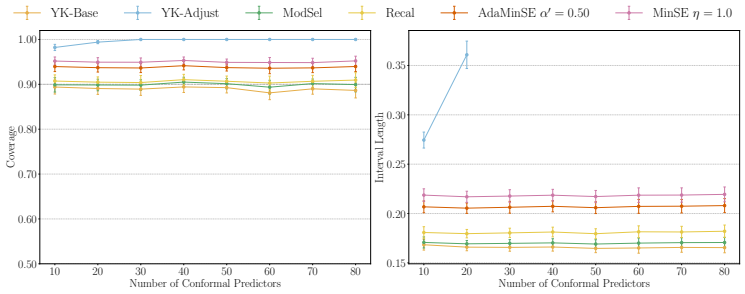
Then, we construct 10 conformal predictors by mixing the two base conformal models in different ways, i.e. each model outputs the predictions of the *clean* for 50% samples and *degraded* for the remaining. We use this structure to where different predictors perform better/worst on different samples.

Using this construction, we compare YK-Base, MinSE, and AdaMinSE, which operate purely at the set level – inspecting only the resulting conformal sets without requiring access to the underlying scores or split-conformal internals. We also evaluate a score-averaging baseline of Luo and Zhou [13] that linearly combines per-class scores via simplex weights learned on held-out indices. Because this baseline requires score access, we perform the mixing at the logit/score level rather than at the set level. We report the results in Figure 9. We note that this setting is one where the performance of the different predictors varies across the input space, which is precisely where our approach is most beneficial.

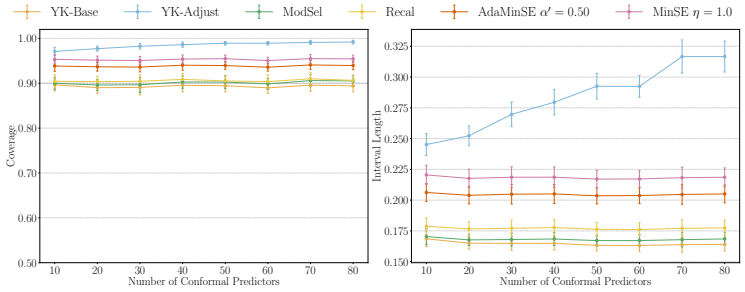
C.2 Online Experiments:

Online Setting Experiments: we tested our online algorithm, AdaCOMA, by constructing an online analogue of our heterogeneous batch setting, where the performance of different predictors vary across time. The key advantage of AdaCOMA lies in its ability to condition selection on the current features X_t, via the observed interval sizes ξ_t used in the stable selection mechanism. In contrast, COMA relies solely on historical performance in determining its weights $w^{(t)}$ (Algorithm 1).

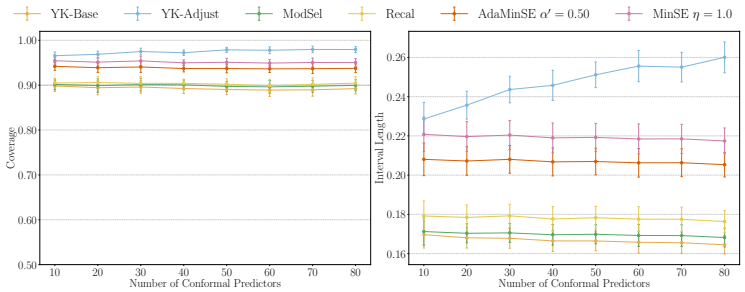
To evaluate this, we designed an environment in which both COMA and AdaCOMA track and assign weights to $K = 10$ distinct forecasters. These forecasters are not fixed models; instead,



(a) $N = 300$



(b) $N = 400$



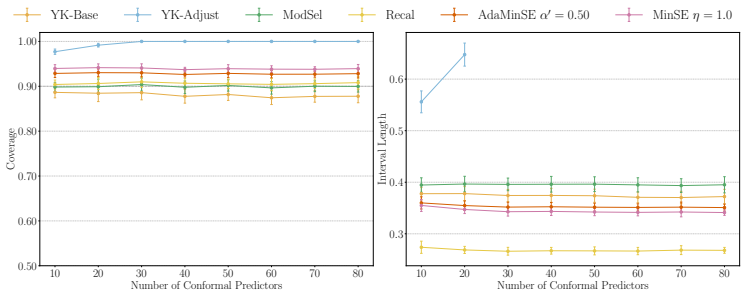
(c) $N = 500$

Figure 7: Homogeneous synthetic results. Each plot shows coverage (left) and interval length (right) for a different number of calibration examples (N).

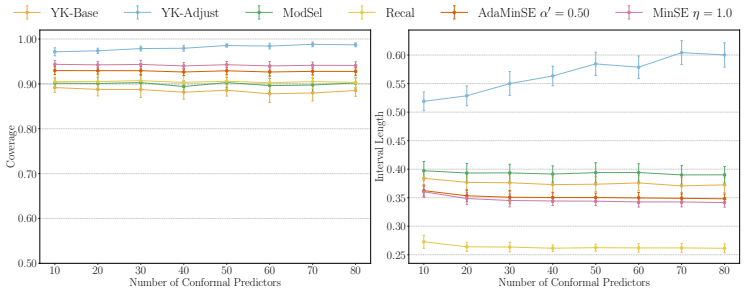
they dynamically generate prediction intervals by drawing from a smaller pool of $M = 6$ diverse base online conformal algorithms. Each base algorithm is an instance of Adaptive Conformal Inference (ACI) [6] applied to an online learning model. For each of the K forecasters, we simulate a heterogeneous environment where the optimal conformal predictor varies over time by partitioning the input data stream conceptually as follows: at the start of the experiment, each of the K forecaster pick one model from the smaller subset of M models. Then each $\tau = 50$ timesteps, the forecasters pick a different models, and so on. This setup, where forecasters change their models each τ timesteps, aims to simulate an environment, which requires the selection algorithm choosing among the forecasters to be strongly adaptive.

Base Online Conformal Algorithms. The $M = 6$ base algorithms were instances of Adaptive Conformal Inference (ACI) [6]. Default ACI parameters were: initialization period of 100 timesteps, we used an adaptive ACI stepsize adapted from the original code of [10]. The underlying online learning models were:

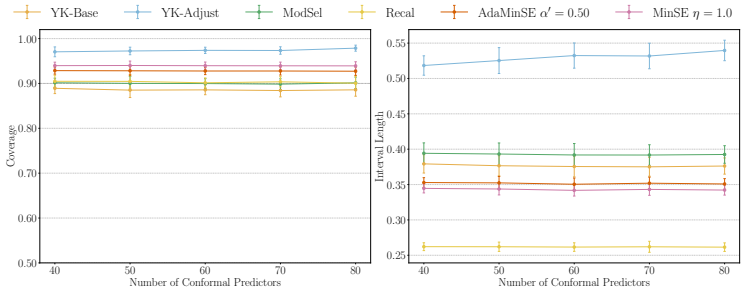
- Two SGD regressors: one with L1 penalty (Lasso, $\eta_0 = 0.001$, penalty $\alpha = 0.1$) and one with L2 penalty (Ridge, $\eta_0 = 0.001$, penalty $\alpha = 0.1$).
- Two SGD regressors (no penalty) with learning rates $\eta_0 \in \{0.001, 0.005\}$.
- Two Rolling Linear Regression models with window sizes of 50 (retrain frequency 12) and 100 (retrain frequency 25).



(a) $N = 300$



(b) $N = 400$



(c) $N = 500$

Figure 8: Heterogeneous synthetic results. Each plot shows coverage (left) and interval length (right) for a different number of calibration examples (N).

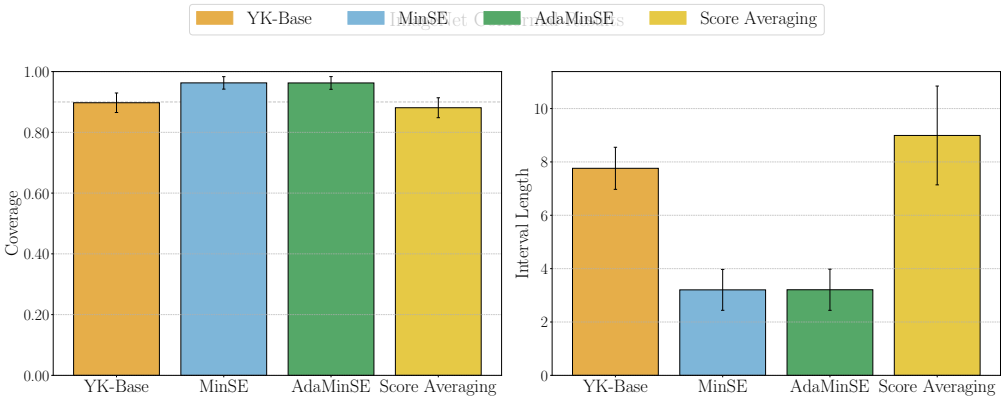


Figure 9: ImageNet classification results

We conducted experiments on two datasets, used for evaluation in Gasparin and Ramdas [10], `ELEC` [39, CC-BY 4.0] and a synthetic data generate according to the ARMA(1,1) model [40, Chapter 2]. For the precise data generation procedure for ARMA(1,1) model, refer to [10, Page 18.]. For both, AdaCOMA and COMA, we used two algorithms for the weights $w^{(t)}$ (Algorithm 1) over the forecasters: AdaHedge and Hedge with learning rate $\eta = 0.1$. We repeated the experiment for 50 seeds. For ARMA(1,1) dataset, 4 runs experienced numerical instability producing interval length multiple orders of magnitude above the rest for both COMA and AdaCOMA. We excluded those runs in calculating the reported results. The results are presented in Table 1. For COMA, we ran the underlying predictors using ACI with nominal miscoverage rate 0.1. For AdaCOMA, we ran ACI with nominal miscoverage rate of 0.09 and used AdaMinSE for the selection. For AdaMinSE selection, we tuned the selection such that COMA and AdaCOMA achieve similar coverage. The results are reported in Table 1. For ELEC dataset, AdaCOMA significantly outperformed COMA. For ARMA(1,1), both methods performed similarly with a small advantage to AdaCOMA.

Table 1: Comparison of COMA and AdaCOMA with different underlying aggregation algorithms (AdaHedge, Hedge) on ELEC and ARMA(1,1) datasets. Values are mean $\pm$ standard deviation (divide by the root of the number of seeds ($1/\sqrt{50}$) for the standard error). The target miscoverage is $\alpha = 0.1$.

Dataset	Method	Avg. Miscoverage	Avg. Length
ELEC	COMA (AdaHedge)	0.0942 ± 0.002	0.69 ± 0.06
	AdaCOMA (AdaHedge)	0.0959 ± 0.001	0.32 ± 0.09
	COMA (Hedge)	0.0942 ± 0.002	0.69 ± 0.06
	AdaCOMA (Hedge)	0.0963 ± 0.001	0.31 ± 0.01
ARMA(1,1)	COMA (AdaHedge)	0.102 ± 0.001	4.40 ± 0.91
	AdaCOMA (AdaHedge)	0.101 ± 0.001	4.23 ± 0.65
	COMA (Hedge)	0.102 ± 0.001	4.40 ± 0.92
	AdaCOMA (Hedge)	0.101 ± 0.001	4.31 ± 1.06