Strategic Conformal Prediction

Daniel Csillag
FGV EMAp

Claudio José Struchiner
FGV EMAp

Guilherme Tegoni Goedert
FGV EMAp

Abstract

When a machine learning model is deployed, its predictions can alter its environment, as better informed agents strategize to suit their own interests. With such alterations in mind, existing approaches to uncertainty quantification break. In this work we propose a new framework, Strategic Conformal Prediction, which is capable of robust uncertainty quantification in such a setting. Strategic Conformal Prediction is backed by a series of theoretical guarantees spanning marginal coverage, training-conditional coverage, tightness and robustness to misspecification that hold in a distribution-free manner. Experimental analysis further validates our method, showing its remarkable effectiveness in face of arbitrary strategic alterations, whereas other methods break.

1 INTRODUCTION

Machine learning models are widely used to manage risks associated with the actions of individuals. However, these individuals can leverage their knowledge about the model to strategize and align the model's predictions to their own interests, causing a distribution shift. This is typically referred to as the *strategic* (or *performative*) setting, and is ubiquitous in machine learning practice, such as in credit scoring, autonomous driving, spam mail classification, and chat-based interactions.

Strategic machine learning is a topic with deep connections to not only statistics, but also game theory, economics and causality, and has been the topic of significant prior research (e.g., [Perdomo et al., 2020, Rosenfeld et al., 2020, Levanon and Rosenfeld, 2021, Izzo et al., 2021]; see also the survey of [Hardt and Mendler-Dünner, 2023]).

However, to the best of our knowledge, there has been no prior work on uncertainty quantification for the strategic setting. This is essential to the safe use of machine learning methods, and especially so in safety-critical systems, where strategic issues can be found aplenty.

In this paper, we present the first method for uncertainty quantification in the strategic setting, based on the ideas of split conformal prediction [Vovk et al., 2005, Papadopoulos et al., 2002]. By considering the underlying ML model and conformity score fixed and leveraging the model-agnostic nature of conformal prediction, our method is able to deal with strategic alterations on arbitrary tasks (e.g. regression, classification, structured prediction) while achieving a strategic marginal guarantee as well as a corresponding training-conditional guarantee [Vovk, 2012, Bian and Barber, 2022].

We also present extensions of our method akin to the ones of standard conformal prediction, in particular ensuring group-conditional and label-conditional coverage. Furthermore, by leveraging change-of-measure inequalities and select modeling assumptions, we prove a series of additional theoretical results on the robustness and efficiency/tightness of our method. Finally, we explore our method empirically in experiments on various datasets, validating the efficacy of our proposed solution.

Our main contributions are:

1. A new method for uncertainty quantification based on conformal prediction that is valid under strategic alterations, with guarantees akin to type-I error control and the usual validity property of conformal prediction. To our knowledge, this is the first instance of uncertainty quantification for the strategic setting, and is directly applicable in diverse settings including regression, classification, structured prediction, etc.

2. Extensions of our proposed method for group-conditional and label-conditional coverage, show-

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

casing that our method is versatile and can be extended similarly to standard conformal prediction methods;

3. We develop a rich supporting theory for our method and framework, including guarantees of precision under model misspecification, as the strategic alterations converge (e.g. due to an equilibrium state) and on the tightness of our predictive intervals, further justifying our framework.

Related work Our setting is closely related to that of distribution shifts: strategic alterations can be seen as shifts in the data distribution, induced by the agents' interests and resources. CP under distribution shift has attracted much attention, perhaps most notably for covariate shifts (e.g., [Tibshirani et al., 2019, Gibbs et al., 2023, Cauchois et al., 2020]). In contrast to this existing literature, our method sidesteps requiring knowledge of density ratios, use of density estimation procedures or unverifiable assumptions such as closeness per some f-divergence, while also having a larger supporting theory tailored to the strategic setting. Additionally, robustness to strategic alterations can also be seen as a more general form of adversarial robustness, where the alterations are driven by e.g. worst-case ℓ_2-bounded noise. Concurrent work of [Zargarbashi and Bojchevski, 2025] has explored similar procedures to ours in this more restricted setting.

2 FRAMEWORK AND METHOD

Suppose that we have a random sample $(Z_i)_{i=1}^n = (X_i, Y_i)_{i=1}^n \subset \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and consider the problem of learning to predict Y_i using X_i.

In standard conformal prediction, we seek to find a set-predictor $C_t : \mathcal{X} \to 2^{\mathcal{Y}}$ (indexed by a threshold $t \in \mathbb{R}$) such that, for any significance level α, there is a corresponding threshold t_α^{std} such that

$$\mathbb{P}[Y \in C_{t_\alpha^{\text{std}}}(X)] \geq 1 - \alpha,$$

with C_t, per split conformal prediction, taking the form

$$C_t(X) = \{y \in \mathcal{Y} : s(x, y) \leq t\} \tag{1}$$

where $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is an arbitrary function (typically referred to as the 'conformity score' in the conformal prediction literature) and t_α^{std} is given by

$$t_\alpha^{\text{std}} = \inf\left\{ t \in \mathbb{R} : \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}[s(X_i, Y_i) > t] + \frac{1}{n+1} \leq \alpha \right\}.$$

In the strategic setting, we must consider that the distribution of the covariates X may suffer alterations in response to our model, as better informed agents strategize to suit their own interests. Such alterations can be denoted as stochastic functions $\Delta \in \mathcal{D}$, with $\Delta : \mathcal{X} \to \mathcal{X}$. We thus seek guarantees of the form

$$\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y \in C_\alpha(\Delta(X))] \geq 1 - \alpha, \tag{2}$$

where the covariates X are altered by these functions $\Delta : \mathcal{X} \to \mathcal{X}$, while the response Y does not change.[1] By considering the probability uniform over $\Delta \in \mathcal{D}$ we can obtain simultaneous validity with regards to multiple plausible alterations, in particular including the case of no alteration ($\Delta(x) = x$).

These alterations will typically be in the efforts of making the algorithm confident of the some outcome, which corresponds to ensuring that some region $\Omega \subset \mathcal{Y}$ of the outcome space is *not* contained in the predicted $C_\alpha(\Delta(x))$ (as it corresponds to the set of labels that, with high probability, *may* be the true one). I.e., by considering the construction in Equation 1, Δ seeks to maximize $\inf_{y \in \Omega} s(\Delta(X), y)$. Indeed, one can even think of the many $\Delta \in \mathcal{D}$ as ranging from no alteration to an 'optimal' alteration, interpolating in terms of the cost of performing these alterations.

Crucially, these alteration functions Δ will generally depend arbitrarily on the conformity score s and on the calibrated set predictor C_t (which is fully specified by threshold t itself).

For tractability purposes, however, we assume that the Δ do not depend on the threshold t (but can depend freely on the conformity score s). Thanks to the for-all over $\Delta \in \mathcal{D}$, this is not too harsh of an assumption: if we have a $\Delta_t^\star$ as the "true" Δ as a function of the learned threshold t, we can consider $\mathcal{D} = \{\Delta_{t'}^* : t' \in \mathbb{R}\}$, and our guarantee will certainly hold for Δ_t^*. Moreover, this sort of construction is robust: we prove that as long as the choice of $\mathcal{D}$ is able to approximately cover the range of the "true" $\Delta_t^\star$, we still get strong coverage guarantees (see Proposition A.2).

With this in mind, our goal becomes to attain some set-predictor $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ satisfying Equation 2, where the $\Delta \in \mathcal{D}$ depend freely on s but are independent from C_α. We show that this can be achieved simply by modifying the computation of t_α^{std}, incorpo-

[1] In the scope of this paper, we consider that the outcome Y does not change. Though restrictive, this is a typical assumption in the strategic ML literature, and not too unreasonable, though perhaps a bit too conservative. Our setup can also be extended to the case where the Δ alters the outcomes Y, but this has additional subtleties and we leave to future work.

rating suprema over the $\Delta \in \mathcal{D}$:

$$t_\alpha^{\text{strat}} = \inf\left\{ t \in \mathbb{R} : \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\left[\sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i) > t \right] + \frac{1}{n+1} \leq \alpha \right\}.$$

Theorem 2.1. *Let $Z_1, \ldots, Z_n, Z^{\text{test}}$ be $n+1$ exchangeable random variables in $\mathcal{X} \times \mathcal{Y}$. Let t_α^{strat} be as above. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(\Delta(X^{\text{test}})) \right] \geq 1 - \alpha.$$

We can also make a training-conditional PAC guarantee for the batch setting, under an i.i.d. assumption:

Theorem 2.2. *For any $\delta \in (0,1)$, with probability of at least $1-\delta$ over the draw of i.i.d. random variables $Z_1, \ldots, Z_n$, it holds that*

$$\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(\Delta(X^{\text{test}})) \;\middle|\; t_\alpha^{\text{strat}} \right] \geq 1 - \alpha - \sqrt{\frac{\log 1/\delta}{2n}}.$$

The threshold t_α^{strat} can be computed efficiently in $O(n \log n)$ time (computing the suprema of the scores takes $O(n)$ time, and the threshold computation can be naively done in $O(n \log n)$ time through sorting) and with a space complexity of $O(n)$ (for storing the computed suprema), making it a practical choice.

2.1 Construction of the Δ

By means of Theorems A.2 and A.4, once we have our set $\mathcal{D}$ of alteration functions Δ, we are able to provably control coverage error in face of these specified alterations. However, modeling these Δs can be nontrivial.

In this section, we outline some practical and pragmatic ways of constructing the Δs to be used, and exhibit how they relate to more usual constructions in the strategic classification and economics literature.

2.1.1 From an utility-cost decomposition

In the economic and game-theoric literature it is typical to model actions as the solution of an optimization problem expressed in terms of utility and cost functions. An utility function $u : \mathcal{X} \to \mathbb{R}$ in which higher values of $u(X)$ correspond to more 'desired/rewarding' values of $X \in \mathcal{X}$, and a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ (which should satisfy $c(X, X) = 0$ for all $X \in \mathcal{X}$) in which a cost $c(X, X')$ corresponds to how 'costly' it would be for an individual to alter themselves from covariates X to covariates X' (higher values indicate higher cost).

The effective actions are then modelled as the maximization[2] of utility 'regularized' by the cost:

$$\Delta^{(u,c)}(X) = \arg\max_{X' \in \mathcal{X}} \left[u(X') - c(X, X') \right].$$

So one can directly recover the usual case in the strategic classification and performative optimization literature by considering a singleton $\mathcal{D}$ containing just $\Delta^{(u,c)}$ for some fixed known choices of u and c.

However, it is advisable to consider not just the 'fully rational' alteration $\Delta^{(u,c)}$ but rather a whole range of potential alterations ranging from no alteration at all to the fully rational one. This can be pragmatically done by altering the cost function:

$$\Delta^{(u,\lambda^{-1}c)}(X) = \arg\max_{X' \in \mathcal{X}} \left[u(X') - \lambda^{-1} c(X, X') \right], \quad \text{for } \lambda \in [0, 1],$$

with the corner case $\lambda = 0$ (being a bit lax with infinities) corresponding to alterations that have zero cost:

$$\begin{aligned} \Delta^{(u,0^{-1}c)}(X) &= \arg\max_{X' \in \mathcal{X}} \left[u(X') - \infty \mathbb{1}[c(X, X') = 0] \right] \\ &= \arg\max_{\substack{X' \in \mathcal{X} \\ c(X,X')=0}} u(X'). \end{aligned}$$

Taking $\lambda = 1$ recovers $\Delta^{(u,c)}$ exactly; meanwhile $\lambda = 0$ (considering that $0^{-1} = +\infty$) gives us zero-cost alterations, which typically amounts to the identity alteration (as the cost $c(X, X')$ is usually taken to be positive unless $X = X'$). Any λ between the two effectively interpolates between them.

It is also worth noting that the Strategic Conformal Prediction setting admits a particularly natural form of utility: since the individual's goal is to fool the predicted set into not containing some particular subset of outcomes $\Omega \subset \mathcal{Y}$ – i.e., maximizing $\inf_{y \in \Omega} s(\Delta(X), y)$, we can consider exactly that as our utility:

$$u(X') = \inf_{y \in \Omega} s(X', y).$$

Nevertheless, our procedure is more general and can work with arbitrary utilities (e.g., ones that take into account the individuals' perceptions), not just this fully-rational one.

2.1.2 As a stochastic simulator

Another way of constructing the Δs is to leverage stochastic simulators. There is a wide literature on modeling populations with agent-based

[2] Assuming that there is an unique maximizer. If there isn't one, we can consider a random tie-breaking rule, making use of the fact that we allow for stochastic Δs.

models and stochastic simulators, ranging from more parsimonious models (e.g., [Epstein, 1999, DeAngelis and Mooij, 2005]) to more complex ones (e.g., [Vargas-Pérez et al., 2023, Genari et al., 2022]), sometimes even leveraging learning procedures underneath.

Let $M : \mathcal{X} \to \mathcal{X}$ represent one call to an underlying stochastic simulator, which gives us the alteration proposed by it. Note that M is generally not idempotent: if we were to compute $M(M(X))$, this would generally not be the same as just $M(X)$ (even in lack of stochasticity). This leads us to define a sequence of altered covariates starting from the initial covariates X:

$$X^{(0)} = X, \qquad X^{(k+1)} = M(X^{(k)}).$$

As we consider $k \to +\infty$, $X^{(k)}$ corresponds to a "higher level of effort" and thus a "higher cost" than the previous ones. This naturally leads to a set $\mathcal{D}^{(M)}$ of Δs:

$$\mathcal{D}^{(M,k_{\max})} = \{(X \mapsto X^{(k)}) : k = 0, \dots, k_{\max}\}, \quad \text{for } k_{\max} \in \mathbb{N}.$$

It is necessary to consider the sequence only up to $k_{\max}$ for computability purposes – if we consider M to be a black-box stochastic function, then we would not be able to compute suprema over all $k \in \mathbb{N}$ (as required for our Strategic Conformal calibration).

One particularly interesting instance of this construction is to use a generative foundation model with agentic capabilities (such as recent LLMs) to simulate actions of the individuals. Such models can not only be prompted to simulate particular characteristics of subpopulations but also *calibrated* to do so while better aligning with the actual populations.

2.1.3 Via iterative random search

An interesting variant of the stochastic simulator setting of Section 2.1.2 is that in which we formulate the stochastic step M as a random search in a local neighborhood. In particular, consider that we have some utility function $u : \mathcal{X} \to \mathbb{R}$ and a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that we further assume to be a metric over $\mathcal{X}$.

Instead of considering the regularized utility maximization problem presented in Section 2.1.1, we can consider an iterative procedure where in each step we evaluate the utility on a few points of the neighborhood of the current point (as determined by the metric c) and take the best one. I.e.,

$$X^{(0)} = X, \qquad X^{(k+1)} = \arg\max_{j=1,\dots,m} u(X + \delta_j^{(k)}),$$
$$\delta_1^{(k)}, \dots, \delta_m^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}_c(0, \sigma^2 I) \quad \forall k \in \mathbb{N},$$

for some σ hyperparameter (and anisotropy given by the choice of metric c).[3] Intuitively, this corresponds to each agent considering m different actions and opting for the one they find best. It is also reasonable to include an additional $\delta_0^{(k)} = 0$ to allow the agent to do nothing at each step.

This approach can be seen as a middle term between the typical utility-cost formulations (Section 2.1.1) and stochastic simulator formulations (Section 2.1.2) – it allows one to leverage the pragmatic nature of the decomposition of actions into utility and cost while being much easier to implement than the optimization-based approach of Section 2.1.1 (which would require solving many optimization problems for every data point in the calibration set, and these problems could be quite tricky [e.g., highly nonconvex]).

2.2 Extensions of Strategic Conformal Prediction

In this section, we present extensions and variants of our base method for group-conditional and label-conditional coverage [Vovk, 2012, Sadinle et al., 2016], demonstrating that the same strategy used to adapt standard split conformal prediction to the strategic setting can also be applied to other conformal prediction algorithms.

For the sake of brevity, we present here only the extension for group-conditional coverage; the label-conditional case can be found in the supplementary material.

For group-conditional guarantees, suppose that we have some (possibly intersecting) subsets $G_1, \dots, G_l \subset \mathcal{X}$ (the groups) such that $\bigcup_{i=1}^{l} G_j = \mathcal{X}$. We want to find some $C_\alpha^{\text{strat-}(G_*)}$ such that

$$\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y \in C_\alpha^{\text{strat-}(G_*)}(\Delta(X)) | X \in G_j] \geq 1 - \alpha, \quad \text{for all } j = 1, \dots, l.$$

Similar to standard Group-Conditional Conformal Prediction [Vovk, 2012], it suffices to do our strategic calibration separately over each group – thus obtaining a threshold for each group – and then using the threshold corresponding to the group the test covariate is in (and, in the case of overlap between groups, taking the highest threshold):

$$C_\alpha^{\text{strat-}(G_*)}(X) = \left\{ y \in \mathcal{Y} : s(X, y) \leq \max_{j; X \in G_j} t_\alpha^{\text{strat-}(G_j)} \right\}$$

[3] This notation suggests the common case that $\mathcal{X} = \mathbb{R}^d$, but the idea is more general; a normal distribution/law can be defined over any Banach space (see, e.g., [Ledoux and Talagrand, 1991]), and everything still follows.

where

$$t_\alpha^{\text{strat-}(G_j)} = \inf\left\{ t \in \mathbb{R} : \sum_{\substack{i=1 \\ X_i \in G_j}}^{n} \mathbb{1}\left[\sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i) > t\right] + 1 \leq \alpha(n_{G_j} + 1) \right\}$$

and $n_{G_j} = |\{i = 1, \ldots, n : X_i \in G_j\}|$.

Theorem 2.3. *Let $Z_1, \ldots, Z_n, Z^{\text{test}}$ be $n+1$ exchangeable random variables in $\mathcal{X} \times \mathcal{Y}$. Let $C_\alpha^{\text{strat-}(G_j)}$ be as above (for $j = 1, \ldots, l$). Then, for any $\alpha \in (0, 1)$,*

$$\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_\alpha^{\text{strat-}(G_j)}(\Delta(X^{\text{test}})) \middle| X^{\text{test}} \in G_j\right] \geq 1 - \alpha, \quad \forall j = 1, \ldots, l.$$

2.3 Additional Guarantees under Modeling Assumptions

In this section, we will present additional theoretical properties enjoyed by our procedure. Throughout, these guarantees will either hold under specific modeling assumptions or involve some quantity that is generally inestimable in practice, but enlightening from a theoretical point of view.

2.3.1 Under misspecification of the Δs

Key to our procedure is the choice of the alteration functions $\Delta \in \mathcal{D}$. As these are specified by the practitioner, there is a significant possibility of misspecification. This is further accentuated by our constraint that the Δ do not depend on the learned threshold t.

Fortunately, our model is reasonably robust to such violations. Suppose that there is a "true" alteration function $\Delta^\star$ – i.e., what we are really interested in is coverage when the covariates have been modified by $\Delta^\star$. In particular, this $\Delta^\star$ can be a function not only of s (as the $\Delta \in \mathcal{D}$ can) but also on the learned threshold t. First, we prove a bound on the difference between the coverage of a set predictor with the true alteration function Δ^* versus with any of our Δs in $\mathcal{D}$ in terms of how much the two deviate on average over X (per total variation distance TV):

Proposition 2.1. *Let $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ be a set predictor. For any $\Delta \in \mathcal{D}$,*

$$\left|\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta(X))] - \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))]\right| \leq \mathbb{E}_X\left[\text{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right].$$

This shows that as long as the two Δs are sufficiently aligned in high-probability regions of X, there will be little difference in the coverage of the two, showcasing a certain robustness to misspecification of our model.

Proposition A.1 bounds the difference in coverage between Δ^* and some particular $\Delta \in \mathcal{D}$. However, we are more interested in the *set* $\mathcal{D}$, not just a single one of its elements. The following proposition considers just this and, for the specific case of our method, bounding the coverage with Δ^* in terms of the infimum of the average deviation terms from Proposition A.1.

Proposition 2.2. *Let $C_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ be a set predictor satisfying Equation 2. Then*

$$\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] \geq 1 - \alpha - \inf_{\Delta \in \mathcal{D}} \mathbb{E}_X\left[\text{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right].$$

Note also that the randomness of the Δ is absolutely essential for these results to be useful, and suggests that it is even advisable to *always* have them be somewhat stochastic. If the $\Delta \in \mathcal{D}$ are fully deterministic, then the total variation distances will equal 1 if there is even a slight misspecification. However, if there is randomness, then even under significant misspecification the total variation distances will be well below 1, meaning a smaller deviation from the true alterations Δ^* and more robustness.

2.3.2 As the Δ converge

Consider that we have a sequence of alteration functions $\Delta_k : \mathcal{X} \to \mathcal{X}$, $k = 0, 1, \ldots$, and that $\mathcal{D} = \{\Delta_k : k = 0, \ldots, k_{\max}\}$ for some $k_{\max} \in \mathbb{N}$ (e.g., as is the case in the stochastic simulation and iterative random search constructions in Sections 2.1.2 and 2.1.3). As mentioned in the sections prior, this $k_{\max}$ is necessary from a computational point-of-view – for all we know, the underlying stochastic simulators process may change behavior arbitrarily and we would be unable to compute the suprema in the calibration procedure for the t_α^{strat}. But what happens for Δ_k beyond this $k_{\max}$?

We show that if this sequence of Δ_ks converges to some $\overline{\Delta}$, then as the $k_{\max}$ used for calibration increases, the coverage under $\overline{\Delta}$ improves, achieving the nominal coverage at the limit.

Proposition 2.3. *Let $(\Delta_k)_{k=0}^\infty$ be a sequence of alteration functions $\Delta_k : \mathcal{X} \to \mathcal{X}$, such that $\lim_{k\to\infty} \Delta_k(X) = \overline{\Delta}(X)$ according to the total variation metric. Then, for any $k_{\max} > 0$, consider the set predictor $C_{t_\alpha^{\text{strat}}}$ generated by $\mathcal{D} = \{\Delta_0, \ldots, \Delta_{k_{\max}}\}$. There exists some $\epsilon_{k_{\max}}$ such that,*

$$\mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\text{strat}}}(\overline{\Delta}(X))] \geq 1 - \alpha - \epsilon,$$

and $\lim_{k_{\max}\to\infty} \epsilon_{k_{\max}} = 0$.

In fact, as ϵ decreases, it can become smaller than the gap between the actual coverage $\mathbb{P}[Y \in C_\alpha(\overline{\Delta}(X))]$

and the nominal coverage $1-\alpha$ – at which point we actually attain the nominal coverage even for Δ_k with $k \gg k_{\max}$. Indeed, this remarkable behavior does happen in practice, as can be seen in Figure 2.

2.3.3 Tightness of the intervals under stability assumptions

So far all of our results have been only about the coverage of our predictions in the face of strategic alterations, $\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y \in C^{\text{strat}}_{t_\alpha}(\Delta(X))]$ (and variants). However, just as important are results about *how large* the sets outputted by C_α will be; after all, one trivial way to attain all the coverage guarantees would have been to simply produce the whole output space $\mathcal{Y}$, but that is hardly useful.

On one hand, we desire short predictive sets $C^{\text{strat}}_{t_\alpha}(X)$. However, if the underlying model is itself vulnerable to strategic perturbations, then these sets will necessarily have to increase accordingly; and of course, if there is uncertainty about the outcome itself, this also adds to the size of the sets. But the question remains: if neither of these are the case, then are our sets tight?

The answer is, fortunately, yes, as the two following propositions show. Such matters fundamentally depend on the choice of conformity score; for the sake of concreteness, we consider here only the case of the simple conformity score $s(x,y) = |\mu(x) - y|$, where $\mu : \mathcal{X} \to \mathcal{Y}$ for $\mathcal{Y} = \mathbb{R}$ is a base regression model, but the same argument can be used for other conformity scores. Additionally, focusing on such a conformity score and the regression setting gives us a natural way to quantify the size of the predicted sets, in the form of their Lebesgue measure.

In what follows, $\text{leb}(A)$ corresponds to the Lebesgue measure of the set A, $\text{leb}_A(B) = \text{leb}(B)/\text{leb}(A)$ and $(W)_{[i]}$ corresponds to the i-th order statistic of the random variable W. First, we show that if the model is already robust against strategic perturbations, then the interval size will be approximately equal that of standard conformal prediction:

Proposition 2.4. *Let $\mu : \mathcal{X} \to \mathbb{R}$ be a possibly-stochastic function representing a base model. Consider the predictive sets $C^{\text{strat}}_{t_\alpha}$ with conformity score $s : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ given by $s(x,y) = |\mu(x) - y|$ with $\mathcal{Y}$ being a compact subset of $\mathbb{R}$ such that $C^{\text{strat}}_{t_\alpha}(\cdot) \subset \mathcal{Y}$ and alteration functions $\Delta \in \mathcal{D}$, and contrast it with the corresponding standard conformal predictive sets $C^{\text{std}}_{t_\alpha}$. Then, almost surely over X,*

$$\mathbb{E}_{C^{\text{strat}}_{t_\alpha}}\left[\text{leb}_{\mathcal{Y}}\left(C^{\text{strat}}_{t_\alpha}(X)\right)\right] \le \mathbb{E}_{C^{\text{std}}_{t_\alpha}}\left[\text{leb}_{\mathcal{Y}}\left(C^{\text{std}}_{t_\alpha}(X)\right)\right] + 2\,\text{TV}\left(\left(s(X,Y)\right)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X),Y)\right)_{[j_{n,\alpha}]}\right),$$

for $j_{n,\alpha} = \lceil (1-\alpha)(1+n) \rceil$.

And, in turn, we can show that if there is little noise in the predictions of the underlying model μ, then the produced intervals are near singletons.

Corollary 2.3.1. *In the same setting as Proposition A.4, further assume that $\mu(X) = Y + \epsilon$ for some random ϵ (arbitrarily dependent on X and Y), with $\epsilon \in [-M, M]$ almost surely. Then, almost surely over X,*

$$\mathbb{E}_{C^{\text{strat}}_{t_\alpha}}\left[\text{leb}_{\mathcal{Y}}\left(C^{\text{strat}}_{t_\alpha}(X)\right)\right] \le 2M + 2\,\text{TV}\left((\epsilon)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X),Y)\right)_{[j_{n,\alpha}]}\right),$$

for $j_{n,\alpha} = \lceil (1-\alpha)(1+n) \rceil$.

Note that as $M \to 0$ and the strategic alterations do not cause the scores to deviate much from the ϵ, the size of the strategic conformal predictive sets approaches zero.

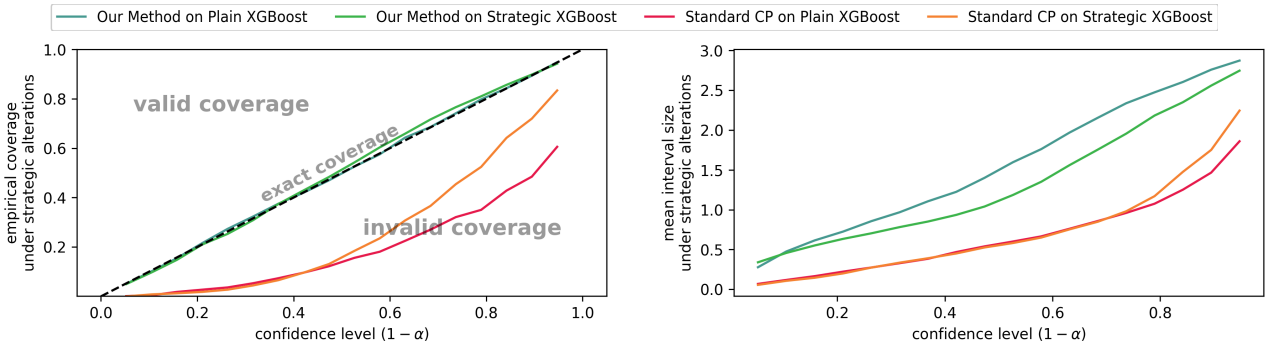
3 EXPERIMENTS

In order to fully assess our method's efficacy and further verify our theoretical guarantees, we conduct an empirical investigation of our method.

Datasets We evaluate our method and baselines on 3 datasets for briefness, with 6 more used in the supplementary material.

- `academic-dropout` [Realinho et al., 2021]: A dataset for prediction of academic dropout in higher education. Students try to exclude the possibility of dropout from the predictive sets, as the possibility of such an outcome may hinder opportunities.
- `spambase` [Hopkins et al., 1999]: A dataset for classification of email into spam and non-spam. Spammers try to exclude the spam outcome from the predictive sets.
- `shoppers` [Sakar and Kastro, 2018]: A dataset for prediction of revenue generated by products. Sellers have the incentive to make the system believe that there will be a larger revenue, as it would increase exposure.

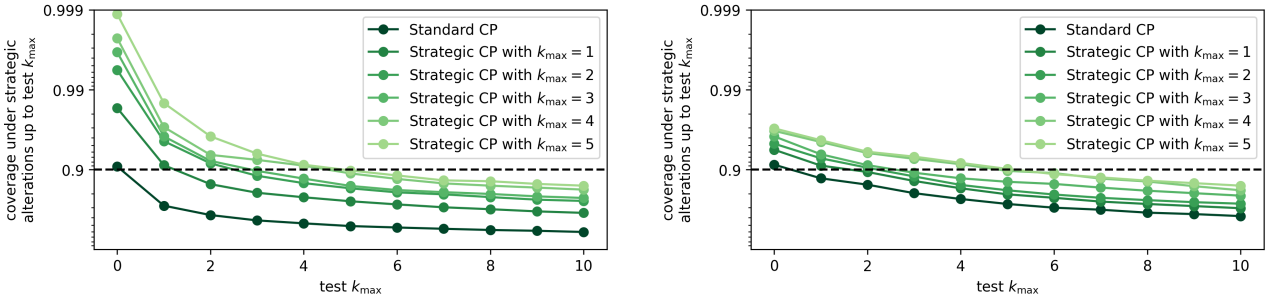
Base models We apply our method atop two XGBoost models: one trained in the usual way ('Plain XGBoost') and another trained by repeated risk minimization [Perdomo et al., 2020] ('Strategic XGBoost'), which is more well-suited to the strategic setting. The pseudocode for the exact repeated risk minimization procedure we use can be found in the supplementary material. For conformal prediction, we then consider the conformity score $s(x,y) = |\mu(x) - y|$ for regression (where $\mu : \mathcal{X} \to \mathbb{R}$ is the base XGBoost model) and $s(x,y) = 1 - p(y|x)$ for classification (where $p(y|\cdot) : \mathcal{X} \to [0,1]$ for each $y \in \mathcal{Y}$ are the conditional probabilities from the base XGBoost model).



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

Figure 1: **Coverage and interval sizes for our method versus standard CP.** Figures 1a and 1b evaluate the coverage and interval sizes under strategic alterations over varying confidence levels $1-\alpha$ on the `academic-dropout` dataset. Note that our method attains exact coverage, while the baselines have (very) invalid coverage. This comes at a cost of larger predictive sets, but using a better model (Strategic XGBoost) reduces this. For more details and more datasets, see the supplementary material.



(a) Strategic Conformal Prediction on top of a Plain XGBoost Model

(b) Strategic Conformal Prediction on top of a Strategic XGBoost Model

Figure 2: **Coverage of our method and standard CP for varying levels of strategic alterations.** The severity ($k_{\max}$) of strategic alterations at test time is indicated in the x-axis, while different curves correspond to conformal calibrations targeting different levels of $k_{\max}$ (along with standard CP) on the `academic-dropout` dataset, all with $\alpha = 0.1$. The y-axis indicates the test miscoverage under the strategic alterations up to the severity indicated on the x-axis. These curves follow the behaviour guaranteed by Theorems A.2 and A.4 as well as Proposition A.3. Note that using a model more well-suited to strategic alterations "stabilizes" these curves, making them closer together and smoother. For more details and more datasets, see the supplementary material.

Strategic Alterations In our experiments, we consider two particular forms of strategic alterations: one based on a utility-cost decomposition (Section 2.1.1) and another based on the iterative random search formulation (Section 2.1.3). In either case, we consider "fully-rational" utility functions of the form $u(X') = \inf_{y \in \Omega} s(X', y)$ for some $\Omega \subset \mathcal{Y}$. The exact choices of Ω are done on a per-dataset basis and can be found in the supplementary material.

Baselines As our baselines we consider Standard Conformal Prediction over (i) the plain XGBoost model, and (ii) the Strategic XGBoost model. The latter (combining a model fitted considering the strategic setting and then using standard conformal prediction) is, to our knowledge, the closest to uncertainty quantification for the strategic setting that one can get with previously existing methods.

Computational budget and code All experiments were conducted on an AMD Ryzen 9 5950X CPU (2.2GHz/5.0GHz, 32 threads) with 64GB of RAM. The code is available at `https://github.com/dccsillag/experiments-strategic-cp`.

As a first target of investigation, we look in Figure 1 at the strategic coverage $\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y \in C_t(\Delta(X))]$ obtained by our method compared to baselines (Figure 1a), along with the corresponding average sizes of the predictive sets (Figure 1b) on the `academic-dropout` dataset. It is immediately evident that standard conformal prediction does not obtain any semblance of coverage. Using a strategically-

Table 1: **Evaluation of our method on multiple datasets and forms of strategic alterations for $\alpha = 0.1$.** Strategic coverages substantially below the specified confidence level of 90% (below 60%) are marked in **purple**. Note how our method is consistently close to the specified coverage, while standard CP never attains it, most of the time substantially so. This improvement from Strategic CP can result in increases of interval size, but this can be managed by using a base model crafted for the strategic setting. Details and an extended version of this table with more datasets can be found in the supplementary material.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE		AVG SET SIZE DIFF
			OURS	STD CP	
academic-dropout	Plain XGBoost	Utility-cost	**90**% ± **2**%	**57**% ± **3**%	+107% ± 5%
		Rand. Search.	**90**% ± **2**%	**48**% ± **3**%	+129% ± 5%
	Strategic XGBoost	Utility-cost	**91**% ± **2**%	76% ± 3%	+46% ± 4%
		Rand. Search.	**90**% ± **2**%	74% ± 3%	+58% ± 4%
spambase	Plain XGBoost	Utility-cost	**90**% ± **2**%	**16**% ± **2**%	+4% ± 1%
		Rand. Search.	**89**% ± **2**%	**28**% ± **3**%	+3% ± 1%
	Strategic XGBoost	Utility-cost	**90**% ± **2**%	**40**% ± **3**%	+38% ± 3%
		Rand. Search.	**88**% ± **2**%	**46**% ± **3**%	+3% ± 1%
shoppers	Plain XGBoost	Utility-cost	**89**% ± **1**%	**44**% ± **2**%	+14% ± 1%
		Rand. Search.	**88**% ± **1**%	**46**% ± **2**%	+13% ± 1%
	Strategic XGBoost	Utility-cost	**89**% ± **1**%	64% ± 2%	+8% ± 1%
		Rand. Search.	**88**% ± **1**%	73% ± 2%	+9% ± 1%

refitted model, which is more well-adapted to the strategic setting, is an improvement, but not nearly enough. Our method, in contrast, consistently and tightly obtains the specified nominal coverage, regardless of how well-adapted the underlying model is.

We see that our method does, unfortunately, lead to larger predicted sets. This is inevitable: the original sets were too small and thus did not cover the outcome under strategic alterations, and so such an increase is justified. Nevertheless, we do see – as was indicated by Proposition A.4 – that using a model more well-suited to the strategic setting (namely, our Strategic XGBoost model) leads to smaller average interval sizes in general, as the underlying conformity score is then more well-adapted to the strategic alterations.

We also perform a similar experiment while varying the dataset and strategic alterations being considered – results of which can be found in Table 1. Our method consistently dominates standard conformal prediction (by a substantial amount!) throughout all of the settings explored, showcasing its reliability and versatility.

Figure 2 considers a setting in which the strategic alterations used for calibration may not match the ones that occur in practice. In particular, it considers the case where both alterations follow the iterative random search scheme of Section 2.1.3, but with a different number of iterations. In practical terms, this corresponds to the actual population putting a "different amount of effort" into breaking the deployed model than anticipated during calibration.

We see that, as designed for in Section 2.1.3, if the actual population does "less effort" than anticipated (i.e., the $k_{\max}$ in practice is *smaller* than the one used for calibration), then we are completely protected. If the population does "more effort" (i.e., the $k_{\max}$ in practice is *larger* than the one used for calibration), then we are not protected in general, *unless* the $k_{\max}$ used for calibration is large enough; if that is so, then, as anticipated in Proposition A.3, we will be sufficiently close to obtaining coverage for the accumulation point of the strategic alterations, and thus protect even against larger amounts of effort than originally foreseen.

Moreover, still in Figure 2 we can clearly see the impact of using a base model that is more well-suited to the strategic setting: such a model "stabilizes" the curves in the plot, making it so that the varying sizes of $k_{\max}$, both for test and calibration, have a smaller impact on the overall curve, and thus leading to more efficient and robust predictive sets.

Finally, we also analyze the behavior of our method under a more strenuous form of model specification where we alter the actual stochastic step used for the stochastic simulations. The figures for this experiment can be found in the supplementary material. While significant misspecification does break our guarantees, small changes approximately maintain coverage, in line with the results of Propositions A.1 and A.2.

4 CONCLUSION

In this work we presented Strategic Conformal Prediction, a variant of the Conformal Prediction framework that is well-suited for the strategic setting, where the individuals' covariates change in response to the deployed model.

Our method is, to our knowledge, the first instance of uncertainty quantification for the strategic setting. We show, both theoretically and empirically, that it properly ensures coverage in spite of strategic alterations, whereas previously available methods (which were not tailored for the strategic setting) do not. This is further supplemented by a series of theoretical results on the behavior of our method under model misspecification, establishing its robustness. For these reasons, we believe our method to be not only a meaningful theoretical contribution the the landscape of strategic ML, but also a reliable tool for real-world application.

References

[dat, 2020] (2020). Productivity Prediction of Garment Employees. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51S6D.

[Bian and Barber, 2022] Bian, M. and Barber, R. F. (2022). Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*.

[Cauchois et al., 2020] Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *ArXiv*, abs/2008.04267.

[Cortez et al., 2009] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

[DeAngelis and Mooij, 2005] DeAngelis, D. L. and Mooij, W. M. (2005). Individual-based modeling of ecological and evolutionary processes. *Annual Review of Ecology, Evolution, and Systematics*, 36:147–168.

[Epstein, 1999] Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complex.*, 4:41–60.

[Fernandes et al., 2015] Fernandes, K., Vinagre, P., Cortez, P., and Sernadela, P. (2015). Online News Popularity. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NS3V.

[Genari et al., 2022] Genari, J., Goedert, G. T., Lira, S., Oliveira, K., de Souza Barbosa, A., Lima, A., Silva, J. A., Oliveira, H., Maciel, M., Ledoino, I., Resende, L., Santos, E. R. D., Marchesin, D., Struchiner, C. J., and Pereira, T. (2022). Quantifying protocols for safe school activities. *PLoS ONE*, 17.

[Gibbs et al., 2023] Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

[Hardt and Mendler-Dünner, 2023] Hardt, M. and Mendler-Dünner, C. (2023). Performative prediction: Past and future. *ArXiv*, abs/2310.16608.

[Hopkins et al., 1999] Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. (1999). Spambase. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53G6X.

[Izzo et al., 2021] Izzo, Z., Ying, L., and Zou, J. Y. (2021). How to learn when data reacts to your model: Performative gradient descent. In *International Conference on Machine Learning*.

[Kohavi, 1996] Kohavi, R. (1996). Census Income. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP7S.

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand, M. (1991). Probability in banach spaces: Isoperimetry and processes.

[Levanon and Rosenfeld, 2021] Levanon, S. and Rosenfeld, N. (2021). Strategic classification made practical. In *International Conference on Machine Learning*.

[Moro et al., 2014] Moro, S., Rita, P., and Cortez, P. (2014). Bank Marketing. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

[Ohnishi and Honorio, 2020] Ohnishi, Y. and Honorio, J. (2020). Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. In *International Conference on Artificial Intelligence and Statistics*.

[Papadopoulos et al., 2002] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*.

[Perdomo et al., 2020] Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. *ArXiv*, abs/2002.06673.

[Realinho et al., 2021] Realinho, V., Martins, M. V., Machado, J., and Baptista, L. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MC89.

[Rosenfeld et al., 2020] Rosenfeld, N., Hilgard, S., Ravindranath, S. S., and Parkes, D. C. (2020). From predictions to decisions: Using lookahead regularization. *ArXiv*, abs/2006.11638.

[Sadinle et al., 2016] Sadinle, M., Lei, J., and Wasserman, L. A. (2016). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114:223 – 234.

[Sakar and Kastro, 2018] Sakar, C. and Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F88Q.

[Tibshirani et al., 2019] Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Neural Information Processing Systems*.

[Vargas-Pérez et al., 2023] Vargas-Pérez, V., Mesejo, P., Chica, M., and Cordón, Ó. (2023). Deep reinforcement learning in agent-based simulations for optimal media planning. *Inf. Fusion*, 91:644–664.

[Vovk, 2012] Vovk, V. (2012). Conditional validity of inductive conformal predictors. *Machine Learning*, 92:349–376.

[Vovk et al., 2005] Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world.

[Yeh, 2009] Yeh, I.-C. (2009). Default of Credit Card Clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

[Zargarbashi and Bojchevski, 2025] Zargarbashi, S. H. and Bojchevski, A. (2025). Robust conformal prediction with a single binary certificate.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable] (In the supplementary material)
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable] (In the supplementary material)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable] (Certain details were relegated to the supplementary material)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**] (All are very permissive and only require citation)
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A PROOFS

We will use the following well-known lemma from the Conformal Prediction literature:

Lemma A.1 (Quantile lemma)**.** *If $V_1, \ldots, V_{n+1}$ are exchangeable random variables, then for any $\beta \in (0,1)$, we have*

$$\mathbb{P}[V_{n+1} \leq \hat{q}_\beta(V_1, \ldots, V_n, +\infty)] \geq \beta,$$

where $\hat{q}_\beta(W_1, \ldots, W_m)$ is the empirical quantile of $W_1, \ldots, W_m$, given by

$$\hat{q}_\beta(W_1, \ldots, W_m) = \inf\left\{ t \in \mathbb{R} : \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[W_i \leq t] \geq \beta \right\}.$$

Proof. See Lemma 1 of [Tibshirani et al., 2019]. □

Theorem A.2 (Theorem 2.1 in the main text)**.** *Let $Z_1, \ldots, Z_n, Z^{\text{test}}$ be $n+1$ exchangeable random variables in $\mathcal{X} \times \mathcal{Y}$. Let t_α^{strat} be as above. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(\Delta(X^{\text{test}}))\right] \geq 1 - \alpha.$$

Proof. By definition of $C_{t_\alpha^{\text{strat}}}$,

$$\begin{aligned}
&\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(\Delta(X^{\text{test}}))] \\
&\quad = \mathbb{P}[\forall \Delta \in \mathcal{D},\ s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}}] \\
&\quad = \mathbb{P}[\sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}}]
\end{aligned}$$

For convenience, let $S_i := \sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i)$ and $S^{\text{test}} := \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})$, and let $S_{(1)}, \ldots, S_{(n)}, S_{(n+1)}$ be the order statistics of $S_1, \ldots, S_n, S^{\text{test}}$. Then

$$\mathbb{P}[\sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}}] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty)].$$

And, by Lemma A.1:

$$\begin{aligned}
&\mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty)] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, S^{\text{test}})] \\
&= \mathbb{P}[S^{\text{test}} \leq S_{(\lceil (1-\alpha)(n+1) \rceil)}] = \mathbb{P}[\text{rank}(S^{\text{test}}) \leq \lceil (1-\alpha)(n+1) \rceil] \\
&= \mathbb{P}_{U \sim \text{Unif}(1..n+1)}[U \leq \lceil (1-\alpha)(n+1) \rceil] = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq \frac{(1-\alpha)(n+1)}{n+1} = 1 - \alpha.
\end{aligned}$$

□

To prove our training-conditional guarantee, we will leverage a generalization of a lemma of [Bian and Barber, 2022]:

Lemma A.3 (Slight generalization of Lemma 1 from [Bian and Barber, 2022])**.** *Let $n \geq 2$ and choose a holdout set A with $\emptyset \subsetneq A \subsetneq [n]$. Let $\varphi_{[n]\setminus A} = \mathcal{A}((X_i, Y_i) : i \in [n] \setminus A) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{A}$ is any algorithm and may be deterministic or randomized. Define*

$$p_{A,\varphi}(x,y) := \frac{1}{|A|}\sum_{i \in A} \mathbb{1}[\varphi_{[n]\setminus A}(X_i, Y_i) \geq \varphi_{[n]\setminus A}(x,y)],$$

and

$$p^\star_{A,\varphi}(x,y) := \mathbb{P}[\varphi_{[n]\setminus A}(X,Y) \geq \varphi_{[n]\setminus A}(x,y) | \varphi_{[n]\setminus A}].$$

Then $p^\star_{A,\varphi}(x,y)$ is a valid p-value conditional on the training data, i.e.,

$$\mathbb{P}[p^\star_{A,\varphi}(x,y) \leq a | \mathcal{D}_n] <= a \text{ for all } a \in [0,1],\ \text{almost surely over } \mathcal{D}_n.$$

Moreover, for any $\epsilon \geq \sqrt{\log(2)/2|A|}$,

$$\mathbb{P}\left[\sup_{x,y}\left(p^\star_{A,\varphi}(x,y) - p_{A,\varphi}(x,y)\right) > \epsilon\right] \leq e^{-2|A|\epsilon^2}.$$

Proof. First, for any fixed function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, define

$$\bar{F}_s(t) = \mathbb{P}_P[s(X, Y) \geq t].$$

In other words, $\bar{F}_\mu(t)$ is the right-tailed CDF of the conformity score $s(X, Y)$ under $(X, Y) \sim P$. We can therefore write

$$p^*_{A,\varphi}(X_{n+1}, Y_{n+1}) = \bar{F}_s(s(X_{n+1}, Y_{n+1})).$$

Since $(X_{n+1}, Y_{n+1}) \sim P$ (and is independent of s), this is clearly a valid p-value by definition of $\bar{F}_s$.

Next, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we can calculate

$$\begin{aligned} p^*_{A,\varphi}(x, y) - p_{A,\varphi}(x, y) &= \bar{F}_s(s(x, y)) - \frac{1}{|A|} \sum_{i \in A} \mathbb{1}[\varphi_{[n]\setminus A}(X_i, Y_i) \geq \varphi_{[n]\setminus A}(x, y)] \\ &\leq \sup_{t \in \mathbb{R}} \left(\bar{F}_s(s(x, y)) - \frac{1}{|A|} \sum_{i \in A} \mathbb{1}[\varphi_{[n]\setminus A}(X_i, Y_i) \geq t] \right). \end{aligned}$$

Finally, since $(X_i, Y_i)_{i \in A}$ are drawn i.i.d. from P and are independent from $\varphi_{[n]\setminus A}$, for any $\epsilon \geq \sqrt{log(2)/2|A|}$ the Dvoretzky-Kiefer-Wolfowitz-Massart (DKW) inequality implies that, conditional on $\varphi_{[n]\setminus A}$,

$$\sup_{t \in RR} \left(\bar{F}_s(t) - \frac{1}{|A|} \sum_{i \in A} \mathbb{1}[\varphi_{[n]\setminus A}(X_i, Y_i) \geq t] \right) \leq \epsilon$$

holds with probability at least $1 - e^{-2|A|\epsilon^2}$. The same bound therefore holds marginally as well. □

We can now prove the Theorem:

Theorem A.4 (Theorem 2.2 in the main text)**.** *For any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the draw of i.i.d. random variables $Z_1, \ldots, Z_n$, it holds that*

$$\begin{aligned} &\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t^{\text{strat}}_\alpha}(\Delta(X^{\text{test}})) \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad \geq 1 - \alpha - \sqrt{\frac{\log 1/\delta}{2n}}. \end{aligned}$$

Proof. By definition of $C_{t^{\text{strat}}_\alpha}$,

$$\begin{aligned} &\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t^{\text{strat}}_\alpha}(\Delta(X^{\text{test}})) \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad = \mathbb{P}\left[\forall \Delta \in \mathcal{D},\ s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t^{\text{strat}}_\alpha \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad = \mathbb{P}\left[\sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t^{\text{strat}}_\alpha \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{1}\left[\sup_{\Delta' \in \mathcal{D}} s(\Delta'(X_i), Y_i) > \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})\right] > (n+1)\alpha - 1 \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{1}\left[\sup_{\Delta' \in \mathcal{D}} s(\Delta'(X_i), Y_i) > \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})\right] > n\alpha - (1-\alpha) \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad \geq \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{1}\left[\sup_{\Delta' \in \mathcal{D}} s(\Delta'(X_i), Y_i) > \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})\right] > n\alpha \;\middle|\; t^{\text{strat}}_\alpha\right] \\ &\quad = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{1}\left[\sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i) > \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})\right] > n\alpha \;\middle|\; t^{\text{strat}}_\alpha\right] = \mathbb{P}\left[p_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) > \alpha \;\middle|\; t^{\text{strat}}_\alpha\right], \end{aligned}$$

for $\varphi(x, y) = \sup_{\Delta \in \mathcal{D}} s(\Delta(x), y)$.

Now, by Lemma A.3, we have that for any $\epsilon \geq \sqrt{\log(2)/2n}$,

$$\mathbb{P}\left[\sup_{x,y}\big(p^*_{A,\varphi}(x,y) - p_{A,\varphi}(x,y)\big) > \epsilon\right] \leq e^{-2n\epsilon^2};$$

solving for $\delta = e^{-2n\epsilon^2}$, we get that, with probability of at least $1-\delta$,

$$\sup_{x,y}\big(p^*_{A,\varphi}(x,y) - p_{A,\varphi}(x,y)\big) \leq \sqrt{\frac{\log 1/\delta}{2n}},$$

and thus that still with probability of at least $1-\delta$,

$$\begin{aligned}
&\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t^{\text{strat}}_\alpha}(\Delta(X^{\text{test}})) \;\middle|\; t^{\text{strat}}_\alpha\right] \geq \mathbb{P}\left[p_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) > \alpha \;\middle|\; t^{\text{strat}}_\alpha\right] \\
&= \mathbb{P}\left[p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) - \left(p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) - p_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}})\right) > \alpha \;\middle|\; t^{\text{strat}}_\alpha\right] \\
&= \mathbb{P}\left[p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) > \alpha + \left(p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) - p_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}})\right) \;\middle|\; t^{\text{strat}}_\alpha\right] \\
&\geq \mathbb{P}\left[p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) > \alpha + \sup_{x,y}\left(p^\star_{[n]\setminus[n_0],\varphi}(x,y) - p_{[n]\setminus[n_0],\varphi}(x,y)\right) \;\middle|\; t^{\text{strat}}_\alpha\right] \\
&\geq \mathbb{P}\left[p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) > \alpha + \sqrt{\frac{\log 1/\delta}{2n}} \;\middle|\; t^{\text{strat}}_\alpha\right] \\
&= 1 - \mathbb{P}\left[p^\star_{[n]\setminus[n_0],\varphi}(X^{\text{test}}, Y^{\text{test}}) \leq \alpha + \sqrt{\frac{\log 1/\delta}{2n}} \;\middle|\; t^{\text{strat}}_\alpha\right] \geq 1 - \alpha - \sqrt{\frac{\log 1/\delta}{2n}},
\end{aligned}$$

where the last step holds by Lemma A.3, finishing the proof. □

The proofs of marginal coverage the group-conditional and label-conditional variants of our method are similar to the proof of Theorem A.2, with modifications analogous to the ones in the non-strategic setting.

Theorem A.5 (Theorem 2.3 in the main text)**.** *Let $Z_1, \ldots, Z_n, Z^{\text{test}}$ be $n+1$ exchangeable random variables in $\mathcal{X} \times \mathcal{Y}$. Let $C^{\text{strat-}(G_j)}_\alpha$ be as above (for $j = 1, \ldots, l$). Then, for any $\alpha \in (0,1)$,*

$$\begin{aligned}
\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C^{\text{strat-}(G_j)}_\alpha(\Delta(X^{\text{test}}))\middle| X^{\text{test}} \in G_j\right] \\
\geq 1-\alpha, \quad \forall j = 1, \ldots, l.
\end{aligned}$$

Proof. Consider any group G_j. By definition of $C^{\text{strat}-(G_j)}_\alpha$,

$$\begin{aligned}
&\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C^{\text{strat}-(G_j)}_\alpha(\Delta(X^{\text{test}}))|X^{\text{test}} \in G_j] \\
&= \mathbb{P}[\forall \Delta \in \mathcal{D},\ s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t^{\text{strat}-(G_j)}_\alpha|X^{\text{test}} \in G_j] \\
&= \mathbb{P}[\sup_{\Delta\in\mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t^{\text{strat}-(G_j)}_\alpha|X^{\text{test}} \in G_j]
\end{aligned}$$

For convenience, let $S_i := \sup_{\Delta\in\mathcal{D}} s(\Delta(X_i), Y_i)$ and $S^{\text{test}} := \sup_{\Delta\in\mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})$, $k = |i = 1, \ldots, n : X_i \in G_j|$, and let $S_{(1)}, \ldots, S_{(n)}, S_{(n+1)}$ be the order statistics of $S_1, \ldots, S_n, S^{\text{test}}$. Then

$$\mathbb{P}[\sup_{\Delta\in\mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t^{\text{strat}-(G_j)}_\alpha|X^{\text{test}} \in G_j] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty|X_* \in G_j)|X^{\text{test}} \in G_j].$$

And, by Lemma A.1:

$$\begin{aligned}
&\mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty|X_* \in G_j)|X^{\text{test}} \in G_j] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, S^{\text{test}}|X_* \in G_j)|X^{\text{test}} \in G_j] \\
&= \mathbb{P}[S^{\text{test}} \leq S_{(\lceil(1-\alpha)(k+1)\rceil)}|X^{\text{test}} \in G_j] = \mathbb{P}[\text{rank}(S^{\text{test}}|X_* \in G_j) \leq \lceil(1-\alpha)(k+1)\rceil] \\
&= \mathbb{P}_{U\sim\text{Unif}(1..k+1)}[U \leq \lceil(1-\alpha)(k+1)\rceil] = \frac{\lceil(1-\alpha)(k+1)\rceil}{k+1} \geq \frac{(1-\alpha)(k+1)}{k+1} = 1-\alpha.
\end{aligned}$$

□

For the label-conditional case, see Section B.1.

Most of our remaining results leverage the total variation distance, defined as such:

Definition A.1 (Total variation distance)**.** *Consider a measurable space $(\Omega, \mathcal{F})$ and probability measures(/distributions) P and Q defined on $(\Omega, \mathcal{F})$. The total variation distance between P and Q is defined as*

$$\mathrm{TV}(P \parallel Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

The definition above defines the total variation distance as an Integral Probability Metric (IPM). We note the well-known fact that it is also an f-divergence:

$$\mathrm{TV}(P \parallel Q) = \frac{1}{2}\mathbb{E}_{Z\sim P}\left[\left|\frac{\mathrm{d}Q}{\mathrm{d}P} - 1\right|\right].$$

Proposition A.1 (Proposition 2.1 in the main text)**.** *Let $C : \mathcal{X} \to 2^{\mathcal{Y}}$ be a set predictor. For any $\Delta \in \mathcal{D}$,*

$$\begin{aligned}&|\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta(X))] - \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))]| \\ &\quad \leq \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right].\end{aligned}$$

Proof. First, we rewrite the probabilities as probabilities over specific joint distributions:

$$\begin{aligned}\mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta(X))] &= \mathbb{P}_{\Delta(X),X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta(X))], \\ \mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta^\star(X))] &= \mathbb{P}_{\Delta^\star(X),X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta^\star(X))].\end{aligned}$$

By the definition of the total variation distance, it then follows that

$$\left|\mathbb{P}_{\Delta(X),X,Y}[Y \in C_t(\Delta(X))] - \mathbb{P}_{\Delta^\star(X),X,Y}[Y \in C_t(\Delta^*(X))]\right| \leq \mathrm{TV}(\Delta(X), X, Y \parallel \Delta^*(X), X, Y),$$

which, by the dual representation of the total variation distance as an f-divergence, can be rewritten as

$$\begin{aligned}&\mathrm{TV}(\Delta(X), X, Y \parallel \Delta^\star(X), X, Y) = \frac{1}{2}\,\mathbb{E}_{\Delta^\star(X),X,Y}\left[\left|\frac{\mathrm{d}P_{\Delta(X),X,Y}}{\mathrm{d}P_{\Delta^\star(X),X,Y}} - 1\right|\right] \\ &= \frac{1}{2}\,\mathbb{E}_{\Delta^\star(X),X,Y}\left[\left|\frac{\mathrm{d}P_{\Delta(X)|X,Y}\cdot \mathrm{d}P_{X,Y}}{\mathrm{d}P_{\Delta^\star(X)|X,Y}\cdot \mathrm{d}P_{X,Y}} - 1\right|\right] = \frac{1}{2}\,\mathbb{E}_{\Delta^\star(X),X,Y}\left[\left|\frac{\mathrm{d}P_{\Delta(X)|X,Y}}{\mathrm{d}P_{\Delta^\star(X)|X,Y}} - 1\right|\right] \\ &= \frac{1}{2}\,\mathbb{E}_{\Delta^\star(X),X,Y}\left[\left|\frac{\mathrm{d}P_{\Delta(X)|X}}{\mathrm{d}P_{\Delta^\star(X)|X}} - 1\right|\right] = \mathbb{E}_{X,Y}\left[\frac{1}{2}\,\mathbb{E}_{\Delta^\star(X)|X,Y}\left[\left|\frac{\mathrm{d}P_{\Delta(X)|X}}{\mathrm{d}P_{\Delta^\star(X)|X}} - 1\right|\middle| X, Y\right]\right] \\ &= \mathbb{E}_X\left[\frac{1}{2}\,\mathbb{E}_{\Delta^\star(X)|X}\left[\left|\frac{\mathrm{d}P_{\Delta(X)|X}}{\mathrm{d}P_{\Delta^\star(X)|X}} - 1\right|\middle| X\right]\right] = \mathbb{E}_X[\mathrm{TV}(\Delta(X)|X \| \Delta^\star(X)|X)],\end{aligned}$$

which concludes the proof. □

Proposition A.2 (Proposition 2.2 in the main text)**.** *Let $C : \mathcal{X} \to 2^{\mathcal{Y}}$ be a set predictor satisfying Equation 2. Then*

$$\begin{aligned}&\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] \\ &\quad \geq 1 - \alpha - \inf_{\Delta \in \mathcal{D}} \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right].\end{aligned}$$

Proof. By Proposition A.1, for any $\Delta \in \mathcal{D}$,

$$|\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta(X))] - \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))]| \leq \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right],$$

which means that

$$\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] \geq \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta(X))] - \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right],$$

and by Theorem A.2 the probability on the right-hand-side is lower-bounded by $1-\alpha$, and thus

$$\mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] \geq 1 - \alpha - \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right],$$

Taking the supremum over $\Delta \in \mathcal{D}$ on both sides, we get

$$\begin{aligned}\sup_{\Delta\in\mathcal{D}} \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] &\geq \sup_{\Delta\in\mathcal{D}} \left(1 - \alpha - \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right]\right)\\ \mathbb{P}_{X,Y}[Y \in C_\alpha(\Delta^\star(X))] &\geq 1 - \alpha - \inf_{\Delta\in\mathcal{D}} \mathbb{E}_X\left[\mathrm{TV}(\Delta(X)|X \parallel \Delta^\star(X)|X)\right],\end{aligned}$$

as desired. □

Proposition A.3 (Proposition 2.3 in the main text)**.** *Let $(\Delta_k)_{k=0}^\infty$ be a sequence of alteration functions $\Delta_k : \mathcal{X} \to \mathcal{X}$, such that $\lim_{k\to\infty} \Delta_k(X) = \overline{\Delta}(X)$ according to the total variation metric. Then, for any $k_{\max} > 0$, consider the set predictor $C_{t_\alpha^{\mathrm{strat}}}$ generated by $\mathcal{D} = \{\Delta_0, \ldots, \Delta_{k_{\max}}\}$. There exists some $\epsilon_{k_{\max}}$ such that,*

$$\mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\overline{\Delta}(X))] \geq 1 - \alpha - \epsilon_{k_{\max}},$$

and $\lim_{k_{\max}\to\infty} \epsilon_{k_{\max}} = 0$.

Proof. Consider $\epsilon_{k_{\max}} = \mathrm{TV}(\Delta_k \parallel \overline{\Delta})$. Since the Δ_k converge to $\overline{\Delta}$ in total variation distance, by definition, it holds that $\lim_{k\in\infty} \mathrm{TV}(\Delta_k \parallel \overline{\Delta}) = 0$. And, by the definition of total variation distance, we have that

$$\begin{aligned}\mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\overline{\Delta}(X))] &= \mathbb{E}_{\overline{\Delta}(X)}[\mathbb{P}_Y[Y \in C_{t_\alpha^{\mathrm{strat}}}(\overline{\Delta}(X))|\overline{\Delta}(X)]]\\ &\geq \mathbb{E}_{\Delta_k(X)}[\mathbb{P}_Y[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta_k(X))|\Delta_k(X)]] + \mathrm{TV}(\Delta_k \parallel \overline{\Delta})\\ &= \mathbb{P}_{X,Y}[Y \in C_{t_\alpha^{\mathrm{strat}}}(\Delta_k(X))] + \mathrm{TV}(\Delta_k \parallel \overline{\Delta})\\ &\geq 1 - \alpha + \mathrm{TV}(\Delta_k \parallel \overline{\Delta}) = 1 - \alpha - \epsilon_{k_{\max}}.\end{aligned}$$

□

To prove the following results we will make use of the following well-known lemma about the total-variation distance:

Lemma A.6. *For any two distributions P and Q over $\mathcal{H}$ and function $\phi : \mathcal{H} \to [0,1]$, it holds that*

$$\left|\mathbb{E}_{X\sim P}[\phi(X)] - \mathbb{E}_{Y\sim Q}[\phi(Y)]\right| \leq \mathrm{TV}(P\|Q).$$

Proof. See, e.g., [Ohnishi and Honorio, 2020], Lemma 4. □

Proposition A.4 (Proposition 2.4 in the main text)**.** *Let $\mu : \mathcal{X} \to \mathbb{R}$ be a possibly-stochastic function representing a base model. Consider the predictive sets $C_{t_\alpha^{\mathrm{strat}}}$ with conformity score $s : \mathcal{X}\times\mathbb{R} \to \mathbb{R}$ given by $s(x,y) = |\mu(x) - y|$ with $\mathcal{Y}$ being a compact subset of $\mathbb{R}$ such that $C_{t_\alpha^{\mathrm{strat}}}(\cdot) \subset \mathcal{Y}$ and alteration functions $\Delta \in \mathcal{D}$, and contrast it with the corresponding standard conformal predictive sets $C_{t_\alpha^{\mathrm{std}}}$. Then, almost surely over X,*

$$\begin{aligned}&\mathbb{E}_{C_{t_\alpha^{\mathrm{strat}}}}\left[\mathrm{leb}_{\mathcal{Y}}\left(C_{t_\alpha^{\mathrm{strat}}}(X)\right)\right] \leq \mathbb{E}_{C_{t_\alpha^{\mathrm{std}}}}\left[\mathrm{leb}_{\mathcal{Y}}\left(C_{t_\alpha^{\mathrm{std}}}(X)\right)\right]\\ &+ 2\,\mathrm{TV}\left(\left(s(X,Y)\right)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X),Y)\right)_{[j_{n,\alpha}]}\right),\end{aligned}$$

for $j_{n,\alpha} = \lceil (1-\alpha)(1+n)\rceil$.

Proof. First we need to connect the size of the conformal sets $C_t(X)$ with the calibrated threshold t. In the case of the conformity score we are considering, $s(x,y) = |\mu(x) - y|$, this is particularly simple, with the interval size corresponding to $2t$ precisely. So we have that

$$\mathbb{E}_{C_{t_\alpha^{\mathrm{strat}}}}[\mathrm{leb}(C_{t_\alpha^{\mathrm{strat}}}(X))] = \mathbb{E}_{t_\alpha^{\mathrm{strat}}}[t_\alpha^{\mathrm{strat}}] \qquad \text{and} \qquad \mathbb{E}_{C_{t_\alpha^{\mathrm{std}}}}[\mathrm{leb}(C_{t_\alpha^{\mathrm{std}}}(X))] = \mathbb{E}_{t_\alpha^{\mathrm{std}}}[t_\alpha^{\mathrm{std}}].$$

Next, note that the calibrated thresholds correspond exactly to the $j_{n,\alpha}$-th order statistic of the conformity scores (resp. suprema of conformity scores in the strategic setting), so

Finally, we will use Lemma A.6. Since $\mathcal{Y}$ is bounded and by definition $C_t(X) \subset \mathcal{Y}\ \forall X$, by monotonicity of measures it must also hold that

$$\begin{aligned}
\text{leb}(\mathcal{Y}) &\geq \text{leb}(C_{t_\alpha^{\text{strat}}}(X)) = 2t_\alpha^{\text{strat}} = 2(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y))_{[j_{n,\alpha}]} \\
\text{leb}(\mathcal{Y}) &\geq \text{leb}(C_{t_\alpha^{\text{std}}}(X)) = 2t_\alpha^{\text{std}} = 2(s(X, Y))_{[j_{n,\alpha}]},
\end{aligned}$$

and thus, by Lemma A.6, taking $\phi(t) = \text{leb}(C_t(X)) = 2t$,

$$\mathbb{E}[\phi(t^{\text{strat}})] = \mathbb{E}[\text{leb}(C_{t_\alpha^{\text{strat}}}(X))] = 2(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y))_{[j_{n,\alpha}]} \leq \text{leb}(C_{t_\alpha^{\text{std}}}(X))$$

□

Corollary A.6.1 (Corollary 2.3.1 in the main text)**.** *In the same setting as Proposition A.4, further assume that $\mu(X) = Y + \epsilon$ for some random ϵ (arbitrarily dependent on X and Y), with $\epsilon \in [-M, M]$ almost surely. Then, almost surely over X,*

$$\begin{aligned}
\mathbb{E}_{C_{t_\alpha^{\text{strat}}}}[\text{leb}_{\mathcal{Y}}\left(C_{t_\alpha^{\text{strat}}}(X)\right)] &\leq 2M \\
&+ 2\,\text{TV}\left((\epsilon)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y)\right)_{[j_{n,\alpha}]}\right),
\end{aligned}$$

for $j_{n,\alpha} = \lceil (1-\alpha)(1+n) \rceil$.

Proof. By Proposition A.4,

$$\mathbb{E}[\text{leb}_{\mathcal{Y}}[C_{t_\alpha^{\text{strat}}}(X)]] \leq \mathbb{E}[\text{leb}_{\mathcal{Y}}[C_{t_\alpha^{\text{std}}}(X)]] + 2\,\text{TV}\left(\left(s(X, Y)\right)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y)\right)_{[j_{n,\alpha}]}\right).$$

Now, note that since $\mu(X) = Y + \epsilon$ almost surely, we have that $s(X, Y) = |\mu(X) - Y| = |Y + \epsilon - Y| = |\epsilon| \leq M$ also almost surely. Thus

$$\mathbb{E}[\text{leb}_{\mathcal{Y}}[C_{t_\alpha^{\text{std}}}(X)]] \leq 2t_\alpha^{\text{std}} = 2\left(s(X, Y)\right)_{[j_{n,\alpha}]} = 2(\epsilon)_{[j_{n,\alpha}]} \leq 2M,$$

and so

$$\begin{aligned}
\mathbb{E}[\text{leb}_{\mathcal{Y}}[C_{t_\alpha^{\text{strat}}}(X)]] &\leq 2M + 2\,\text{TV}\left(\left(s(X, Y)\right)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y)\right)_{[j_{n,\alpha}]}\right). \\
&\leq 2M + 2\,\text{TV}\left((\epsilon)_{[j_{n,\alpha}]} \,\middle\|\, \left(\sup_{\Delta\in\mathcal{D}} s(\Delta(X), Y)\right)_{[j_{n,\alpha}]}\right).
\end{aligned}$$

□

B ADDITIONAL RESULTS

B.1 Label-Conditional Guarantee

We can get a Label-Conditional Guarantee [Sadinle et al., 2016] for the case where $\mathcal{Y}$ is discrete. We want to find a $C_\alpha^{\text{strat-}(\mathcal{Y})}$ such that

$$\begin{aligned}
\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y \in C_\alpha^{\text{strat-}(\mathcal{Y})}(\Delta(X)) | Y = y] \geq 1 - \alpha, \\
\text{for all } y \in \mathcal{Y}.
\end{aligned}$$

Again, we can do strategic calibration separately for each label, and use the corresponding thresholds.

$$C_\alpha^{\text{strat-}(\mathcal{Y})}(X) = \left\{ y \in \mathcal{Y} : s(X, y) \leq t_\alpha^{\text{strat-}(y)} \right\}$$

$$t_\alpha^{\text{strat-}(y)} = \inf\Bigg\{ t \in \mathbb{R} : \sum_{\substack{i=1 \\ X_i \in G_j}}^{n} \mathbb{1}\left[\sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i) > t \right] + 1 \leq \alpha(n_{G_j} + 1) \Bigg\}.$$

Theorem B.1. *Let $Z_1, \ldots, Z_n, Z^{\text{test}}$ be $n+1$ exchangeable random variables in $\mathcal{X} \times \mathcal{Y}$. Let $C_\alpha^{\text{strat-}(\mathcal{Y})}$ be as above. Then, for any $\alpha \in (0, 1)$,*

$$\mathbb{P}\left[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_\alpha^{\text{strat-}(\mathcal{Y})}(\Delta(X^{\text{test}})) \middle| Y^{\text{test}} = y \right] \geq 1 - \alpha, \quad \forall y \in \mathcal{Y}.$$

Proof. Consider any label $y \in \mathcal{Y}$. By definition of $C_\alpha^{\text{strat}-(\mathcal{Y})}$,

$$\begin{aligned}
&\mathbb{P}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_\alpha^{\text{strat}-(\mathcal{Y})}(\Delta(X^{\text{test}}))|Y^{\text{test}} = y] \\
&= \mathbb{P}[\forall \Delta \in \mathcal{D},\ s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}-(y)}|Y^{\text{test}} = y] \\
&= \mathbb{P}[\sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}-(y)}|Y^{\text{test}} = y]
\end{aligned}$$

For convenience, let $S_i := \sup_{\Delta \in \mathcal{D}} s(\Delta(X_i), Y_i)$ and $S^{\text{test}} := \sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}})$, $k = |i = 1, \ldots, n : Y_i = y|$, and let $S_{(1)}, \ldots, S_{(n)}, S_{(n+1)}$ be the order statistics of $S_1, \ldots, S_n, S^{\text{test}}$. Then

$$\mathbb{P}[\sup_{\Delta \in \mathcal{D}} s(\Delta(X^{\text{test}}), Y^{\text{test}}) \leq t_\alpha^{\text{strat}-(y)}|Y^{\text{test}} = y] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty|Y_* = y)|Y^{\text{test}} = y].$$

And, by Lemma A.1:

$$\begin{aligned}
&\mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, +\infty|Y_* = y)|Y^{\text{test}} = y] = \mathbb{P}[S^{\text{test}} \leq \widehat{q}_{1-\alpha}(S_1, \ldots, S_n, S^{\text{test}}|Y_* = y)|Y^{\text{test}} = y] \\
&= \mathbb{P}[S^{\text{test}} \leq S_{(\lceil (1-\alpha)(k+1) \rceil)}|Y^{\text{test}} = y] = \mathbb{P}[\text{rank}(S^{\text{test}}|Y_* = y) \leq \lceil (1-\alpha)(k+1) \rceil] \\
&= \mathbb{P}_{U \sim \text{Unif}(1..k+1)}[U \leq \lceil (1-\alpha)(k+1) \rceil] = \frac{\lceil (1-\alpha)(k+1) \rceil}{k+1} \geq \frac{(1-\alpha)(k+1)}{k+1} = 1 - \alpha.
\end{aligned}$$

□

Theorem B.1 holds true even if $\mathcal{Y}$ is not discrete. However, the predicted intervals will contain infinitely many points, since the thresholds for the infinitely many values of $y \in \mathcal{Y}$ that did not appear in the calibration set will be infinite.

C EXPERIMENT DETAILS

C.1 Datasets

We list our datasets along with their corresponding "fully-rational" utilities $u(X')$. In these utilities, $\mu(x)$ and $p(\dot{|}x)$ correspond to the trained base model, as described in the main text.

- `academic-dropout` [Realinho et al., 2021]: A dataset for prediction of academic dropout in higher education. Students try to exclude the possibility of dropout from the predictive sets, as the possibility of such an outcome may hinder opportunities. $u(X') = -p(\text{will dropout}|X')$.
- `spambase` [Hopkins et al., 1999]: A dataset for classification of email into spam and non-spam. Spammers try to exclude the spam outcome from the predictive sets. $u(X') = -p(\text{is spam}|X')$.

- `shoppers` [Sakar and Kastro, 2018]: A dataset for prediction of sales of products. Sellers have the incentive to make the system believe that there will be a larger amount of sales, as it would increase exposure. $u(X') = -p(\text{will not be a sale}|X')$.
- `news` [Fernandes et al., 2015]: A dataset for prediction of the number of 'shares' for news articles (e.g., for recommendation purposes). News sources have an incentive to have the algorithm believe there will be a larger number of shares, as it would increase exposure. $u(X') = \sup_{0 \leq y \leq 1000} |\mu(X') - y|$.
- `wine` [Cortez et al., 2009]: A dataset for assessment of wine quality, with samples composed of physicochemical tests of red and white vinho verde wine samples from the north of Portugal. Wineries have an incentive to obtain predictions of higher quality, as it would increase the value of their wine. $u(X') = \sup_{0 \leq y \leq 7} |\mu(X') - y|$.
- `productivity` [dat, 2020]: A dataset for inferring the productivity of workers of a garmet manufacturing process. Workers naturally wish to appear more productive. $u(X') = \sup_{0 \leq y \leq 0.5} |\mu(X') - y|$.
- `taiwan` [Yeh, 2009]: A dataset for credit scoring over a population of customers in Taiwan. Individuals attempt to have positive credit scores. $u(X') = -p(\text{will default}|X')$.
- `bank-marketing` [Moro et al., 2014]: A dataset for the prediction of the success of marketing campaigns of a Portuguese banking institution, which can then be leveraged to make calls. Malicious individuals may seek to fool these predictions for their own benefit, e.g., to maximize opportunities. $u(X') = -p(\text{will churn}|X')$.
- `census-income` [Kohavi, 1996]: A dataset for predicting whether individuals' income exceeds \$50K/yr based on census data. Such predictions can then be used, e.g., for credit scoring, and so there is significant incentive to make the system predict a higher income. $u(X') = -p(\text{is under \$50k}|X')$.

C.2 Base models and conformity score

We apply our method atop two XGBoost models: one trained in the usual way ('Plain XGBoost') and another trained by repeated risk minimization [Perdomo et al., 2020] ('Strategic XGBoost'), which is more well-suited to the strategic setting.

Let $\text{train}(X_{1:n}, Y_{1:n}) : \mathcal{X} \to \mathcal{Y}$ denote the XGBoost model trained on supervised data $X_{1:n} \in \mathcal{X}^n$ and $Y_{1:n} \in \mathcal{Y}^n$, and let $X^{\text{tr}}_{1:n}, Y^{\text{tr}}_{1:n}$ be the complete training dataset. Furthermore, let Δ_h denote the strategic alterations as a function of the trained model h.

The 'Plain XGBoost' model is given simply by $\text{train}(X^{\text{tr}}_{1:n}, Y^{\text{tr}}_{1:n})$. The 'Strategic XGBoost' model corresponds to $\text{train}(X^{(10)}_{1:n}, Y^{\text{tr}}_{1:n})$,where

$$X^{(0)}_{1:n} = X^{\text{tr}}_{1:n} \qquad\qquad X^{(k+1)}_{1:n} = \Delta_{\text{train}(X^{(k)}_{1:n}, Y)}(X^{\text{tr}}_{1:n}).$$

For conformal prediction, we then consider the conformity score $s(x, y) = |\mu(x) - y|$ for regression (where $\mu : \mathcal{X} \to \mathbb{R}$ is the base XGBoost model) and $s(x, y) = 1 - p(y|x)$ for classification (where $p(y|\cdot) : \mathcal{X} \to [0, 1]$ for each $y \in \mathcal{Y}$ are the conditional probabilities from the base XGBoost model).

C.3 Strategic alterations

Throughout, we consider two kinds of strategic alterations:

Utility + cost Following the scheme described in Section 2.1.1, using the utility specified in Section C.1 for the corresponding datasets, and cost given by the Malahanobis distance $c(X, X') = X^T \Sigma^{-1} X'$, where Σ is the estimated covariance matrix of the covariates. The set $\mathcal{D}$ is generated by regularization scales $\lambda \in \{10^{-7}, 1, 5\}$, and optimization is done by random search over the domain (whose zeroth-order nature allows for usage over non-differentiable base models).

Iterative random search Following the scheme described in Section 2.1.2 and 2.1.3, using the utility specified in Section C.1 for the corresponding datasets, while sampling the δ from a multivariate normal distribution with mean 0 and covariance matching the one estimated from the data rescaled by 0.2, with just two δ being sampled per iteration.

C.4 Figure 1: coverage and interval sizes for our method versus standard CP

On the left, the x-axis corresponds to the confidence level $1-\alpha$ for the conformal calibration, while the y-axis corresponds to the empirical coverage under strategic alterations, $\widehat{\mathbb{P}}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(X^{\text{test}})]$, estimated on the test set. All instances of Figure 1 use the iterative random search alterations.

Ideally, both will match up, leading to the 'exact coverage' diagnoal line. Below the line we get 'invalid coverage', meaning that the probability is below the specified $1-\alpha$, and above the line we get 'valid coverage', meaning that the probability is above the specified $1-\alpha$ (though possibly excessively so).

On the right, the x-axis is the same, but the y-axis reports the mean interval size under strategic alterations, $\widehat{\mathbb{E}}[\sup_{\Delta\in\mathcal{D}} \text{size}(C_{t_\alpha^{\text{strat}}}(X^{\text{test}}))]$, where size is the Lebesgue measure in the regression case and the counting measure for the classification case, similarly estimated on the test set.

C.5 Figure 2: coverage of our method andstandard CP for varying levels of strategic alterations

All instances of Figure 2 use the iterative random search alterations. The x-axis corresponds to the size of the $\mathcal{D}$ used for metrics on the test set, while the color of the curves indicate the size of the $\mathcal{D}$ used for calibration (as indicated in the legend). The y-axis corresponds to the value of the strategic coverage under the level of alteration indicated on the x-axis.

All conformal calibrations were done with $1-\alpha = 90\%$, which is marked on the plots by the dashed line. (Above the dashed line is valid coverage, while below is invalid coverage.)

The left plot is for the Plain XGBoost model, while the right one is for the Strategic XGBoost one.

C.6 Table 1: evaluation of our method on multiple datasets and forms of strategic alterations for $\alpha = 0.1$

Table columns:

- **DATASET**: the dataset being considered.
- **UNDERLYING MODEL**: the base model – either Plain XGBoost or Strategic XGBoost
- **Δs**: the strategic alterations being considered – either Utility-cost (the utlity + cost construction) or Rand. Search. (the iterative random search construction).
- **STRATEGIC COVERAGE: OURS**: the strategic coverage of our method (Strategic Conformal Prediction): $\widehat{\mathbb{P}}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{strat}}}(X^{\text{test}})]$, estimated on the test set;
- **STRATEGIC COVERAGE: STD**: the strategic coverage of standard conformal prediction: $\widehat{\mathbb{P}}[\forall \Delta \in \mathcal{D},\ Y^{\text{test}} \in C_{t_\alpha^{\text{std}}}(X^{\text{test}})]$, estimated on the test set;
- **AVG SET SIZE DIFF**: the average strategic size difference: $\widehat{\mathbb{E}}[\inf_{\Delta\in\mathcal{D}} \text{size}(C_{t_\alpha^{\text{std}}}(X^{\text{test}}))]$, estimated on the test set.

All $\pm$ correspond to 95% bootstrapped confidence intervals. Black bold coverages indicate the best method for that row (incidentally always ours), while purple coverages indicate substantially invalid coverages (more than 30% away from attaining valid coverage).

C.7 Extra figure: coverage under change of the stochastic step

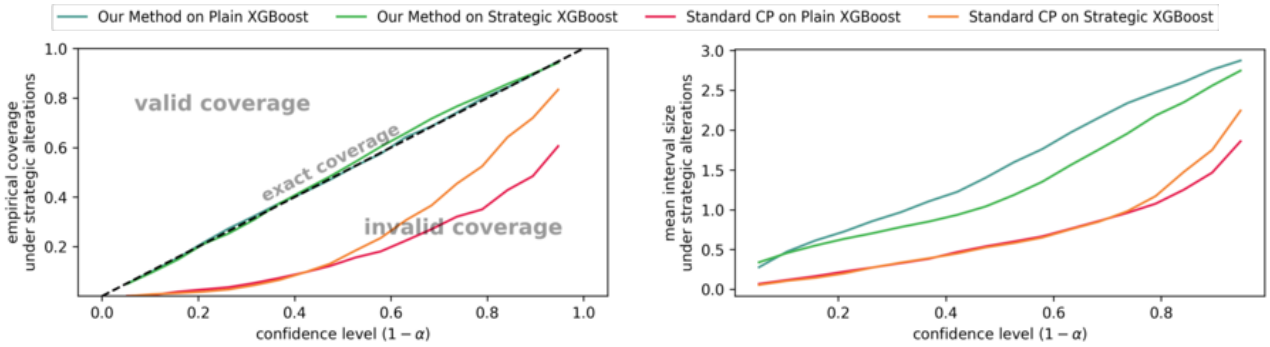
With this experiment, our goal is to analyze the behavior of our method under a more strenuous form of model specification where we alter the actual stochastic step used for the stochastic simulations. To do so, we consider the strategic alterations based on iterative random search. Let $\mathcal{D} = \{\Delta_0, \ldots, \Delta_k\}$ denote the set of these strategic alterations. We consider a family of 'modified strategic alterations' $\widetilde{\Delta}_k(X) = \Delta_k(X) + \mathcal{N}(0, \lambda\Sigma)$, where $\lambda > 0$ and Σ is the estimated covariance matrix of the data. This way, by taking $\lambda = 0$ (at the limit) we recover the original Δs, while by increasing λ we stray away from it, increasing $\text{TV}(\widetilde{\Delta}_k(X) \parallel \Delta_k(X))$.

The figures plot on the x-axis the values of this λ, and on the y-axis the strategic coverage. The blue line corresponds to Plain XGBoost, and the orange one to Strategic XGBoost.

Interestingly, using Strategic XGBoost increases our reliance on the correct specification of the Δs, and so actually *decreases* our robustness.

D MORE EXPERIMENT RESULTS

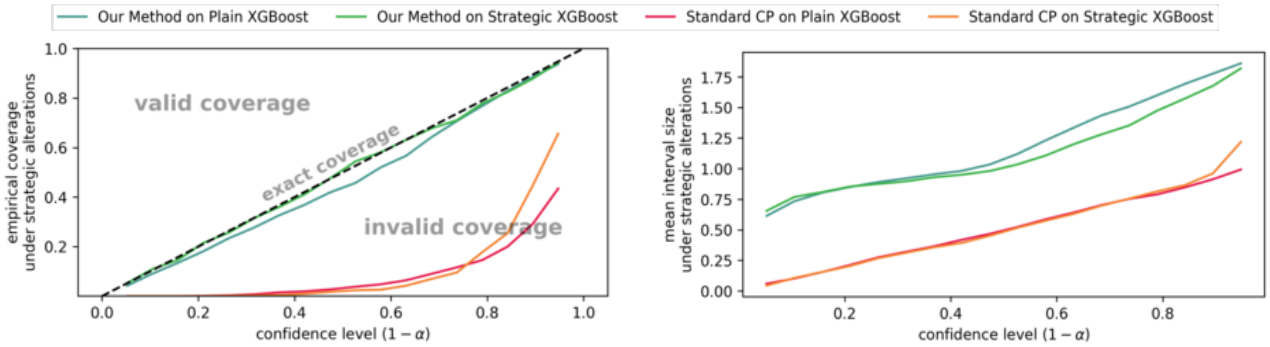
D.1 Figure 1: coverage and interval sizes for our method versus standard CP



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

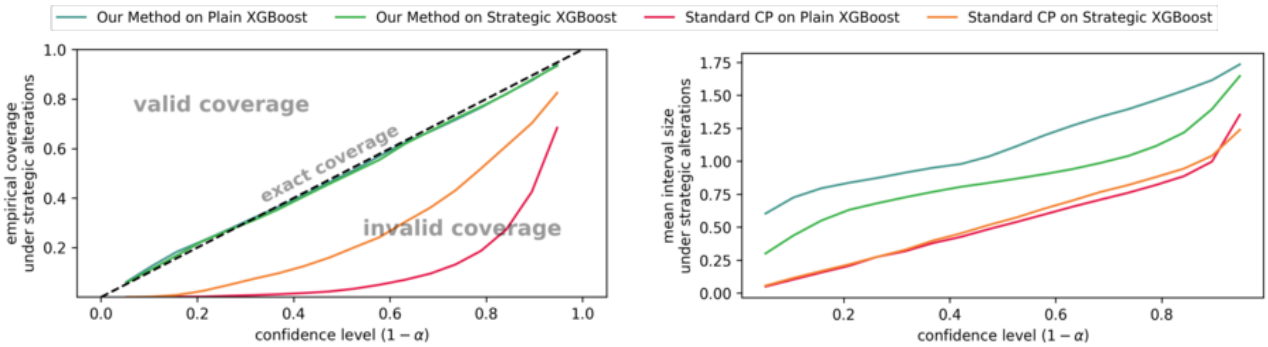
Figure 3: xgboost on `academic-dropout`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

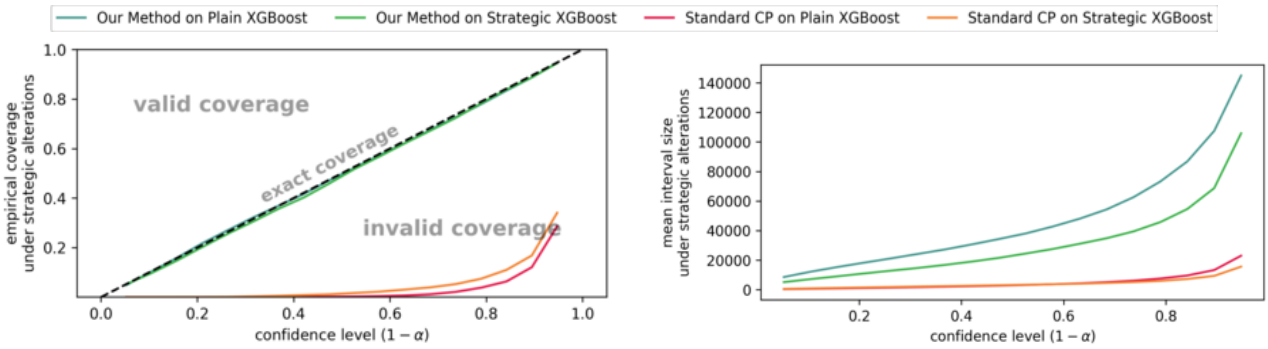
Figure 4: xgboost on `spambase`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

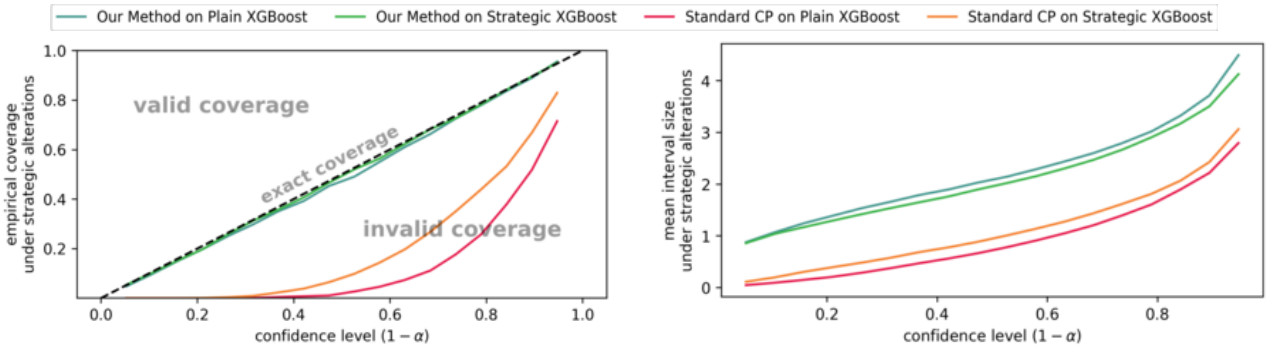
Figure 5: xgboost on `shoppers`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

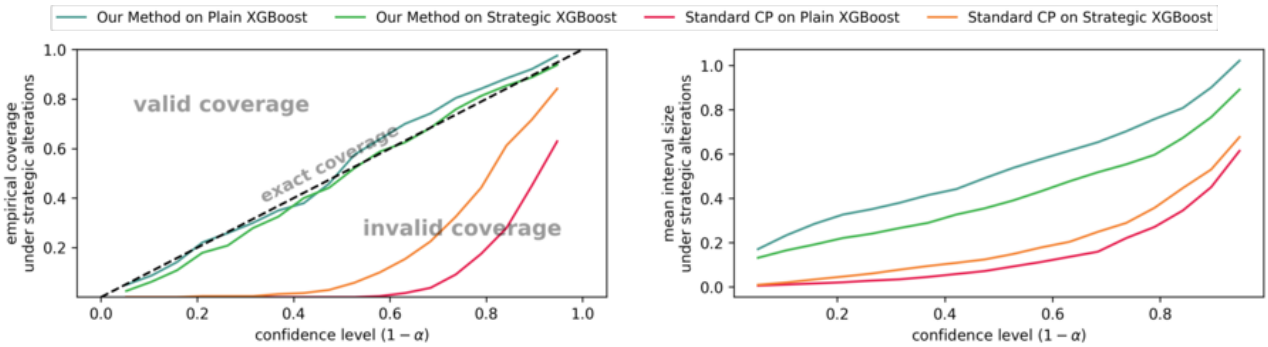
Figure 6: xgboost on `news`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

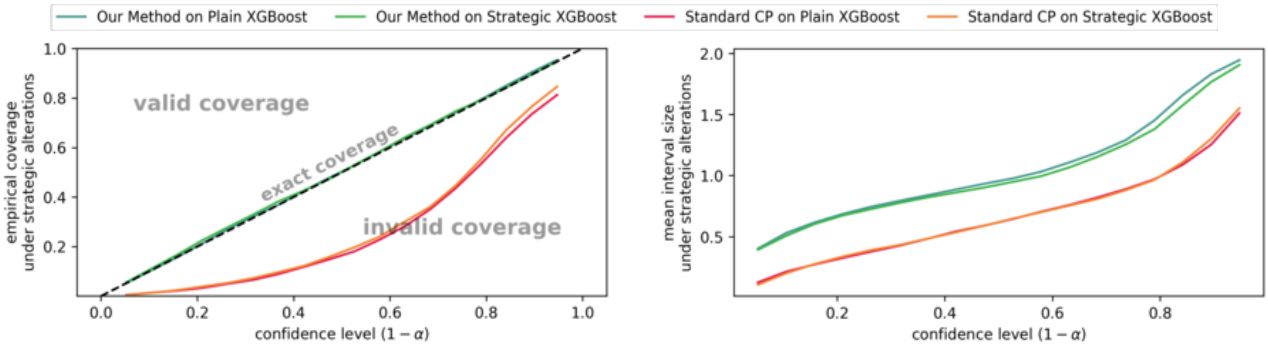
Figure 7: xgboost on `wine`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

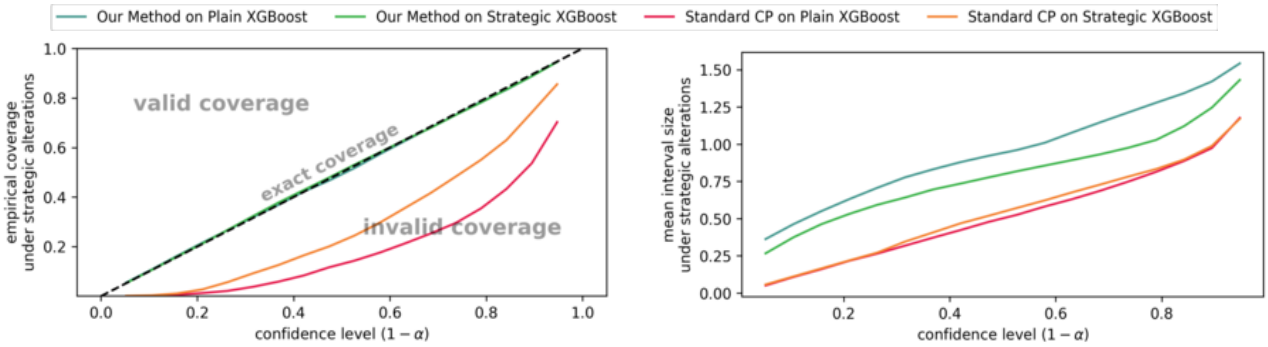
Figure 8: xgboost on `productivity`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

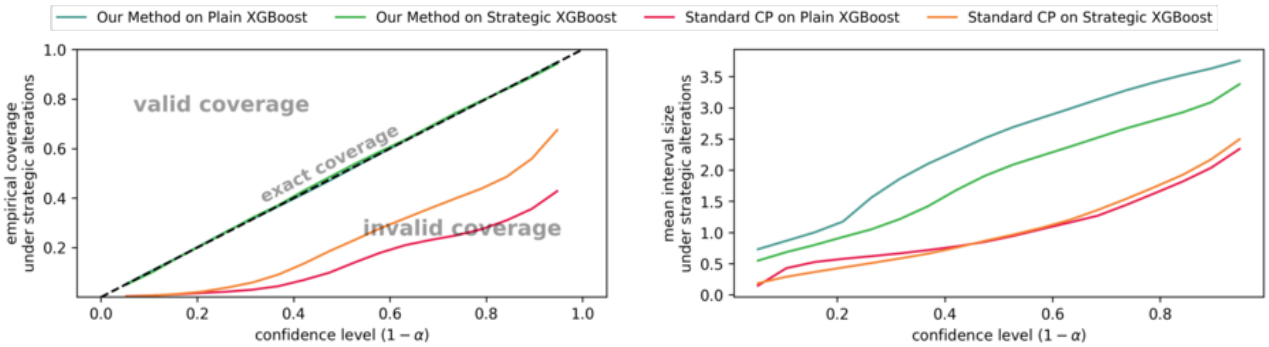
Figure 9: xgboost on `taiwan`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

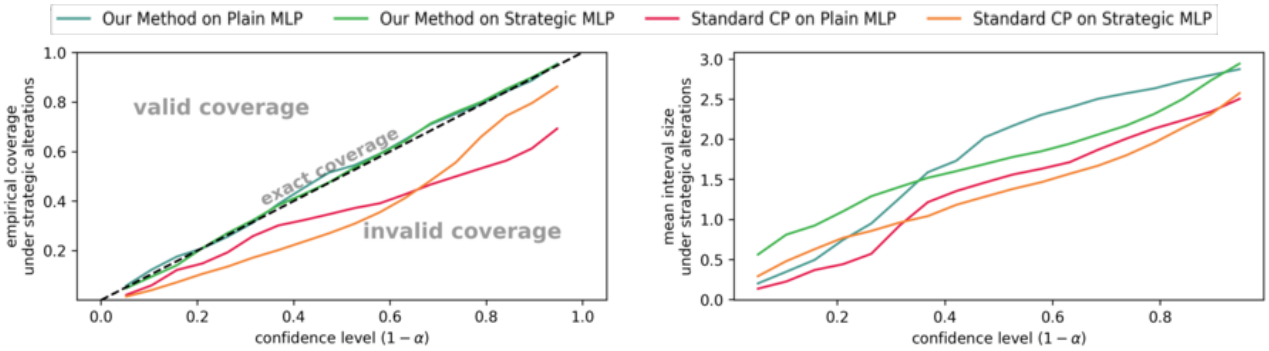
Figure 10: xgboost on `bank-marketing`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

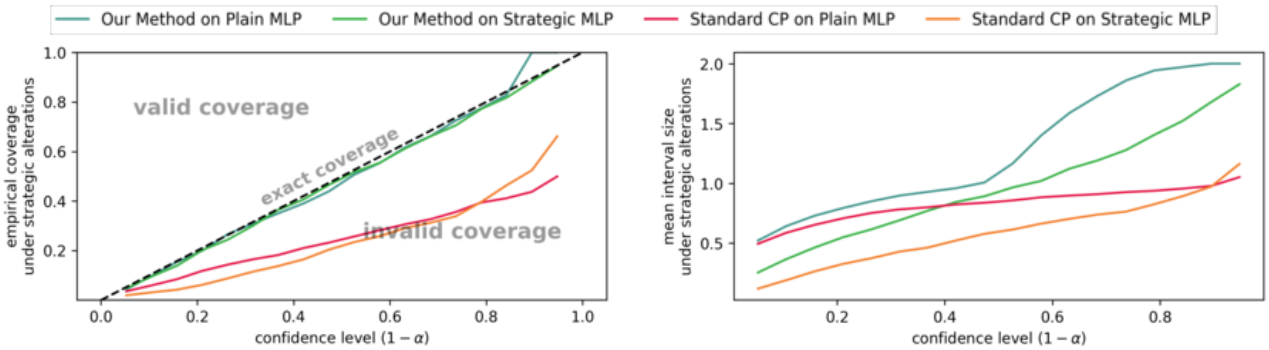
Figure 11: xgboost on `census-income`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

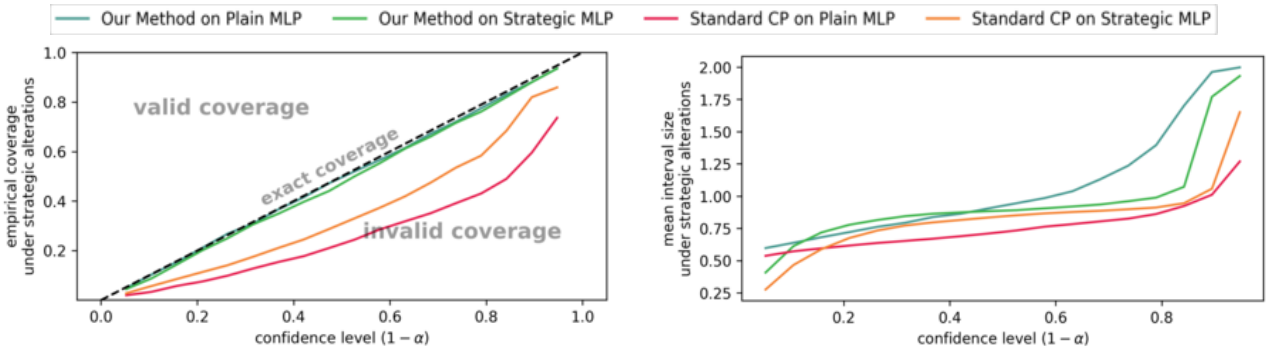
Figure 12: mlp on `academic-dropout`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

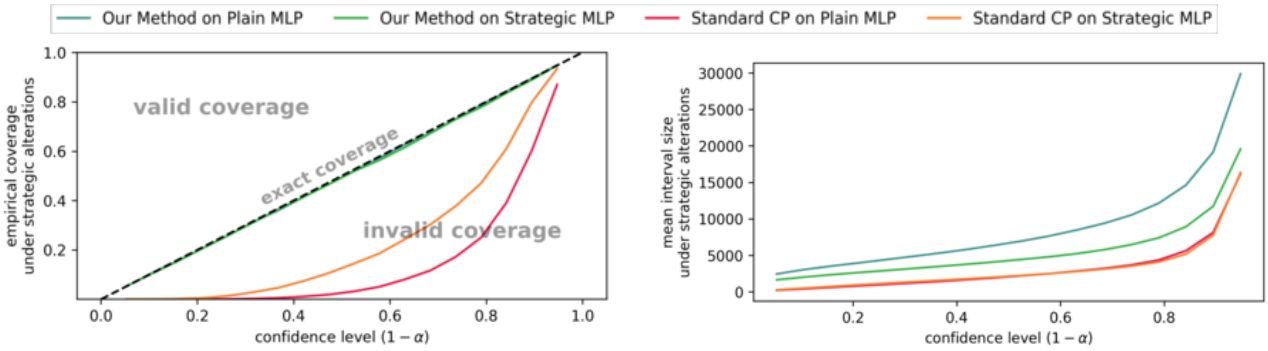
Figure 13: mlp on `spambase`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

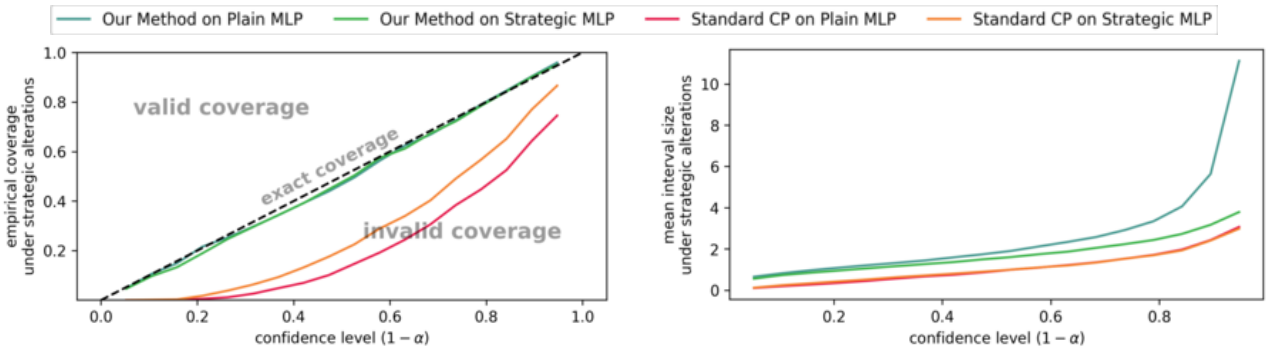
Figure 14: mlp on `shoppers`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

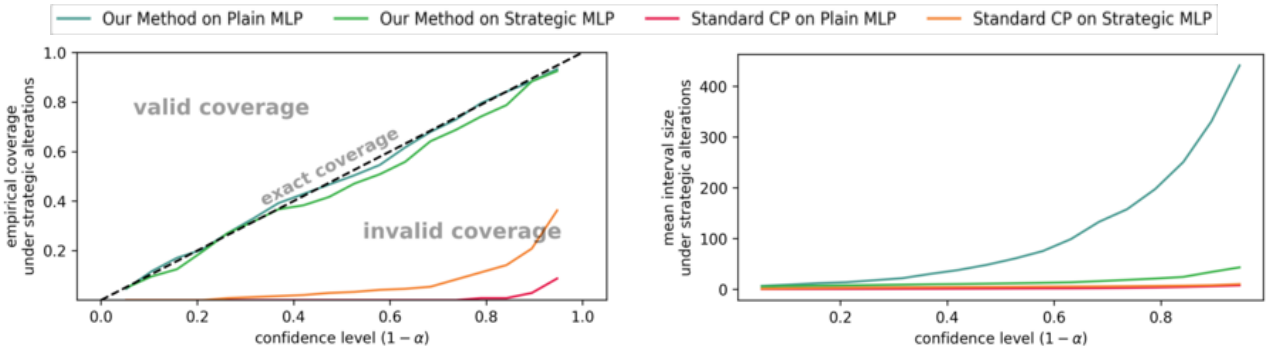
Figure 15: mlp on `news`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

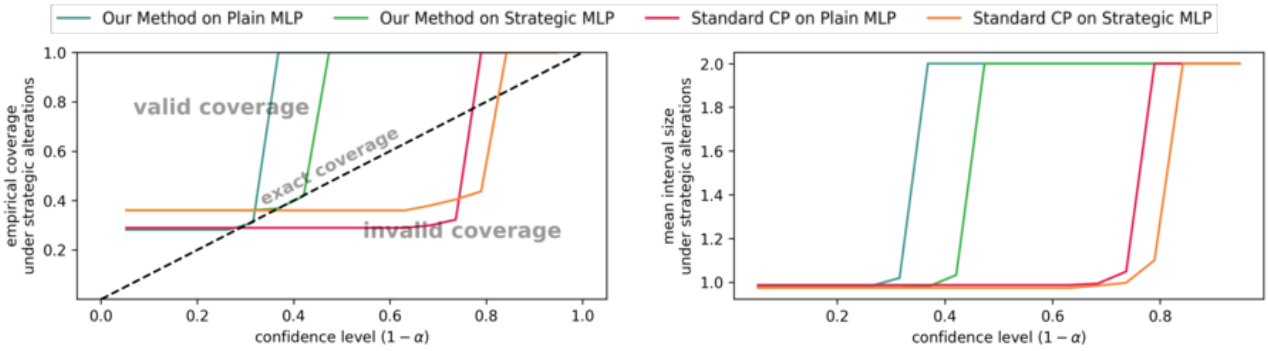
Figure 16: mlp on `wine`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

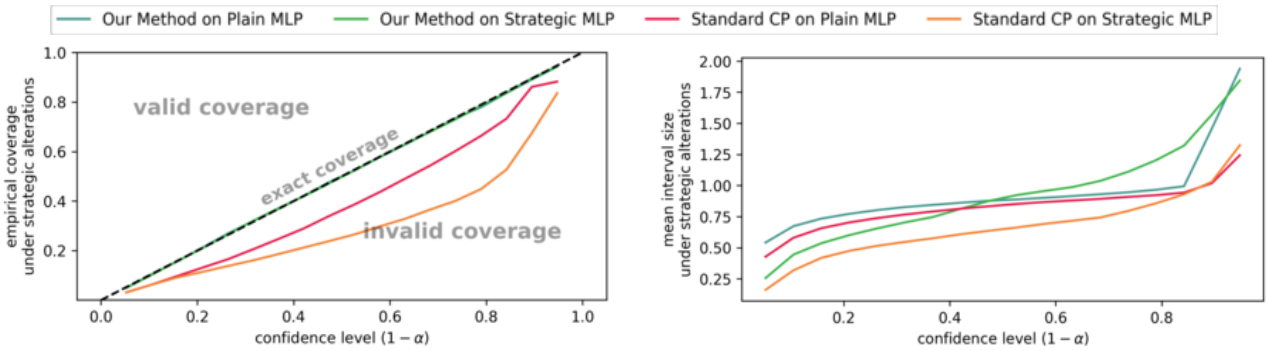
Figure 17: mlp on `productivity`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

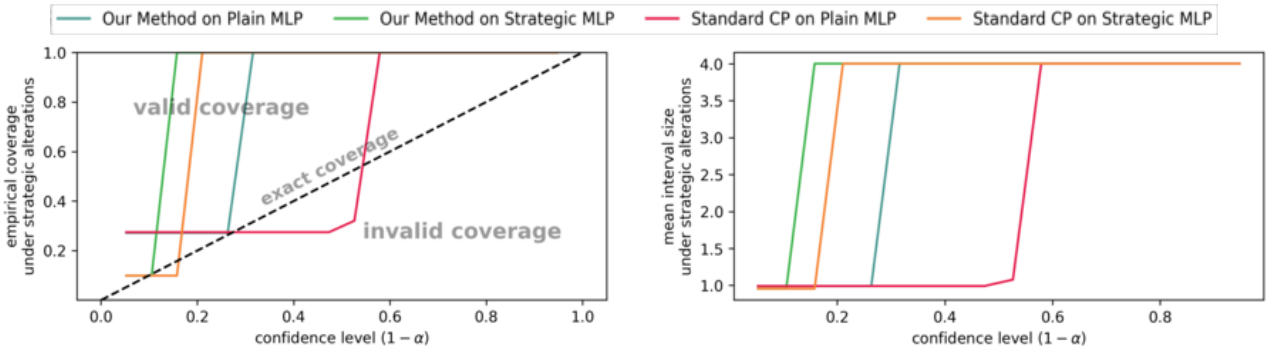
Figure 18: mlp on `taiwan`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

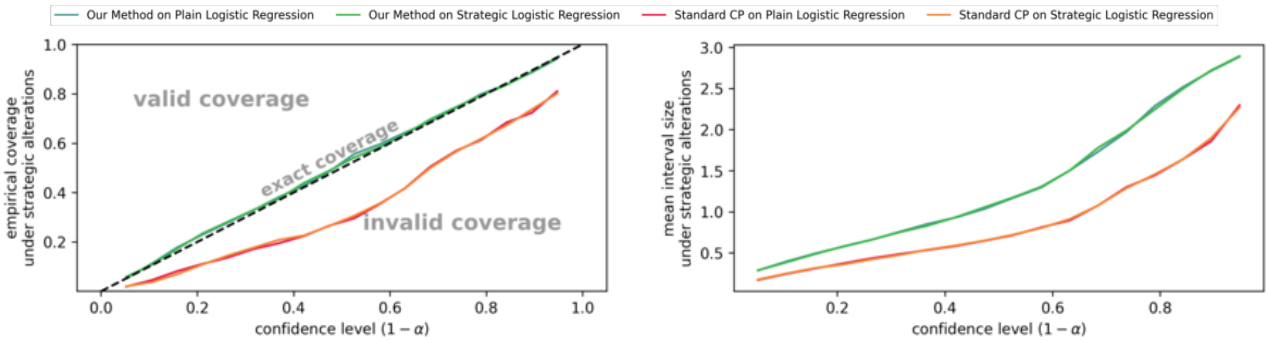
Figure 19: mlp on `bank-marketing`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

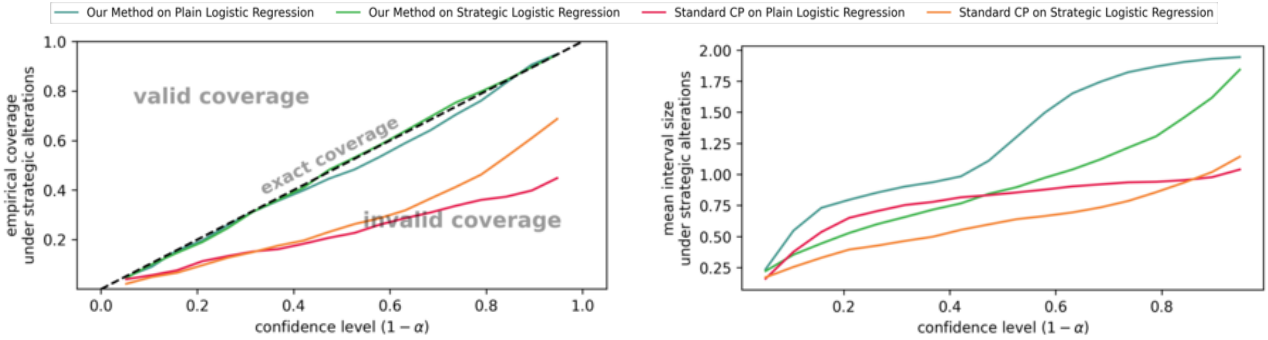
Figure 20: mlp on `census-income`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

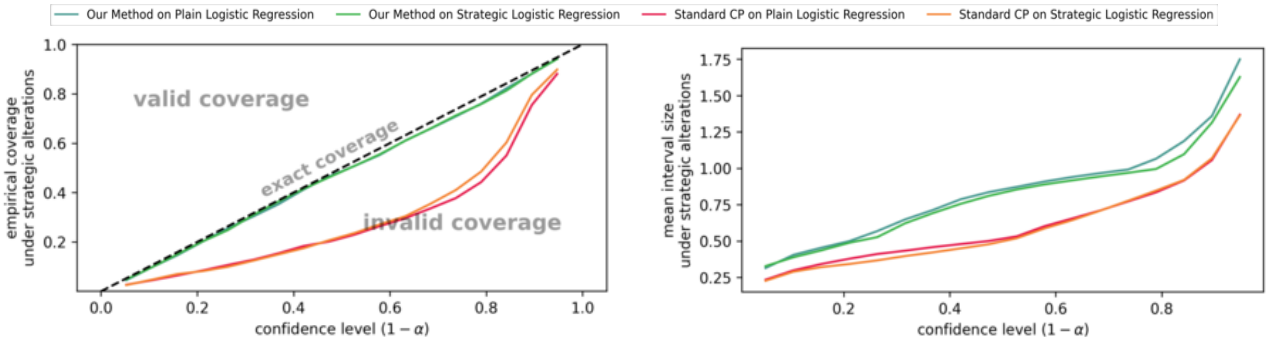
Figure 21: linear on `academic-dropout`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

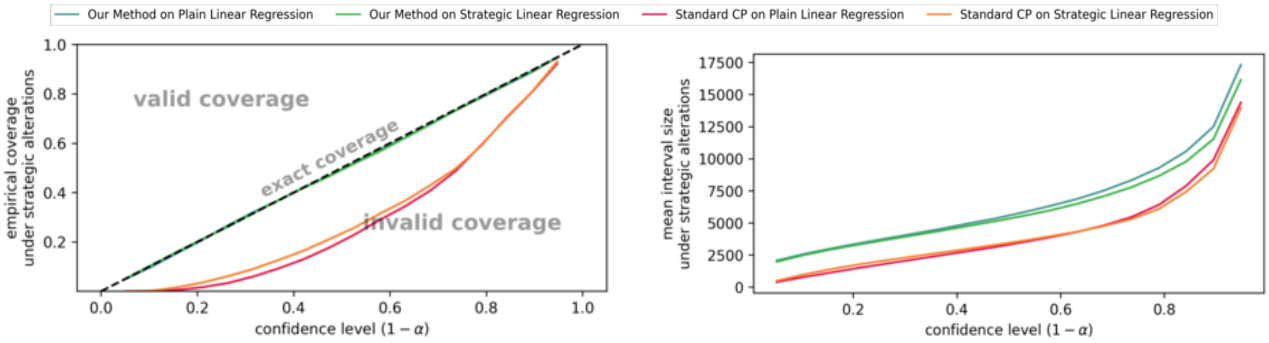
Figure 22: linear on `spambase`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

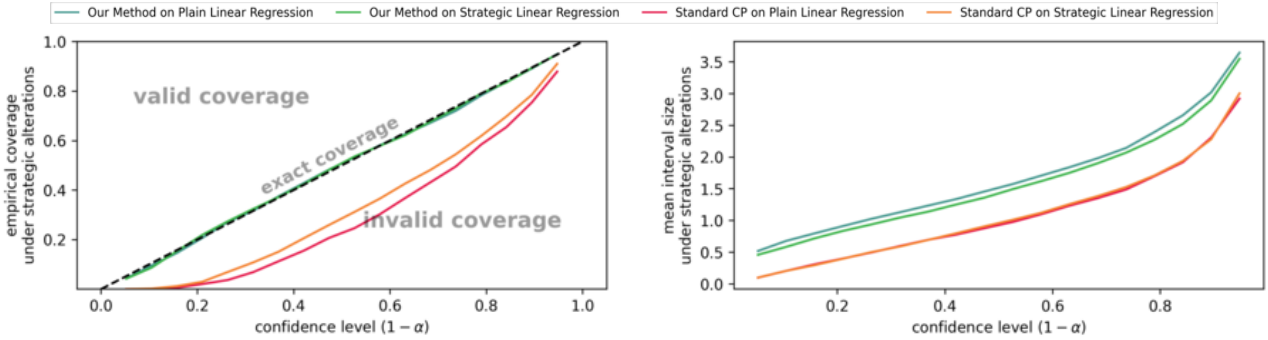
Figure 23: linear on `shoppers`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

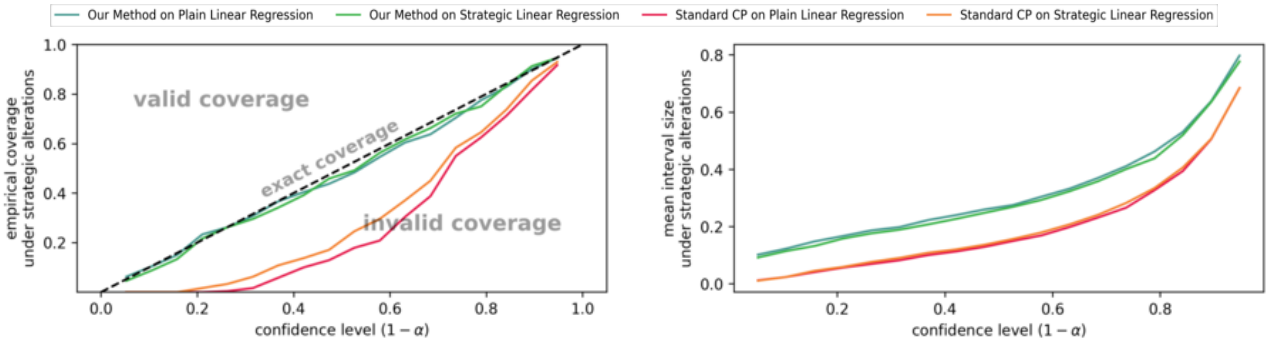
Figure 24: linear on `news`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

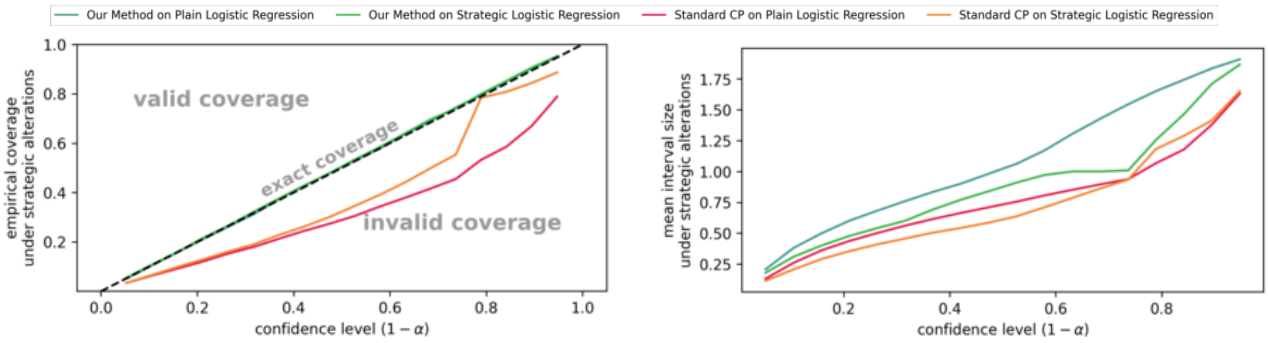
Figure 25: linear on `wine`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

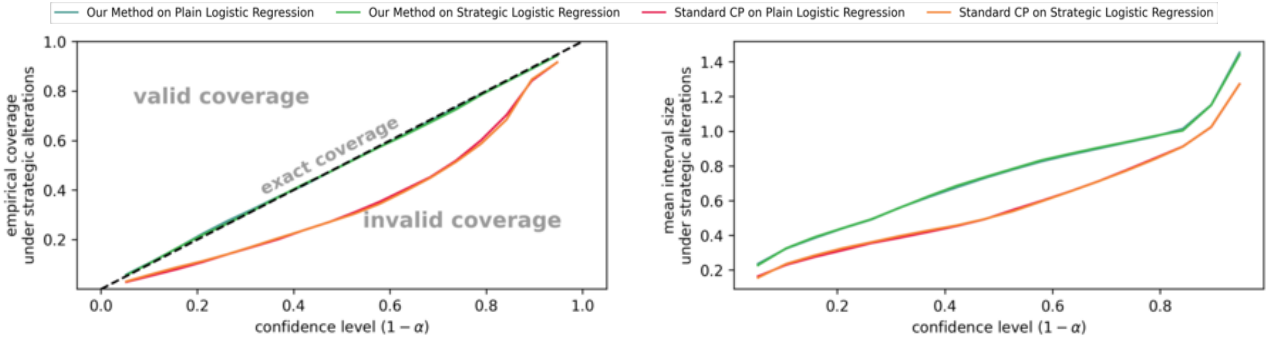
Figure 26: linear on `productivity`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

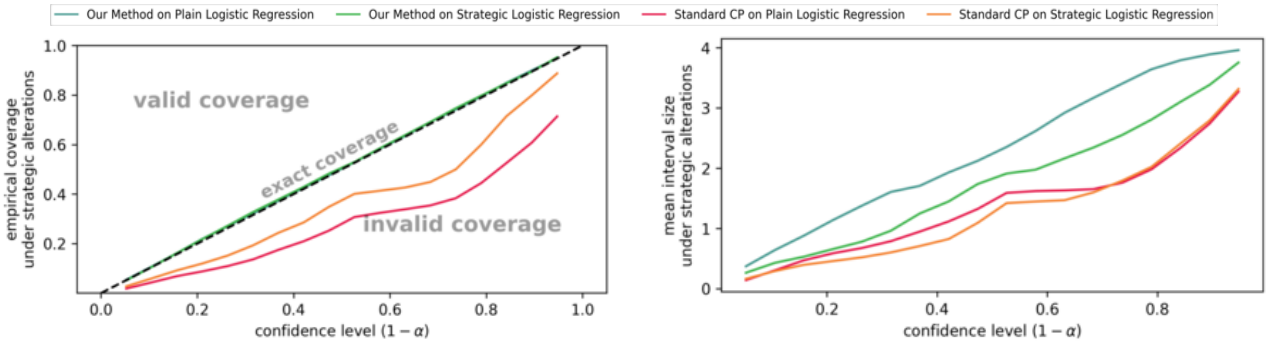
Figure 27: linear on `taiwan`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

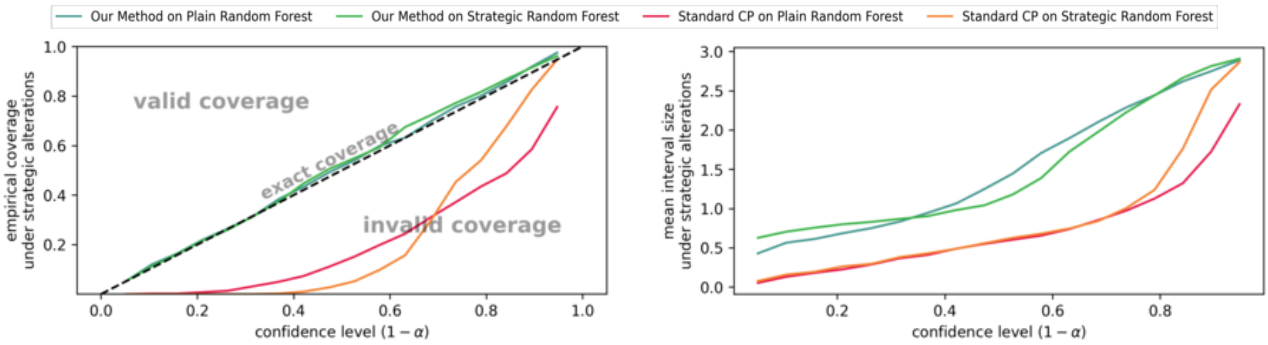
Figure 28: linear on `bank-marketing`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

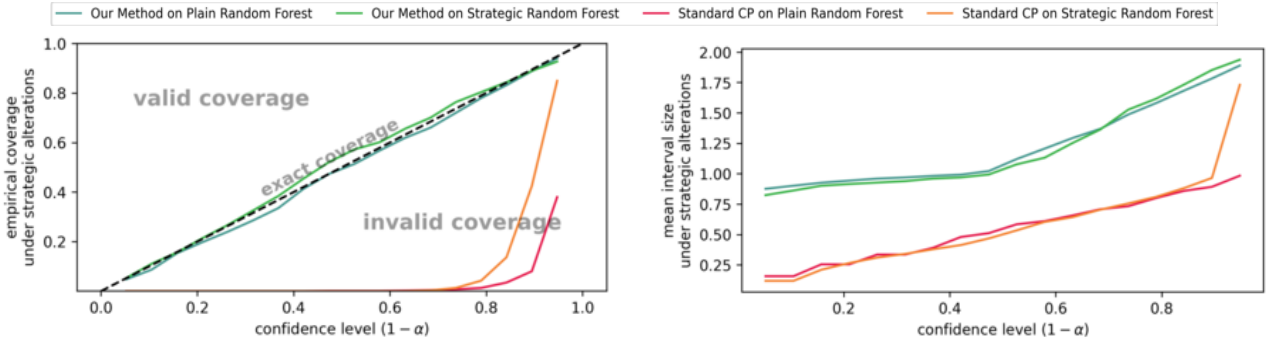
Figure 29: linear on `census-income`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

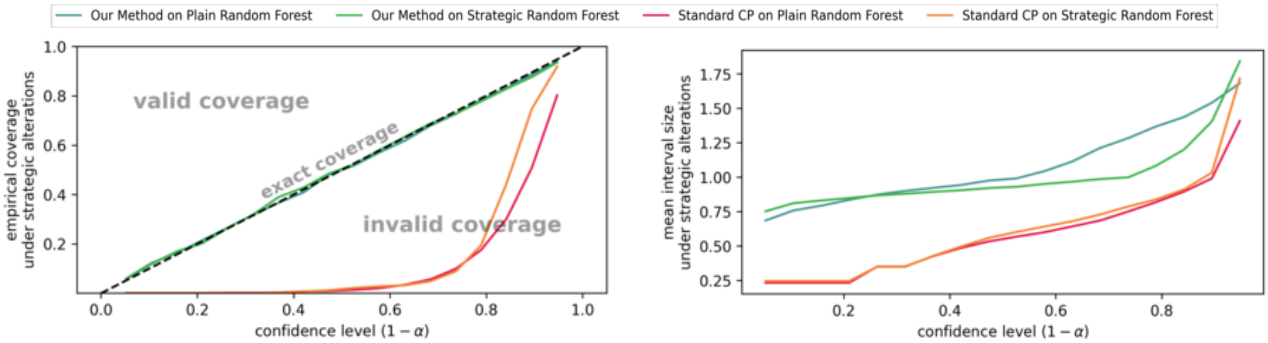
Figure 30: random-forest on `academic-dropout`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

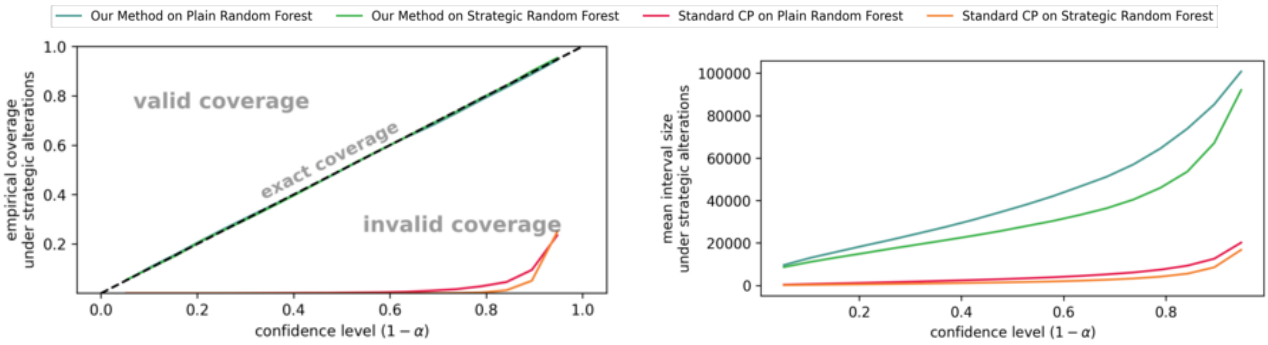
Figure 31: random-forest on `spambase`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

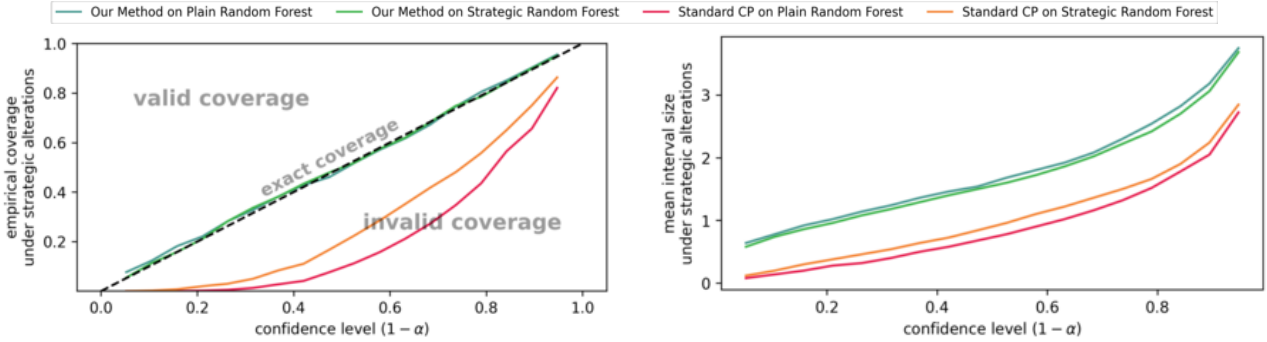
Figure 32: random-forest on `shoppers`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

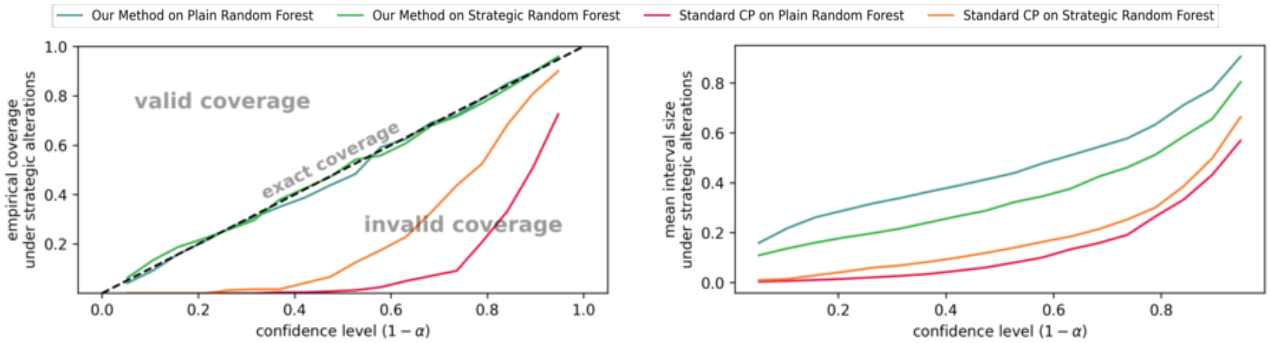
Figure 33: random-forest on `news`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

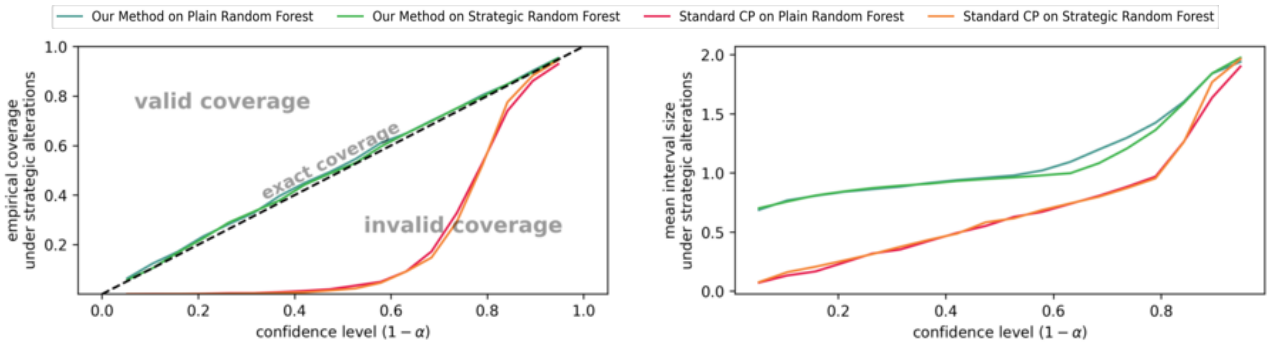
Figure 34: random-forest on `wine`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

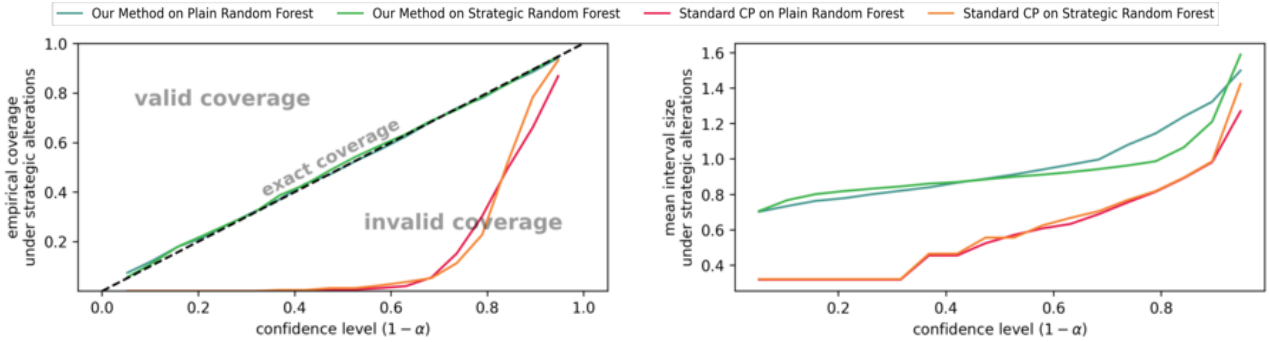
Figure 35: random-forest on `productivity`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

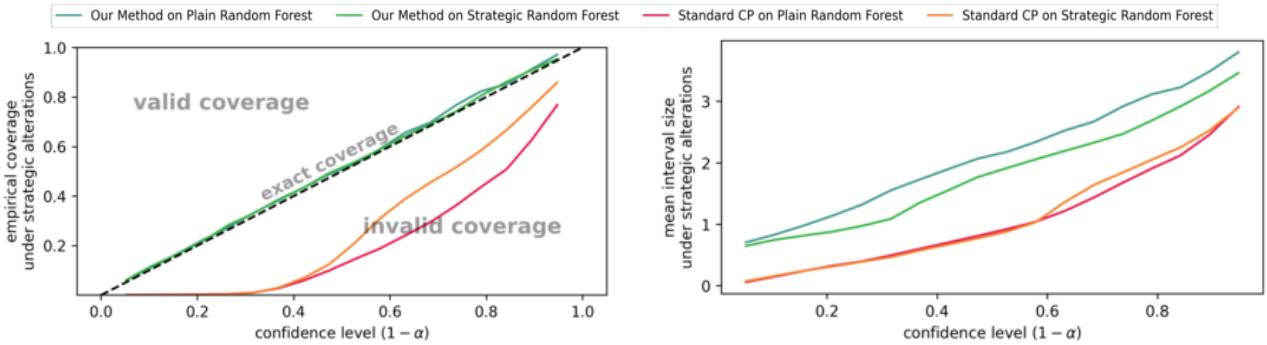
Figure 36: random-forest on `taiwan`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

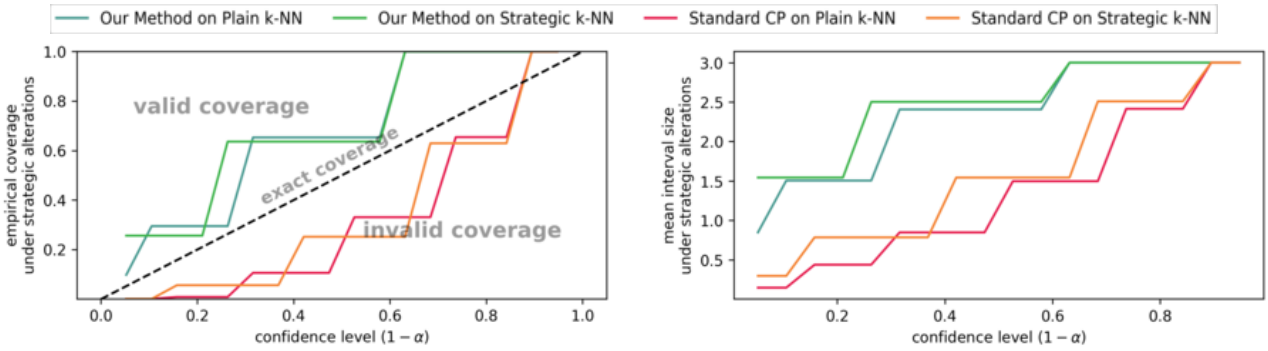
Figure 37: random-forest on `bank-marketing`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

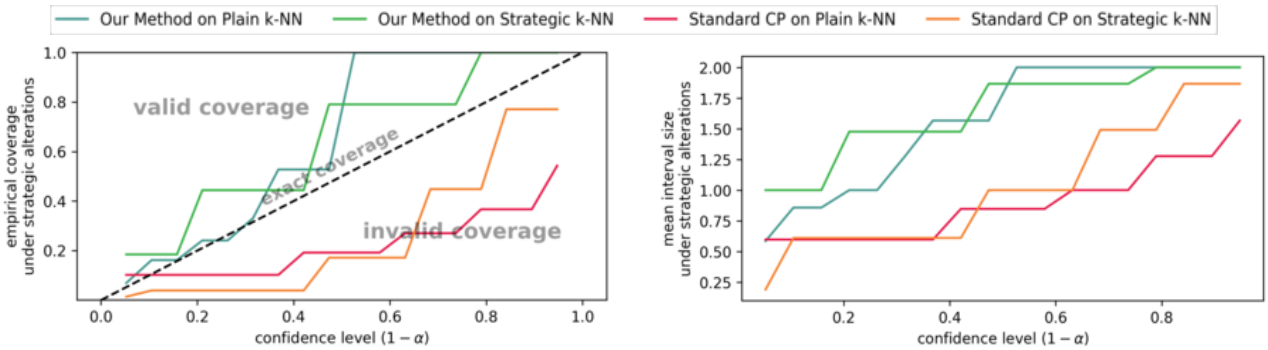
Figure 38: random-forest on `census-income`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

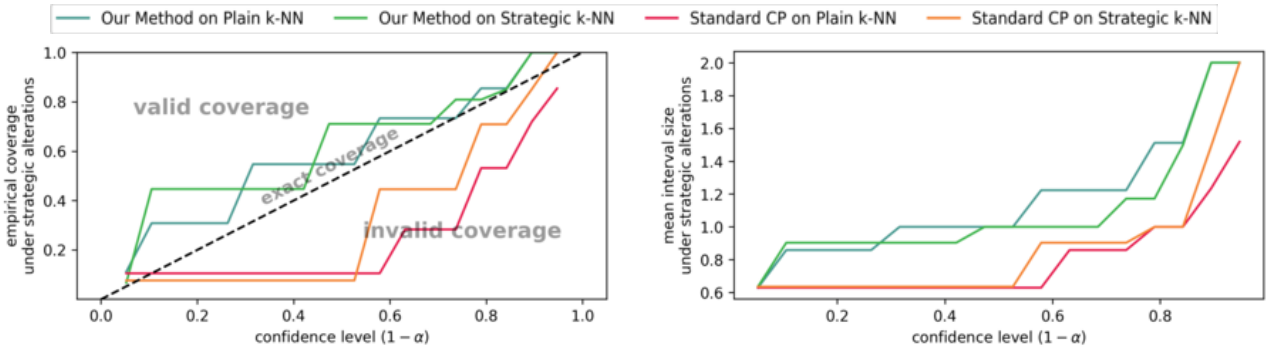
Figure 39: knn on `academic-dropout`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

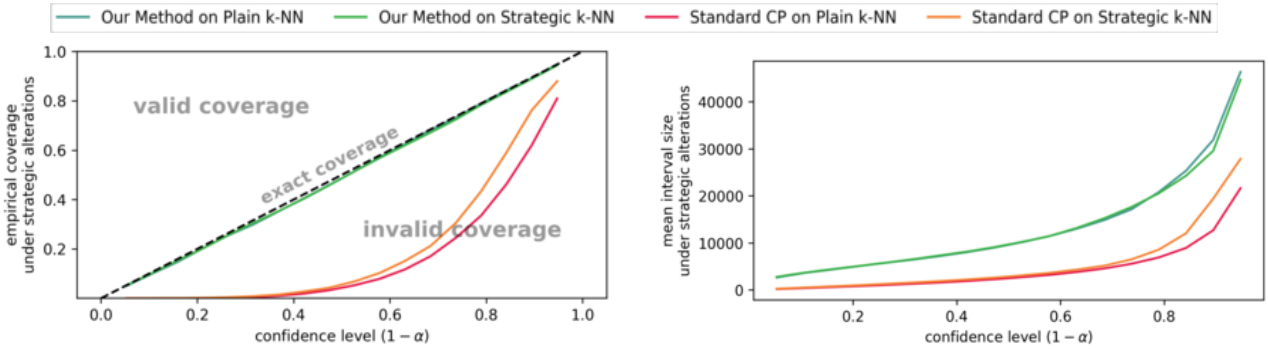
Figure 40: knn on `spambase`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

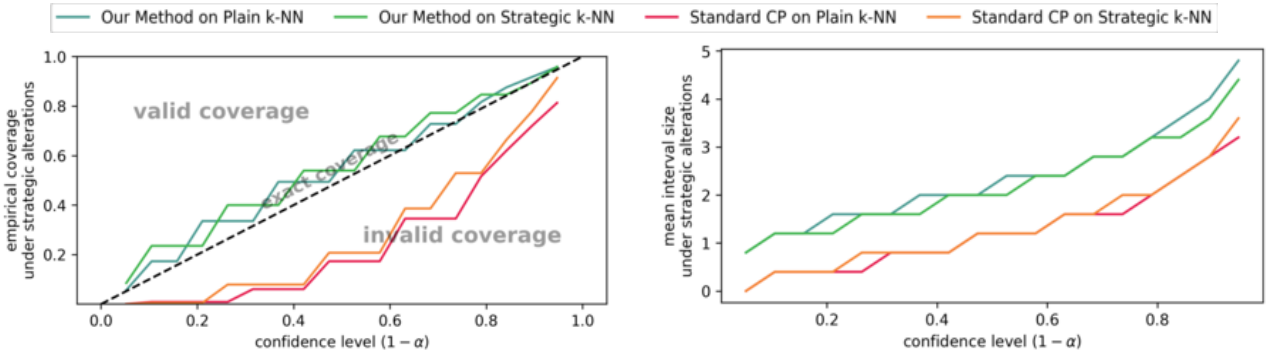
Figure 41: knn on `shoppers`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

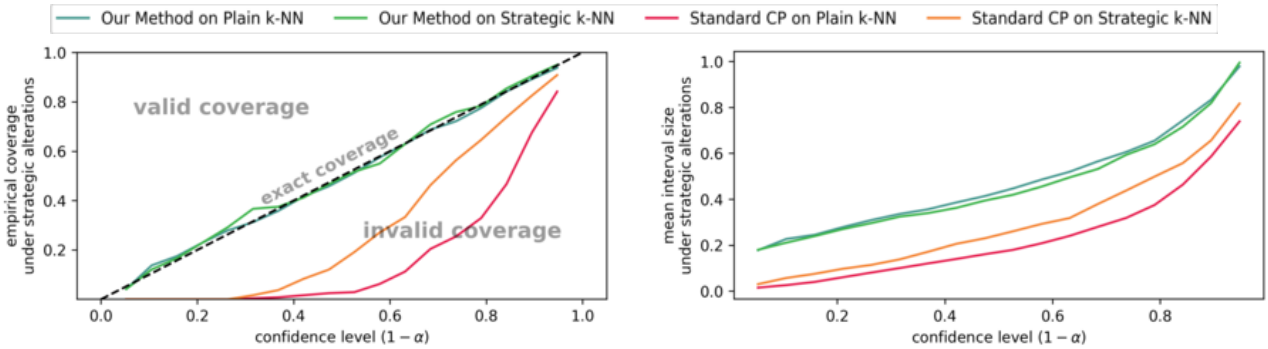
Figure 42: knn on `news`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

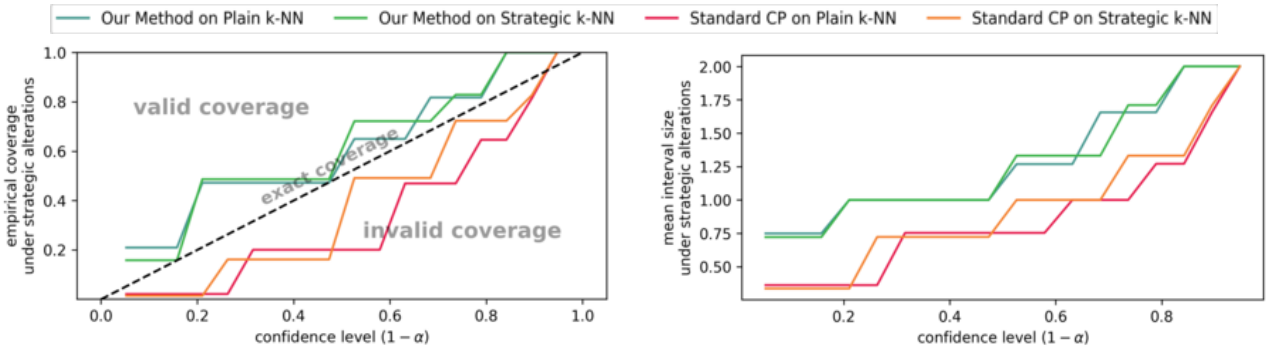
Figure 43: knn on `wine`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

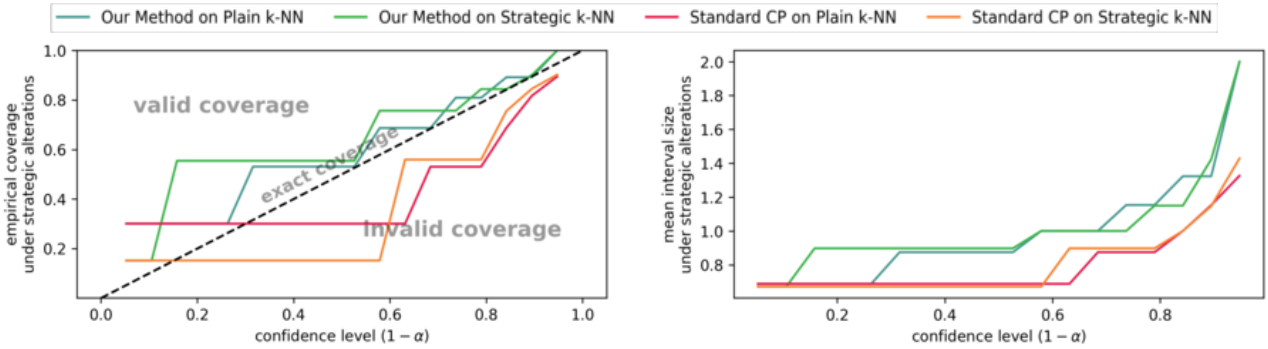
Figure 44: knn on `productivity`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

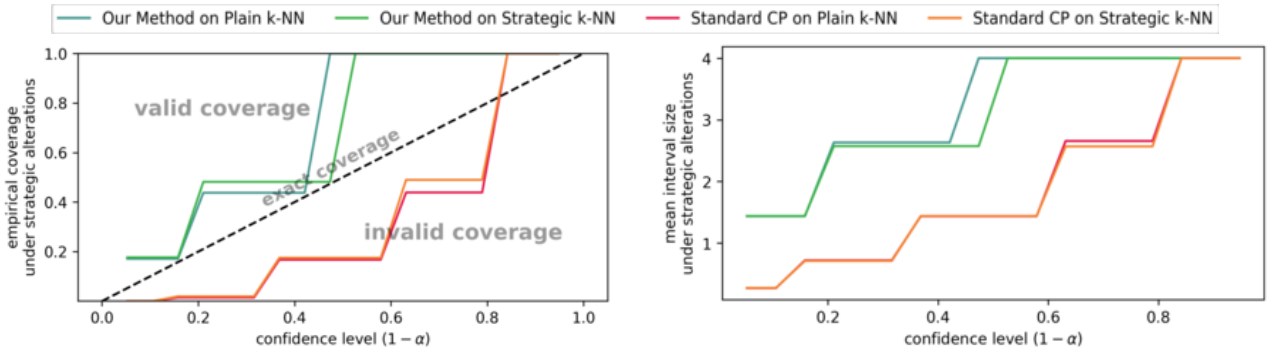
Figure 45: knn on `taiwan`



(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

Figure 46: knn on `bank-marketing`

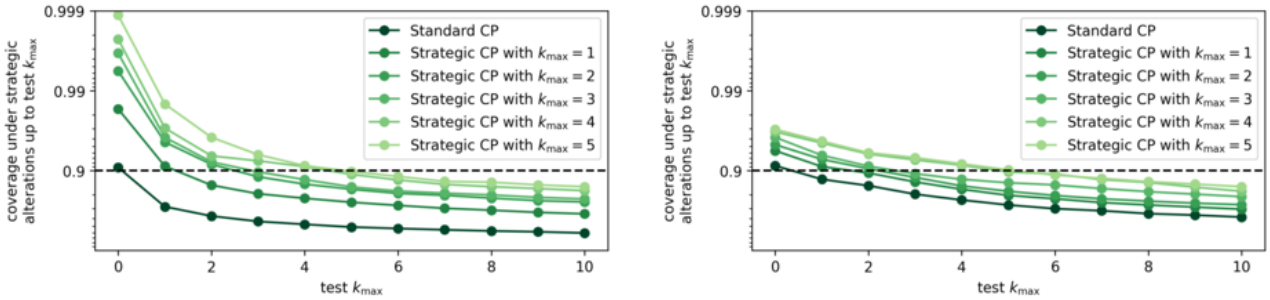


(a) Coverage of standard CP & our method under strategic alterations over varying values of α.

(b) Average sizes of the predictive sets of standard CP & our method under strategic alterations.

Figure 47: knn on `census-income`

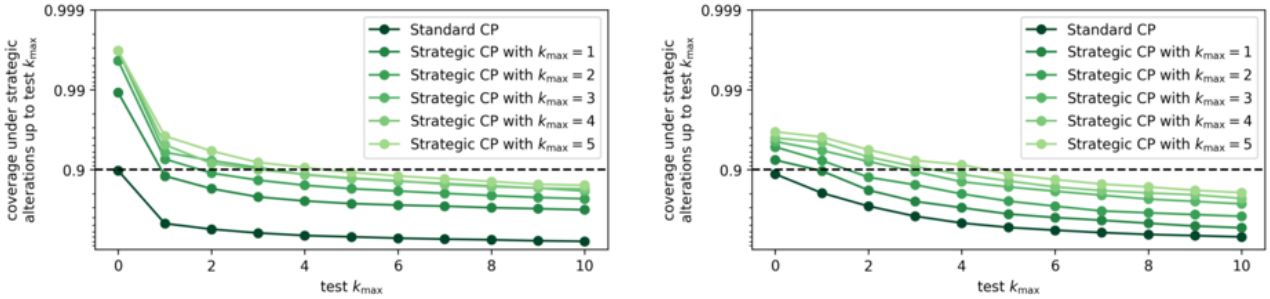
D.2 Figure 2: coverage of our method andstandard CP for varying levels of strategic alterations



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

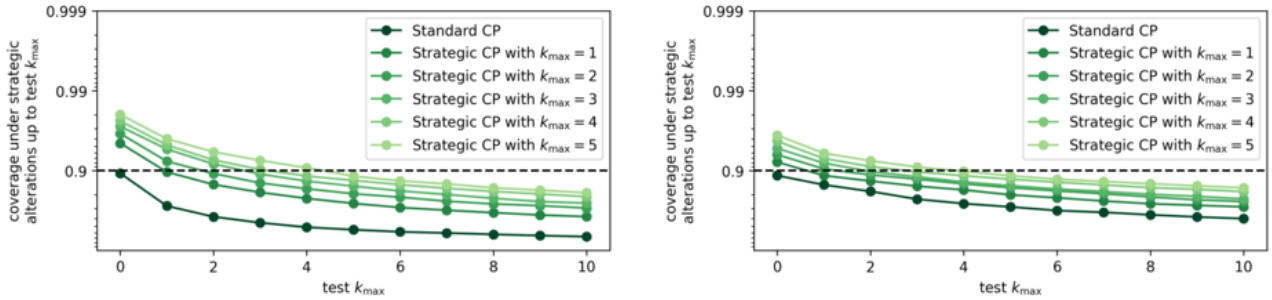
Figure 48: xgboost on `academic-dropout`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

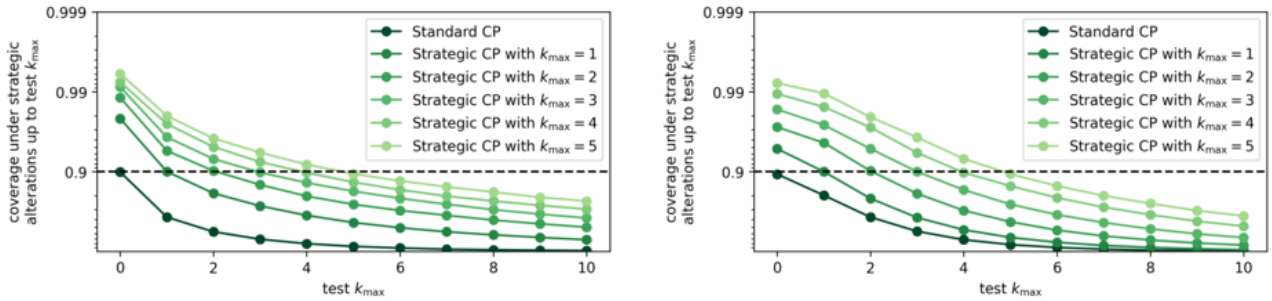
Figure 49: xgboost on `spambase`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

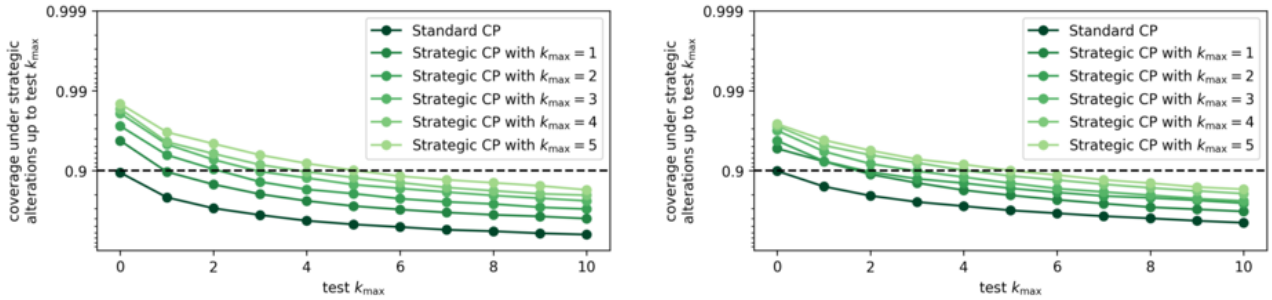
Figure 50: xgboost on `shoppers`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

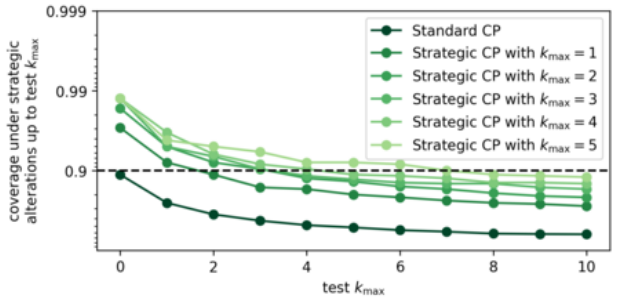
Figure 51: xgboost on `news`



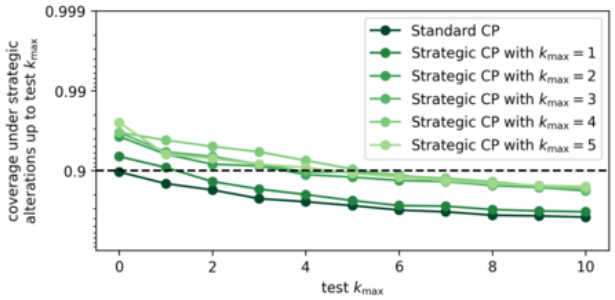
(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 52: xgboost on `wine`

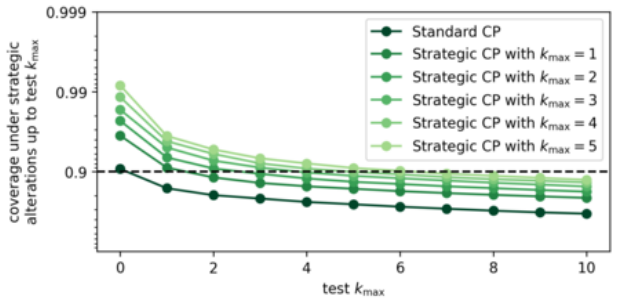


(a) Strategic Conformal Prediction on top of a Plain Model

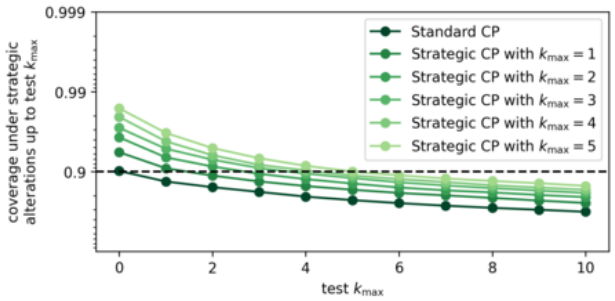


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 53: xgboost on `productivity`

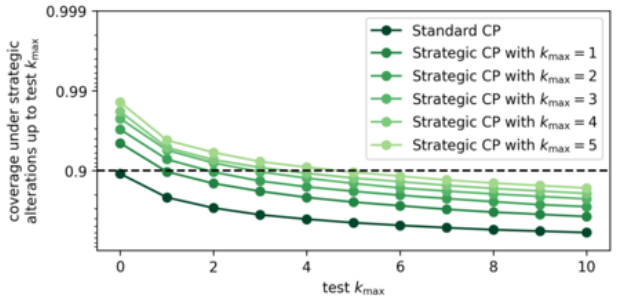


(a) Strategic Conformal Prediction on top of a Plain Model

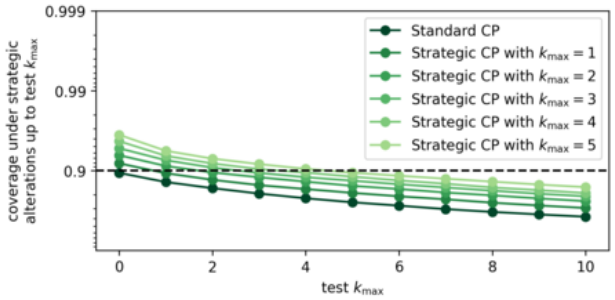


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 54: xgboost on `taiwan`

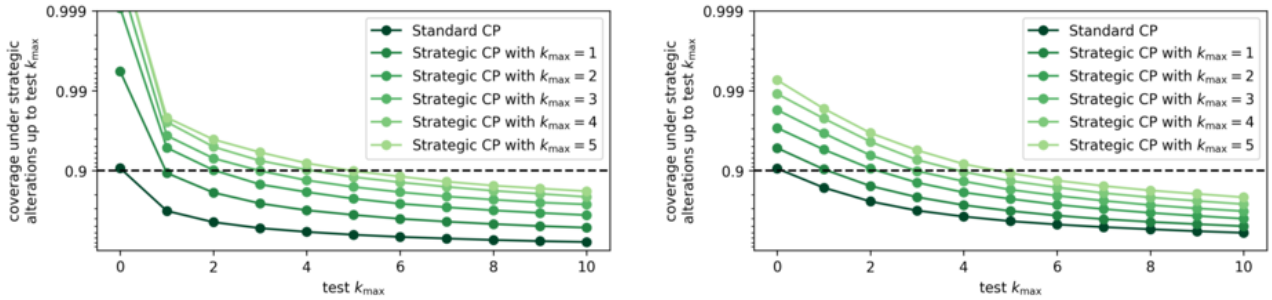


(a) Strategic Conformal Prediction on top of a Plain Model



(b) Strategic Conformal Prediction on top of a Strategic Model

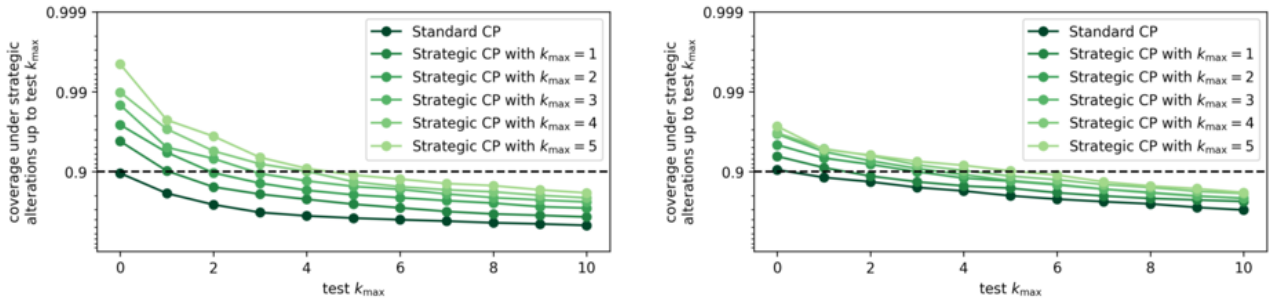
Figure 55: xgboost on `bank-marketing`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

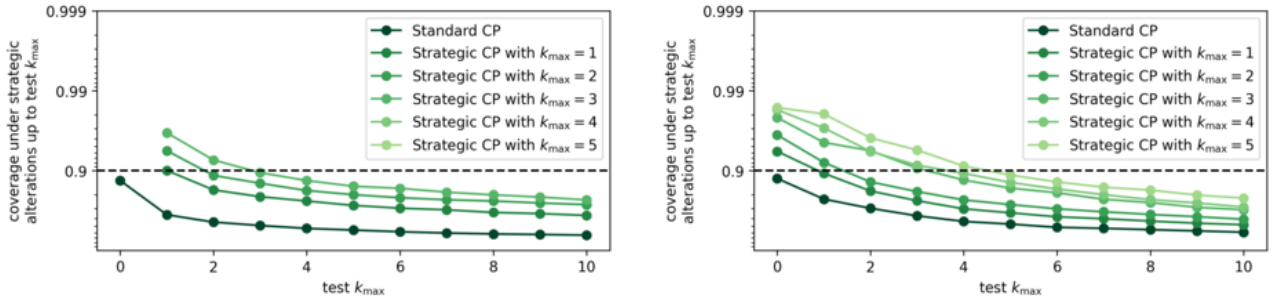
Figure 56: xgboost on `census-income`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

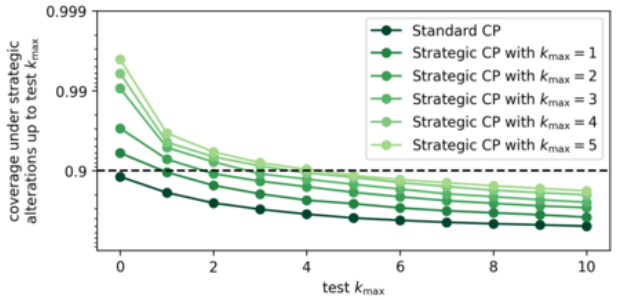
Figure 57: mlp on `academic-dropout`



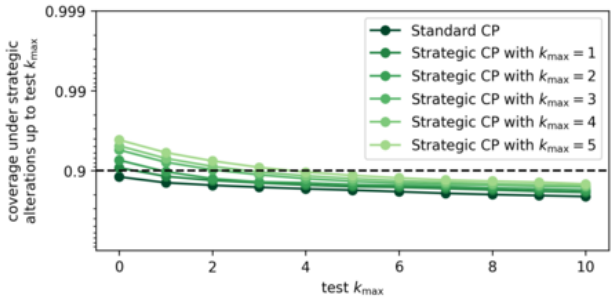
(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 58: mlp on `spambase`

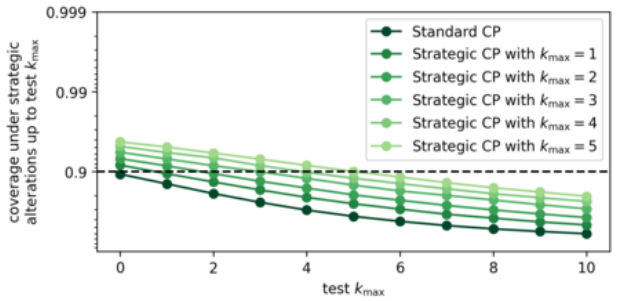


(a) Strategic Conformal Prediction on top of a Plain Model

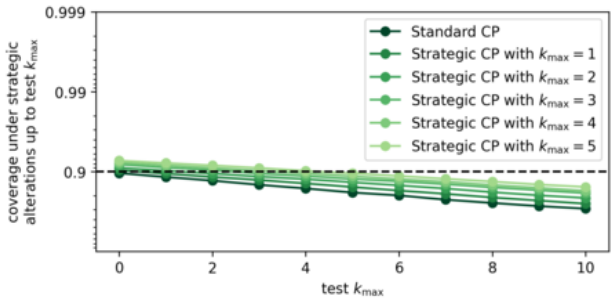


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 59: mlp on `shoppers`

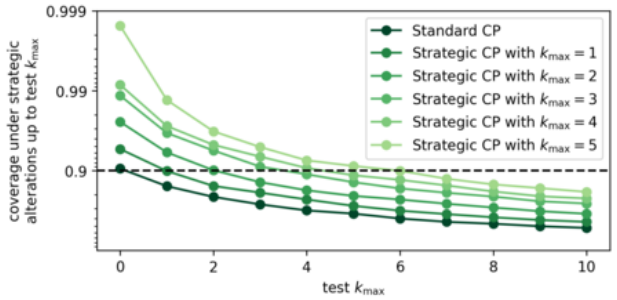


(a) Strategic Conformal Prediction on top of a Plain Model

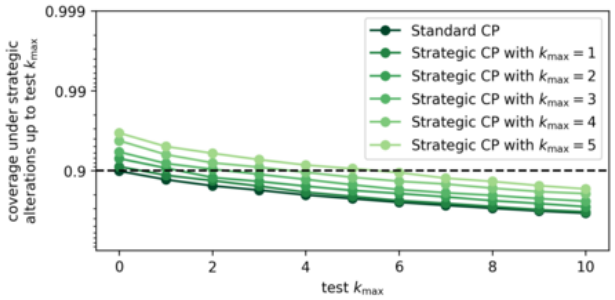


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 60: mlp on `news`

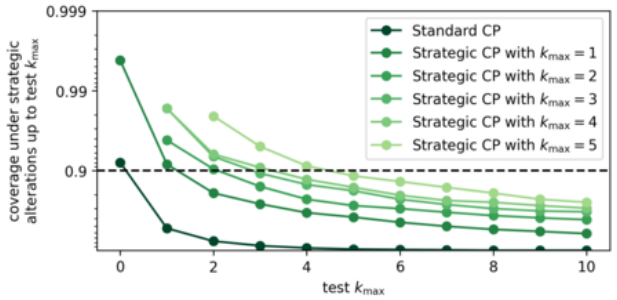


(a) Strategic Conformal Prediction on top of a Plain Model

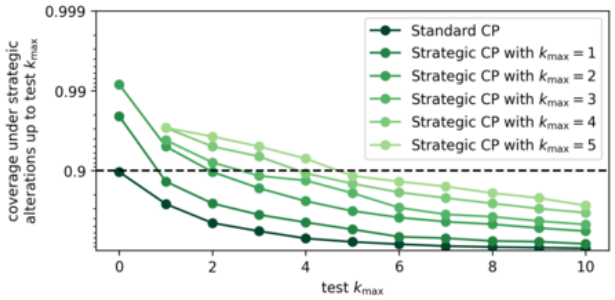


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 61: mlp on `wine`

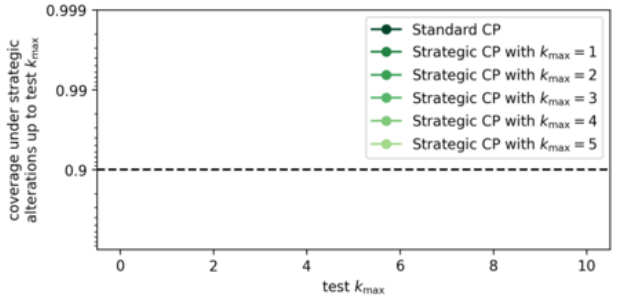


(a) Strategic Conformal Prediction on top of a Plain Model

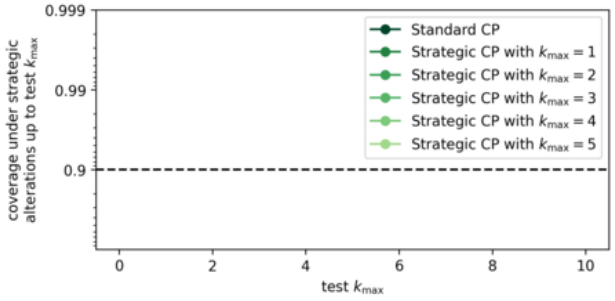


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 62: mlp on `productivity`

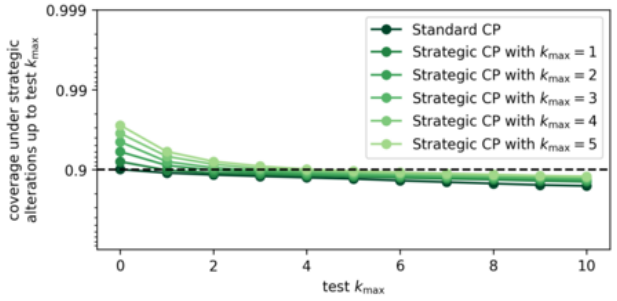


(a) Strategic Conformal Prediction on top of a Plain Model

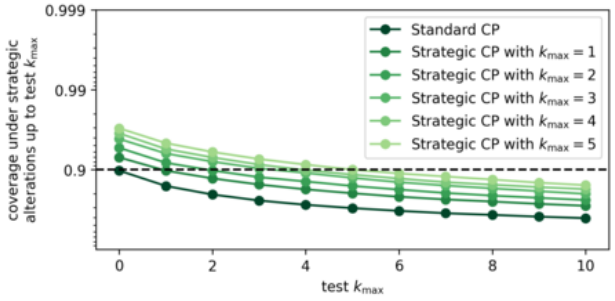


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 63: mlp on `taiwan`

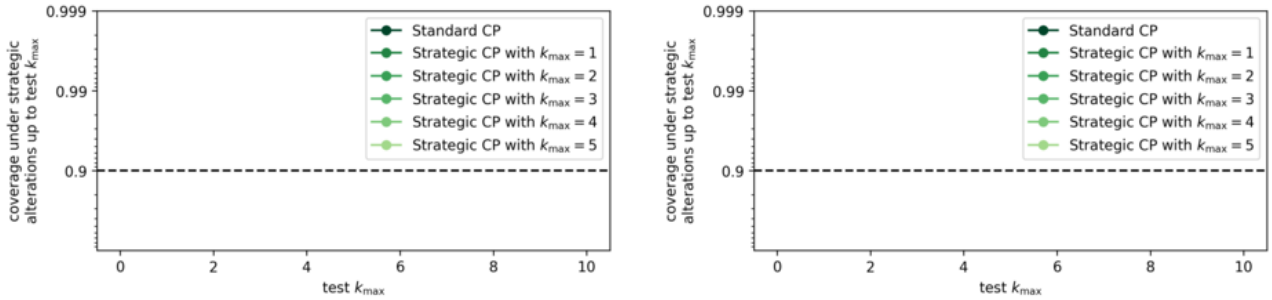


(a) Strategic Conformal Prediction on top of a Plain Model



(b) Strategic Conformal Prediction on top of a Strategic Model

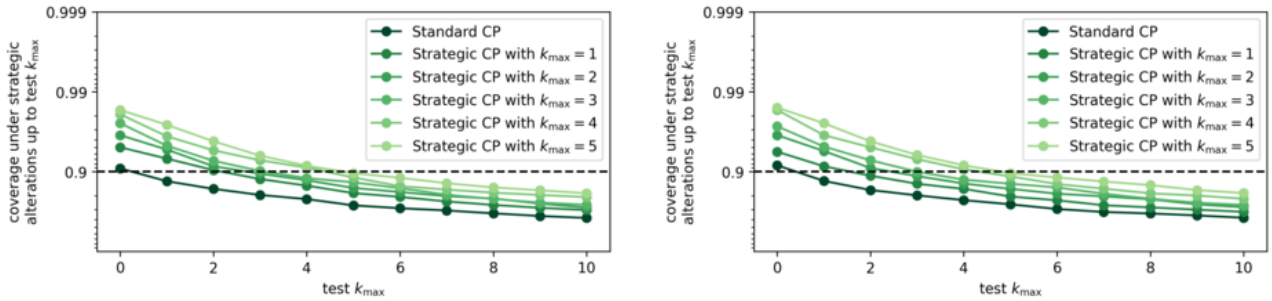
Figure 64: mlp on `bank-marketing`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

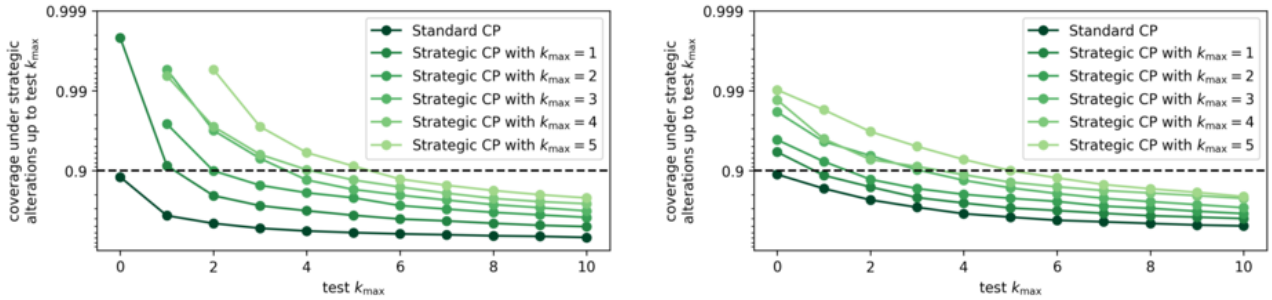
Figure 65: mlp on `census-income`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

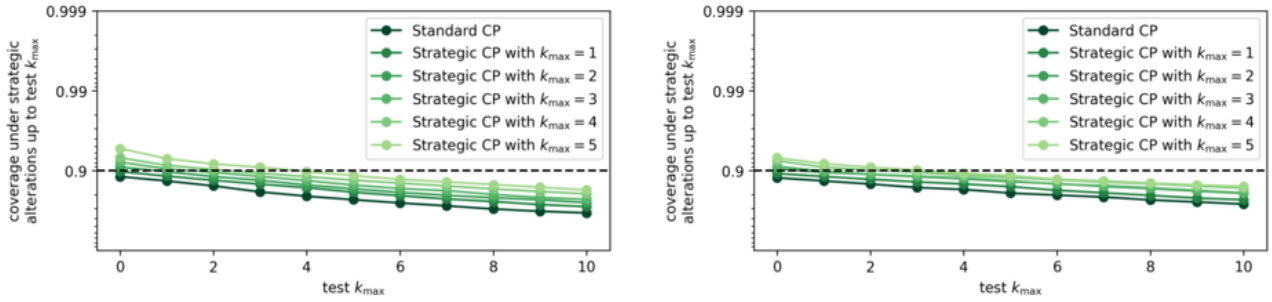
Figure 66: linear on `academic-dropout`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

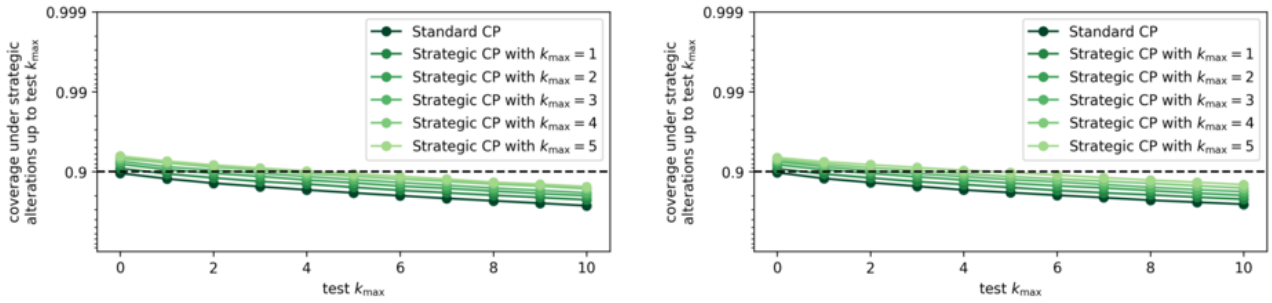
Figure 67: linear on `spambase`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

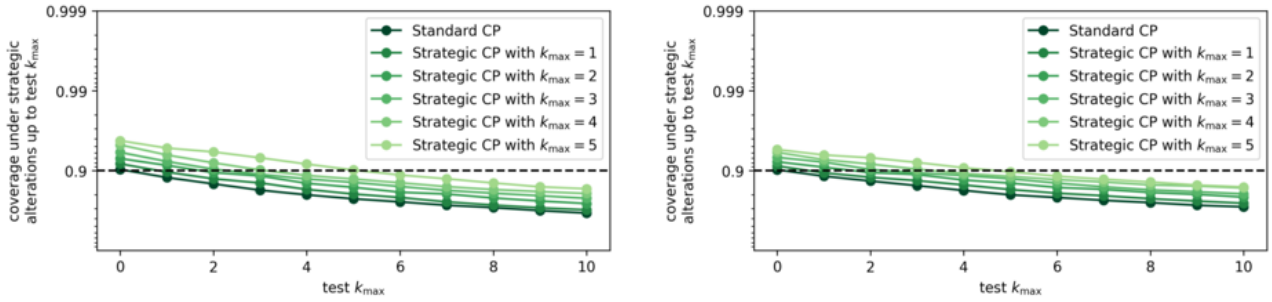
Figure 68: linear on `shoppers`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

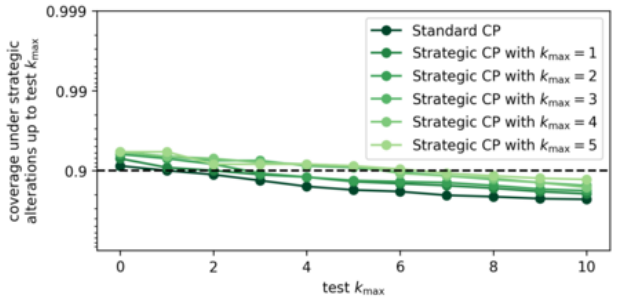
Figure 69: linear on `news`



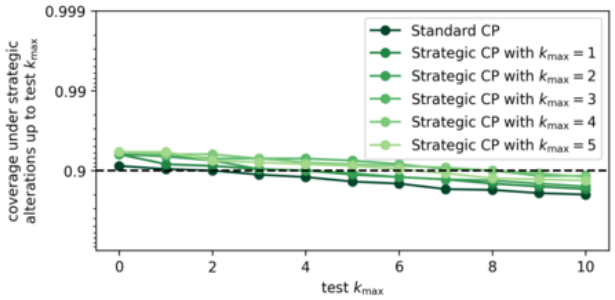
(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 70: linear on `wine`

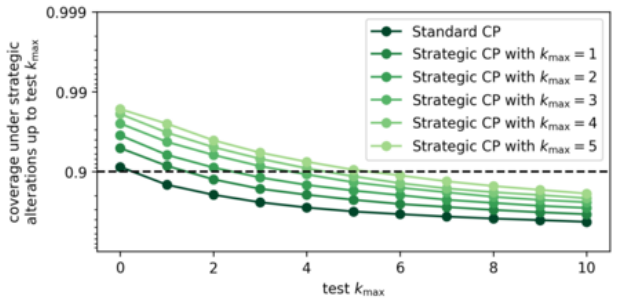


(a) Strategic Conformal Prediction on top of a Plain Model

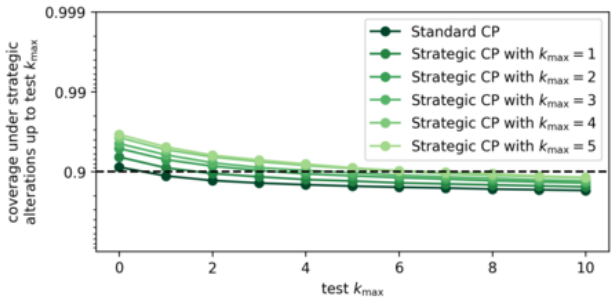


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 71: linear on `productivity`

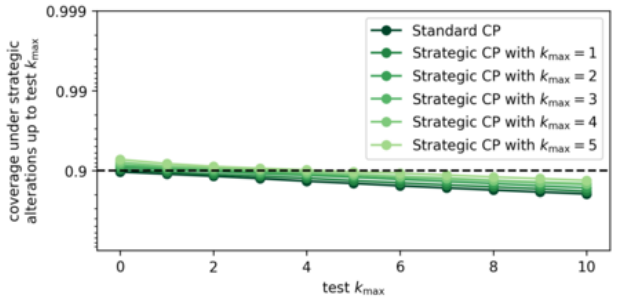


(a) Strategic Conformal Prediction on top of a Plain Model

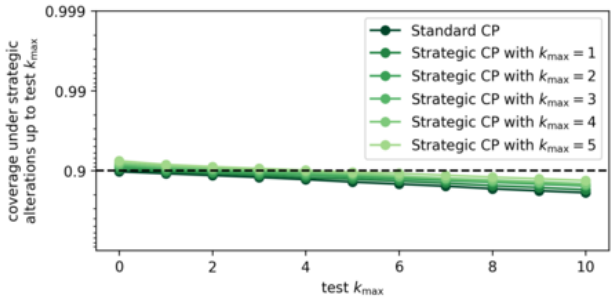


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 72: linear on `taiwan`

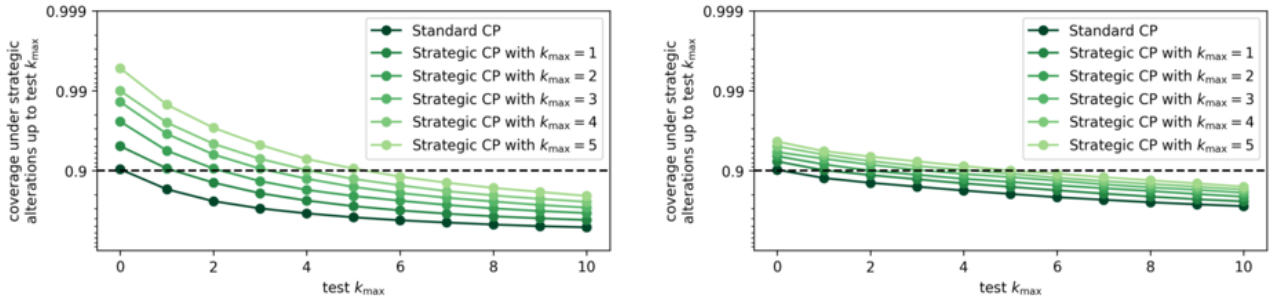


(a) Strategic Conformal Prediction on top of a Plain Model



(b) Strategic Conformal Prediction on top of a Strategic Model

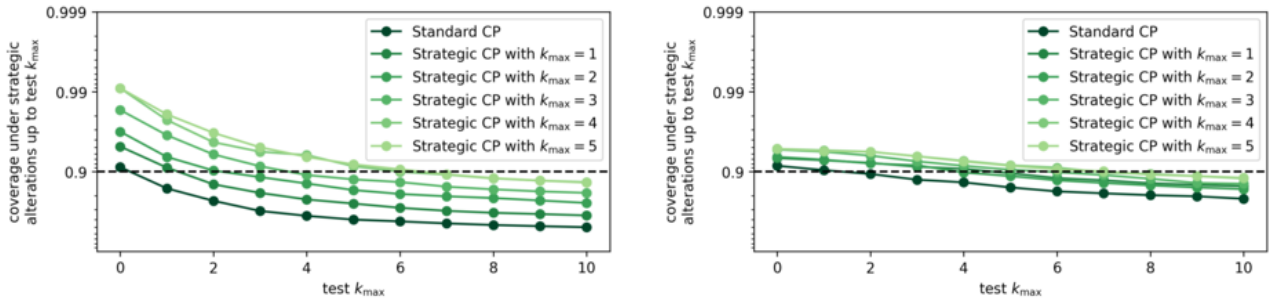
Figure 73: linear on `bank-marketing`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

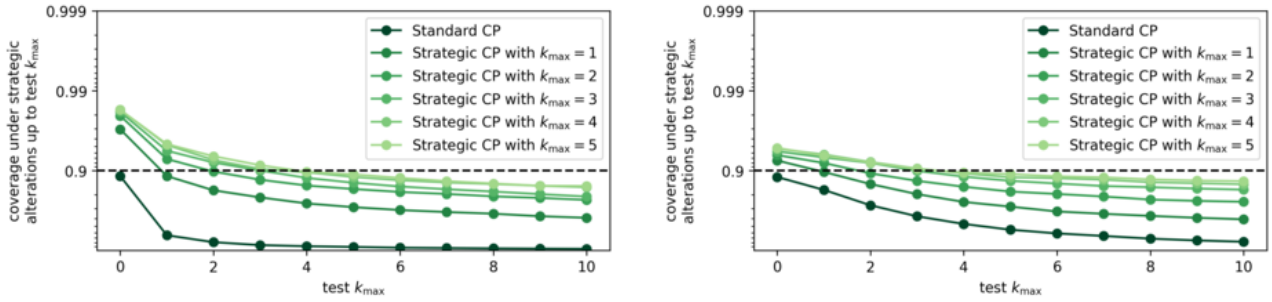
Figure 74: linear on `census-income`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

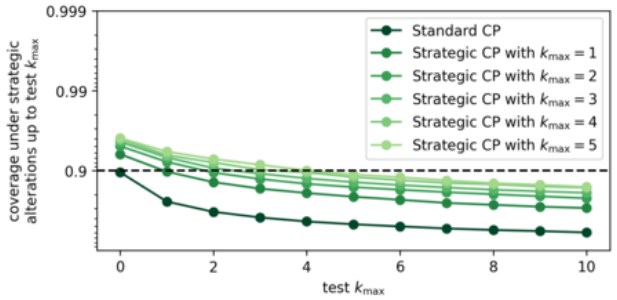
Figure 75: random-forest on `academic-dropout`



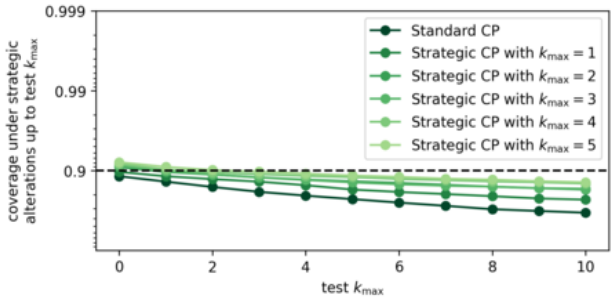
(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 76: random-forest on `spambase`

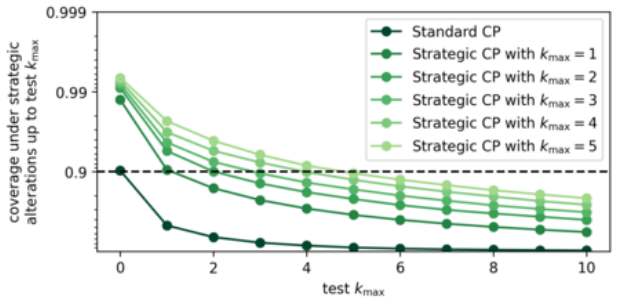


(a) Strategic Conformal Prediction on top of a Plain Model

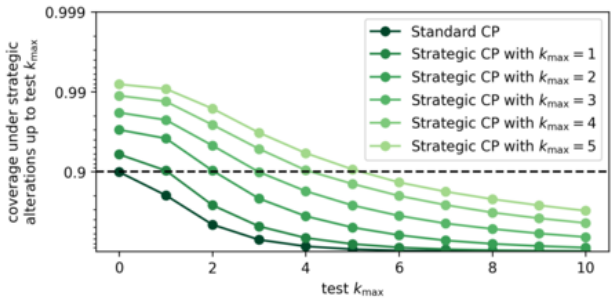


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 77: random-forest on `shoppers`

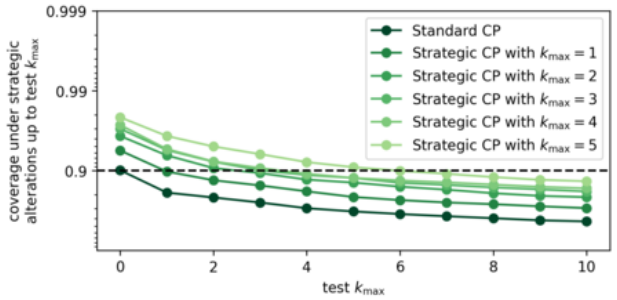


(a) Strategic Conformal Prediction on top of a Plain Model

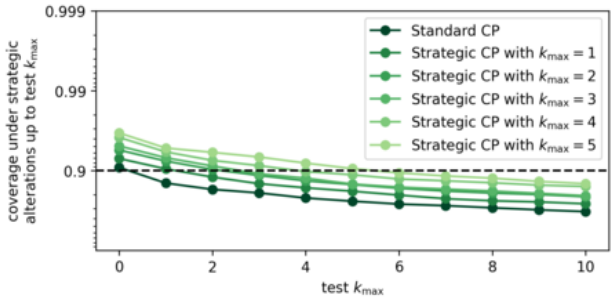


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 78: random-forest on `news`

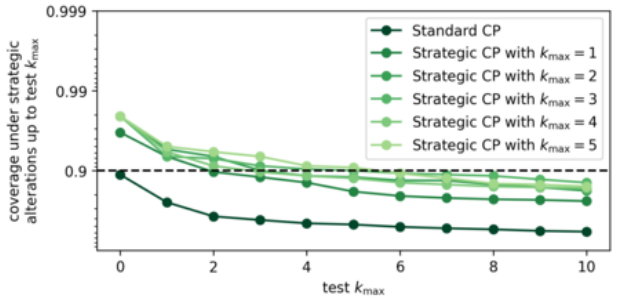


(a) Strategic Conformal Prediction on top of a Plain Model

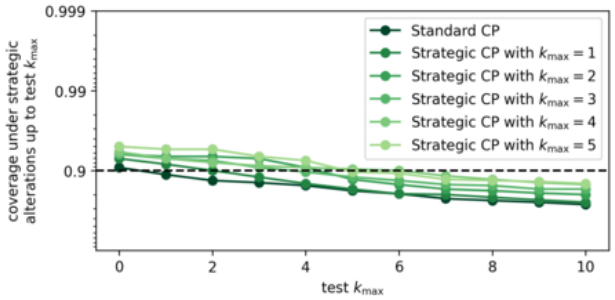


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 79: random-forest on `wine`

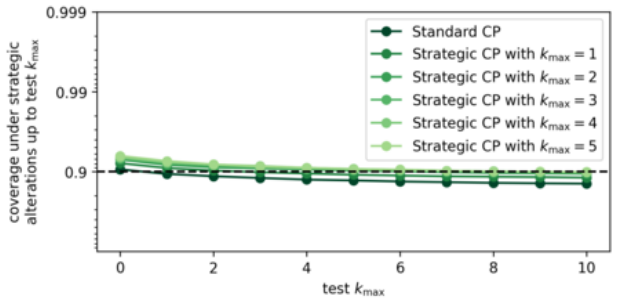


(a) Strategic Conformal Prediction on top of a Plain Model

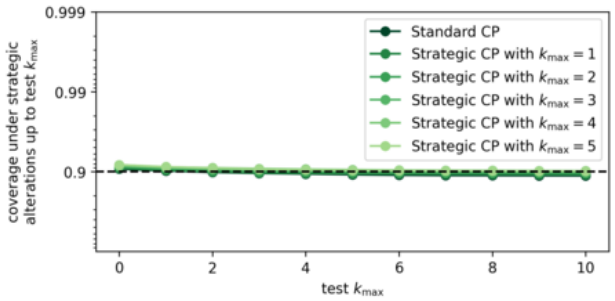


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 80: random-forest on `productivity`

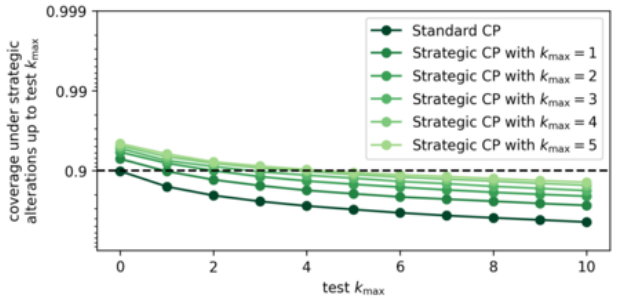


(a) Strategic Conformal Prediction on top of a Plain Model

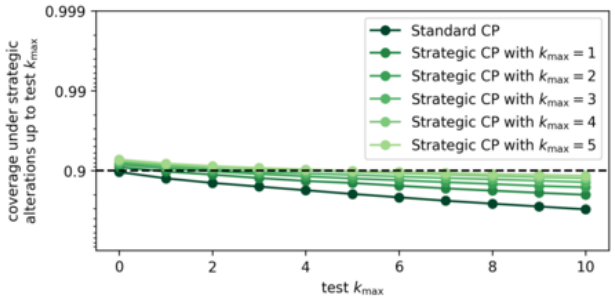


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 81: random-forest on `taiwan`

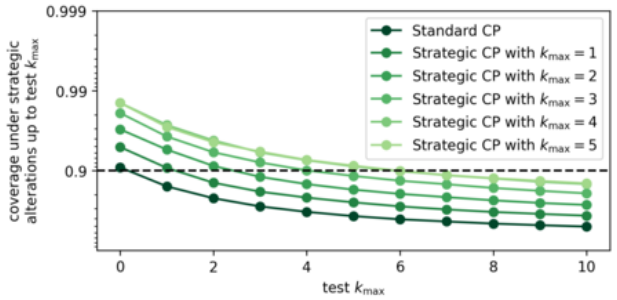


(a) Strategic Conformal Prediction on top of a Plain Model

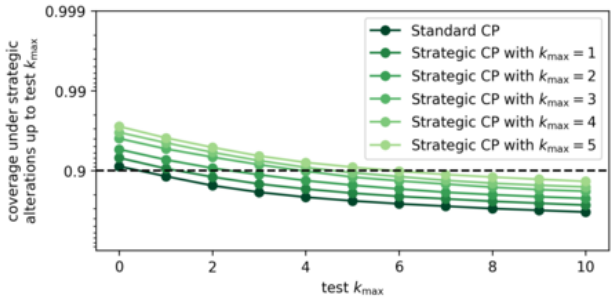


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 82: random-forest on `bank-marketing`

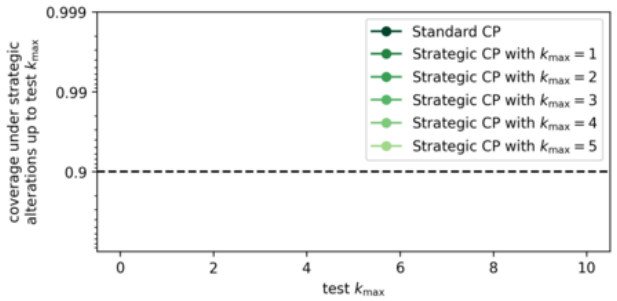


(a) Strategic Conformal Prediction on top of a Plain Model

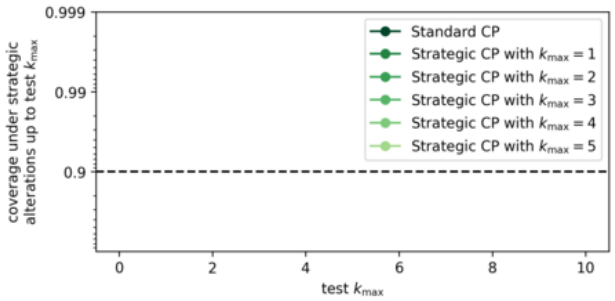


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 83: random-forest on `census-income`

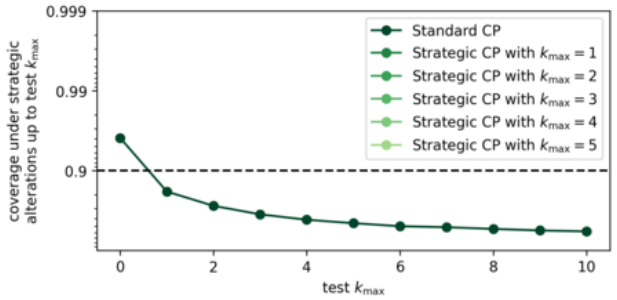


(a) Strategic Conformal Prediction on top of a Plain Model

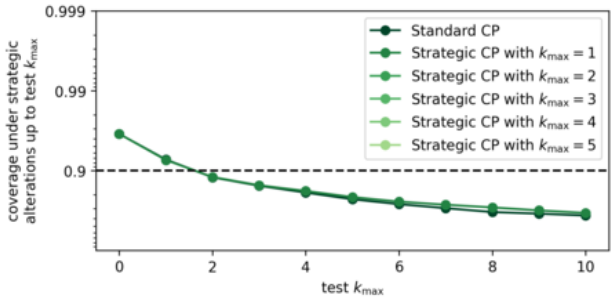


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 84: knn on `academic-dropout`

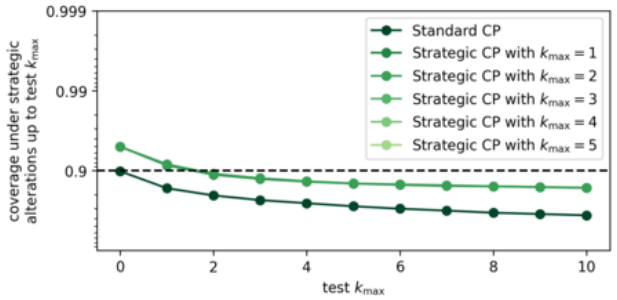


(a) Strategic Conformal Prediction on top of a Plain Model

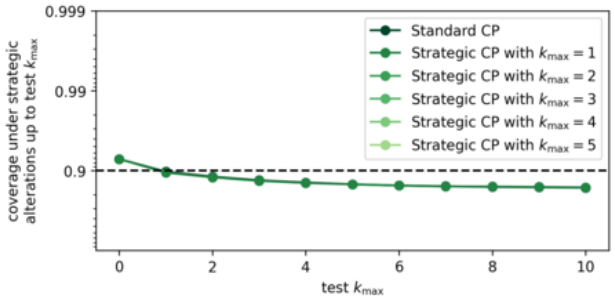


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 85: knn on `spambase`

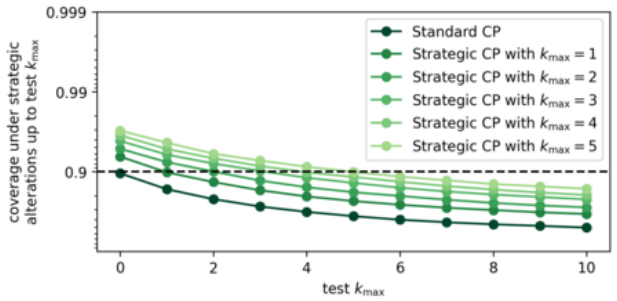


(a) Strategic Conformal Prediction on top of a Plain Model

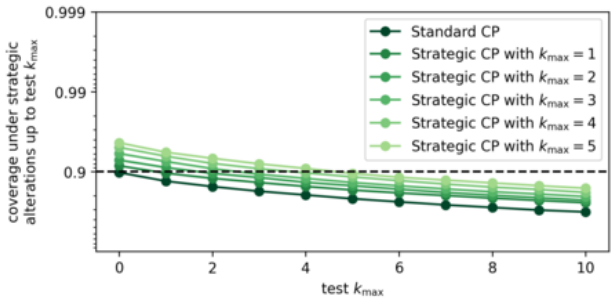


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 86: knn on `shoppers`

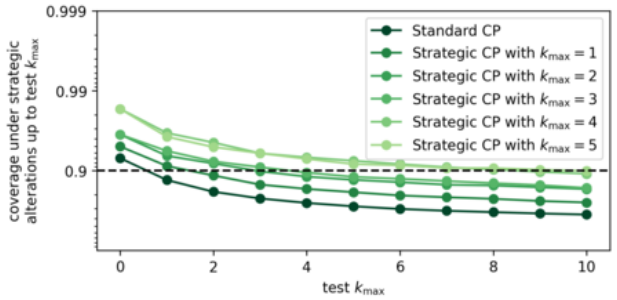


(a) Strategic Conformal Prediction on top of a Plain Model

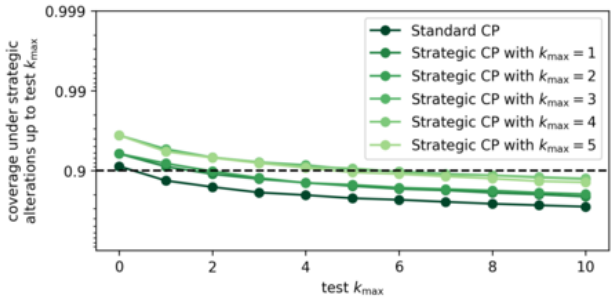


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 87: knn on `news`

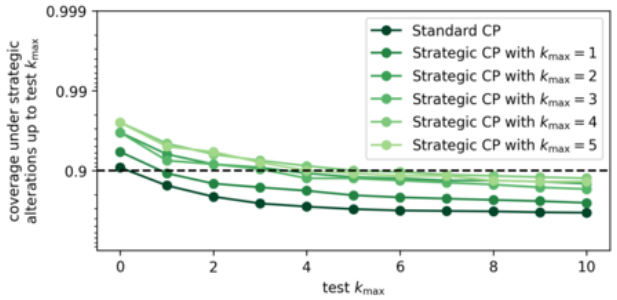


(a) Strategic Conformal Prediction on top of a Plain Model

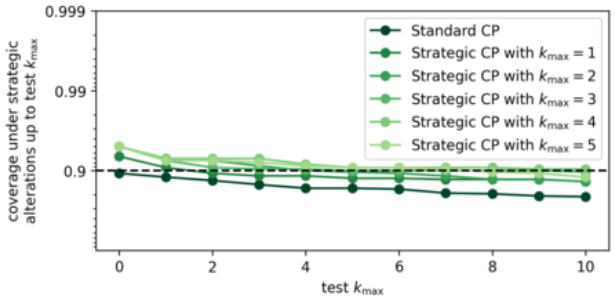


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 88: knn on `wine`

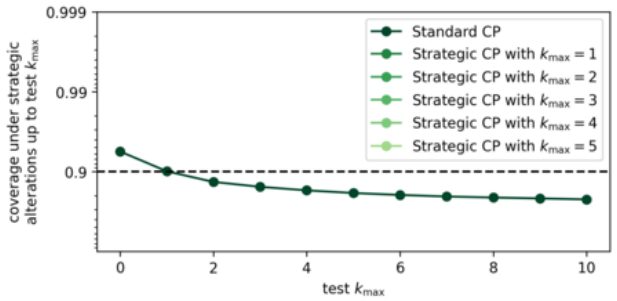


(a) Strategic Conformal Prediction on top of a Plain Model

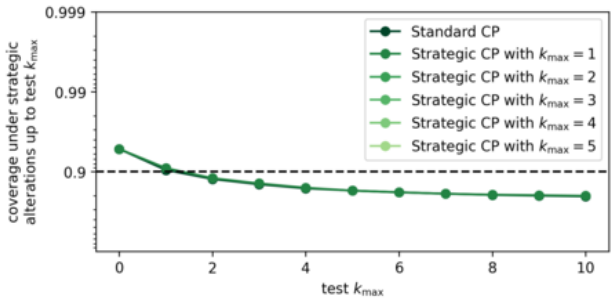


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 89: knn on `productivity`

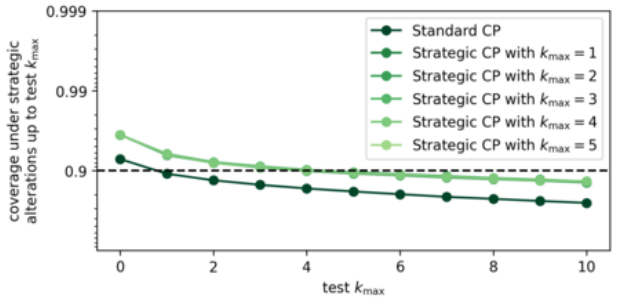


(a) Strategic Conformal Prediction on top of a Plain Model

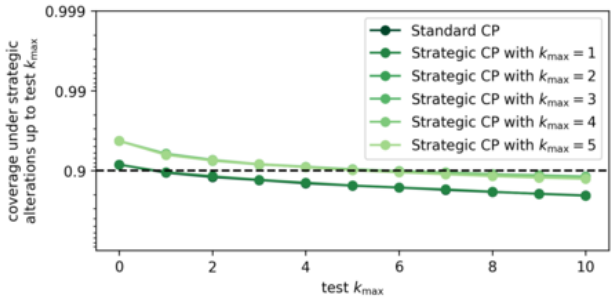


(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 90: knn on `taiwan`

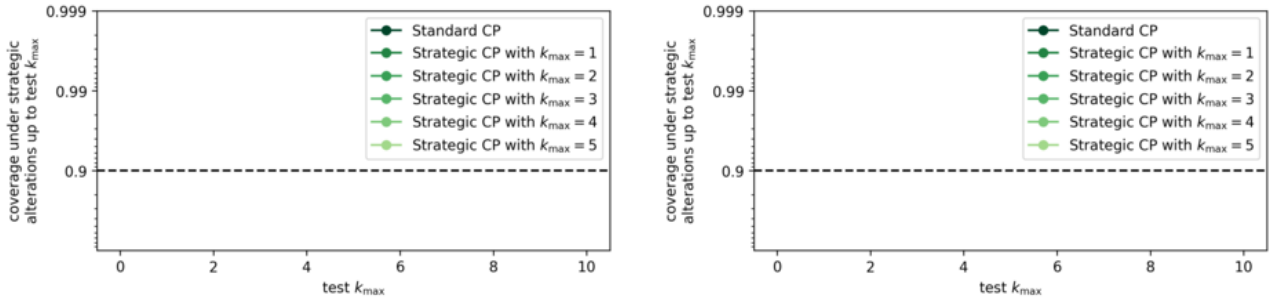


(a) Strategic Conformal Prediction on top of a Plain Model



(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 91: knn on `bank-marketing`



(a) Strategic Conformal Prediction on top of a Plain Model

(b) Strategic Conformal Prediction on top of a Strategic Model

Figure 92: knn on `census-income`

D.3 Table 1: evaluation of our method on multiple datasets and forms of strategic alterations for $\alpha = 0.1$

Table 2: Extended version of Table 1 for the XGBoost model. Note that the news dataset has an average set size difference on another order of magnitude mainly due to the labels of its regression task being of much larger magnitude as well.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE		AVG SET SIZE DIFF
			OURS	STD CP	
academic-dropout	Plain XGBoost	Utility-cost	**90**% ± **2**%	**57**% ± **3**%	+107% ± 5%
		Rand. Search.	**89**% ± **2**%	**49**% ± **3**%	+129% ± 5%
	Strategic XGBoost	Utility-cost	**91**% ± **2**%	76% ± 3%	+46% ± 4%
		Rand. Search.	**90**% ± **2**%	74% ± 3%	+58% ± 4%
spambase	Plain XGBoost	Utility-cost	**91**% ± **2**%	**16**% ± **2**%	+4% ± 1%
		Rand. Search.	**89**% ± **2**%	**27**% ± **3**%	+3% ± 1%
	Strategic XGBoost	Utility-cost	**90**% ± **2**%	**40**% ± **3**%	+38% ± 3%
		Rand. Search.	**88**% ± **2**%	**46**% ± **3**%	+3% ± 1%
shoppers	Plain XGBoost	Utility-cost	**89**% ± **1**%	**44**% ± **2**%	+14% ± 1%
		Rand. Search.	**88**% ± **1**%	**46**% ± **2**%	+13% ± 1%
	Strategic XGBoost	Utility-cost	**89**% ± **1**%	64% ± 2%	+8% ± 1%
		Rand. Search.	**88**% ± **1**%	73% ± 2%	+9% ± 1%
news	Plain XGBoost	Utility-cost	**90**% ± **1**%	**6**% ± **1**%	+948% ± 0%
		Rand. Search.	**89**% ± **1**%	**13**% ± **1**%	+695% ± 0%
	Strategic XGBoost	Utility-cost	**89**% ± **1**%	**7**% ± **1**%	+776% ± 0%
		Rand. Search.	**89**% ± **1**%	**16**% ± **1**%	+644% ± 0%
wine	Plain XGBoost	Utility-cost	**90**% ± **2**%	**48**% ± **3**%	+73% ± 0%
		Rand. Search.	**90**% ± **2**%	**53**% ± **3**%	+70% ± 0%
	Strategic XGBoost	Utility-cost	**90**% ± **2**%	61% ± 3%	+69% ± 0%
		Rand. Search.	**90**% ± **2**%	68% ± 2%	+44% ± 0%
productivity	Plain XGBoost	Utility-cost	**91**% ± **4**%	**38**% ± **6**%	+100% ± 0%
		Rand. Search.	**92**% ± **3**%	**43**% ± **7**%	+91% ± 0%
	Strategic XGBoost	Utility-cost	**91**% ± **4**%	**46**% ± **7**%	+75% ± 0%
		Rand. Search.	**89**% ± **4**%	74% ± 6%	+44% ± 0%
taiwan	Plain XGBoost	Utility-cost	**90**% ± **1**%	77% ± 1%	+46% ± 1%
		Rand. Search.	**91**% ± **1**%	74% ± 1%	+41% ± 1%
	Strategic XGBoost	Utility-cost	**90**% ± **1**%	75% ± 1%	+40% ± 1%
		Rand. Search.	**90**% ± **1**%	77% ± 1%	+32% ± 1%
bank-marketing	Plain XGBoost	Utility-cost	**90**% ± **1**%	**52**% ± **1**%	+6% ± 0%
		Rand. Search.	**90**% ± **1**%	**55**% ± **1**%	+6% ± 0%
	Strategic XGBoost	Utility-cost	**90**% ± **1**%	73% ± 1%	+4% ± 0%
		Rand. Search.	**89**% ± **1**%	76% ± 1%	+4% ± 0%
census-income	Plain XGBoost	Utility-cost	**89**% ± **1**%	**37**% ± **1**%	+149% ± 2%
		Rand. Search.	**90**% ± **1**%	**36**% ± **1**%	+158% ± 2%
	Strategic XGBoost	Utility-cost	**90**% ± **1**%	**60**% ± **1**%	+82% ± 2%
		Rand. Search.	**89**% ± **1**%	**58**% ± **1**%	+87% ± 2%

Table 3: Extended version of Table 1 for an MLP model.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE		AVG SET SIZE DIFF
			OURS	STD CP	
academic-dropout	Plain MLP	Utility-cost	**91**% ± **2**%	**50**% ± **3**%	+97% ± 5%
		Rand. Search.	**89**% ± **2**%	62% ± 3%	+63% ± 4%
	Strategic MLP	Utility-cost	**91**% ± **2**%	61% ± 3%	+56% ± 4%
		Rand. Search.	**90**% ± **2**%	80% ± 3%	+23% ± 3%
spambase	Plain MLP	Utility-cost	**90**% ± **2**%	**40**% ± **3**%	+53% ± 3%
		Rand. Search.	*nan*% ± *nan*%	**40**% ± **3**%	+80% ± 3%
	Strategic MLP	Utility-cost	**91**% ± **2**%	**38**% ± **3**%	+64% ± 3%
		Rand. Search.	**88**% ± **2**%	**53**% ± **3**%	+20% ± 3%
shoppers	Plain MLP	Utility-cost	**91**% ± **1**%	**59**% ± **2**%	+33% ± 2%
		Rand. Search.	**88**% ± **1**%	**60**% ± **2**%	+29% ± 2%
	Strategic MLP	Utility-cost	**91**% ± **1**%	**57**% ± **2**%	+48% ± 2%
		Rand. Search.	**88**% ± **1**%	82% ± 2%	+19% ± 1%
news	Plain MLP	Utility-cost	**90**% ± **1**%	67% ± 1%	+87% ± 0%
		Rand. Search.	**90**% ± **1**%	64% ± 1%	+129% ± 0%
	Strategic MLP	Utility-cost	**90**% ± **1**%	82% ± 1%	−*nan*% ± *nan*%
		Rand. Search.	**90**% ± **1**%	82% ± 1%	−*nan*% ± *nan*%
wine	Plain MLP	Utility-cost	**91**% ± **2**%	63% ± 3%	+144% ± 0%
		Rand. Search.	**91**% ± **2**%	65% ± 3%	+139% ± 0%
	Strategic MLP	Utility-cost	**91**% ± **2**%	82% ± 2%	+24% ± 0%
		Rand. Search.	**90**% ± **2**%	78% ± 2%	+32% ± 0%
productivity	Plain MLP	Utility-cost	**85**% ± **5**%	**14**% ± **4**%	−*nan*% ± *nan*%
		Rand. Search.	**88**% ± **4**%	**4**% ± **2**%	+5622% ± 0%
	Strategic MLP	Utility-cost	**92**% ± **3**%	**36**% ± **6**%	+251% ± 0%
		Rand. Search.	**87**% ± **4**%	**25**% ± **6**%	−*nan*% ± *nan*%
taiwan	Plain MLP	Utility-cost	*nan*% ± *nan*%	*nan*% ± *nan*%	−− 0% ± 0%
		Rand. Search.	*nan*% ± *nan*%	*nan*% ± *nan*%	−− 0% ± 0%
	Strategic MLP	Utility-cost	*nan*% ± *nan*%	*nan*% ± *nan*%	+0% ± 0%
		Rand. Search.	*nan*% ± *nan*%	*nan*% ± *nan*%	−− 0% ± 0%
bank-marketing	Plain MLP	Utility-cost	*nan*% ± *nan*%	88% ± 1%	+100% ± 0%
		Rand. Search.	**90**% ± **1**%	87% ± 1%	+4% ± 0%
	Strategic MLP	Utility-cost	**90**% ± **1**%	86% ± 1%	+4% ± 0%
		Rand. Search.	**90**% ± **1**%	69% ± 1%	+18% ± 1%
census-income	Plain MLP	Utility-cost	*nan*% ± *nan*%	*nan*% ± *nan*%	+0% ± 0%
		Rand. Search.	*nan*% ± *nan*%	*nan*% ± *nan*%	+0% ± 0%
	Strategic MLP	Utility-cost	*nan*% ± *nan*%	*nan*% ± *nan*%	+0% ± 0%
		Rand. Search.	*nan*% ± *nan*%	*nan*% ± *nan*%	+0% ± 0%

Table 4: Extended version of Table 1 for linear models.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE		AVG SET SIZE DIFF
			OURS	STD CP	
academic-dropout	Plain Logistic Regression	Utility-cost	**90**% ± **2**%	75% ± 3%	+61% ± 4%
		Rand. Search.	**90**% ± **2**%	73% ± 3%	+69% ± 5%
	Strategic Logistic Regression	Utility-cost	**91**% ± **2**%	77% ± 3%	+56% ± 4%
		Rand. Search.	**90**% ± **2**%	73% ± 3%	+67% ± 4%
spambase	Plain Logistic Regression	Utility-cost	**90**% ± **2**%	**37**% ± **3**%	+57% ± 3%
		Rand. Search.	**91**% ± **2**%	**39**% ± **3**%	+45% ± 3%
	Strategic Logistic Regression	Utility-cost	**91**% ± **2**%	**51**% ± **3**%	+55% ± 3%
		Rand. Search.	**90**% ± **2**%	63% ± 3%	+28% ± 3%
shoppers	Plain Logistic Regression	Utility-cost	**90**% ± **1**%	62% ± 2%	+31% ± 2%
		Rand. Search.	**88**% ± **1**%	77% ± 2%	+10% ± 1%
	Strategic Logistic Regression	Utility-cost	**88**% ± **1**%	83% ± 1%	+8% ± 1%
		Rand. Search.	**88**% ± **1**%	81% ± 2%	+8% ± 1%
news	Plain Logistic Regression	Utility-cost	**90**% ± **1**%	79% ± 1%	+34% ± 0%
		Rand. Search.	**89**% ± **1**%	82% ± 1%	$-nan\% \pm nan\%$
	Strategic Logistic Regression	Utility-cost	**90**% ± **1**%	77% ± 1%	+35% ± 0%
		Rand. Search.	**90**% ± **1**%	83% ± 1%	+25% ± 0%
wine	Plain Logistic Regression	Utility-cost	**90**% ± **2**%	79% ± 2%	+30% ± 0%
		Rand. Search.	**90**% ± **2**%	78% ± 2%	+30% ± 0%
	Strategic Logistic Regression	Utility-cost	**90**% ± **2**%	81% ± 2%	+25% ± 0%
		Rand. Search.	**89**% ± **2**%	80% ± 2%	+25% ± 0%
productivity	Plain Logistic Regression	Utility-cost	**88**% ± **4**%	71% ± 6%	+36% ± 0%
		Rand. Search.	**91**% ± **4**%	82% ± 5%	+24% ± 0%
	Strategic Logistic Regression	Utility-cost	**90**% ± **4**%	82% ± 5%	+26% ± 0%
		Rand. Search.	**91**% ± **4**%	84% ± 5%	+23% ± 0%
taiwan	Plain Logistic Regression	Utility-cost	**91**% ± **1**%	78% ± 1%	+47% ± 1%
		Rand. Search.	**90**% ± **1**%	68% ± 1%	+51% ± 1%
	Strategic Logistic Regression	Utility-cost	**91**% ± **1**%	78% ± 1%	+46% ± 1%
		Rand. Search.	**91**% ± **1**%	85% ± 1%	+25% ± 1%
bank-marketing	Plain Logistic Regression	Utility-cost	**89**% ± **1**%	80% ± 1%	+7% ± 0%
		Rand. Search.	**90**% ± **1**%	86% ± 1%	+4% ± 0%
	Strategic Logistic Regression	Utility-cost	**90**% ± **1**%	81% ± 1%	+5% ± 0%
		Rand. Search.	**90**% ± **1**%	86% ± 1%	+3% ± 0%
census-income	Plain Logistic Regression	Utility-cost	**90**% ± **1**%	**59**% ± **1**%	+141% ± 2%
		Rand. Search.	**91**% ± **1**%	61% ± 1%	+57% ± 1%
	Strategic Logistic Regression	Utility-cost	**90**% ± **1**%	61% ± 1%	+130% ± 2%
		Rand. Search.	**90**% ± **1**%	80% ± 1%	+28% ± 1%

Table 5: Extended version of Table 1 for random forests.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE OURS	STRATEGIC COVERAGE STD CP	AVG SET SIZE DIFF
academic-dropout	Plain Random Forest	Utility-cost	**90**% ± **2**%	72% ± 3%	+54% ± 4%
		Rand. Search.	**92**% ± **2**%	61% ± 3%	+84% ± 5%
	Strategic Random Forest	Utility-cost	**91**% ± **2**%	82% ± 3%	+24% ± 3%
		Rand. Search.	**92**% ± **2**%	84% ± 2%	+25% ± 3%
spambase	Plain Random Forest	Utility-cost	**90**% ± **2**%	**8**% ± **2**%	− − 0% ± 0%
		Rand. Search.	**89**% ± **2**%	**9**% ± **2**%	− − 0% ± 0%
	Strategic Random Forest	Utility-cost	**90**% ± **2**%	**49**% ± **3**%	+18% ± 2%
		Rand. Search.	**89**% ± **2**%	**46**% ± **3**%	+3% ± 1%
shoppers	Plain Random Forest	Utility-cost	**90**% ± **1**%	**46**% ± **2**%	+7% ± 1%
		Rand. Search.	**89**% ± **1**%	**54**% ± **2**%	+8% ± 1%
	Strategic Random Forest	Utility-cost	**89**% ± **1**%	79% ± 2%	+4% ± 1%
		Rand. Search.	**88**% ± **1**%	77% ± 2%	+6% ± 1%
news	Plain Random Forest	Utility-cost	**90**% ± **1**%	**4**% ± **0**%	$-nan\% \pm nan\%$
		Rand. Search.	**89**% ± **1**%	**11**% ± **1**%	+560% ± 0%
	Strategic Random Forest	Utility-cost	**89**% ± **1**%	**1**% ± **0**%	+509% ± 0%
		Rand. Search.	**91**% ± **1**%	**6**% ± **0**%	$-nan\% \pm nan\%$
wine	Plain Random Forest	Utility-cost	**90**% ± **2**%	68% ± 2%	+51% ± 0%
		Rand. Search.	**91**% ± **2**%	68% ± 2%	+53% ± 0%
	Strategic Random Forest	Utility-cost	**90**% ± **2**%	74% ± 2%	+46% ± 0%
		Rand. Search.	**90**% ± **2**%	76% ± 2%	+36% ± 0%
productivity	Plain Random Forest	Utility-cost	**90**% ± **4**%	**33**% ± **6**%	+109% ± 0%
		Rand. Search.	**90**% ± **4**%	**48**% ± **6**%	+82% ± 0%
	Strategic Random Forest	Utility-cost	**92**% ± **3**%	**58**% ± **6**%	+58% ± 0%
		Rand. Search.	**89**% ± **4**%	77% ± 5%	$-nan\% \pm nan\%$
taiwan	Plain Random Forest	Utility-cost	**90**% ± **1**%	82% ± 1%	+29% ± 1%
		Rand. Search.	**91**% ± **1**%	87% ± 1%	+14% ± 1%
	Strategic Random Forest	Utility-cost	**91**% ± **1**%	82% ± 1%	+31% ± 1%
		Rand. Search.	**90**% ± **1**%	89% ± 1%	+6% ± 1%
bank-marketing	Plain Random Forest	Utility-cost	**89**% ± **1**%	63% ± 1%	+4% ± 0%
		Rand. Search.	**89**% ± **1**%	68% ± 1%	+4% ± 0%
	Strategic Random Forest	Utility-cost	**90**% ± **1**%	78% ± 1%	+3% ± 0%
		Rand. Search.	**90**% ± **1**%	80% ± 1%	+2% ± 0%
census-income	Plain Random Forest	Utility-cost	**91**% ± **1**%	64% ± 1%	+39% ± 1%
		Rand. Search.	**91**% ± **1**%	63% ± 1%	+40% ± 1%
	Strategic Random Forest	Utility-cost	**91**% ± **1**%	75% ± 1%	+26% ± 1%
		Rand. Search.	**91**% ± **1**%	77% ± 1%	+25% ± 1%

Table 6: Extended version of Table 1 for a k-NN model.

DATASET	UNDERLYING MODEL	Δs	STRATEGIC COVERAGE		AVG SET SIZE DIFF
			OURS	STD CP	
academic-dropout	Plain k-NN	Utility-cost	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$--0\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$--0\% \pm 0\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$--0\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$--0\% \pm 0\%$
spambase	Plain k-NN	Utility-cost	$nan\% \pm nan\%$	$\mathbf{38}\% \pm \mathbf{3}\%$	$+100\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$\mathbf{53}\% \pm \mathbf{3}\%$	$+95\% \pm 1\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$\mathbf{43}\% \pm \mathbf{3}\%$	$+92\% \pm 2\%$
		Rand. Search.	$nan\% \pm nan\%$	$79\% \pm 3\%$	$+55\% \pm 3\%$
shoppers	Plain k-NN	Utility-cost	$\mathbf{90}\% \pm \mathbf{1}\%$	$77\% \pm 2\%$	$+19\% \pm 2\%$
		Rand. Search.	$nan\% \pm nan\%$	$74\% \pm 2\%$	$+100\% \pm 0\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$--0\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$85\% \pm 1\%$	$+94\% \pm 1\%$
news	Plain k-NN	Utility-cost	$\mathbf{90}\% \pm \mathbf{1}\%$	$\mathbf{55}\% \pm \mathbf{1}\%$	$-nan\% \pm nan\%$
		Rand. Search.	$\mathbf{90}\% \pm \mathbf{1}\%$	$63\% \pm 1\%$	$-nan\% \pm nan\%$
	Strategic k-NN	Utility-cost	$\mathbf{90}\% \pm \mathbf{1}\%$	$70\% \pm 1\%$	$+81\% \pm 0\%$
		Rand. Search.	$\mathbf{89}\% \pm \mathbf{1}\%$	$78\% \pm 1\%$	$+51\% \pm 0\%$
wine	Plain k-NN	Utility-cost	$\mathbf{95}\% \pm \mathbf{1}\%$	$73\% \pm 3\%$	$+57\% \pm 0\%$
		Rand. Search.	$\mathbf{91}\% \pm \mathbf{2}\%$	$73\% \pm 2\%$	$+43\% \pm 0\%$
	Strategic k-NN	Utility-cost	$\mathbf{91}\% \pm \mathbf{1}\%$	$75\% \pm 2\%$	$+37\% \pm 0\%$
		Rand. Search.	$\mathbf{90}\% \pm \mathbf{2}\%$	$77\% \pm 2\%$	$+29\% \pm 0\%$
productivity	Plain k-NN	Utility-cost	$\mathbf{91}\% \pm \mathbf{4}\%$	$77\% \pm 5\%$	$+35\% \pm 0\%$
		Rand. Search.	$\mathbf{89}\% \pm \mathbf{4}\%$	$72\% \pm 5\%$	$+39\% \pm 0\%$
	Strategic k-NN	Utility-cost	$\mathbf{89}\% \pm \mathbf{4}\%$	$78\% \pm 5\%$	$+16\% \pm 0\%$
		Rand. Search.	$\mathbf{90}\% \pm \mathbf{3}\%$	$81\% \pm 5\%$	$+24\% \pm 0\%$
taiwan	Plain k-NN	Utility-cost	$nan\% \pm nan\%$	$79\% \pm 1\%$	$+96\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$82\% \pm 1\%$	$+83\% \pm 1\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$78\% \pm 1\%$	$+97\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$83\% \pm 1\%$	$+82\% \pm 1\%$
bank-marketing	Plain k-NN	Utility-cost	$nan\% \pm nan\%$	$82\% \pm 1\%$	$+100\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$81\% \pm 1\%$	$+100\% \pm 0\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$82\% \pm 1\%$	$+100\% \pm 0\%$
		Rand. Search.	$\mathbf{90}\% \pm \mathbf{1}\%$	$84\% \pm 1\%$	$+5\% \pm 0\%$
census-income	Plain k-NN	Utility-cost	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$+0\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$+0\% \pm 0\%$
	Strategic k-NN	Utility-cost	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$+0\% \pm 0\%$
		Rand. Search.	$nan\% \pm nan\%$	$nan\% \pm nan\%$	$+0\% \pm 0\%$

D.4 Extra figure: coverage under change of the stochastic step

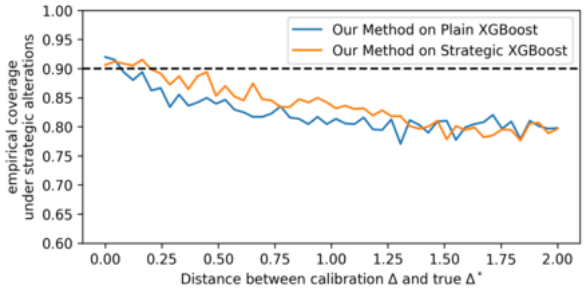


Figure 93: xgboost on `academic-dropout`

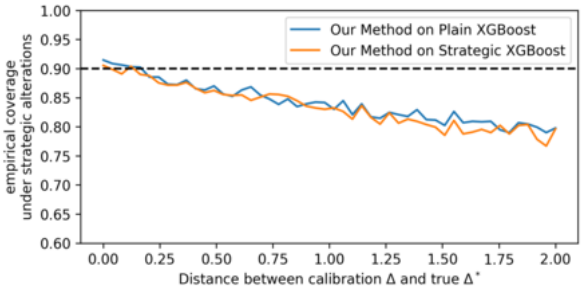


Figure 97: xgboost on `wine`

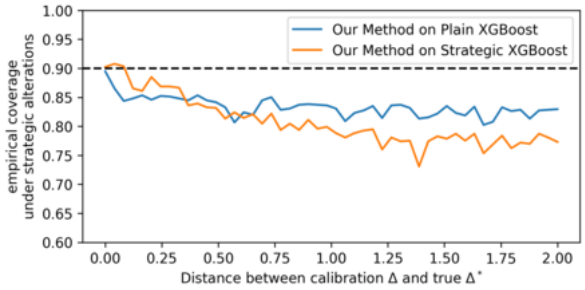


Figure 94: xgboost on `spambase`

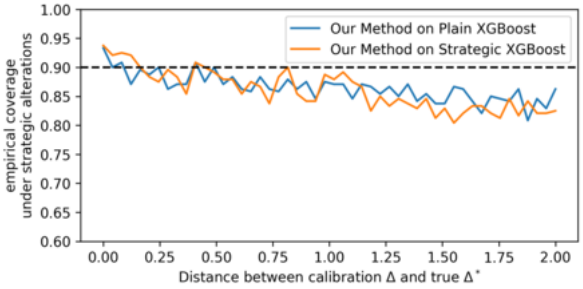


Figure 98: xgboost on `productivity`

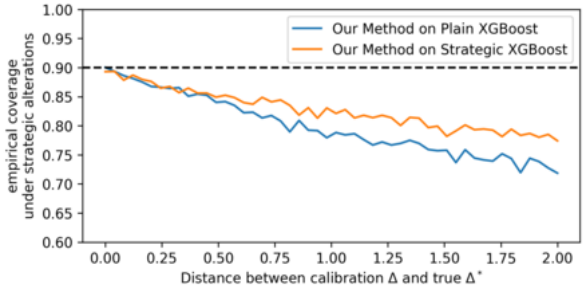


Figure 95: xgboost on `shoppers`

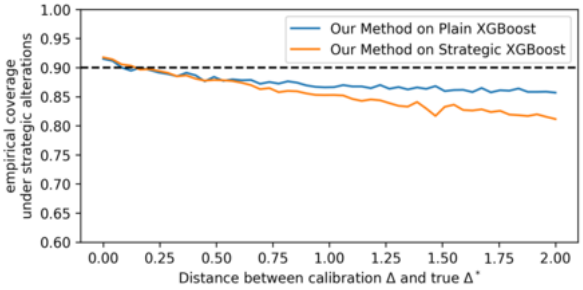


Figure 99: xgboost on `taiwan`

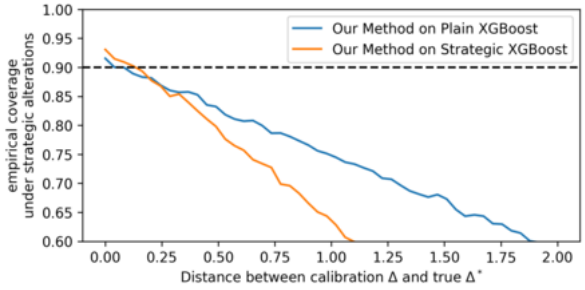


Figure 96: xgboost on `news`

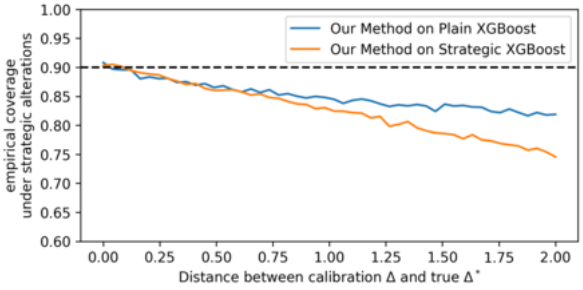


Figure 100: xgboost on `bank-marketing`

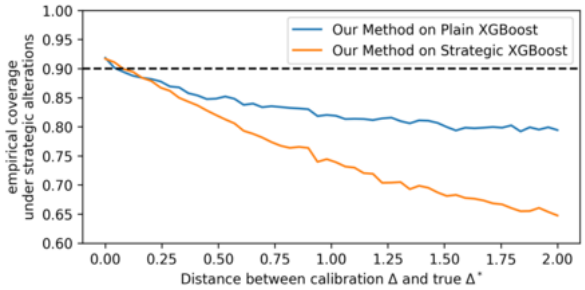


Figure 101: xgboost on `census-income`

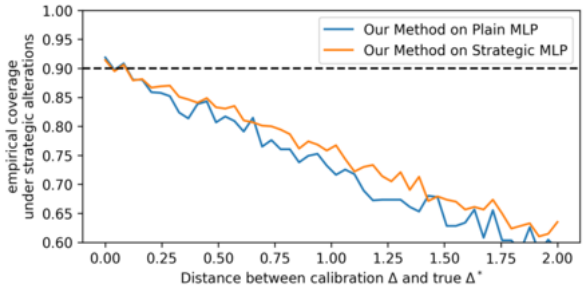


Figure 102: mlp on `academic-dropout`

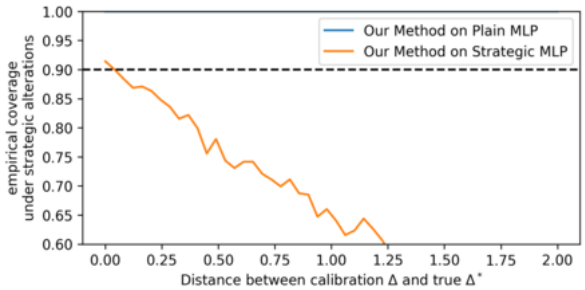


Figure 103: mlp on `spambase`

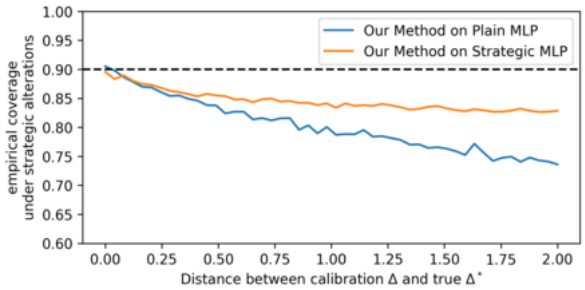


Figure 104: mlp on `shoppers`

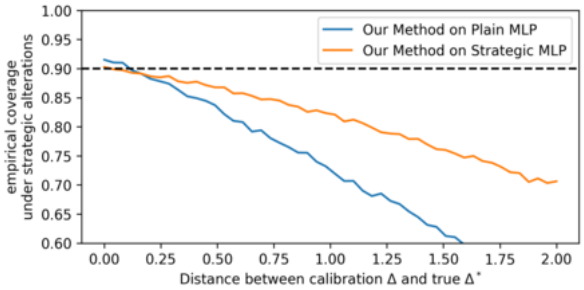


Figure 105: mlp on `news`

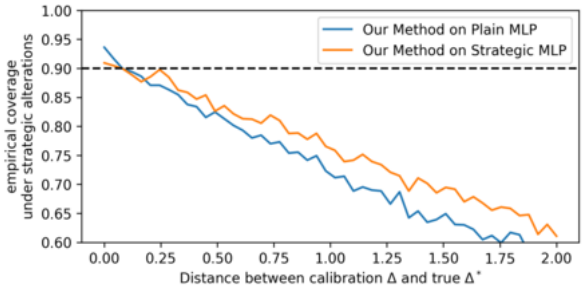


Figure 106: mlp on `wine`

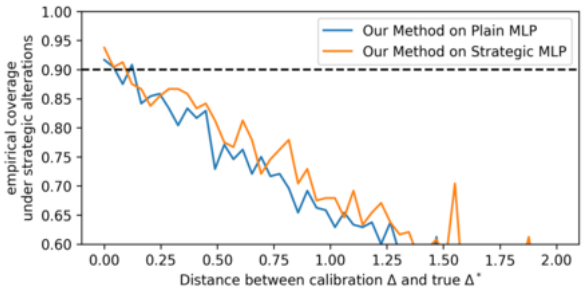


Figure 107: mlp on `productivity`

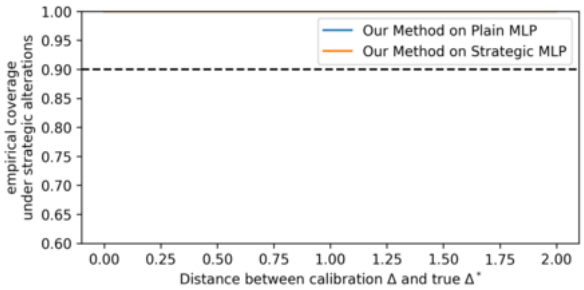


Figure 108: mlp on `taiwan`

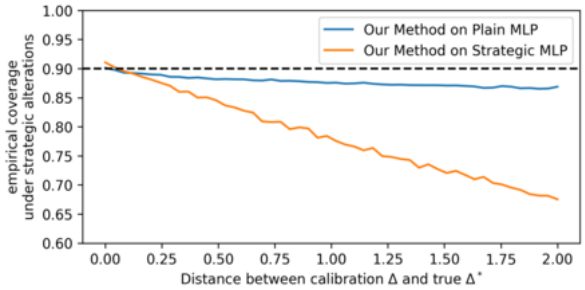


Figure 109: mlp on `bank-marketing`

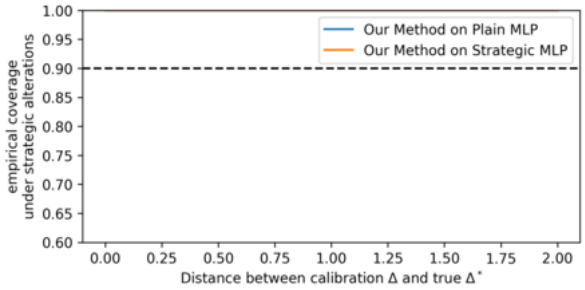


Figure 110: mlp on `census-income`

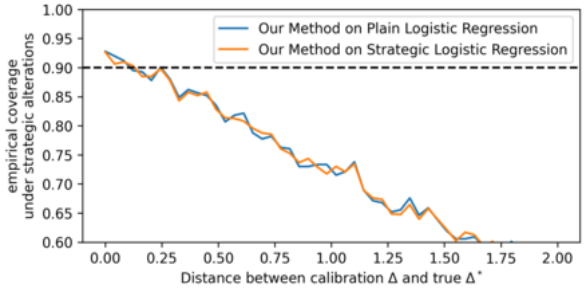


Figure 111: linear on `academic-dropout`

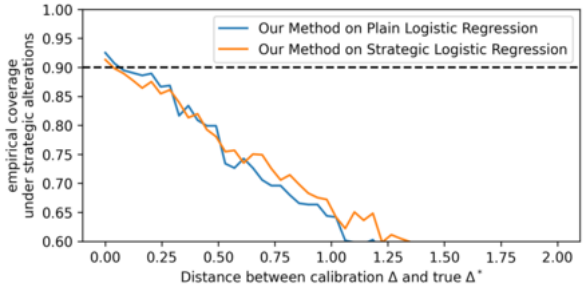


Figure 112: linear on `spambase`

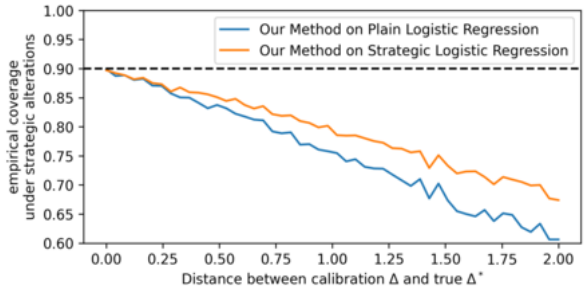


Figure 113: linear on `shoppers`

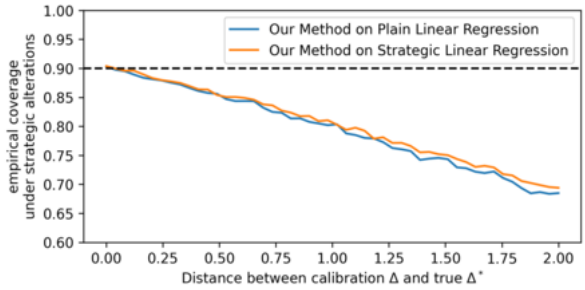


Figure 114: linear on `news`

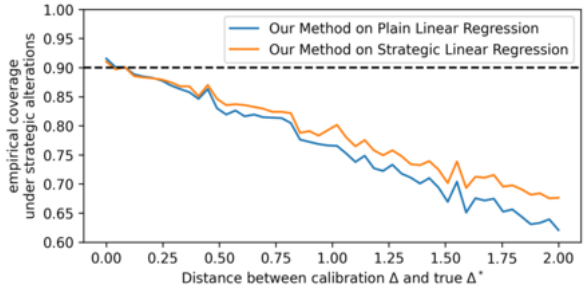


Figure 115: linear on `wine`

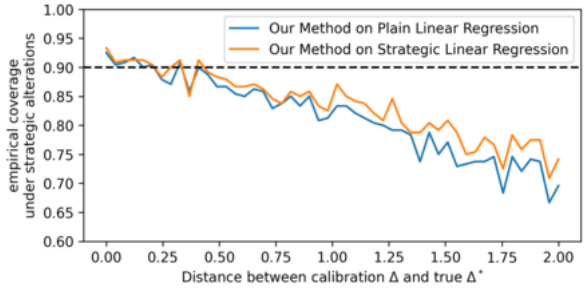


Figure 116: linear on `productivity`

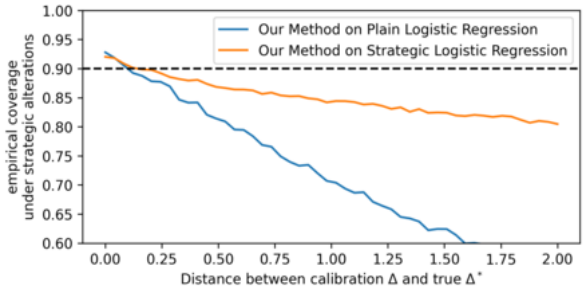


Figure 117: linear on `taiwan`

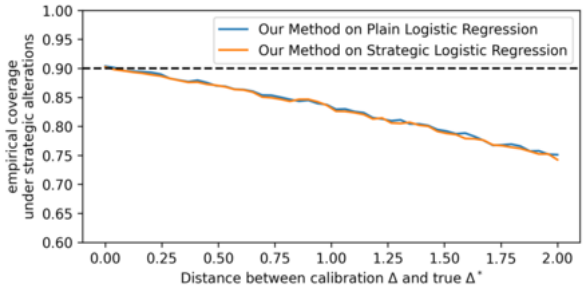


Figure 118: linear on `bank-marketing`

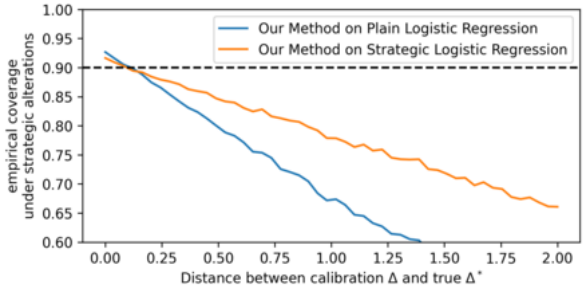


Figure 119: linear on `census-income`

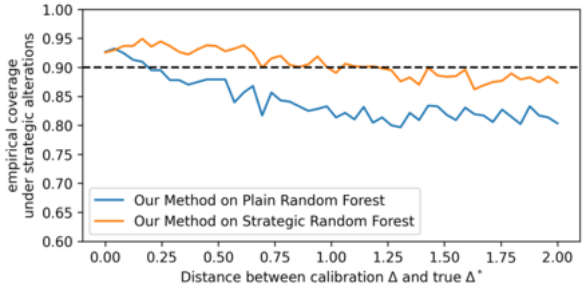


Figure 120: random-forest on `academic-dropout`

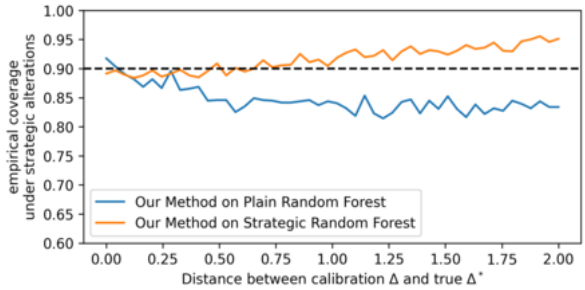


Figure 121: random-forest on `spambase`

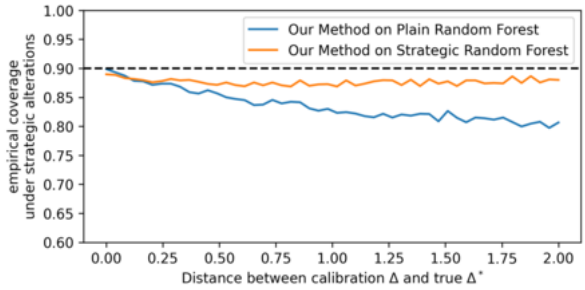


Figure 122: random-forest on `shoppers`

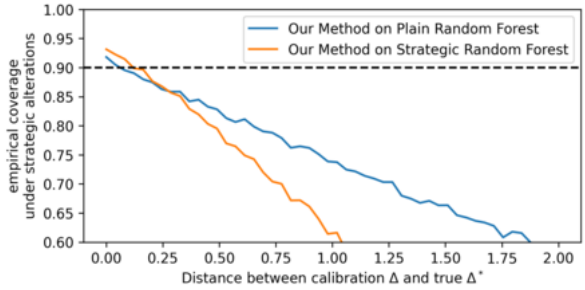


Figure 123: random-forest on `news`

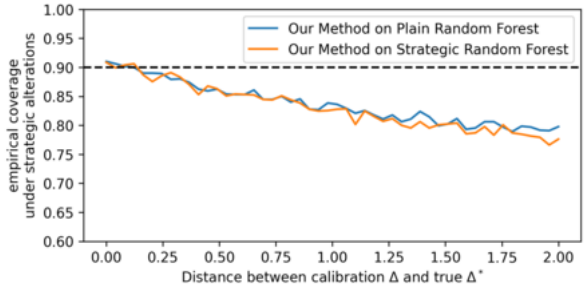


Figure 124: random-forest on `wine`

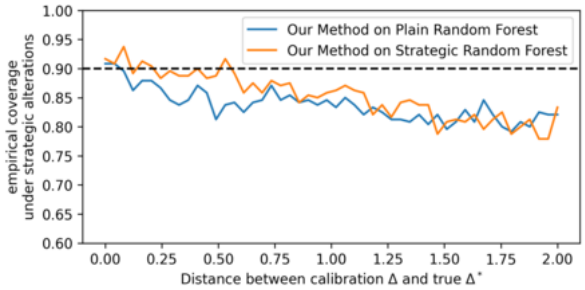


Figure 125: random-forest on `productivity`

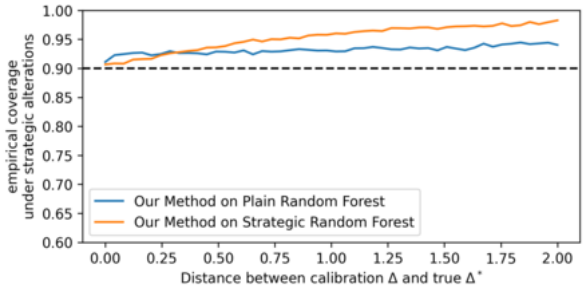


Figure 126: random-forest on `taiwan`

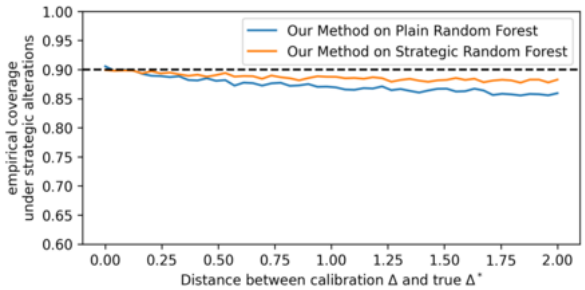


Figure 127: random-forest on `bank-marketing`

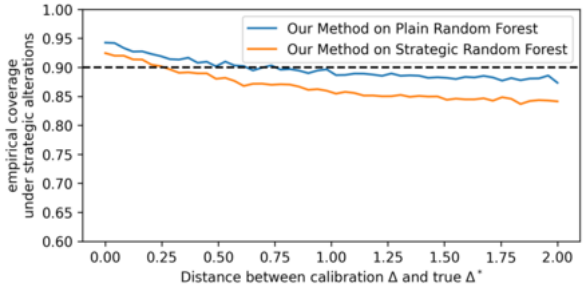


Figure 128: random-forest on `census-income`

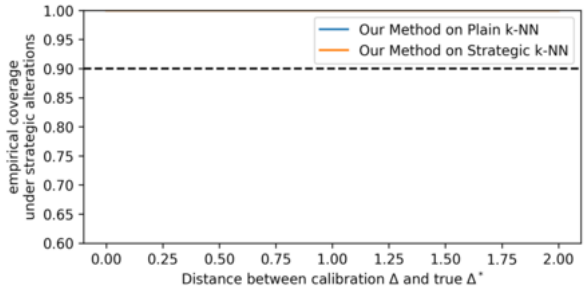


Figure 129: knn on `academic-dropout`

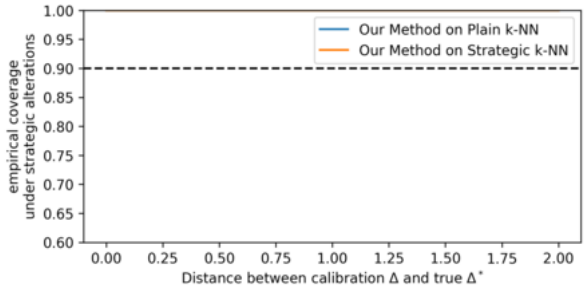


Figure 130: knn on `spambase`

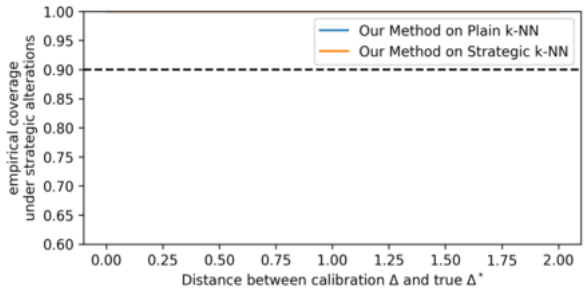


Figure 131: knn on `shoppers`

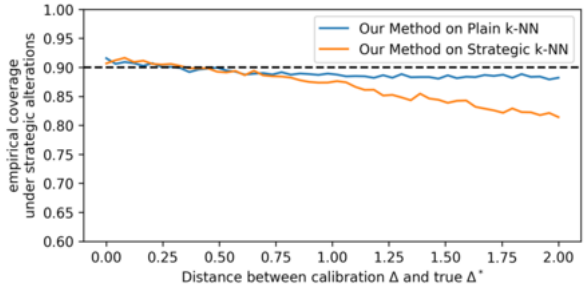


Figure 132: knn on `news`

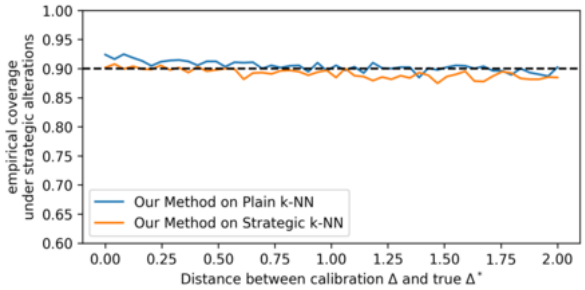


Figure 133: knn on `wine`

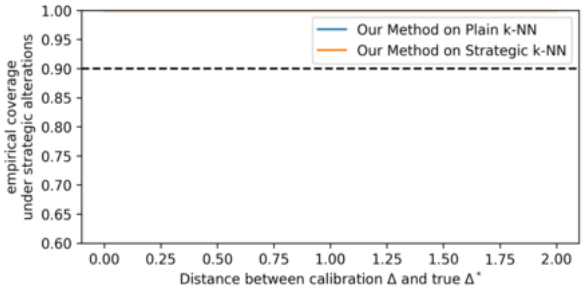


Figure 137: knn on `census-income`

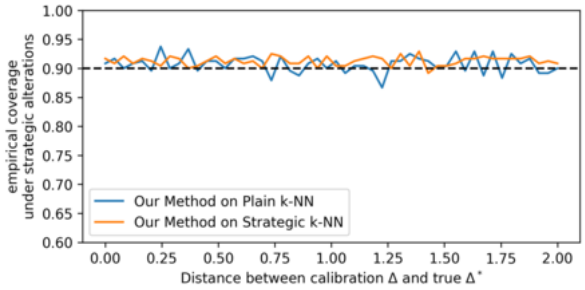


Figure 134: knn on `productivity`

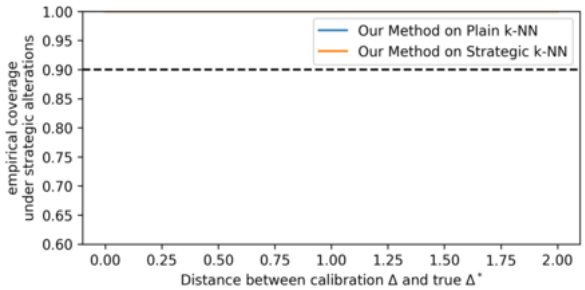


Figure 135: knn on `taiwan`

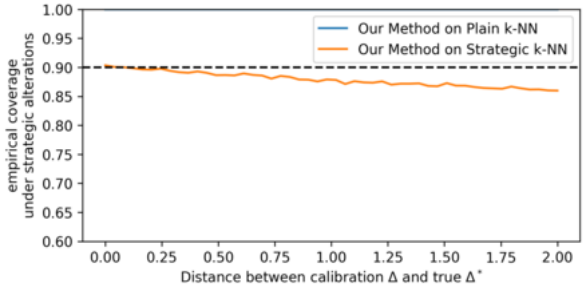


Figure 136: knn on `bank-marketing`