Received: 9 January 2023 | Revised: 24 May 2023 | Accepted: 26 May 2023

DOI: 10.1002/qre.3392

SPECIAL ISSUE ARTICLE

WILEY

Robust online active learning

Davide Cacciarelli[1,2] | **Murat Kulahci**[1,3] | **John Sølve Tyssedal**[2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

[3]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

Correspondence
Davide Cacciarelli, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs, Lyngby, Denmark.
Email: dcac@dtu.dk

Funding information
DTU Strategic Alliances Fund

Abstract

In many industrial applications, obtaining labeled observations is not straightforward as it often requires the intervention of human experts or the use of expensive testing equipment. In these circumstances, active learning can be highly beneficial in suggesting the most informative data points to be used when fitting a model. Reducing the number of observations needed for model development alleviates both the computational burden required for training and the operational expenses related to labeling. Online active learning, in particular, is useful in high-volume production processes where the decision about the acquisition of the label for a data point needs to be taken within an extremely short time frame. However, despite the recent efforts to develop online active learning strategies, the behavior of these methods in the presence of outliers has not been thoroughly examined. In this work, we investigate the performance of online active linear regression in contaminated data streams. Our study shows that the currently available query strategies are prone to sample outliers, whose inclusion in the training set eventually degrades the predictive performance of the models. To address this issue, we propose a solution that bounds the search area of a conditional D-optimal algorithm and uses a robust estimator. Our approach strikes a balance between exploring unseen regions of the input space and protecting against outliers. Through numerical simulations, we show that the proposed method is effective in improving the performance of online active learning in the presence of outliers, thus expanding the potential applications of this powerful tool.

KEYWORDS

active learning, data stream, optimal experimental design, outliers, robust regression, unlabeled data

1 | INTRODUCTION

Predictive models often need to be trained on a large amount of labeled data before being deployed. However, in industrial applications data is often abundant only in an unlabeled form. Active learning strategies provide a solution to this problem

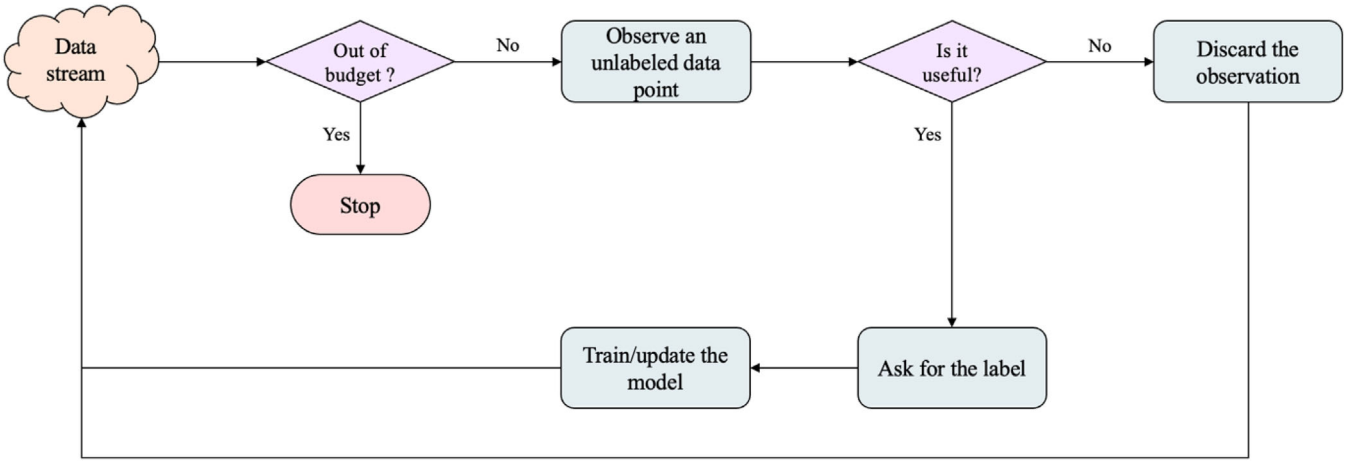

FIGURE 1 General online active learning flowchart.

by prioritizing the labeling of the most useful instances for building the model, thus accelerating the convergence of its learning curve.[1] Active learning problems can be classified into three macro-scenarios.[2] The first and most studied scenario is the pool-based scenario, where the learner can select the most useful instances to be labeled by maximizing an evaluation criterion over a closed set of observations. The second scenario is referred to as membership query synthesis, and it allows the learner to query the labels of synthetically generated instances rather than those sampled from the process distribution. Finally, the third scenario is online, or stream-based, active learning.[3] In this case, the unlabeled observations are drawn sequentially by the learner, which must immediately decide whether to keep the instance and query its label or discard it. While many researchers have been working on active learning in the recent years, the pool-based scenario has received the most attention.[4,5]

Although online active learning has become more popular in the last few years,[6–10] the majority of the methods have been developed for classification tasks.[11] An interesting approach to online active learning for fuzzy regression models has been proposed by Lughofer.[12] Other researchers tried to adapt the optimality criteria of the experimental design theory to the online active linear regression framework.[13–16] Linear regression models are still very useful in industrial applications as they can be efficiently trained on a small number of observations. They are able to offer a straightforward interpretation, along with the possibility of constructing confidence intervals on the parameter estimates.[17,18] They can also be easily coupled with variable selection and robust estimation methods. Furthermore, whereas many pool-based active learning approaches employ ensemble methods or complex models, linear models can support online active learning due to the decreased computational cost associated with model training and updating.

Figure 1 depicts a general online active learning flowchart. The main difference among the query strategies lies in how they assess the usefulness of an unlabeled instance when the learner samples it from the data stream. Another important aspect is the assumptions on the input distribution. Indeed, despite the increased interest in the online active linear regression framework, the performance of the sampling strategies in the presence of outliers has not been thoroughly explored. The few works we are aware of that analyze this issue, are related to the pool-based scenario. Deldossi et al.[19] highlighted how sampling methods based on D-optimality are affected by outliers and high leverage points. Zhao et al.[20] focused on robust active representations based on the $\ell_{2,p}$-norm constraints for selecting highly representative data. Finally, He et al.[21] emphasized the problem of being prone to sample outliers while proposing a semi-supervised active learning strategy for multivariate time series classification, using uncertainty and local density.

In this paper, we study the problem of learning from contaminated data streams with limited sampling resources. We first investigate the effects of outliers on the sampling decisions made by state-of-the-art online active learning approaches for linear regression, and successively propose a solution for this issue. It should be noted that the presence of outliers considered in this work cannot be tackled using traditional anomaly detection methods. Indeed, most unsupervised anomaly detection strategies rely on the assumption that a large training set free from outliers, usually referred to as phase I data in the statistical process control literature, is available beforehand.[22–25] However, this assumption is violated in many practical applications,[26] especially in label-scarce scenarios where few to no labels are available before the beginning of the active learning routine. The proposed strategy for online active learning utilizes a double-threshold approach to limit the search area of a conditional D-optimality (CDO) algorithm. By using two thresholds, the strategy aims to identify

informative data points while excluding outliers. In cases of highly contaminated environments, robust estimators based on the Huber and Tukey bisquare loss are employed.

The remainder of this paper is organized as follows. In Section 2, we introduce the terminology and describe the sampling strategies that are used as the baseline in our analysis. Section 3 offers a review on the use of robust estimators and introduces ways of modifying the CDO algorithm. In Section 4, we test our approach using numerical simulations in four scenarios, using different contamination ratios. Section 5 offers a discussion on the results obtained. Finally, Section 6 provides some conclusions.

2 | BACKGROUND AND RELATED WORK

The labeled observations that are collected from the contaminated data stream are used to fit a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector representing the zero-mean Gaussian noise. Here, n represents the total number of observations and p the number of variables. Before starting the active learning routine and the collection of additional labels, it is commonly assumed to have at our disposal an initial set of labeled observations.[5,27,28] This set is used to obtain an initial estimate $\hat{\boldsymbol{\beta}}$ for the coefficients $\boldsymbol{\beta}$. Using an ordinary least squares (OLS) estimator, we have that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{y}$. Then, the fitted linear regression model is $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and the residuals are obtained as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. When the variables are highly correlated, a pre-whitening might be performed to avoid an ill-conditioned problem when computing $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. It should be noted that the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is important to obtain information about the design geometry. In particular, for a design composed of n runs, the moment matrix, $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$, plays a central role in the definition of optimal experimental designs. The two most commonly employed optimality criteria, which have been adapted for the online active learning scenario, are A-optimality and D-optimality. An A-optimal design is achieved by minimizing the trace of the inverse of the moment matrix $\mathbf{M}$. It can be shown how this corresponds to minimizing the individual variances of the estimated coefficients. This approach has been adapted for the online active linear regression framework by Riquelme et al.[14] They proposed a norm-thresholding algorithm that only selects observations $\mathbf{x}$ with large, scaled norm by estimating a threshold Γ as

$$\mathrm{P}_{\mathrm{D}}\left(\|\mathbf{x}\| \geq \Gamma\right) = \alpha \tag{2}$$

where α is the ratio of observations we are willing to label out of the incoming data stream. The probability distribution of the norms can be approximated using kernel density estimation (KDE) on a set of unlabeled observations $\mathbf{C}$, which can be regarded as a warm-up or calibration set and can either be retrieved from historical data or by observing the data stream for a while. Using this thresholding approach, we would be sampling, with high probability, observations that help achieve A-optimality. Given n statistics, $(s_1, \ldots, s_n)$, KDE can be used to estimate the shape of an unknown distribution f using

$$\hat{f}(s) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{s - s_i}{h}\right) \tag{3}$$

where the bandwidth h is a positive number that is used to control the amount of smoothing, and the kernel K is a smooth function such that $K(s) \geq 0$, $\int K(s)ds = 1$, $\int sK(s)ds = 0$ and $\sigma_K^2 \equiv \int s^2K(s)ds > 0$. In this paper, the Gaussian (Normal) kernel, $K(s) = (2\pi)^{-1/2}\, e^{-s^2/2}$ is used.

D-optimality is another fundamental criterion,[29] which takes both the variances and covariances of the model coefficients into account by maximizing the determinant of the moment matrix $\mathbf{M}$. As in the case of A-optimality, D-optimality has been adapted to the online active learning scenario with the proposal of a CDO algorithm.[16] CDO suggests setting a threshold Γ by using

$$\mathrm{P}_{\mathrm{D}}\left(\mathbf{x}_{l+1}^{\mathrm{T}}\left(\mathbf{X}_l^{\mathrm{T}}\mathbf{X}_l\right)^{-1}\mathbf{x}_{l+1} \geq \Gamma\right) = \alpha \tag{4}$$

where $\mathbf{X}_l$ is the model matrix with the l labeled observations currently available and $\mathbf{x}_{l+1}$ is the unlabeled data point that is under evaluation. It can be shown that by selecting observations that maximize $\mathbf{x}_{l+1}^{\mathrm{T}}(\mathbf{X}_l^{\mathrm{T}}\mathbf{X}_l)^{-1}\mathbf{x}_{l+1}$, we are at the same time seeking D-optimality and labeling observations with a large unscaled prediction variance (UPV),[30] which is generally defined as

$$\mathrm{UPV}(\mathbf{x}) = \mathbf{x}^{(m)\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}^{(m)} \tag{5}$$

where $\mathbf{x}^{(m)}$ represents the data point where the UPV is being estimated, expanded to the model form (e.g., if polynomial features are added to the model). To estimate the threshold Γ, we use KDE after computing the UPV of all the observations in $\mathbf{C}$. The CDO intuition is coherent with the idea that a point for which we have a large UPV value represents a less explored region of the input space and will help, with high probability, attaining D-optimality, conditional on the already collected observations. The equivalence between sampling data points with high UPV and D-optimality is demonstrated in our previous work.[16]

Given these preliminaries, we now propose methods that are robust to the presence of outliers in the data stream.

3 | METHODS

When training a linear regression model on a dataset corrupted by the presence of outliers, a simple yet effective solution is to resort to the use of robust estimators. An extensive overview of robust regression has been provided by Fox and Weisberg.[31] In general, robust estimation methods attempt to estimate the coefficients $\hat{\boldsymbol{\beta}}$ by minimizing a particular loss function given by

$$\mathcal{I} = \sum_{i=1}^{n}\rho(e_i) = \sum_{i=1}^{n}\rho\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right) \tag{6}$$

where ρ is a function that regulates the contribution of each residual to the loss, and e_i is the residual for the ith observation $(\mathbf{x}_i,\ y_i)$. The function ρ is nonnegative, equal to zero when the argument is zero, symmetrical and monotone in $|e|$. In the case of an OLS estimator, the loss is given by

$$\rho_{LS} = e^2 \tag{7}$$

It can be seen how the objective function minimized by an OLS estimator is equally affected by all the observations for which we measure the residuals. Instead, robust estimators try to reduce the impact of observation with very large residuals on the estimation of $\hat{\boldsymbol{\beta}}$. One of the most popular robust loss functions is the Huber loss,[32] which is defined as

$$\rho_H = \begin{cases} e^2 & for\ |e| \le k \\ 2k|e| - k^2 & for\ |e| > k \end{cases} \tag{8}$$

where k is a tuning parameter, which is usually set to 1.345σ to achieve 95% efficiency when the errors are normally distributed, while keeping a good protection against outliers.[31] It can be seen how the contribution of each observation is reduced based on the magnitude of the corresponding residual. However, despite being much more robust than the OLS estimator, the Huber loss is still proportional to the magnitude of the residuals even when the absolute errors are larger than k. Conversely, the Tukey bisquare loss function[33] sets a threshold for the residuals, above which the value of the residuals does not influence the loss.

The Tukey loss function is given by

$$\rho_T = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & for\ |e| \le k \\ \frac{k^2}{6} & for\ |e| > k \end{cases} \tag{9}$$

ALGORITHM 1 Bounded CDO

Input: data stream $\mathbf{S}$; initial random design $\mathbf{X}$; warm-up length m; budget B

Output: an augmented design $\mathbf{Z}$

Set $\mathbf{C} = \emptyset$ // calibration set to estimate $\boldsymbol{\Sigma}$, Γ_1, Γ_2
$i \leftarrow 1, b \leftarrow 0$ // b represents the currently used budget
while $i \leq m$ **do**
  Observe the ith data point $\mathbf{x}_i \in \mathbf{S}$
  Select $\mathbf{x}_i : \mathbf{C} = \mathbf{C} \cup \mathbf{x}_i$
  $i \leftarrow i + 1$
end while
Estimate the covariance matrix $\boldsymbol{\Sigma}$ from $\mathbf{C}$ and perform eigendecomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$
Whiten the initial design by computing $\mathbf{Z} = \boldsymbol{\Lambda}^{-1/2}\ \mathbf{U}^{\mathrm{T}}\mathbf{X}$
Whiten the calibration set by computing $\mathbf{V} = \boldsymbol{\Lambda}^{-1/2}\ \mathbf{U}^{\mathrm{T}}\mathbf{C}$
Estimate Γ_1, Γ_2 by estimating the UPV of the model trained on $\mathbf{Z}$ on the points in $\mathbf{V}$
while $b \leq B$ **and** $i \leq |\mathbf{S}|$ **do**
  Observe the ith data point $\mathbf{x}_i \in \mathbf{S}$
  Whiten $\mathbf{x}_i$ by computing $\mathbf{z}_i = \boldsymbol{\Lambda}^{-1/2}\ \mathbf{U}^{\mathrm{T}}\mathbf{x}_i$
  if $\Gamma_1 \leq \mathbf{z}_i^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{z}_i \leq \Gamma_2$ **then**
    Ask for the label y_i and augment the labeled dataset $\mathbf{Z} = \mathbf{Z} \cup \mathbf{z}_i$
    $b \leftarrow b + 1$
    Update thresholds Γ_1, Γ_2 using the augmented design
  else
    Discard $\mathbf{x}_i$
  $i \leftarrow i + 1$
  end if
end while
return $\mathbf{Z}$

where the value of the tuning constant k is usually set up to 4.685σ.[31] Besides using a Huber or Tukey loss to obtain a robust estimator, we consider the possibility of filtering out outliers while selecting the most informative observations from the data stream. To this extent, we propose an adaptation of the CDO algorithm, where instead of estimating a threshold, we define a bounded area of interest for the unscaled prediction variance of an observation as

$$\mathrm{P_D}\left(\Gamma_1 \leq \mathbf{x}_{l+1}^{\mathrm{T}}\left(\mathbf{X}_l^{\mathrm{T}}\mathbf{X}_l\right)^{-1}\mathbf{x}_{l+1} \leq \Gamma_2\right) = \alpha \tag{10}$$

This approach is hereinafter referred to as bounded CDO. The idea is coherent with the method proposed by Hoaglin and Welsch.[34,35] of considering as potential outliers observations for which $\mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_i \geq 2p/n$ is verified. The filtering approach suggested by Hoaglin and Welsch is also used by Deldossi et al.,[19] in the offline scenario. Here, instead of opting for a fixed value for Γ_2, we use KDE with a Gaussian kernel to estimate Γ_1 and Γ_2. The upper limit Γ_2 is selected by determining a cut-off value c, which is related to the amount of protection against outliers that we would like to achieve. This value is a tuning constant similar to the k used by robust estimators and, when possible, should be selected by exploiting previous knowledge of the process. Given the cut-off value c and the sampling rate α, Γ_2 is given by the $100(1-c)\%$ percentile, and Γ_1 by the $100(1-c-\alpha)\%$ percentile. As anticipated in Section 2, the threshold estimation is based on a set of unlabeled data, which is also used to estimate the covariance matrix $\boldsymbol{\Sigma}$ and whitening the observations to remove dependencies and facilitate the estimation of $\hat{\boldsymbol{\beta}}$. At this stage, semi-supervised methods might also be considered to perform tasks like feature extraction and exploit all the information available in the unlabeled data.[21,36–39]

Algorithm 1 provides a detailed explanation of how to implement the bounded CDO strategy for online active learning in a fixed-budget setting. The strategy involves collecting new labels and incorporating them into the design until a specified budget constraint B is reached. In some cases, it might be beneficial to anticipate the stop of the active learning routine if the marginal improvement of the model is no longer significant.[40] Previous studies have proposed various stopping criteria to enhance the efficiency of data collection schemes based on active learning.[41–45] Appendix A explores how

some of these approaches could be adapted to the regression framework. From a computational standpoint, the update of $\hat{\boldsymbol{\beta}}$ is done by means of a complete retraining each time a new labeled example is added to the design. However, if the data matrix becomes considerably large and the time required for model updates increases, one may opt to update the model and estimate new thresholds when a batch of new observations is collected, aligning with the principles of batch-mode active learning.[46] Additionally, incremental and recursive updating techniques can also be considered for improving computational efficiency.

The estimation of the UPV can be modified by taking into account the weight matrix obtained from the robust estimators. The weighted UPV (UPV_w) is estimated as follows

$$\text{UPV}_w(\mathbf{x}) = \mathbf{x}^{(m)\text{T}} \left(\mathbf{X}^{\text{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{x}^{(m)} \tag{11}$$

where $\mathbf{W}$ represents the weight matrix used to downweigh the influence of outliers in the estimation of the regression parameters.[31] Each element of the weight matrix $\mathbf{W}$ is a positive number that determines the weight given to each observation in the regression analysis. Larger weights correspond to observations with less outlier-like behavior, while smaller weights correspond to observations with more outlier-like behavior. The weight matrix $\mathbf{W}$ is a diagonal matrix, where each diagonal element corresponds to the weight assigned to a particular observation. In the case of an OLS estimator, we have $\mathbf{W} = \mathbf{I}_k$, as the weight given to each observation is not sensitive to the residual. In other words, $w_{LS}(e) = 1$, regardless of the specific residual observed. With a Huber estimator, $w_H(e) = 1$ if $|e| \leq k$ and $w_H(e) = k/|e|$ if $|e| > k$. Finally, with a Tukey model, $w_T(e) = 0$ if $|e| > k$ and to $w_T(e) = [1 - (e/k)^2]^2$ if $|e| \leq k$. Then, to select the most informative observations while seeking protection against outliers, instead of estimating a single threshold, we define a bounded area of interest for the unscaled prediction variance of an observation as follows

$$\text{P}_\text{D}\left(\Gamma_1 \leq \mathbf{x}_{l+1}^{\text{T}}\left(\mathbf{X}_l^{\text{T}}\mathbf{W}\mathbf{X}_l\right)^{-1}\mathbf{x}_{l+1} \leq \Gamma_2\right) = \alpha \tag{12}$$

4 | EXPERIMENTS

In the experiments, we evaluate the performance of the active learning strategies in four scenarios, according to the percentage of outliers affecting the data stream. We compare the bounded CDO strategy, coupled with OLS and robust estimators, to the norm-thresholding approach, standard CDO, and random sampling. When using random sampling, each time a new sample arrives, a number $r \sim U(0,\ 1)$ is generated and the data point is only selected if $r \geq 1 - \alpha$, where α represents the labeling or sampling rate. The sampling strategies based on the use of robust estimators select the most informative data points using the standard UPV, as in Equation (10). The results obtained with the weighted prediction variance, UPV_w, were very similar and are included in the Appendix B for completeness. All the approaches receive as input the same random design and then they iteratively collect labeled observations until the budget constraint B is met. The number of observations contained in the initial design is equal to $p + 2$, where p is the number of process variables. We analyzed both the case of the initial design being outliers-free and contaminated. The results assuming the presence of outliers also in the initial design are included in the Appendix C. For each simulated scenario, the ith observation for the process variables, here considered a row vector, is generated according to a joint multivariate normal distribution

$$\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0},\ \boldsymbol{\Sigma}_0) \tag{13}$$

where $\boldsymbol{\Sigma}_0$ is given by $\sigma_\mathbf{x}^2\mathbf{I}$. The corresponding response is obtained using

$$y_i = \mathbf{x}_i\,\boldsymbol{\beta} + \varepsilon_i,\ \text{where}\ \varepsilon_i \sim \mathcal{N}\left(0,\ \sigma_\varepsilon^2\right) \tag{14}$$

For normal data points, we used $\sigma_\mathbf{x} = \sigma_\varepsilon = 1$ for both input and output variables, and, for simulating outliers, we set $\sigma_\mathbf{x} = \sigma_\varepsilon = 3$. Moreover, for each of the true coefficients of the underlying model, we assumed $\beta \sim U(-5,\ 5)$ for normal data points and $\beta \sim U(10,\ 15)$ for outliers. As in Deldossi et al.,[19] the outliers are introduced in the data stream in the form of isolated covariate and concept shifts. That is, an anomalous data point is a point for which we have both a larger variation in the input space, and a different relationship with the corresponding response variable. In the simulated scenarios, outliers are randomly distributed in the data stream according to a pre-defined percentage describing the contamination

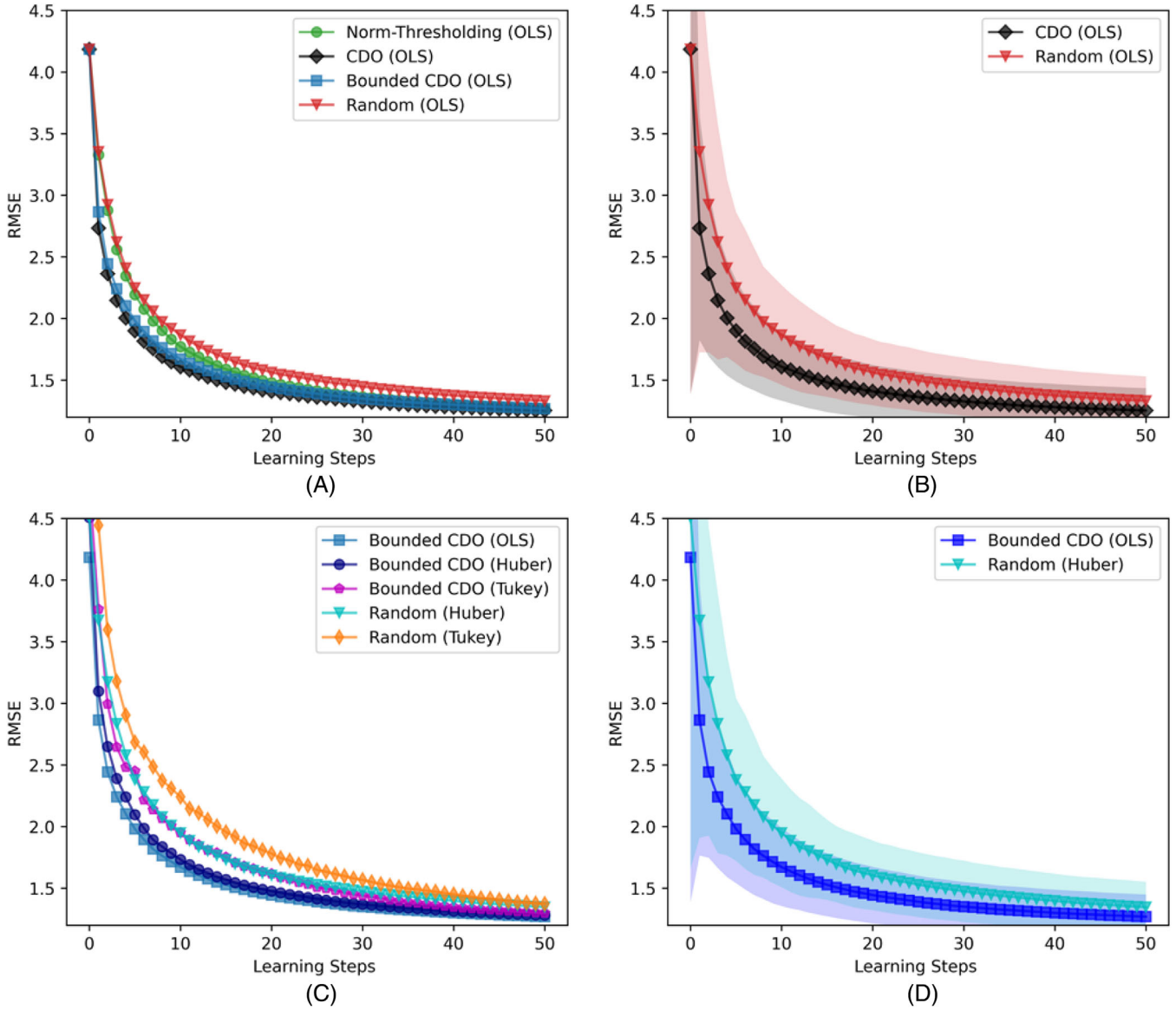

FIGURE 2 Comparing query strategies in the absence of outliers: results from 1000 simulations. Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

level of the environment. The performance of the models is expressed, in predictive terms, by the root mean squared error (RMSE) of the predictions on a separate test set, only composed of normal observations. This is coherent with the objective of trying to understand the true underlying relationship between predictors and response, and not the erroneous one that could be derived from the outliers.

The effectiveness of the proposed approach is evaluated by comparing the learning curves reporting the average RMSE values for each learning step, which are obtained using 1000 simulations for each scenario. A learning step indicates the acquisition of a new labeled observation and its inclusion in the training set. Hence, at each step, we are comparing models that are trained using the same number of labeled examples. We set the number of process variables equal to 20, the budget constraint B equal to 50, and the warm-up length m to 500. The warm-up length indicates the number of unlabeled observations that are used to estimate the covariance matrix $\mathbf{\Sigma}$ that is used for pre-whitening the observations. With regards to the sampling rate, we used $\alpha = 5\%$ for all the sampling strategies, and $c = 5\%$ for the protection cut-off value used by the bounded CDO algorithm. We selected 5% as it is a commonly employed value, especially when no previous specific knowledge is available.

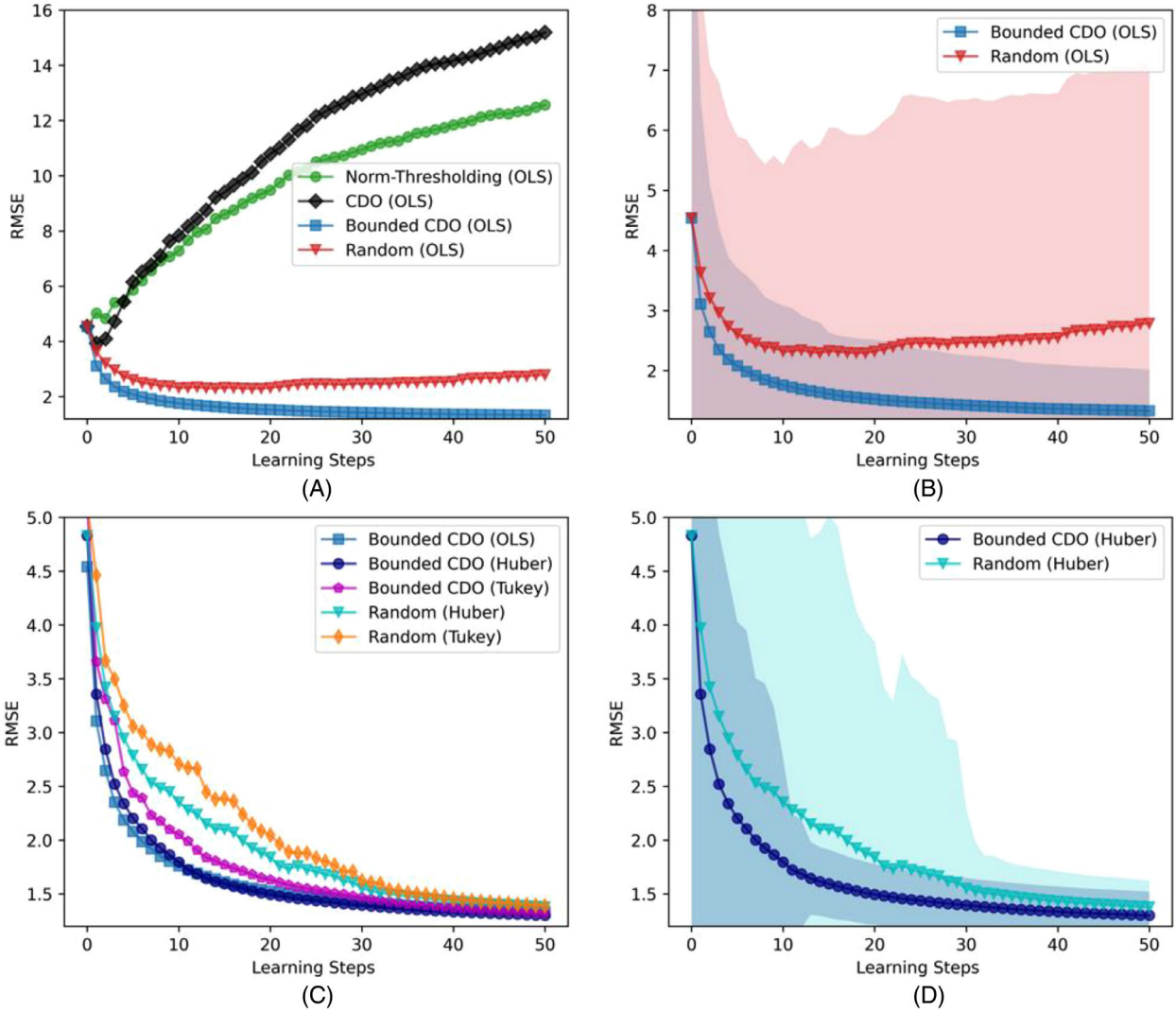

FIGURE 3 Comparing query strategies with 0.275% outliers (1000 simulations). Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

4.1 | No outliers

We first evaluated the query strategies to assess their performance in the absence of outliers. Consistently with the findings reported in our previous work,[16] our results in Figure 2 indicate that the standard CDO algorithm performs best when there are no outliers in the data stream. The use of robust estimators does not provide any added value in this scenario. Both the Huber and Tukey estimators are unable to outperform the bounded CDO strategy with the OLS model, which in turn is only marginally worse than the standard CDO. In Figure 2, plots (A) and (B) represent the strategies that rely on the OLS models, while plots (C) and (D) show the strategies that use robust models, with the bounded CDO based on OLS included for comparison.

4.2 | 0.275% outliers

The second scenario depicts a circumstance where only a modest fraction of the data stream is represented by outliers. We can see from the plot (A) of Figure 3 how the performance of the norm-thresholding and the CDO algorithm is dramatically worsened, as they are both prone to sample outliers. The random strategy seems to be a better option and the bounded CDO strategy offers the best results. In the plots (C) and (D) of the same figure, we can see the comparison with the

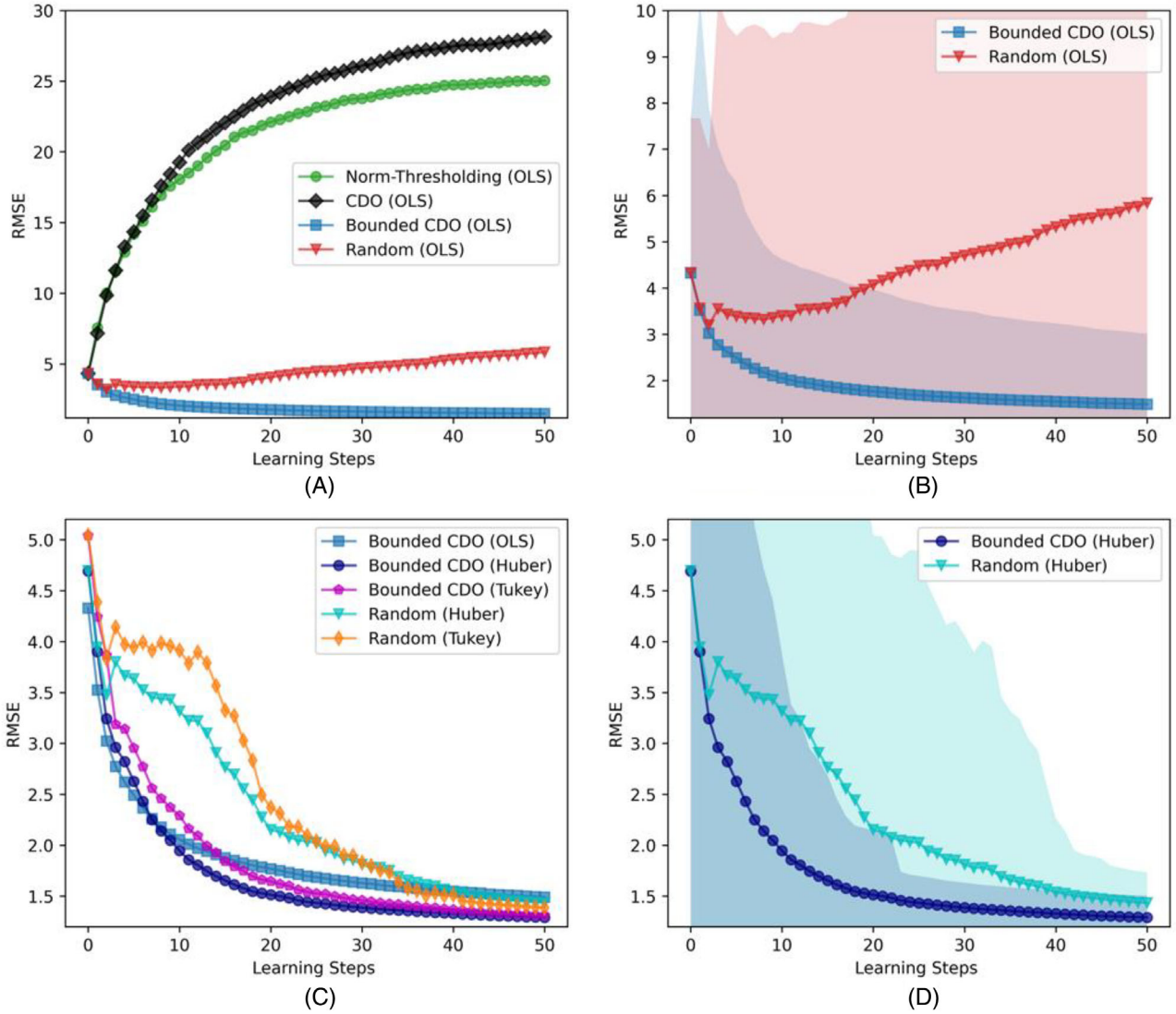

FIGURE 4 Comparing query strategies with 1% outliers (1000 simulations): results from 1000 simulations. Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

results obtained from the robust estimators. In this scenario, using a robust estimator does not seem to offer a significant improvement over the bounded CDO strategy based on OLS. Indeed, the learning curves obtained with the bounded strategy employing the OLS estimator and the Huber estimator are very similar.

4.3 | 1% outliers

The third scenario reports a worse situation, where the process is affected by a large number of outliers, that is, 1% of the total number of observations. The results in Figure 4 are similar to the ones from the previous scenario, with the exception that now the gap between bounded CDO and random sampling is much wider. This should be due to the fact that uniformly sampling observations with $\alpha = 5\%$ would most certainly lead to the inclusion of a greater number of outliers in the training set.

As per the robust estimators shown in the plots (C) and (D) of Figure 4, it is possible to see how the use of robust estimators now offers an evident value-added, also when compared to the OLS-based bounded CDO. While the learning curves are more or less overlapping in the first five learning steps, the models fitted using the Huber and Tukey losses are yielding a lower prediction error in the remaining steps.

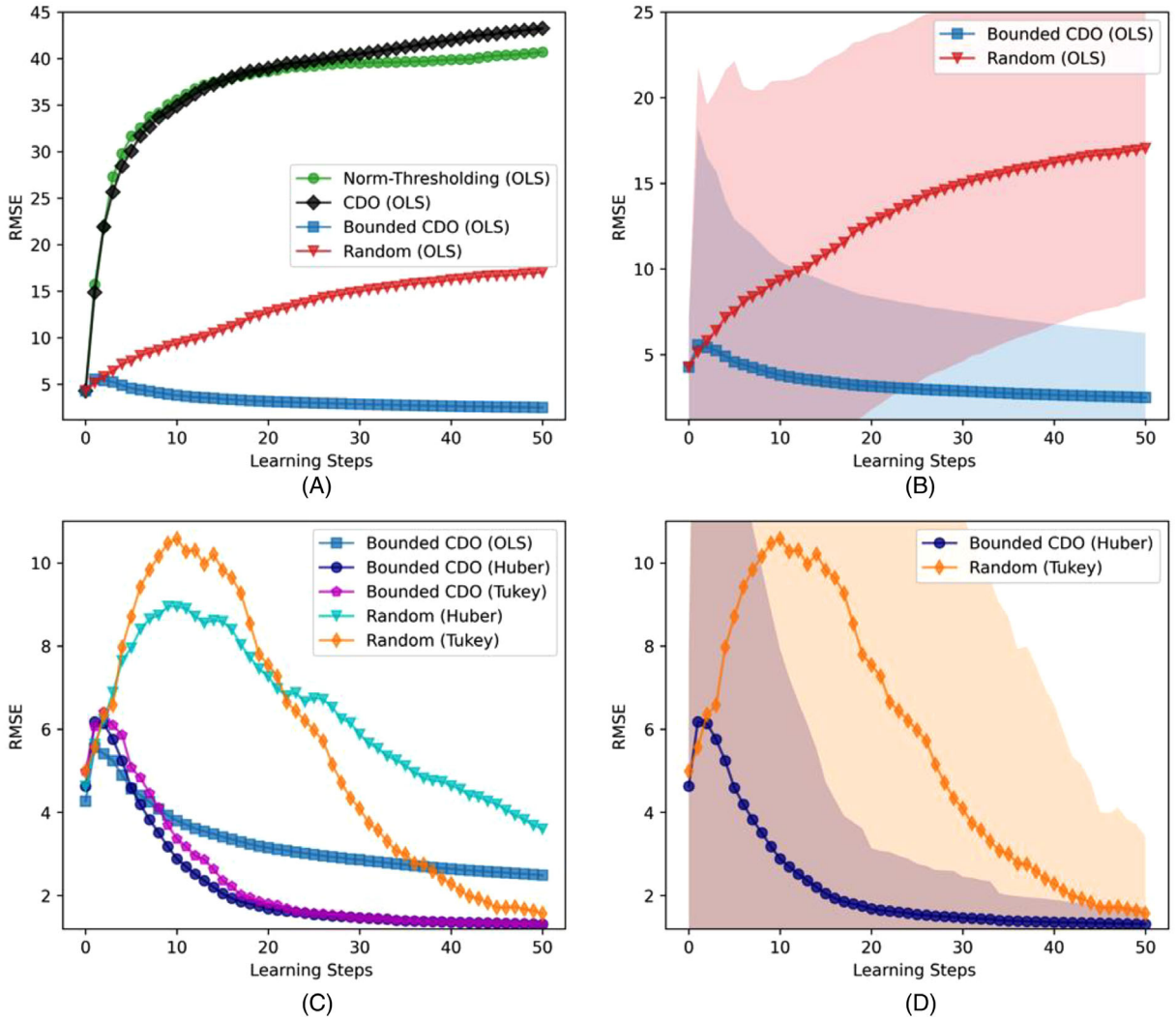

FIGURE 5 Comparing query strategies with 5% outliers (1000 simulations): results from 1000 simulations. Plots (B) and (D) offer a closer view on the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

4.4 | 5% outliers

The final scenario simulates a pathological case, where 5% of the observations from the data stream are outliers. The results from the third scenario are exacerbated here. In the case of the OLS estimators, the bounded CDO is still the best strategy, being the only one with a descending learning curve (plots (A) and (B) of Figure 5). Instead, from the plots (C) and (D) of Figure 5 we can see how the robust estimators are able to improve the results obtained with the bounded CDO strategy. In this circumstance, there is not a clear distinction between the Huber and the Tukey models.

5 | DISCUSSION

The experiments presented in this study aimed to evaluate the performance of different active learning strategies in the presence of outliers in a data stream. The results showed that the standard CDO algorithm performed best in the absence of outliers, while the bounded CDO strategy coupled with OLS and robust estimators provided better results when outliers were present. In scenarios where an initial training set free from outliers is available and only a modest fraction of the data stream is represented by outliers, the bounded CDO strategy employing an OLS estimator seems to be the better option. Conversely, in the case of a larger contamination level, sampling strategies based on robust estimators yield the best results.

When using robust estimators, for our datasets we did not find solid evidence that using a weighted prediction variance is an advantage. Another interesting observation is that, in the presence of outliers, the standard OLS methods (random, norm-thresholding, and CDO) never converge to the results obtained with the robust query strategies. This is because they tend to accumulate outliers in the training set, which degrade the predictive performance as the model is not allowed to forget old or redundant data. The findings from this study have important consequences for practical applications of active learning strategies, especially in contexts where the data stream is contaminated by outliers. The results suggest that the choice of the active learning strategy should depend on the level of contamination of the data stream. When the data stream is free from outliers, the standard CDO is a good strategy. However, even when a modest fraction of the observations is corrupted, bounding the search area of the active learning algorithm or using robust estimators might be necessary. Overall, this study provides valuable insights into the performance of active learning strategies in the presence of outliers and can inform the development of more effective approaches for real-world applications. However, it is worth noting that the simulations were based on specific assumptions about the data generation process and may not fully capture the complexity of real-world data streams. Further research is needed to validate these findings on real-world datasets and to investigate the generalizability of the proposed approach.

6 | CONCLUSIONS

In many real-world problems, data is only available in an unlabeled form, and acquiring the labels is often an expensive and time-consuming task. In these circumstances, active learning is able to reduce the computational burden required to achieve compelling predictive performance by selecting the most informative data points to query. In this paper, we analyze the online active learning framework when the data stream is corrupted by the presence of outliers. In general, we show how the presence of outliers dramatically worsens the performance of the currently proposed methods for active linear regression. To tackle this issue, we propose a modification of the CDO algorithm that filters the outliers, while still focusing on the most promising observations based on the concepts of D-optimality and prediction variance. The analysis shows how this solution is sufficient to make the CDO strategy robust to a modest presence of outliers. When the percentage of outliers in the data stream is higher, the best results are obtained by coupling the bounded CDO strategy with a robust estimator. In general, the proposed approaches can effectively solve the problem of outliers contaminating the data stream, without adding computational complexity compared to the original CDO strategy.

ACKNOWLEDGMENTS

This work has been funded by the DTU Strategic Alliances Fund.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Davide Cacciarelli https://orcid.org/0000-0001-6664-9038
Murat Kulahci https://orcid.org/0000-0003-4222-9631
John Sølve Tyssedal https://orcid.org/0000-0003-1628-4725

REFERENCES

1. Kumar P, Gupta A. Active learning query strategies for classification, regression, and clustering: a survey. *J Comput Sci Technol*. 2020;35:913-945. doi:10.1007/s11390-020-9487-4
2. Settles B. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Science. 2009.
3. Cacciarelli D, Kulahci M. A survey on online active learning. 2023. arXiv preprint 10.48550/arXiv.2302.08893
4. Chan LLT, Wu QY, Chen J. Dynamic soft sensors with active forward-update learning for selection of useful data from historical big database. *Chemom Intell Lab Syst*. 2018;175:87-103. doi:10.1016/j.chemolab.2018.01.015
5. Ge Z. Active learning strategy for smart soft sensor development under a small number of labeled data samples. *J Process Control*. 2014;24:1454-1461. doi:10.1016/j.jprocont.2014.06.015
6. Liu D, Zhang P, Zheng Q. An efficient online active learning algorithm for binary classification. *Pattern Recognit Lett*. 2015;68:22-26. doi:10.1016/j.patrec.2015.08.010

7. Bouguelia M-R, Belaïd Y, Belaïd A. An adaptive streaming active learning strategy based on instance weighting. *Pattern Recognit Lett*. 2016;70:38-44. doi:10.1016/j.patrec.2015.11.010
8. Lughofer E. Single-pass active learning with conflict and ignorance. *Evol Syst*. 2012;3:251-271. doi:10.1007/s12530-012-9060-7
9. Shan J, Zhang H, Liu W, Liu Q. Online active learning ensemble framework for drifted data streams. *IEEE Trans Neural Netw Learn Syst*. 2019;30:486-498. doi:10.1109/TNNLS.2018.2844332
10. Krawczyk B. Active and adaptive ensemble learning for online activity recognition from data streams. *Knowl Based Syst*. 2017;138:69-78. doi:10.1016/j.knosys.2017.09.032
11. Lughofer E. On-line active learning: a new paradigm to improve practical useability of data stream modeling methods. *Inf Sci (N Y)*. 2017;415-416:356-376. doi:10.1016/j.ins.2017.06.038
12. Lughofer E, Pratama M. Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models. *IEEE Trans Fuzzy Syst*. 2018;26:292-309. doi:10.1109/TFUZZ.2017.2654504
13. Riquelme C, Ghavamzadeh M, Lazaric A. Active learning for accurate estimation of linear models. Proceedings of the 34th International Conference on Machine Learning; 2017.
14. Riquelme C, Johari R, Zhang B. Online active linear regression via thresholding. Thirty-First AAAI Conference on Artificial Intelligence; 2017. www.aaai.org
15. Fontaine X, Perrault P, Valko M, Perchet V. Online a-optimal design and active linear regression. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.
16. Cacciarelli D, Kulahci M, Tyssedal JS. Stream-based active learning with linear models. *Knowl Based Syst*. 2022;254:109664. doi:10.1016/j.knosys.2022.109664
17. Melis DA, Jaakkola T. Towards robust interpretability with self-explaining neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Adv Neural Inf Process Syst*. Curran Associates, Inc; 2018. https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf
18. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32. doi:10.1214/009053604000000067
19. Deldossi L, Pesce E, Tommasi C. A sub-sampling algorithm preventing outliers. 2022. arXiv preprint http://arxiv.org/abs/2208.06218
20. Zhao J, Yi S, Liang Y, Liu W, Cao X. Robust active representation via ℓ2,p-norm constraints[Formula presented]. *Knowl Based Syst*. 2022;235:107639. doi:10.1016/j.knosys.2021.107639
21. He G, Li Y, Zhao W. An uncertainty and density based active semi-supervised learning scheme for positive unlabeled multivariate time series classification. *Knowl Based Syst*. 2017;124:80-92. doi:10.1016/j.knosys.2017.03.004
22. Cacciarelli D, Kulahci M. A novel fault detection and diagnosis approach based on orthogonal autoencoders. *Comput Chem Eng*. 2022;163:107853. doi:10.1016/j.compchemeng.2022.107853
23. Nguyen QP, Lim KW, Divakaran DM, Low KH, Chan MC. GEE: A gradient-based explainable variational autoencoder for network anomaly detection. IEEE Conference on Communications and Network Security (CNS); 2019. doi:10.1109/CNS.2019.8802833
24. Zhou C, Paffenroth RC, Autoencoders ADRD. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA; 2017. doi:10.1145/3097983.3098052
25. Ruff L, Kauffmann JR, Vandermeulen RA, et al. A unifying review of deep and shallow anomaly detection. *Proc IEEE*. 2021;109:756-795. doi:10.1109/JPROC.2021.3052449
26. Qiu C, Li A, Kloft M, Rudolph M, Mandt S. Latent outlier exposure for anomaly detection with contaminated data. Proceedings of the 39th International Conference on Machine Learning, PMLR 162; 2022. http://arxiv.org/abs/2202.08088
27. Burbidge R, Rowland JJ, King RD. Active learning for regression based on query by committee. Intelligent Data Engineering and Automated Learning; 2007:209-218.
28. Ge Z. Active probabilistic sample selection for intelligent soft sensing of industrial processes. *Chemom Intell Lab Syst*. 2016;151:181-189. doi:10.1016/j.chemolab.2016.01.003
29. st John RC, Draper NR. D-Optimality for regression designs: a review. *Technometrics*. 1975;17:15-23. doi:10.1080/00401706.1975.10489266
30. Myers RH, Montgomery D, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. Wiley Series in Probability and Statistics 2016. ISBN: 978-1-118-91601-8
31. Fox J, Weisberg S. Robust regression. An R and S-Plus Companion to Applied Regression. 2013.
32. Huber PJ. Robust estimation of a location parameter. *Ann Math Statist*. 1964;35:73-101.
33. Beaton AE, Tukey JW. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*. 1974;16:147-185. doi:10.1080/00401706.1974.10489171
34. Hoaglin DC, Welsch RE. The hat matrix in regression and ANOVA. *Am Stat*. 1978;32:17. doi:10.2307/2683469
35. Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. *Stat Sci*. 1986;1(3):379-393.
36. Fernandes MC, Covões TF, Pereira ALV. Improving evolutionary constrained clustering using Active Learning. *Knowl Based Syst*. 2020;209:106452. doi:10.1016/j.knosys.2020.106452
37. Leng Y, Xu X, Qi G. Combining active learning and semi-supervised learning to construct SVM classifier. *Knowl Based Syst*. 2013;44:121-131. doi:10.1016/j.knosys.2013.01.032
38. Frumosu FD, Kulahci M. Big data analytics using semi-supervised learning methods. *Qual Reliab Eng Int*. 2018;34:1413-1423. doi:10.1002/qre.2338
39. Cacciarelli D, Kulahci M, Tyssedal J. Online active learning for soft sensor development using semi-supervised autoencoders. *ICML 2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*. 2022. https://arxiv.org/abs/2212.13067

40. Pullar-Strecker Z, Dost K, Frank E, Wicker J. Hitting the target: stopping active learning at the cost-based optimum. *Mach Learn*. 2022. doi:10.1007/s10994-022-06253-1
41. Zhang Y, Cai W, Wang W, Zhang Y. Stopping criterion for active learning with model stability. *ACM Trans Intell Syst Technol*. 2017;9:1-26. doi:10.1145/3125645
42. Ishibashi H, Hino H. Stopping criterion for active learning based on deterministic generalization bounds. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 108; 2020.
43. Ghayoomi M. Using variance as a stopping criterion for active learning of frame assignment. Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing. 2010:1-9.
44. Laws F, Schütze H, Schütze S. Stopping criteria for active learning of named entity recognition. Proceedings of the 22nd International Conference on Computational Linguistics. 2008;108:465-472
45. Zhu J, Wang H, Hovy E. Multi-criteria-based strategy to stop active learning for data annotation. Proceedings of the 22nd International Conference on Computational Linguistics. 2008;108:1129-1136.
46. Ren P, Xiao Y, Chang X, et al. A survey of deep active learning. *ACM Comput Surv*. 2022;54:1-40. doi:10.1145/3472291
47. Bloodgood M, Vijay-Shanker K. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), 2009:39-47.
48. Tomanek K, Hahn U. Approximating learning curves for active-learning-driven annotation, n.d. http://www.ncbi.nlm.nih.gov/
49. Farquhar S, Gal Y, Rainforth T. On statistical bias in active learning: how and when to fix it. International Conference on Learning Representations (ICLR); 2021.

How to cite this article: Cacciarelli D, Kulahci M, Tyssedal JS. Robust online active learning. *Qual Reliab Engng Int*. 2024;40:277–296. https://doi.org/10.1002/qre.3392

APPENDIX A: STOPPING CRITERION

In real-world applications of active learning, if we do not have an explicit operational budget on the number of experiments that can be run, it can be challenging to determine when to stop collecting new labels due to the unavailability of the true learning curves. To address this problem, it is beneficial to approximate the learning curve using proxy measures. In this study, we investigate the use of two proxy measures. Firstly, we propose monitoring the slope of the stabilization score, drawing inspiration from the stabilizing predictions[47] and validation set agreement[48] methods employed in classification. In the regression framework, we calculate the stabilization of predictions by averaging the sum of squares of the differences between the predictions of the w most recent pairs of models. Similarly to Bloodgood and Vijay-Shanker,[47] we utilize a window size of 3 ($w = 3$). The values being compared are the predicted values of the calibration set **C**, obtained through successive models. As the examples in **C** are not used in the annotation process, this curve is solely influenced by the impact of selected and labeled examples on training new models. Essentially, this curve monitors when the predictions from models trained with newly included observations start producing highly similar results. The stopping rule can then be determined through visual inspection of the curve, by setting a tolerance for the sum of squares not improving or

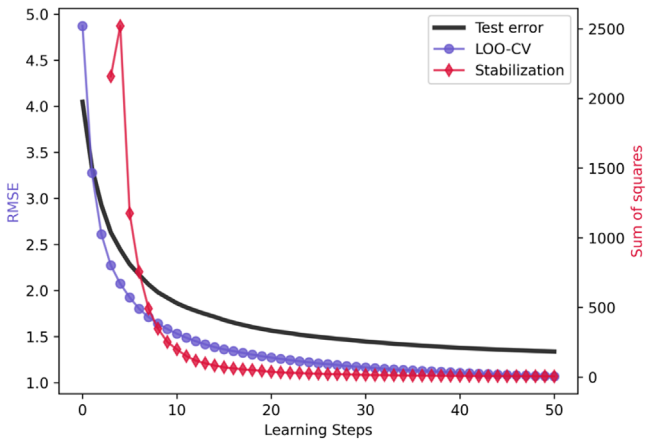

FIGURE 6 Approximating the learning curve: random sampling with no outliers (1000 simulations). The left axis reports the RMSE value for the curves related to the test error and the LOO-CV. The right axis shows the average sum of squares related to the stabilization score.

approaching zero, or by applying a hypothesis testing procedure. Another performance-based metric we consider is the leave-one-out cross-validation (LOO-CV) score obtained by the model on the currently available labeled observations. While this technique relies on ground-truth labels and may appear advantageous, it may not be the optimal choice if the collected training set is biased or does not accurately represent the real data distribution.[49] On the other hand, the stabilization score, despite not relying on real labels, could be more reliable if the calibration set **C** follows the population distribution. Figure 6 demonstrates the effectiveness of the two proposed methods in approximating the true test error curve, offering valuable insights for determining when to halt the active learning routine.

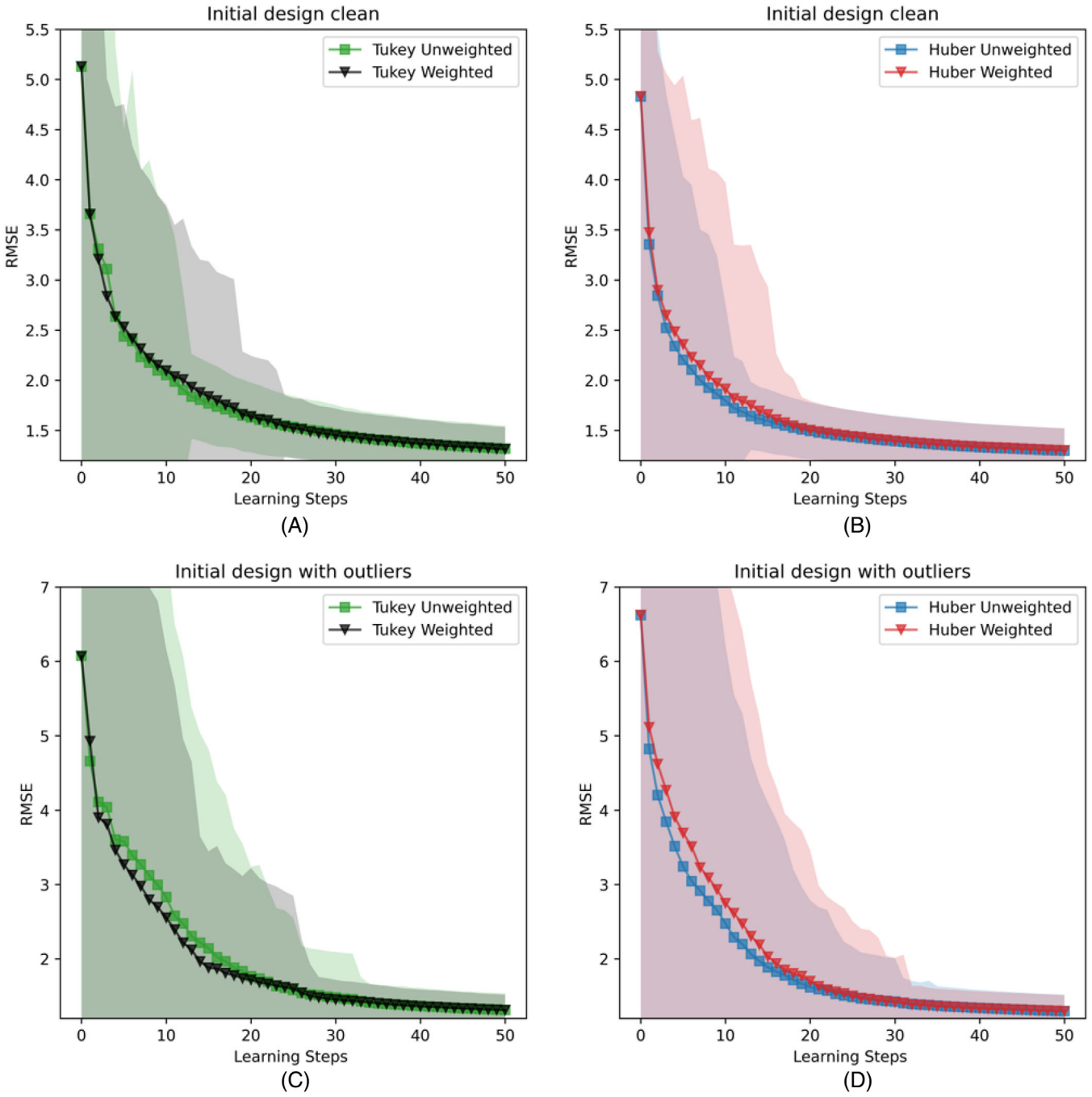

FIGURE 7 Comparing UPV and UPV_w in the scenario with 0.275% outliers (1000 simulations).

APPENDIX B: WEIGHTED PREDICTION VARIANCE

In this section, we examine the impact of switching from the standard UPV to its weighted version on the learning curves of the robust bounded CDO strategies. While it may seem reasonable to use a weighted prediction variance from a theoretical standpoint, we found little compelling evidence that it improves performance even with the use of robust estimators (Figures 7–9). In fact, we observed that using the UPV_w actually worsens results when the initial design is free from outliers. This could be because the robust models mistakenly identify some observations as outliers, resulting in $\mathbf{W} \neq \mathbf{I}_k$.

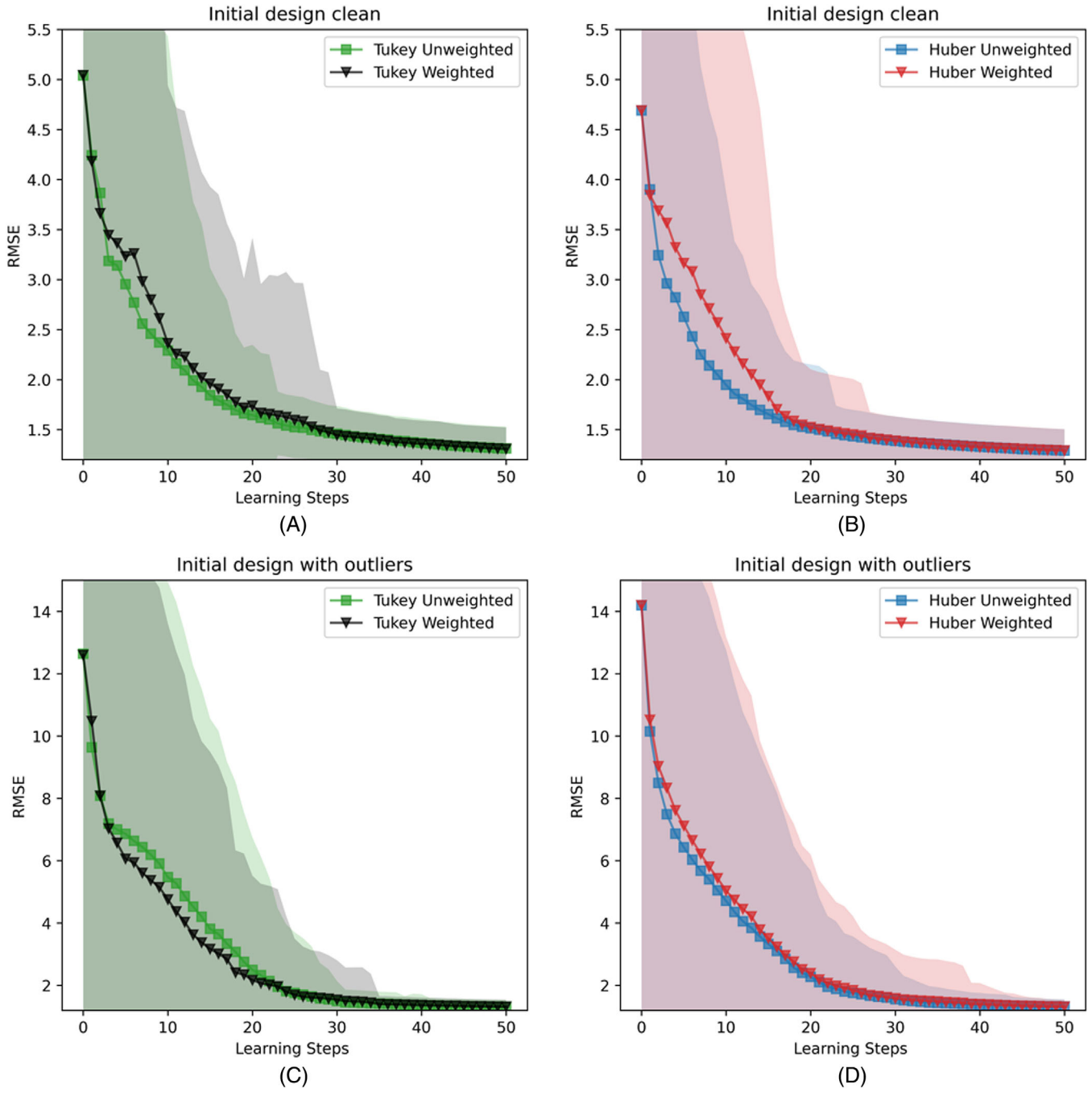

FIGURE 8 Comparing UPV and UPV_w in the scenario with 1% outliers (1000 simulations).

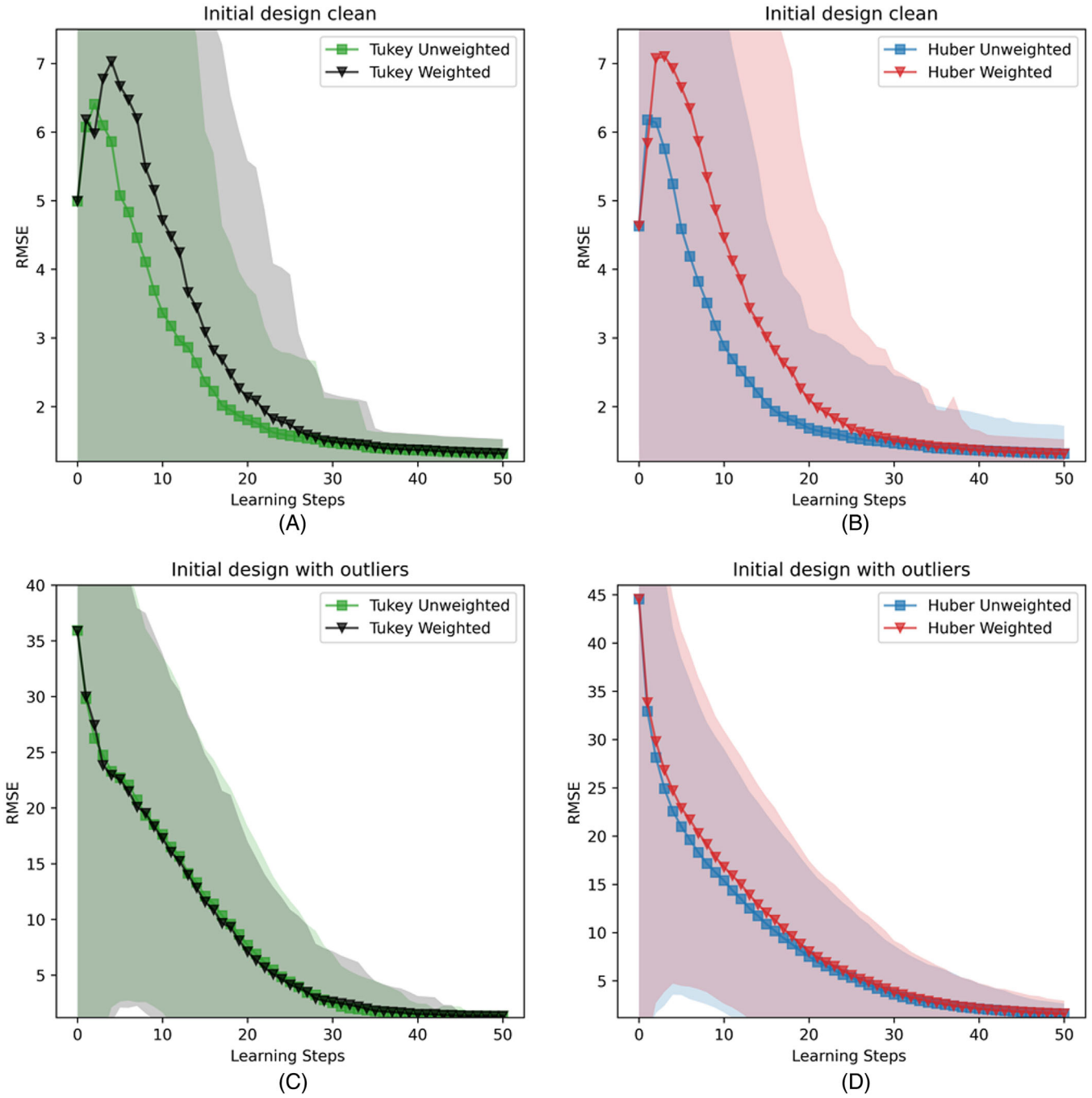

FIGURE 9 Comparing UPV and UPV_w in the scenario with 5% outliers (1000 simulations).

APPENDIX C: PRESENCE OF OUTLIERS IN THE INITIAL DESIGN

In Figures 10–12, we investigate the impact of removing the assumption that the initial design is free from outliers on the sampling strategies. Despite the small size of the initial design when $p = 20$, we observed several notable behaviors. One of the most noticeable differences is that the learning curves start with higher errors, as there are outliers forcibly included in the data. However, over time, the learning curves of the robust strategies are able to converge to satisfactory predictive performance as they can minimize the impact of these observations on the model training. In contrast, the OLS-based bounded CDO performs significantly worse in this scenario. This is because estimating the cutoff value Γ_2 using a contaminated set does not provide adequate protection against the inclusion of outliers in the design.

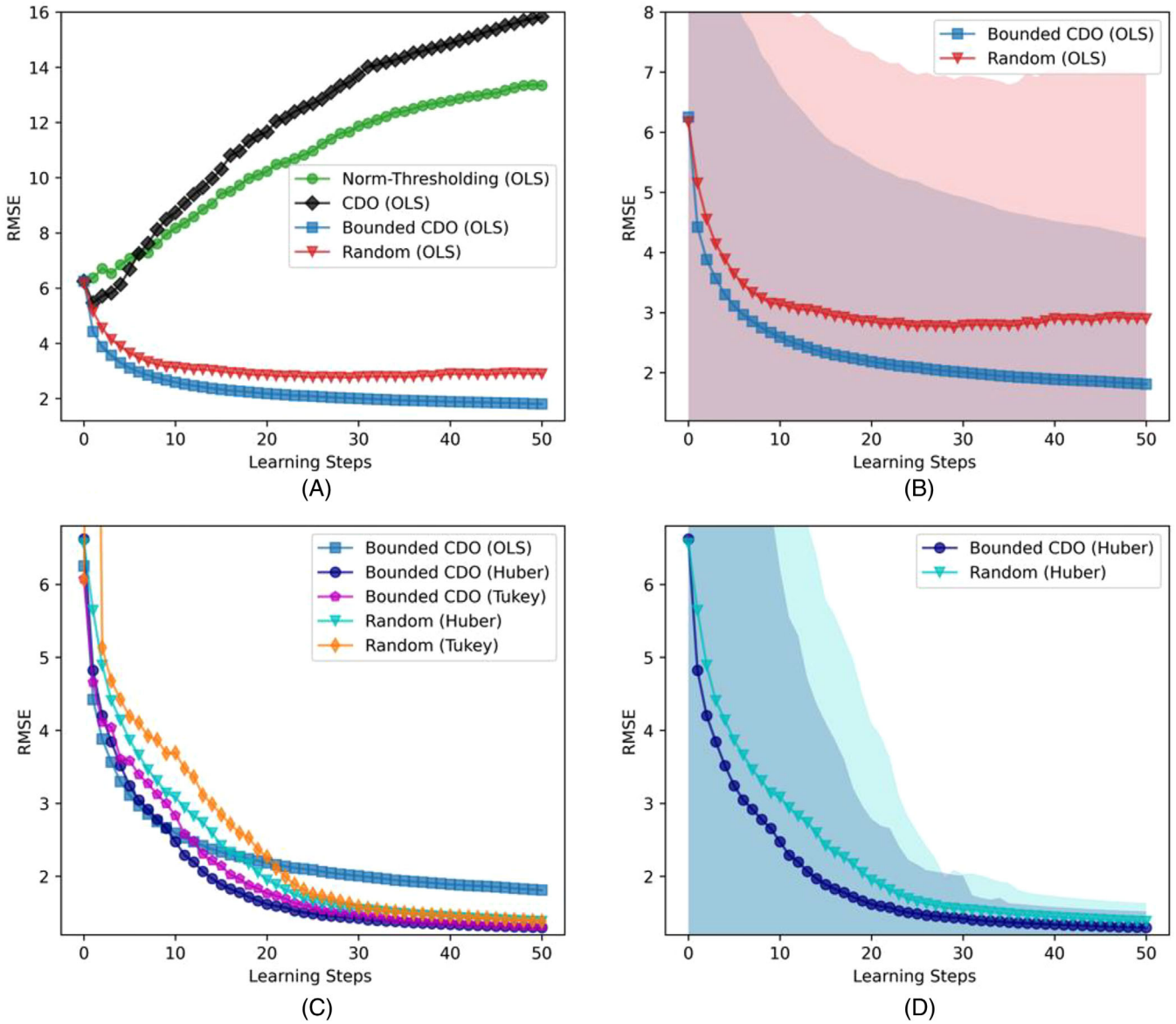

FIGURE 10 Comparing query strategies with 0.275% outliers (1000 simulations): results from 1000 simulations. Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

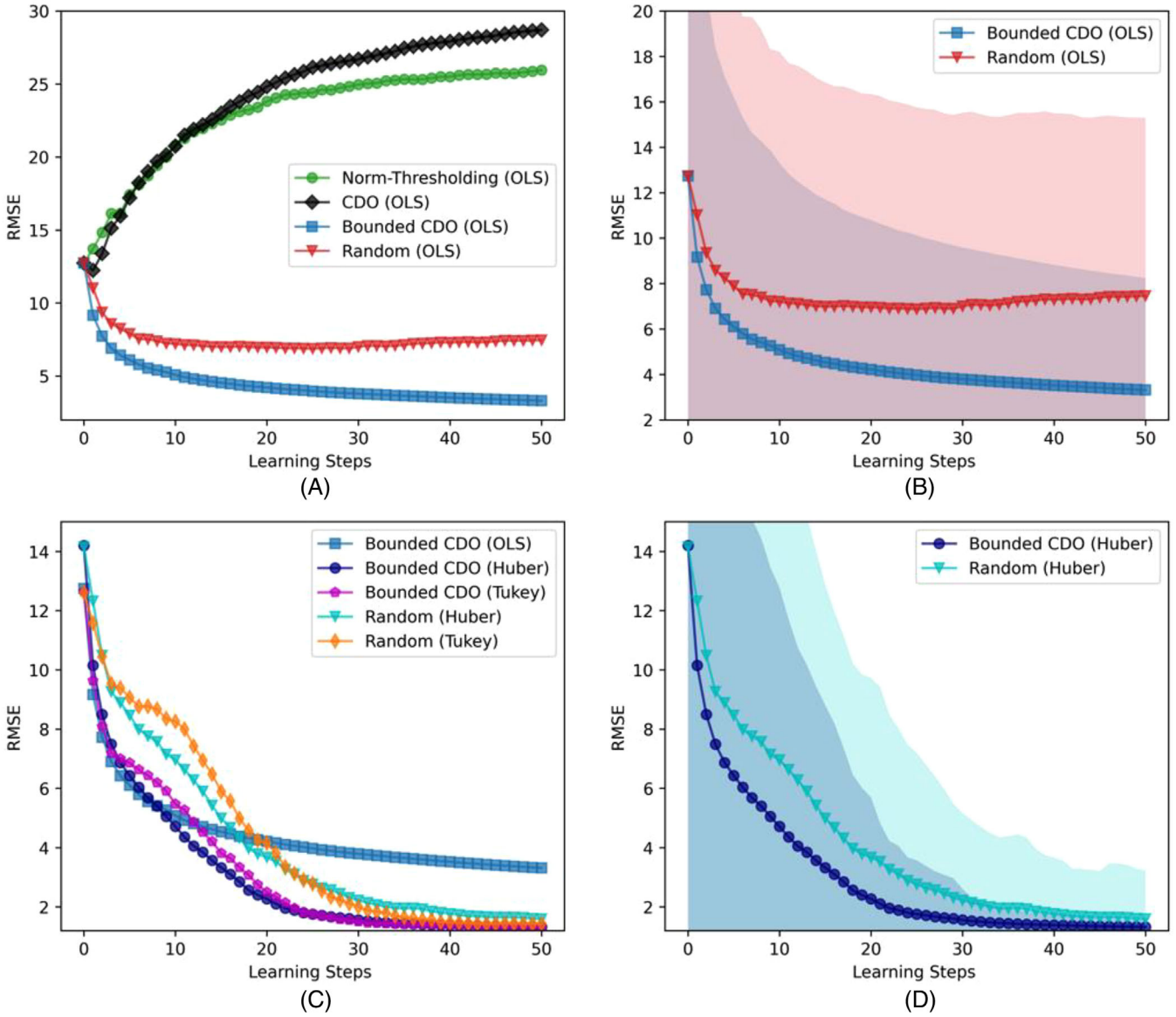

FIGURE 11 Comparing query strategies with 1% outliers (1000 simulations): results from 1000 simulations. Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

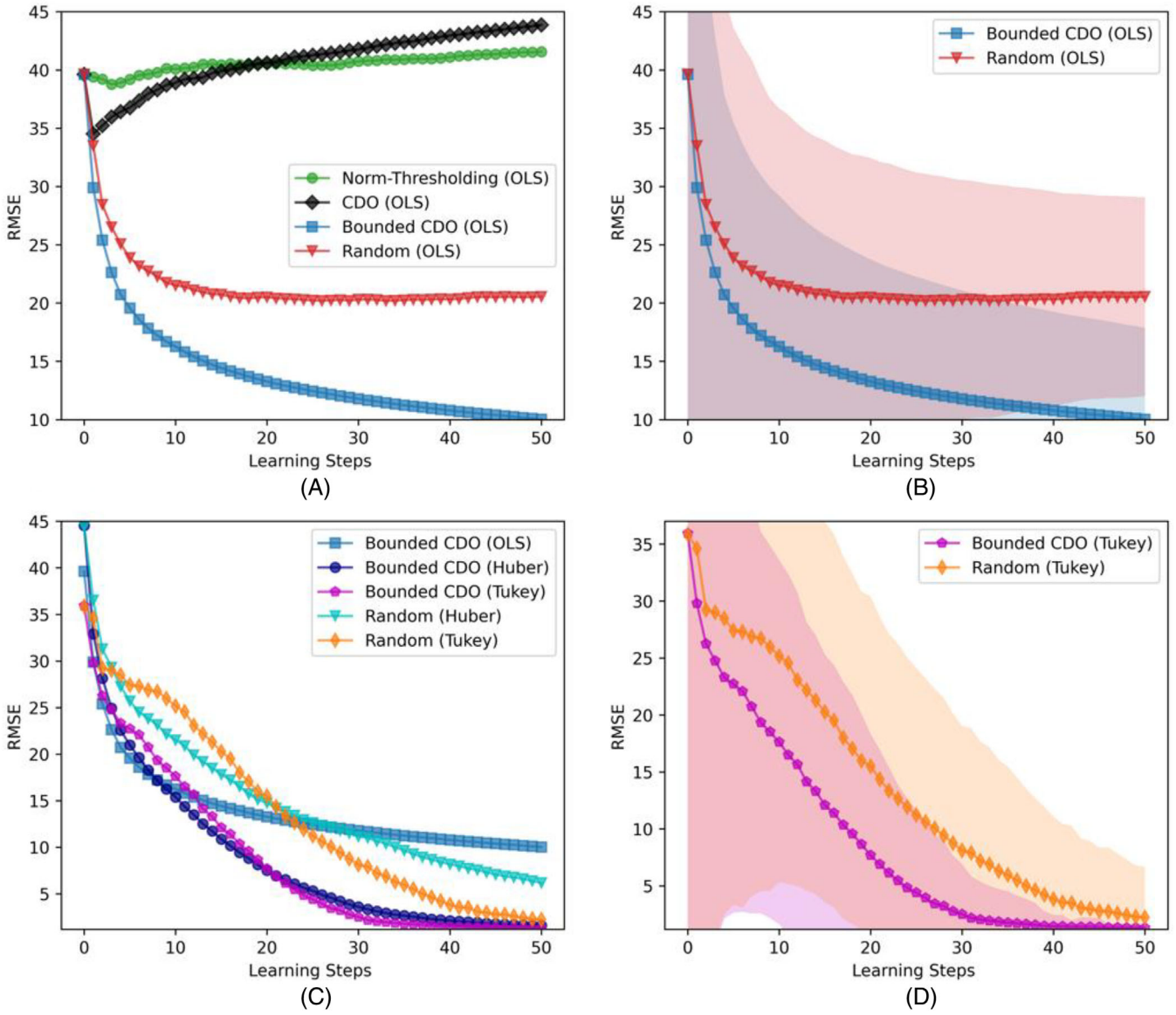

FIGURE 12 Comparing query strategies with 5% outliers (1000 simulations): results from 1000 simulations. Plots (B) and (D) offer a closer view of the two best strategies from plots (A) and (C), respectively, with shaded regions indicating the standard deviation across the simulations.

AUTHOR BIOGRAPHIES

Davide Cacciarelli is a PhD student at the Technical University of Denmark and the Norwegian University of Science and Technology. His research is related to active learning and statistical process monitoring.

Murat Kulahci is a Professor at the Technical University of Denmark and Luleå University of Technology. His research focuses on the design of physical and computer experiments, statistical process monitoring, time series analysis and forecasting, and financial engineering.

John Sølve Tyssedal is a Professor at the Norwegian University of Science and Technology. His research interests include design of experiments, statistical process control and time series analysis.