JAWS-X: Addressing Efficiency Bottlenecks of Conformal Prediction Under Standard and Feedback Covariate Shift

Drew Prinster [1] **Suchi Saria** [1 2] **Anqi Liu** [1]

Abstract

We study the efficient estimation of predictive confidence intervals for black-box predictors when the common data exchangeability (e.g., i.i.d.) assumption is violated due to potentially feedback-induced shifts in the input data distribution. That is, we focus on standard and feedback covariate shift (FCS), where the latter allows for feedback dependencies between train and test data that occur in many decision-making scenarios like experimental design. Whereas prior conformal prediction methods for this problem are in general either extremely computationally demanding or make inefficient use of labeled data, we propose a collection of methods based on the jackknife+ that achieve a practical balance of computational and statistical efficiency. Theoretically, our proposed JAW-FCS method extends the rigorous, finite-sample coverage guarantee of the jackknife+ to FCS. We moreover propose two tunable relaxations to JAW-FCS's computation that maintain finite-sample guarantees: one using only K leave-one-out models (JAW-KLOO) and a second building on K-fold cross validation+ (WCV+). Practically, we demonstrate that JAW-FCS and its computational relaxations outperform state-of-the-art baselines on a variety of real-world datasets under standard and feedback covariate shift, including for biomolecular design and active learning tasks.

1. Introduction

To safely and effectively deploy machine learning (ML) systems in high-stakes decision making, a promising approach is to communicate to users whether a given prediction can be trusted via reliable predictive uncertainty quantification. The use of standard approaches to ML uncertainty quantification (UQ) in practice, however, is often hindered by at least two key challenges. Firstly, real-world scenarios often involve data shifts that violate the common ML assumption that "the future will look like the past" (Finlayson et al., 2021)—indeed, even the mere use of ML-generated insights to inform future decisions can induce feedback-loop shifts between training (e.g., development) and test (e.g., deployment) data that invalidate standard UQ methods (Fannjiang et al., 2022). Secondly, even if data shift is accounted for, many UQ methods are too resource-demanding to implement without sacrificing overall model performance. For instance, many UQ methods have extreme or prohibitive computational demands (e.g., extensive retraining of large ML predictors), while others impose strict data-availability requirements that can be unrealistic (e.g., requiring infinite data in "asymptopia") or harm model performance (e.g., requiring sample splitting to form a "holdout" UQ dataset that cannot be used in training, which degrades accuracy relative to if all labeled data were used to learn parameters).

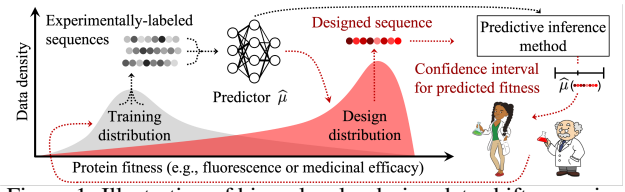

Figure 1: Illustration of biomolecular design data shift scenario.

For a concrete example, take an experimental design scenario, where a biomolecular engineer aims to propose a novel protein sequence with high "fitness"—strong expression of a desired property such as fluorescence or medicinal efficacy—using predictions of an ML model trained on a protein dataset with experimentally-labeled fitness values. This biomolecular design problem often requires leveraging UQ to balance exploring novel sequences that are "far" from the training data with exploiting "close" sequences whose fitness values are already estimated with high confidence. However, by selecting (for downstream experimental design) novel protein sequences according to the models' predicted fitness, the engineer induces a dependency between the training and test (designed sequence) distributions (Fig-

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA [2]Bayesian Health, New York, NY, USA. Correspondence to: Drew Prinster <drew@cs.jhu.edu>, Suchi Saria <ssaria@cs.jhu.edu>, Anqi Liu <aliu@cs.jhu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Table 1: Summary of key properties for JAWS-X and baseline predictive inference methods for standard and feedback covariate shift.

"Proposed" or reference if baseline	**Method name**	**Finite-sample coverage guarantee for standard & feedback covariate shift**	**Statistical efficiency: No sample splitting**	**Computational cost: Total # trained black-box predictors for n train & m test points, $\mathcal{Y}$ label space**
Tibshirani et al. (2019)	weighted split CP	✓	✗	1
Fannjiang et al. (2022)	full CP-FCS	✓	✓	$(n+1)\cdot m\cdot \lvert\mathcal{Y}\rvert$
Proposed	JAW-FCS	✓	✓	n
Proposed	JAW-KLOO	✓	✓	$K \leq n$
Proposed	WCV+	✓	✓	$K \leq n$

ure 1), which violates the common UQ assumption of data being independent and identically distributed (i.i.d.). Computational and statistical (data-use) efficiency are critical in this setting: the scale of modern, nonlinear ML models imposes computational budget constraints, while the expensive process of labeling new protein sequences via experimentation makes paramount the economical use of available data. Similar examples could be given in other decision-making scenarios, including active learning, ML for scientific discovery, and safe exploration in reinforcement learning.

Distribution-free predictive inference under standard and feedback covariate shift Within UQ we focus on wrapper methods for *distribution-free predictive inference*, by which we refer to computing predictive confidence intervals (or sets, more generally) around black-box ML predictions without any assumptions about the parametric family of the data distribution. The first necessary property for a predictive interval is reliable *coverage*, for example meaning that a 90% predictive confidence interval actually "covers" the true target label with at least 90% frequency. Secondarily, without sacrificing coverage, predictive intervals are more useful when they are smaller (i.e., more informative), which is often a byproduct of statistical efficiency.

Conformal prediction (CP) (Vovk et al., 2005) is a framework for distribution-free predictive inference that achieves finite-sample coverage guarantees, traditionally for any i.i.d. (or exchangeable, more generally) samples. Full CP (Vovk et al., 2005) is the most data-efficient CP variant but with notorious computational demands, whereas split CP (Papadopoulos et al., 2002) is the computationally cheapest but least statistically efficient CP variant due to sample splitting. Tibshirani et al. (2019) extend full and split CP to standard covariate shift (SCS). Fannjiang et al. (2022) further allow for a type of feedback-loop data shift they call feedback covariate shift (FCS), which characterizes the biomolecular design problem as a special case (see Section 2.4). In particular, the main contribution of Fannjiang et al. (2022) extends full CP to FCS, and secondarily Fannjiang et al. (2022) describe how the weighted split CP method of Tibshirani et al. (2019) de factor maintains its guarantee under FCS. However, these two prior methods with coverage guarantees for distribution-free predictive inference under FCS—full CP-FCS (Fannjiang et al., 2022) and weighted split CP (Tibshirani et al., 2019)—inherit corresponding computational and statistical limitations from standard CP: For an arbitrary predictor, full CP-FCS is extremely (and often prohibitively) burdensome to compute, while weighted split CP suffers from sample-splitting statistical inefficiencies (see Table 1).

As an alternative to the computational-statistical tradeoff poles of (weighted) full and split CP under SCS, Prinster et al. (2022) develop a collection of methods based on the jackknife+ (Barber et al., 2021) called JAWS, which can offer favorable computational and statistical efficiency compromises (see Sections 2.3 & 2.6). JAWS is based on the JAW method, the **ja**ckknife+ **w**eighted with likelihood-ratio weights for SCS (hereon JAW-SCS for clarity), which extends the finite-sample coverage guarantee of the jackknife+ to SCS. The JAWS framework does not allow for the train-test dependencies present in FCS, however, and moreover the JAW-SCS method can still be expensive to compute.

In this work, we propose JAWS-X, a collection of methods for distribution-free predictive inference with finite-sample guarantees under SCS or FCS that flexibly and favorably balance statistical and computational efficiency. Building on the JAWS framework (Prinster et al., 2022), the letter "X" partly alludes to the "cross" of our weighted *cross*-validation+ method, along with the e**x**tension to FCS. Table 1 summarizes key properties of our methods and baselines.

Our contributions can be summarized as follows:

- We propose JAW-FCS, a first method for distribution-free predictive inference under FCS that can favorably balance statistical and computational efficiency. This method generalizes both the jackknife+ (Barber et al., 2021) and JAW-SCS (Prinster et al., 2022) methods to feedback covariate shift (FCS) while achieving the same rigorous, finite-sample coverage guarantee.
- We propose two computational relaxations of JAW-FCS that apply to both the SCS and FCS settings. The first approach, JAW-KLOO, leverages only $K \leq n$ leave-one-out models while maintaining the JAW-FCS guarantee. The second approach, K-fold WCV+, generalizes

K-fold cross validation+ (CV+) developed in (Barber et al., 2021) to SCS and FCS with a slightly weaker finite-sample coverage guarantee.

- Empirically, we demonstrate that JAW-FCS and its computational relaxations outperform state-of-the-art baselines on a variety of real-world datasets under SCS and FCS, including for protein design and active learning tasks. In particular, JAW-FCS and K-fold WCV+ maintain target coverage levels under SCS and FCS with better performing black-box predictors and sharper (more informative) predictive intervals than baselines.

2. Background and Related Work

2.1. Predictive Inference Preliminaries

We assume a multiset of training data $Z_{1:n} = \{Z_1, ..., Z_n\} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ and a test point $Z_{n+1} = (X_{n+1}, Y_{n+1})$ with unknown label Y_{n+1}, where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for all $i \in \{1, ..., n+1\}$ (and for a standard regression setup $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$). Moreover, let $\widehat{\mu} = \mathcal{A}(\{(X_1, Y_1), ..., (X_n, Y_n)\})$ denote a black-box predictor of interest, where $\mathcal{A}$ is a model-fitting algorithm. Then, a predictive interval is a function $\widehat{C}_{n,\alpha} : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$ that maps a test point X_{n+1} to an interval $\widehat{C}_{n,\alpha}(X_{n+1})$ around the prediction $\widehat{\mu}(X_{n+1})$, for some significance level $\alpha \in (0, 1)$. A coverage guarantee states that $\widehat{C}_{n,\alpha}(X_{n+1})$ is guaranteed to contain the true label Y_{n+1} with high probability, such as satisfying

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha \tag{1}$$

for all $\alpha \in (0, 1)$. It is important to note that we focus on marginal rather than conditional coverage (see Foygel Barber et al. (2021) for more on this distinction).

2.2. Standard Conformal Prediction

Conformal prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2008) is a principled and increasingly popular framework for distribution-free predictive inference; see Angelopoulos & Bates (2021) for a gentle introduction. Standard CP methods rely on the assumption of *exchangeability*, meaning that the distribution of the training and test data is invariant to permutations (i.i.d. is a special case); additionally, non-holdout-set CP methods require that the fitting algorithm $\mathcal{A}$ treat the training data symmetrically (Barber et al., 2022). CP methods use a fitted score function $\widehat{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to quantify the extent to which labeled points "conform" to previous data (e.g., the residual score $\widehat{S}(x, y) = |y - \widehat{\mu}(x)|$), and CP intervals are then constructed from subsets of $\mathcal{Y}$ whose scores lie within a (conservative) quantile on the empirical distribution of score values.

Full (or transductive) CP (Vovk et al., 2005) and split (or inductive) CP (Papadopoulos et al., 2002; Papadopoulos, 2008) are the two main types of standard CP. Together these methods represent polar-opposite ends on the computational-statistical efficiency tradeoff spectrum, with full CP being the most statistically efficient but computationally burdensome, and split CP being the computationally cheapest but with the least efficient use of available data (due to sample splitting to form a holdout set that cannot be trained on).

2.3. Jackknife+ and Cross Validation+

The jackknife+ and cross validation+ (CV+) methods of Barber et al. (2021), which are closely related to cross-conformal prediction (Vovk, 2015; Vovk et al., 2018), offer a range of intermediate, often beneficial compromises between the computational-statistical tradeoff extremes of full and split CP. The jackknife+ is a modified version of classic leave-one-out or "jackknife" resampling (Miller, 1974; Steinberger & Leeb, 2018; 2016), which requires rerunning the training algorithm $\mathcal{A}$ a total of n times, once for each leave-one-out predictor. Assuming exchangeable data and a symmetric algorithm $\mathcal{A}$ (as in standard full CP), Barber et al. (2021) prove that jackknife+ satisfies a slightly weaker coverage guarantee than standard CP methods, namely

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha. \tag{2}$$

However, jackknife+ typically achieves target $(1 - \alpha)$ empirical coverage for exchangeable data (Barber et al., 2021).

CV+ offers a computational relaxation of jackknife+ to allow for retraining $K \leq n$ predictors that each withhold $\frac{n}{K}$ datapoints from training, where n is assumed to be divisible by K (Barber et al., 2021); CV+ can thus be understood as a modification to K-fold cross validation resampling. For each training point $i \in \{1, ..., n\}$, let $k(i) \in \{1, ..., K\}$ denote the index of a cross-validation fold, where the multisets of points in each fold are denoted $\{S_1, ..., S_K\}$. Let $\widehat{\mu}_{-S_k} = \mathcal{A}\big((X_i, Y_i) : i \in \{1, ..., n\} \backslash S_k\big)$ denote the predictor trained with the k-th cross-validation fold S_k removed, and denote the residuals for the model $\widehat{\mu}_{-S_k}$ applied to points in its left-out fold S_k as $R^{CV}_i = \big|\widehat{\mu}_{-S_{k(i)}}(X_i) - Y_i\big|$ for $i : k(i) = k$. Then, the CV+ interval is defined as

$$\begin{aligned}
&\widehat{C}^{\text{CV+}}_{n,K,\alpha}(x) = \\
&\Big[Q_\alpha\Big(\sum_{i=1}^{n} \big[\tfrac{1}{n+1}\delta_{\widehat{\mu}_{-S_{k(i)}}(x) - R^{CV}_i}\big] + \tfrac{1}{n+1}\delta_{-\infty}\Big), \\
&Q_{1-\alpha}\Big(\sum_{i=1}^{n} \big[\tfrac{1}{n+1}\delta_{\widehat{\mu}_{-S_{k(i)}}(x) + R^{CV}_i}\big] + \tfrac{1}{n+1}\delta_{+\infty}\Big)\Big],
\end{aligned} \tag{3}$$

where δ_v denotes a point mass at v and $Q_\beta(\cdot)$ denotes the level β empirical quantile function. Note that the special case of $K = n$ recovers jackknife+. Barber et al. (2021) provide a finite-sample coverage guarantee for the CV+ (under exchangeability) that depends on K, where the strength of the guarantee approaches that of jackknife+ (2) as $K \to n$.

2.4. Standard and Feedback Covariate Shift

Under the standard covariate shift (SCS) assumption, the conditional $Y \mid X$ distribution is assumed to be the same between training and test data but the marginal X distributions may change (Sugiyama et al., 2007; Shimodaira, 2000). Crucially, SCS assumes that the test data distribution is independent of the training data:

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_X \times P_{Y|X}, i = 1, ..., n$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P}_X \times P_{Y|X}, \textit{independently}. \quad (4)$$

Feedback covariate shift (FCS) as described by Fannjiang et al. (2022), however, can be understood as a generalization of SCS where the marginal distribution of the test inputs may depend on the realization of the training data $Z_{1:n}$:

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_X \times P_{Y|X}, i = 1, ..., n$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P}_{X;Z_{1:n}} \times P_{Y|X}. \quad (5)$$

FCS thus characterizes the biomolecular design scenario from the introduction (Figure 1), where a predictor trained on i.i.d. protein training data $Z_{1:n}$ shifts the distribution of designed sequences $\widetilde{P}_{X;Z_{1:n}}$, and where, as a property determined by nature, $Y \mid X$ is expected to remain invariant.

2.5. Conformal Prediction for SCS and FCS

Tibshirani et al. (2019) develop a weighted generalization of exchangeability that the authors then used to extend full and split CP to the SCS setting using likelihood-ratio weights. However, Tibshirani et al. (2019) do not account for dependencies between the train and test data, and in later work Fannjiang et al. (2022) describe how full CP for SCS thus loses its formal guarantee under FCS (though empirically, Fannjiang et al. (2022) find that full CP-SCS can maintain target coverage but with overly conservative interval widths). The main result of Fannjiang et al. (2022) extends full CP to FCS (full CP-FCS) while maintaining a guarantee of the form (1). However, for an arbitrary (potentially nonlinear) predictor with n train points, m test points, and $|\mathcal{Y}|$ as the cardinality of the label space, full CP-FCS demands the often prohibitive computational price of training $(n+1) \cdot m \cdot |\mathcal{Y}|$ distinct predictors (in regression, $\mathcal{Y}$ must be approximated by a fine grid of values). Only in special cases such as for linear models can full CP-FCS's computation be reduced to $(n+1) \cdot m$ runs of the training algorithm $\mathcal{A}$.

Secondarily, Fannjiang et al. (2022) also describes how Tibshirani et al. (2019)'s split CP for SCS (hereon "weighted split CP") de facto maintains its guarantee under FCS: due to sample splitting, the test distribution depends on weighted split CP's "proper" training data but not on its holdout calibration set, which reduces the holdout-to-test shift from FCS to SCS. However, weighted split CP remains statistically inefficient due to its sample splitting requirement, which (as we will demonstrate) results in reduced model performance and overly wide (and thus less informative) predictive intervals relative to our proposed methods.

2.6. JAW: Jackknife+ Under Standard Covariate Shift

Prinster et al. (2022) propose the **ja**ckknife+ **w**eighted with likelihood ratio weights for SCS, or JAW (JAW-SCS for clarity), which relaxes the jackknife+'s assumption of data exchangeability to allow for SCS while achieving the same finite-sample coverage guarantee (2). However, JAW-SCS does not allow for train-test dependencies and thus loses its guarantee under FCS. Prinster et al. (2022) also propose a computationally fast JAW **a**pproximation (JAWA) that uses higher-order influence functions to estimate JAW's leave-one-out models and thus avoid retraining, but with an asymptotic guarantee that requires regularity conditions. In our current work, we extend JAW-SCS to FCS and moreover propose computational relaxations to JAW under either SCS or FCS that maintain finite-sample coverage guarantees without any regularity assumptions.

3. JAW-FCS: Jackknife+ Weighted for FCS

Our first proposed method for efficient distribution-free predictive inference under FCS generalizes both the jackknife+ (Barber et al., 2021) and JAW-SCS (Prinster et al., 2022) methods to FCS. Let training data $Z_{1:n} = \{Z_1, ..., Z_n\}$ and a test point Z_{n+1} be generated from FCS (5), and denote $w(x; D) = \mathrm{d}\widetilde{P}_{X;D}(x)/\mathrm{d}P_X(x)$ as the likelihood ratio function for the data that depends on D for $D \subseteq Z_{1:n}$. Then, for each $j \in \{1, ..., n+1\}$, let $Z_{-j} = Z_{1:n} \backslash Z_j$ denote the training data with point j removed (where $Z_{-(n+1)} = Z_{1:n}$), and define the normalized weight function

$$\tilde{w}_{n+1,j}(x) = \frac{w(x; Z_{-j}) w(X_j; Z_{-j})}{\sum_{j'=1}^{n} \left[w(x; Z_{-j'}) w(X_{j'}; Z_{-j'}) \right] + w(x; Z_{1:n})^2}. \quad (6)$$

Given X_{n+1} as an argument, $\tilde{w}_{n+1,j}(X_{n+1})$ can be thought of as a weight applied to the training point X_j that is carefully normalized with respect to the other training data and the test point X_{n+1} (see Appendix A.2 for further details). To condense notation slightly, for $j = n+1$ we will also write $\tilde{w}_{n+1,n+1}(x)$ as $\tilde{w}_{(n+1)^2}(x)$. We then define the predictive interval for JAW-FCS, as follows:

$$\widehat{C}_{n,\alpha}^{\text{JAW-FCS}}(x) = \Big[Q_\alpha\Big(\sum_{j=1}^{n} \tilde{w}_{n+1,j}(x) \delta_{\widehat{\mu}_{-j}(x) - R_j^{LOO}} + \tilde{w}_{(n+1)^2}(x) \delta_{-\infty} \Big), \quad (7)$$
$$Q_{1-\alpha}\Big(\sum_{j=1}^{n} \tilde{w}_{n+1,j}(x) \delta_{\widehat{\mu}_{-j}(x) + R_j^{LOO}} + \tilde{w}_{(n+1)^2}(x) \delta_{\infty} \Big) \Big],$$

where δ_v denotes a point mass at v and $Q_\beta(\cdot)$ denotes the level β empirical quantile function. The following theorem presents the finite-sample coverage guarantee for the JAW-FCS interval (7), which relaxes the assumption in Prinster et al. (2022) from standard to feedback covariate shift.

Theorem 3.1. *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0,1)$, the JAW-FCS predictive interval in* (7) *satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\textit{JAW-FCS}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha. \tag{8}$$

We defer the proof to Appendix A.2, but here note two key differences from the JAW-SCS proof in Prinster et al. (2022): Firstly, weights are selected to maintain "pseudo-exchangeability", rather than weighted exchangeability, in a weighted comparison matrix; secondly, in our first proof step we establish a bound on an expectation term whose analog in Prinster et al. (2022) is deterministic.

4. Further Computational Relaxations

4.1. Relaxation of JAW with K Leave-One-Out Models

While the calculation of the JAW-FCS prediction interval (7) is in general computationally efficient relative to full CP-FCS from Fannjiang et al. (2022) (see Table 1), the retraining requirements of JAW-FCS can be relaxed even further to enable rerunning the model-training algorithm $\mathcal{A}$ only $K \leq n$ times. As our first proposed computational relaxation, we demonstrate that using only $K \leq n$ of the leave-one-out models required by the JAW-FCS method still achieves the same coverage guarantee, though often at the cost of wider or more variable predictive intervals. We call this computational relaxation JAW-KLOO. The K training points used for JAW-KLOO's leave-one-out retraining can be selected in a range of ways, either deterministically or randomly. For simplicity, we mainly focus on JAW-KLOO deterministically selecting the K points with largest weights, though in the Appendix A.3 we discuss adjustments for random sampling and in Section 5.3 we compare empirical performance of deterministic versus random sampling in terms of interval width and coverage variance.

Let $S_{\text{LOO}} \subseteq \{1, ..., n\}$ denote a subset of the training data where $|S_{\text{LOO}}| = K$. Then, we define the JAW-KLOO normalized weights similarly as for JAW-FCS, except only using leave-one-out models for points $j : j \in S_{\text{LOO}}$:

$$\tilde{w}^{\text{KO}}_{n+1,j}(x) = \frac{w(x; Z_{-j})w(X_j; Z_{-j})}{\sum_{j' \in S_{\text{LOO}}} \left[w(x; Z_{-j'})w(X_{j'}; Z_{-j'})\right] + w(x; Z_{1:n})^2}, \tag{9}$$

where to condense notation slightly we also write $\tilde{w}^{KO}_{(n+1)^2}(x)$ to denote $\tilde{w}^{KO}_{n+1,n+1}(x)$. Then, we define the JAW-KLOO prediction interval as follows

$$\widehat{C}^{\text{JAW-}K\text{LOO}}_{n,\alpha}(x) = \Big[Q_\alpha\Big(\sum_{j \in S_{\text{O}}} \tilde{w}^{\text{KO}}_{n+1,j}(x)\delta_{\widehat{\mu}_{-j}(x) - R^{LOO}_j} + \tilde{w}^{\text{KO}}_{(n+1)^2}(x)\delta_{-\infty}\Big), \; Q_{1-\alpha}\Big(\sum_{j \in S_{\text{LOO}}} \tilde{w}^{\text{KO}}_{n+1,j}(x)\delta_{\widehat{\mu}_{-j}(x) + R^{LOO}_j} + \tilde{w}^{\text{KO}}_{(n+1)^2}(x)\delta_{\infty}\Big)\Big]. \tag{10}$$

The JAW-KLOO model then satisfies the same coverage guarantee as the full JAW-FCS model, which we state formally in the following theorem (proof in Appendix A.3).

Theorem 4.1. *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0,1)$, the JAW-KLOO predictive interval in* (10) *satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\textit{JAW-}K\textit{LOO}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha. \tag{11}$$

4.2. K-fold Weighted Cross-Valdation+

We now propose an alternative computational relaxation of JAW-FCS that relies on a weighted K-fold cross validation resampling procedure with K leave-$\frac{n}{K}$-out models (in contrast to JAW-KLOO that uses K leave-*one*-out models). In particular, this second computational relaxation generalizes the K-fold cross validation+ (CV+) method of Barber et al. (2021) to allow for feedback or standard covariate shift—we call this method K-fold WCV+FCS or WCV+ for short.

Prior to defining the WCV+ predictive interval, we first need to generalize the normalized weights defined in (6) using likelihood ratio functions $w(x; D) = \mathrm{d}\widetilde{P}_{X;D}(x)/\mathrm{d}P_X(x)$ that depend on all the training data aside from a specific fold (rather than depending on leave-one-out subsets Z_{-j} as in (6)). For $k(j) \in \{1, ..., K\}$ denoting the index of the cross validation fold $S_{k(j)} \subseteq Z_{1:n}$ that point j belongs to, let $Z_{-S_{k(j)}} = Z_{1:n} \backslash S_{k(j)}$. Then, we define

$$\tilde{w}^{CV}_{n+1,j}(x) = \frac{w(x; Z_{-S_{k(j)}})w(X_j; Z_{-S_{k(j)}})}{\sum_{j'=1}^{n} \left[w(x; Z_{-S_{k(j')}})w(X_{j'}; Z_{-S_{k(j')}})\right] + w(x; Z_{1:n})^2} \tag{12}$$

for training data $j \in \{1, ..., n\}$. We can now define the K-fold WCV+ predictive interval as follows:

$$\widehat{C}^{\text{WCV+}}_{n,K,\alpha}(x) = \Big[Q_\alpha\Big(\sum_{j=1}^{n} \tilde{w}^{CV}_{n+1,j}(x)\delta_{\widehat{\mu}_{-S_{k(i)}}(x) - R^{CV}_i} + \tilde{w}^{CV}_{(n+1)^2}(x)\delta_{-\infty}\Big), \; Q_{1-\alpha}\Big(\sum_{j=1}^{n} \tilde{w}^{CV}_{n+1,j}(x)\delta_{\widehat{\mu}_{-S_{k(i)}}(x) + R^{CV}_i} + \tilde{w}^{CV}_{(n+1)^2}(x)\delta_{\infty}\Big), \Big], \tag{13}$$

where the $\tilde{w}^{CV}_{n+1,j}(x)$ are defined in (12), and where $\tilde{w}^{CV}_{(n+1)^2}(x) = \tilde{w}^{CV}_{n+1,n+1}(x)$. We now present the coverage guarantee for WCV+, which is weaker than the JAW-FCS guarantee by a term that approximately represents the expected normalized weight of a cross-validation fold minus one point, but with a more technical formulation that we defer with the WCV+ coverage proof to Appendix A.4.

Theorem 4.2. *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0, 1)$, the K-fold WCV+ predictive interval in* (13) *satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \widehat{C}^{\text{WCV+FCS}}_{n,K,\alpha}(X_{n+1})\big\} \geq 1 - 2\alpha - \mathbb{E}\Bigg[\sum_{j \in S_{k(i)} \setminus i} \tilde{w}^{CV}_{ij}(X_{n+1})\Bigg]. \quad (14)$$

Remark 4.3. JAW-KLOO and WCV+ also extend to SCS as a special case, which to our best knowledge is also a novel contribution—that is, the first computational relaxations of JAW-SCS (Prinster et al., 2022) without requiring a holdout dataset and maintaining finite-sample guarantees.

5. Experiments

To demonstrate the practical performance of our JAW-FCS method and its computational relaxations in a real-world FCS scenario, we first focus on the protein design problem (Figure 1), where a common goal is to design a novel protein sequence with high functional fitness (Yang et al., 2019; Sinai & Kelsic, 2020; Wu et al., 2021), such as strong fluorescence. We next consider an active learning task to evaluate how our methods perform when the FCS likelihood-ratio weights need to be estimated, and we lastly evaluate our two computational relaxations in the SCS setting.

5.1. Protein Design Experiments under FCS

Datasets In protein engineering, the labels for designed sequences are usually unknown, so we follow Fannjiang et al. (2022) by using the workaround offered by combinatorially complete protein datasets. In particular, we use data from Poelwijk et al. (2019) where each sequence was measured for both a "red" and a "blue" wavelength fluorescence, thus resulting in two combinatorially complete datasets corresponding to the two distinct fitness functions.

Creation of Feedback Covariate Shift In line with Fannjiang et al. (2022) and other biomolecular engineering papers (Biswas et al., 2021; Zhu et al., 2021), we sample design or "test point" protein sequences from a distribution over fluorescent protein sequences (Poelwijk et al., 2019) with log-likelihood proportional to the regression model's prediction. That is, we sample design sequences from $\widetilde{P}_{X;Z_{1:n}}(X_{n+1}) \propto \exp(\lambda \cdot \widehat{\mu}(X_{n+1}))$, where the hyperparameter $\lambda \geq 0$ is the "inverse temperature". Thus, larger λ values correspond to larger shift magnitudes of FCS. Artificial measurement noise was also added to the sampling procedure as in Fannjiang et al. (2022) to simulate a real experimental scenario where repeat measurements of the same sequence can result in different observations.

Oracle Weights Are Often Known in the Design Problem The one-shot design problem is a special case of FCS where the distribution of the inputs is usually known or can be reliably simulated, which substantially reduces or removes altogether the challenge of likelihood-ratio weight estimation (Fannjiang et al., 2022). The intuition is that the training data are selected by a known procedure defined by a domain expert (e.g., random substitutions to a known wild-type sequence as in Brookes et al. (2019); Biswas et al. (2021); Bryant et al. (2021)), and the (shifted) test distribution is a *designed* distribution shift, intentionally selected so that "test" protein sequences that are expected to have high fitness. In particular, the biomolecular engineering literature commonly uses optimization procedures (such as training a generative model) where the test (design) distribution is explicitly known (Brookes et al., 2022; Fannjiang & Listgarten, 2020; Popova et al., 2018; Kang & Cho, 2018; Russ et al., 2020; Wu et al., 2020; Hawkins-Hooker et al., 2021; Shin et al., 2021) or that produce an implicit test distribution that can be easily simulated, and thus estimated (Killoran et al., 2017; Gómez-Bombarelli et al., 2018; Linder et al., 2020; Sinai et al., 2020; Bashir et al., 2021; Bryant et al., 2021), relative to naive density estimation with unknown shifts. For our protein design experiments we thus evaluate our methods and relevant baselines using oracle weights.

Protein design experimental details We used the scikit-learn package (Pedregosa et al., 2011) MLPRegressor method (with L-BFGS solver and logistic activation function) for the neural network $\widehat{\mu}$ and the package's RandomForestRegressor method (with 20 trees and the absolute error criterion) for the random forest $\widehat{\mu}$. The baselines are jackknife+ (orange squares), weighted split conformal (green triangles), and traditional split conformal (pink diamonds); the predicted fitness values for JAW-FCS and jackknife+ are identical, as are those for weighted and traditional split CP. All values are for 20 repeated experiments, which each have a distinct random seed, 192-sample training set, and set of 200 test points (for 4000 total test points per plotted value).

5.1.1. Protein Design Results for JAW-FCS

A predictive inference method is reliable if its confidence intervals achieve coverage at the target level $1 - \alpha$. Beyond this necessary criterion, however, for protein design a predictive inference method is most useful if it moreover enables a protein engineer to identify promising candidate

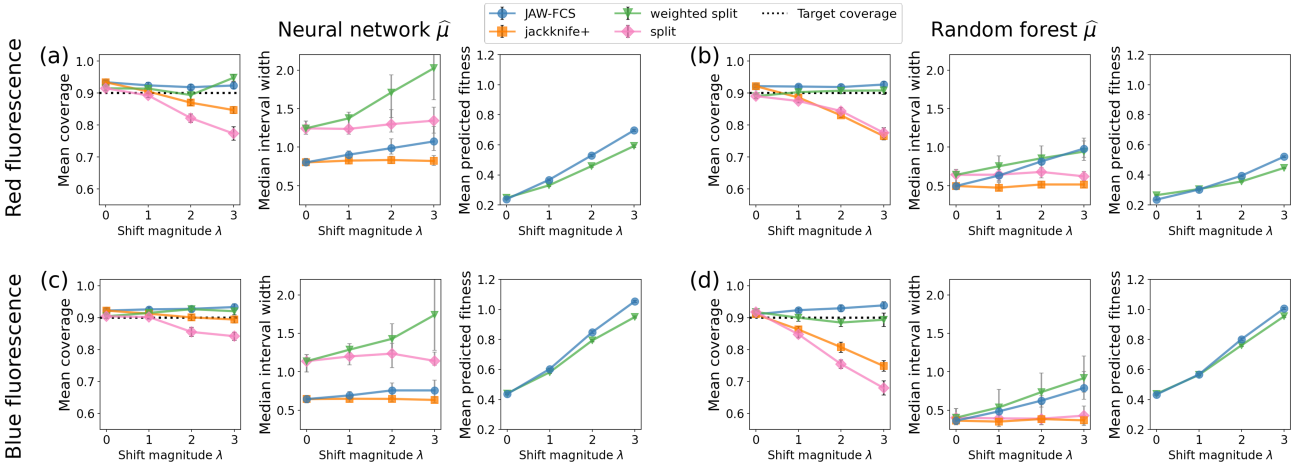

Figure 2: Mean predictive interval coverage, median interval width, and mean predicted fitness (i.e., mean predicted fluorescence for test point "designed" protein sequence) for our proposed JAW-FCS method (blue circles) and its baselines, across four datasets and predictor conditions: (a) red fluorescence with neural network $\widehat{\mu}$, (b) red fluorescence with random forest $\widehat{\mu}$, (c) blue fluorescence with neural network $\widehat{\mu}$, (d) blue fluorescence with random forest $\widehat{\mu}$. For coverage, at or above the target level (black dotted line at $1-\alpha = 0.9$) is best; for median interval width, smaller is better (more informative) so long as target coverage is achieved; for predicted fitness, higher is better. Black error bars for coverage and predicted fitness represent standard error, and gray error bars for median interval width represent upper and lower quartiles. JAW-FCS maintains coverage at the target level regardless of shift magnitudes λ, with higher mean predicted fitness and smaller, more informative prediction intervals than weighted split conformal, the only baseline that also maintains target coverage.

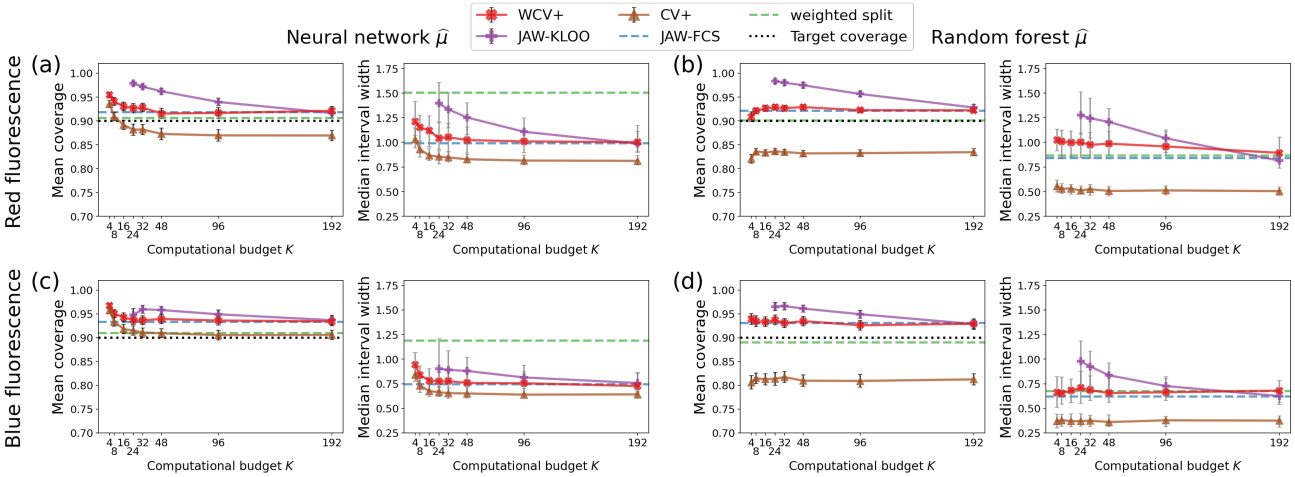

Figure 3: Mean coverage and median interval width for computational relaxations of JAW-FCS—that is, WCV+ (red Xs) and JAW-KLOO (violet plus shape) methods—relative to baselines across a range of computational budgets for training K distinct predictors. The other baseline that depends on K is standard CV+ (brown triangles); meanwhile, the proposed JAW-FCS (blue line) and weighted split conformal (green line) methods are also included as baselines, though with dashed lines indicating that their computational budgets are fixed independently of K (see Table 1). For coverage, at or above the target level (black dotted line at $1-\alpha = 0.9$) is best, while for median interval width, smaller is better (more informative) only for methods that achieve coverage at or above target level. Shift magnitude of FCS is fixed at $\lambda = 2$. Black error bars for coverage represent standard error and gray bars for median interval width represent upper and lower quartiles. Both WCV+ and JAW-KLOO maintain coverage above the target level of $1-\alpha = 0.9$ for all tested computational budget values K; WCV+ interval widths are comparable to JAW-FCS, while JAW-KLOO are overly wide for smaller K.

sequences that are predicted, with minimal uncertainty (minimal interval width), to have maximum fitness. Accordingly, we consider a predictive inference method to have ideal statistical performance if it maintains coverage at or above the user-specified target level $1-\alpha$, if its interval widths are as small and thus as informative as possible, and if the predicted fitness of its designed sequences are as high as possible. In Figure 2 we thus plot these three performance criteria (coverage, interval width, and predicted fitness) for JAW-FCS and several baselines across a grid of shift magnitudes $\lambda \in \{0, 1, 2, 3\}$, for both the red and blue fluorescence datasets with both a neural net and random forest predictor.

For all dataset $\times$ predictor conditions in Figure 2, JAW-FCS and weighted split conformal prediction maintain coverage

at the target level of $1 - \alpha = 0.9$ regardless of data shift magnitude λ, whereas the traditional, exchangeable-data versions of each of the two methods (jackknife+ and split conformal) lose coverage. However, for all dataset $\times$ predictor conditions, the JAW-FCS prediction intervals are generally smaller and thus more informative than the weighted split intervals, and designed sequences from the JAW-FCS's $\widehat{\mu}$ predictor have higher mean predicted fitness than those of the weighted split method. The superior predicted fitness and interval width values for JAW-FCS relative to weighted split CP largely reflect the former's greater statistical efficiency. That is, by avoiding sample splitting, JAW-FCS maintains a $\widehat{\mu}$ predictor that is trained on more data and therefore more "competent" than that of weighted split, and JAW-FCS is also able to efficiently use all of its labeled data for the construction of more precise predictive intervals.

5.1.2. Results for Protein Design with JAW-FCS Computational Relaxations

We now turn to evaluate WCV+ and JAW-KLOO, our two proposed computational relaxations of JAW-FCS, relative to JAW-FCS, standard CV+, and weighted split CP. In Figure 3, we compare these methods' coverage and interval width values across a grid of computational budgets $K \in \{4, 8, 16, 24, 32, 48, 96, 192\}$ denoting the number of retrained predictors, across the same dataset $\times$ predictor conditions. We omit the predicted fitness values, which would be largely overlapping for non-holdout-set methods. We observe that both WCV+ and JAW-KLOO maintain coverage at or above the target $1 - \alpha = 0.9$ level for all evaluated computational budgets, whereas standard CV+ loses coverage. Notably, the prediction interval widths for WCV+ are largely comparable to those of the full JAW-FCS method for all evaluated computational budgets K, but for smaller values of K, JAW-KLOO has overly conservative prediction intervals. Values for JAW-KLOO are not plotted for $K < 24$, where the method's intervals can often become uninformative for small K. These results favor WCV+ as a practical computational relaxation of JAW-FCS that empirically appears to avoid degraded coverage or interval widths even for fairly small computational budgets K. Meanwhile, in cases with severe miscoverage penalties, JAW-KLOO could potentially be used as a conservative computational relaxation to JAW-FCS that achieves the same guarantee.

5.2. Experiments with Estimated FCS Weights

Active Learning Exploration with Probabilistic Bounds While in Section 5.1 the input distributions (and thus the FCS weights) are known for the protein design task, in other settings of FCS the weights may need to be estimated. For instance, take high-stakes or resource-constrained exploration in active learning, where predictive intervals for query (test) points can help (probabilistically) bound risks

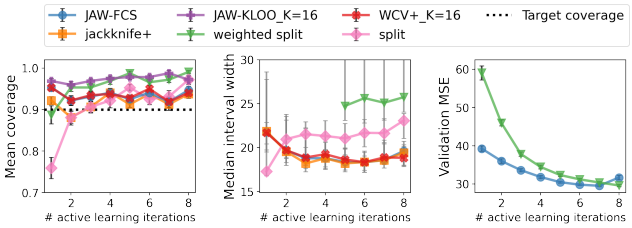

Figure 4: Mean coverage (left), median interval width (middle), and validation-set mean squared error (right), for active learning predictive inference experiments on the airfoil dataset (Dua & Graff, 2017). Coverage and MSE error bars are standard error, width error bars are interquartile range. Values are for 30 experimental replicates each with 8 active learning query iterations, where 16 points were queried in each iteration. Weighted split interval widths are not shown when small sample size in early iterations results in median widths being uninformative (infinite length). The results show that our proposed methods (JAW-FCS, JAW-KLOO, WCV+) can achieve target coverage ($1 - \alpha = 0.9$) even with estimated rather than oracle weights while maintaining smaller (more informative) interval widths and lower MSE.

or costs associated with the "annotation" procedure (e.g., invasive medical diagnostic procedures or expensive lab experiments). In each active learning iteration, the training data are updated with newly labeled "query" points based on a systematic querying strategy, which will usually result in the training distribution in the early stage differing drastically from the distribution at a later iteration. Meanwhile, the distribution of the query (test) points can also change as the model becomes more "informed". However, with common active learning setups such as pool-based active learning with uncertainty sampling, the query or test distribution shift is a direct function of an attribute of the model predictions (e.g., least confidence, prediction entropy) (Nguyen et al., 2022; Brochu et al., 2010; Settles, 2009), which means that the test distribution is known. Therefore, FCS likelihood-ratio weight estimation in active learning often reduces to density estimation of only the training data.

To empirically evaluate our proposed predictive inference methods with estimated rather than oracle weights, we implemented a pool-based active learning task with the NASA Airfoil Self-Noise Dataset (Dua & Graff, 2017). We used kernel density estimation (with a Gaussian kernel) to estimate the density of the labeled training data, and query/test data were sampled from a larger pool of unlabeled points with likelihoods proportional to the predicted variance at each point from a Gaussian process (with a dot product kernel and added white noise), which is a common practice (Yue et al., 2020). Figure 4 firstly demonstrates that our proposed (JAW-FCS, JAW-KLOO, and WCV+) methods can achieve predictive interval coverage above the target level of $1 - \alpha = 0.9$ even with estimated rather than oracle weights. Moreover, in Figure 4 our methods generally maintain smaller (and thus more informative) interval widths than the weighted split conformal prediction baseline as

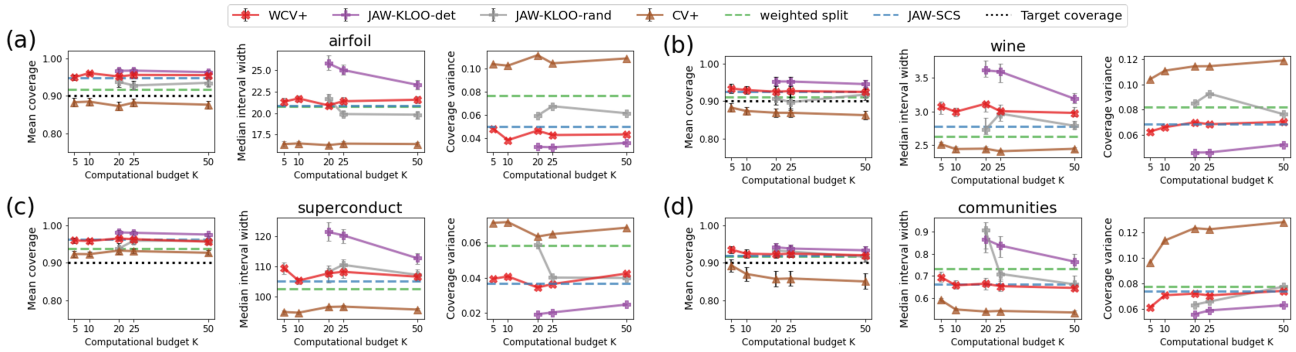

Figure 5: Mean coverage, median interval width, and coverage variance for proposed WCV+ (red Xs), deterministic JAW-KLOO (violet plus shape), and random JAW-KLOO (gray plus shape) under standard covariate shift (SCS), compared to JAW-SCS (blue line), weighted split (green line), and standard CV+ (brown triangles) baselines, for a range of computational budgets K. For coverage, at or above the target level of $1 - \alpha = 0.9$ is best; for interval width, smaller is better (more informative) so long as target coverage is reached; for coverage variance, lower is better (more reliable). Results are across 20 repeated experiments each with a neural net predictor with 200 training points and 200 test points. WCV+ and both JAW-KLOO sampling approaches maintain coverage at the target level of $1 - \alpha = 0.9$ across all tested conditions: WCV+ does so while performing comparable to JAW-SCS, while deterministic JAW-KLOO-det has overly conservative interval widths and random-sampling JAW-KLOO-rand has high coverage variance.

well as achieve smaller validation-set mean squared error.

5.3. Computational Relaxations Under SCS

To evaluate our proposed computational relaxations in the SCS setting, we use four UCI datasets (Dua & Graff, 2017) used for experiments in Prinster et al. (2022): airfoil self-noise, red wine quality prediction (Cortez et al., 2009), superconductivity (Hamidieh, 2018), and communities and crime (Redmond & Baveja, 2002), which represent a range of different dimensionalities. We follow the procedure described in Prinster et al. (2022) for the creation of SCS, and we refer to that work for details. Figure 5 shows the mean coverage, median interval width, and coverage variance results for our proposed WCV+ method, as well as for JAW-KLOO with both deterministic and random sampling when $K \geq 20$ (as JAW-KLOO often becomes uninformative for smaller K). WCV+ and both JAW-KLOO sampling methods maintain coverage above the target level of $1 - \alpha = 0.9$ for all datasets and predictor functions $\widehat{\mu}$, along with the other methods with coverage guarantees under standard covariate shift (JAW-SCS and weighted split conformal prediction). These results suggest that our proposed WCV+ and JAW-KLOO methods, which relax the computational requirements of JAW-FCS, achieve similar target coverage performance under SCS as in the FCS setting. Moreover, across datasets, the predictive intervals for WCV+ have comparable width to the JAW-SCS method, which suggests that WCV+ is a practical computational relaxation of JAW-SCS that appears to avoid sacrificing either empirical coverage or interval sharpness. Meanwhile, JAW-KLOO with deterministic sampling tends to have overly large interval widths for small K, while JAW-KLOO with random sampling tends to have more variable coverage, which indicates less reliability (see Appendix A.3 for details on JAW-KLOO variants).

6. Conclusion

In this paper we propose JAWS-X, a collection of distribution-free predictive inference methods for standard and feedback covariate shift that can flexibly and favorably balance computational and statistical efficiency. We provide rigorous, finite-sample coverage guarantees and validate them in real-world feedback covariate shift problems including protein design, active learning with estimated weights, and standard covariate shift scenarios with real-world datasets. Our methods achieve a substantial speedup in computational demands relative to full conformal for FCS as well as a considerable improvement to predictor performance and interval sharpness relative to weighted split conformal, without losing coverage. We moreover demonstrate that our proposed methods achieve target performance even with estimated rather than oracle weights on an active learning task, although the fact that our guarantees assume oracle weights is a limitation of our work. Promising future directions include examining how weight estimation error impacts guarantees, exploring if stronger guarantees are possible for the WCV+ method, improving the selection of leave-one-out models for the JAW-KLOO method, and generalizations to other distribution shift settings.

Software and Data

GitHub repository with code for all experiments: `https://github.com/drewprinster/jaws-x`

Acknowledgments

Drew Prinster and Suchi Saria are supported by the National Science Foundation grant IIS-1840088. Anqi Liu is partially supported by the JHU-Amazon AI2AI Faculty Award.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., Davis, G., Gong, Q., Armstrong, Z., Jang, J., et al. Machine learning guided aptamer refinement and discovery. *Nature Communications*, 12(1):2366, 2021.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Brochu, E., Cora, V. M., and De Freitas, N. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Brookes, D., Park, H., and Listgarten, J. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pp. 773–782. PMLR, 2019.

Brookes, D. H., Aghazadeh, A., and Listgarten, J. On the sparsity of fitness functions and implications for learning. *Proceedings of the National Academy of Sciences*, 119 (1):e2109649118, 2022.

Bryant, D. H., Bashir, A., Sinai, S., Jain, N. K., Ogden, P. J., Riley, P. F., Church, G. M., Colwell, L. J., and Kelsic, E. D. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Fannjiang, C. and Listgarten, J. Autofocused oracles for model-based design. *Advances in Neural Information Processing Systems*, 33:12945–12956, 2020.

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Hamidieh, K. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. *PLoS computational biology*, 17(2):e1008736, 2021.

Kang, S. and Cho, K. Conditional molecular design with deep generative models. *Journal of chemical information and modeling*, 59(1):43–52, 2018.

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. Generating and designing dna with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

Landau, H. On dominance relations and the structure of animal societies: Iii the condition for a score structure. *The bulletin of mathematical biophysics*, 15(2):143–148, 1953.

Linder, J., Bogard, N., Rosenberg, A. B., and Seelig, G. A generative neural network for maximizing fitness and diversity of synthetic dna and protein sequences. *Cell systems*, 11(1):49–62, 2020.

Miller, R. G. The jackknife-a review. *Biometrika*, 61(1): 1–15, 1974.

Nguyen, V.-L., Shaker, M. H., and Hüllermeier, E. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer, 2002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Poelwijk, F. J., Socolich, M., and Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1):1–11, 2019.

Popova, M., Isayev, O., and Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7): eaap7885, 2018.

Prinster, D., Liu, A., and Saria, S. Jaws: Auditing predictive uncertainty under covariate shift. *Thirty-sixth Conference on Neural Information Processing Systems*, pp. arXiv:2207.10716, 2022.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369 (6502):440–445, 2020.

Settles, B. Active learning literature survey. 2009.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):2403, 2021.

Sinai, S. and Kelsic, E. D. A primer on model-guided exploration of fitness landscapes for biological sequence design. *arXiv preprint arXiv:2010.10614*, 2020.

Sinai, S., Wang, R., Whatley, A., Slocum, S., Locane, E., and Kelsic, E. D. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020.

Steinberger, L. and Leeb, H. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

Steinberger, L. and Leeb, H. Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412*, 2018.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pp. 37–51. PMLR, 2018.

Wu, Z., Yang, K. K., Liszka, M. J., Lee, A., Batzilla, A., Wernick, D., Weiner, D. P., and Arnold, F. H. Signal peptides generated by attention-based neural networks. *ACS Synthetic Biology*, 9(8):2154–2161, 2020.

Wu, Z., Johnston, K. E., Arnold, F. H., and Yang, K. K. Protein sequence design with deep generative models. *Current opinion in chemical biology*, 65:18–27, 2021.

Yang, K. K., Wu, Z., and Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

Yue, X., Wen, Y., Hunt, J. H., and Shi, J. Active learning for gaussian process considering uncertainties with application to shape control of composite fuselage. *IEEE Transactions on Automation Science and Engineering*, 18 (1):36–46, 2020.

Zhu, D., Brookes, D. H., Busia, A., Carneiro, A., Fannjiang, C., Popova, G., Shin, D., Donohue, K. C., Chang, E. F., Nowakowski, T. J., et al. Optimal trade-off control in machine learning-based library design, with application to adeno-associated virus (aav) for gene therapy. *bioRxiv*, pp. 2021–11, 2021.

A. Proofs for theoretical results.

A.1. Preliminaries

Data from feedback covariate shift (FCS) as in (5) are a special case of what Fannjiang et al. (2022) call *pseudo-exchangeable* random variables.

Definition A.1. *Random variables $V_1, ..., V_{n+1}$ are* pseudo-exchangeable *with factor functions $w_1, ..., w_{n+1}$ and core function h if the density, f, of their joint distribution can be factorized as*

$$f(v_1, ..., v_{n+1}) = \prod_{i=1}^{n+1} w_i(v_i; v_{-i}) \cdot h(v_1, ..., v_{n+1}) \tag{15}$$

where $v_{-i} = v_{1:(n+1)} \backslash v_i$, each $w_i(\cdot; v_{-i})$ is a function that depends on the multiset v_{-i} (that is, on the values in v_{-i} but not on their ordering), and h is a function that does not depend on the ordering of its $n+1$ inputs.

The proofs for our theoretical results leverage the observation that any subsequence of pseudo-exchangeable random variables is itself pseudo-exchangeable, which we state formally in the following lemma.

Lemma A.2. *Let $(V_1, ..., V_{n+1})$ be a sequence of pseudo-exchangeable random variables with factor functions $w_1, ..., w_{n+1}$. For any $J = \{j_1, ..., j_m\} \subseteq \{1, ..., n+1\}$, the subsequence $(V_{j_1}, ..., V_{j_m})$ is pseudo-exchangeable.*

Proof. Let $J = \{j_1, ..., j_m\} \subseteq \{1, ..., n+1\}$ denote an arbitrary set of indices so that $(V_{j_1}, ..., V_{j_m})$ is a subsequence of $(V_1, ..., V_{n+1})$, and let $J^C = \{j'_1, ..., j'_{n+1-m}\}$. Then, we can integrate (15) over all $v_{j'}$ such that $j' \in J^C$:

$$\int_{v_{j'_1}} \dots \int_{v_{j'_{n+1-m}}} f(v_1, ..., v_{n+1})\, dv_{j'_1} \dots dv_{j'_{n+1-m}} = \int_{v_{j'_1}} \dots \int_{v_{j'_{n-m}}} \prod_{i=1}^{n+1} w_i(v_i; v_{-i}) \cdot h(v_1, ..., v_n)\, dv_{j'_1} \dots dv_{j'_{n+1-m}} \tag{16}$$

such that the right-hand side of (16) no longer depends on any specific value of $v_{j'}$ for $j' \in J^C$. That is, letting $v_J = \{v_j : j \in J\}$, we can write (16) as

$$f_J(v_{j_1}, ..., v_{j_m}) = \prod_{j \in J} w_j(v_j; v_J \backslash v_j) \cdot g_J(v_{j_1}, ..., v_{j_m}), \tag{17}$$

for some weight functions $w_j(\cdot; v_J \backslash v_j)$ and some core function g_J that does not depend on the ordering of its inputs. Therefore, the subsequence $(V_{j_1}, ..., V_{j_m})$ is pseudo-exchangeable. □

A.2. Proof for JAW-FCS coverage under feedback covariate shift

We first restate the JAW-FCS coverage guarantee before proceeding with the proof.

Theorem 3.1 *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0, 1)$, the JAW-FCS predictive interval in* (7) *satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{JAW-FCS}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha.$$

Our proof technique generalizes the proof for JAW-SCS presented in Prinster et al. (2022) from standard covariate shift to feedback covariate shift. Since the proof for JAW-SCS in Prinster et al. (2022) is itself a generalization of the jackknife+ coverage proof (for exchangeable data) in Barber et al. (2021), the proof we present here is thus also a generalization of the jackknife+ coverage proof.

We use (a) - (d) to denote four setup steps, and after the setup we use 1-3 to denote the main steps in the proof. Our first two initial setup steps (a) and (b) are identical to the corresponding setup steps in the proof for both Theorem 1 in Prinster et al. (2022) and Theorem 1 in Barber et al. (2021):

(a) First, we suppose the hypothetical case where in addition to the training data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, we also have access to the labeled test point (X_{n+1}, Y_{n+1}). For each pair of indices $i, j \in \{1, ..., n+1\}$ with $i \neq j$, we define $\tilde{\mu}_{-(i,j)}$ as the regression function fitted on the training and potentially on the test data except with the points i and j removed. (We follow the notation in Barber et al. (2021) where $\tilde{\mu}$ rather than $\widehat{\mu}$ reminds us that the former is fit on a subset of data $1, ..., n+1$ that may contain the test point $n+1$.) We note that $\tilde{\mu}_{-(i,j)} = \tilde{\mu}_{-(j,i)}$ for any $i \neq j$, and $\tilde{\mu}_{-(i,n+1)} = \widehat{\mu}_{-i}$ for any $i = 1, ..., n$.

(b) We also define the same matrix of residuals in Barber et al. (2021), $R \in \mathbb{R}^{(n+1)\times(n+1)}$, with entries

$$R_{ij} = \begin{cases} +\infty & i = j, \\ |Y_i - \tilde{\mu}_{-(i,j)}(X_i)| & i \neq j \end{cases}$$

such that the off-diagonal entries R_{ij} represent the residual for the ith datapoint where both i and j are not seen by the regression fitting.

At this point we begin to introduce some changes to the proof techniques in both Prinster et al. (2022) and Barber et al. (2021) (see footnotes for additional comparison to prior proofs):

(c) We define a weighted comparison matrix $\hat{A}^w \in \mathbb{R}^{(n+1)\times(n+1)}$. To describe $\hat{A}^w$, we first define A as an unweighted comparison matrix[1] with entries $A_{ij} = \mathbb{1}\{R_{ij} > R_{ji}\}$—that is, each entry A_{ij} is an indicator for the event that, when i and j are excluded from the regression fitting, the prediction on point i yields a larger residual than that for point j. Next, we also define $\hat{W}$ as the weight matrix[2] with entries $\hat{W}_{ij} = w(X_i; z_{-\{i,j\}})$. We are now able to define $\hat{A}^w = \hat{W} \odot A \odot \hat{W}^\top$ where $\odot$ denotes pointwise multiplication, so that $\hat{A}^w$ has entries $\hat{A}^w_{ij} = w(X_i; z_{-\{i,j\}}) \cdot w(X_j; z_{-\{j,i\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\}$, where both weights depend on the same data observations $z_{-\{j,i\}} = z_{-\{i,j\}}$. For any $i, j \in \{1, ..., n+1\}$, note that $\hat{A}^w_{ij} > 0$ implies $\hat{A}^w_{ji} = 0$ for any $i, j \in \{1, ..., n+1\}$. Note that in the special case of standard covariate shift where $w(X_i; D) = w(X_i)$ for any $D \subseteq \{Z_1, ..., Z_{n+1}\}$, that $\hat{A}^w$ is equivalent (up to a row-specific normalization constant) to the weight matrix used in the JAW proof in Prinster et al. (2022); moreover, in the further specialized case of exchangeable data we have $w(X_i; z_{-\{i,j\}}) = w(X_j; z_{-\{i,j\}}) = 1$ for all $i, j \in \{1, ..., n+1\}$ and thus $\hat{A}^w$ recovers the unweighted comparison matrix A used in the jackknife+ proof in Barber et al. (2021).

(d) Next, as in Prinster et al. (2022) and Barber et al. (2021) we are interested in identifying points that have unusually large residuals and are thus hard to predict. Barber et al. (2021) defined such points with unusually large residuals as points i where $\mathbb{1}\{R_{ij} > R_{ji}\}$ for a sufficiently large fraction of other points j. However, in the feedback covariate shift (similarly as in the standard covariate shift case in Prinster et al. (2022)) we need to account for the fact that the informativeness of the comparison $\mathbb{1}\{R_{ij} > R_{ji}\}$ depends on relative weight or likelihood of the points i and j in the test distribution relative to the training distribution. In other words, we are interested in identifying points i where $\mathbb{1}\{R_{ij} > R_{ji}\}$ for a sufficiently large *weighted portion* of other points j that we might compare point i to, and we can thus reference the ith row of our weighted comparison matrix $\hat{A}^w$ for this information. In particular, we define the set of "strange" points $\mathcal{S}(\hat{A}^w) \subseteq \{1, ..., n+1\}$ as the set of points i where the sum of the ith row in $\hat{A}^w$ is at least a $1-\alpha$ portion of the sum of the ith row in $\hat{W} \odot \hat{W}^\top$:

$$\begin{aligned} \mathcal{S}(\hat{A}^w) &= \Big\{ i \in [n+1] \ : \ \sum_{j=1}^{n+1} \hat{A}^w_{ij} \geq (1-\alpha) \sum_{j'=1}^{n+1} [\hat{W} \odot \hat{W}^\top]_{ij'} \Big\} \\ &= \Big\{ i \in [n+1] \ : \ \sum_{j=1}^{n+1} \big[w(X_i; z_{-\{i,j\}}) w(X_j; z_{-\{i,j\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\} \big] \\ &\qquad \geq (1-\alpha) \sum_{j'=1}^{n+1} \big[w(X_i; z_{-\{i,j'\}}) w(X_{j'}; z_{-\{i,j'\}}) \big] \Big\}. \end{aligned} \tag{18}$$

[1] This is the same unweighted comparison matrix A as in the jackknife+ proof in Barber et al. (2021).

[2] The proof for JAW under standard covariate shift in Prinster et al. (2022) defined a similar weight matrix, but with two main differences from ours here: firstly, unlike in Prinster et al. (2022), here the unnormalized weights $w(\cdot; D)$ depend on data $D \subseteq Z_{1:n}$; and secondly whereas Prinster et al. (2022) define a *normalized* weight matrix, here our current proof is more straightforward if we defer the normalization step until later.

Dividing both sides of the inequality in our definition of $\mathcal{S}(\hat{A}^w)$ by $\sum_{j'=1}^{n+1}\left[w(X_i; z_{-\{i,j'\}})w(X_{j'}; z_{-\{i,j'\}})\right]$ yields

$$\mathcal{S}(\hat{A}^w) = \left\{ i \in [n+1] \ : \ \sum_{j=1}^{n+1} \left(\frac{w(X_i; z_{-\{i,j\}})w(X_j; z_{-\{i,j\}})}{\sum_{j'=1}^{n+1}\left[w(X_i; z_{-\{i,j'\}})w(X_{j'}; z_{-\{i,j'\}})\right]} \cdot \mathbb{1}\{R_{ij} > R_{ji}\} \right) \geq 1-\alpha \right\},$$

and to further simplify notation we denote the resulting fraction on the left side of the inequality as $\tilde{w}_{i,j}(X_{n+1})$, that is

$$\tilde{w}_{i,j}(X_{n+1}) = \frac{w(X_i; z_{-\{i,j\}})w(X_j; z_{-\{i,j\}})}{\sum_{j'=1}^{n+1}\left[w(X_i; z_{-\{i,j'\}})w(X_{j'}; z_{-\{i,j'\}})\right]}, \tag{19}$$

so our strange point definition becomes

$$\mathcal{S}(\hat{A}^w) = \left\{ i \in [n+1] \ : \ \sum_{j=1}^{n+1} \left(\tilde{w}_{i,j}(X_{n+1}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\} \right) \geq 1-\alpha \right\}. \tag{20}$$

We will call the quantity $\tilde{w}_{i,j}(X_{n+1})$ defined in (19) a "normalized weight for point j, normalized with respect to i"[3] since it is normalized with respect to the sum over the ith row in the matrix $\hat{W} \odot \hat{W}^\top$. Note that letting $i = n+1$ in (19) yields the normalized weights $\tilde{w}_{n+1,j}(X_{n+1})$ defined in the main paper (6) (given argument X_{n+1}) for the JAW-FCS predictive interval. We moreover point out that under standard covariate shift where $w(X_i; z_{-\{i,j'\}}) = w(X_i) = w(X_i; z_{-\{i,j\}})$ for all $j, j' \in \{1, ..., n+1\}$, the normalized weights in (19) and our definition of strange points $\mathcal{S}(\hat{A}^w)$ reduce to the analogous quantities in Prinster et al. (2022); additionally, for exchangeable data where $w(X_i; z_{-\{i,j\}}) = 1$ for all $i, j \in \{1, ..., n+1\}$, the weights reduce to uniform weights $\frac{1}{n+1}$ and the set of strange points reduces to that in the jackknife+ coverage proof in Barber et al. (2021).

We now proceed to the main steps of our proof, which generalize the corresponding proof steps in Prinster et al. (2022) and Barber et al. (2021) to allow for feedback covariate shift:

- Step 1: Establish that $\mathbb{E}\big[\sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{i,j}(X_{n+1})\big] \leq 2\alpha$. That is, $\sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{i,j}(X_{n+1})$, the total normalized weight of strange points in any row i of any comparison matrix $\hat{A}^w$, is in expectation no more than 2α.
- Step 2: Using the fact that the datapoints are pseudo-exchangeable, show that the probability that the test point $n+1$ is strange (i.e., $n+1 \in \mathcal{S}(\hat{A}^w)$) is thus bounded by 2α.
- Step 3: Lastly, verify that the JAW interval can only fail to cover the test label value Y_{n+1} if $n+1$ is a strange point.

Step 1: Bounding the expected total normalized weight of the strange points. This proof step follows the corresponding proof step for Theorem 1 in Barber et al. (2021) (for jackknife+ with exchangeable data) and for Theorem 1 in Prinster et al. (2022) (for JAW with standard covariate shift), which rely on Landau's theorem for tournaments (Landau, 1953). The analogous proof step in Barber et al. (2021) derives a bound on the *number* of strange points from a bound on the *number of pairs* of strange points, and Prinster et al. (2022) extend this step to bound the *total weight* of the strange points from a bound on the sum of the *product of weights* for two strange points in a pair. Here, however, we use more general weights than in Prinster et al. (2022) that allow for feedback covariate shift. In addition, there are a few technical and stylistic differences from our

[3] A reader might question why in our definition $\tilde{w}_{i,j}(X_{n+1})$ takes X_{n+1} as an argument, since the point $n+1$ does not appear to have special privilege in the definition. This is to maintain some notational similarity between $\tilde{w}_{i,j}(X_{n+1})$ and the analogous "normalized weight" quantities in related work—that is denoted $p_i^w(X_{n+1})$ in Tibshirani et al. (2019) and Prinster et al. (2022), and denoted $w_i^y(X_{\text{test}})$ in Fannjiang et al. (2022)—where in cases of weighted exchangeability more general than standard covariate shift these normalized weight functions can take as input any subset of data $D \subseteq \{Z_1, ..., Z_{n+1}\}$. Similarly, in pseudo exchangeability cases more general than feedback covariate shift we could have defined more general quantities $\tilde{w}_{i,j}(\cdot)$ that take as input any subset of data $D \subseteq \{Z_1, ..., Z_{n+1}\}$ and are defined based on multiple other weight or factor functions $\{w_i\}$ other than the one we use in the FCS case, but we focus on FCS for simplicity. We also note that we choose to use the character $\tilde{w}$ rather than p^w to also maintain at least a loose connection with the analogous normalized weight terms in Fannjiang et al. (2022) (which they denote with w^y). Our notation thus attempts to balance simplicity with the flexibility of notation used in related work.

current proof step to that in Prinster et al. (2022), most notable of which is that here we establish a bound on *expected* total normalized weight of strange points (whereas Prinster et al. (2022) obtain a deterministic bound in this step). [4]

Consider a tournament between pairs of points: for each pair of points i and j, we say that i "wins" its game against point j if $\hat{A}^w_{ij} > 0$, that is if both i and j have nonzero density in the test distribution and if there is a higher residual on point i than on point j for the regression model $\tilde{\mu}_{-(i,j)}$. We say that i loses its game with j otherwise (so, if $R_{ij} = R_{ji}$ then we say both i and j lose, and moreover in this construction a point i plays against itself, but this "game" is counted as a loss). We furthermore "weight" the importance of the game between points i and j by the product $w(X_i; z_{-\{i,j\}})w(X_j; z_{-\{i,j\}}) = [\hat{W} \odot \hat{W}^\top]_{ij}$, so that games between points with larger (unnormalized) weights are given greater "importance" than games between points with smaller (unnormalized) weights. To further aid with intuition, it may also be helpful to think of the product $w(X_i; z_{-\{i,j\}})w(X_j; z_{-\{i,j\}})$ as the area of a rectangle with width $w(X_i; z_{-\{i,j\}})$ and height $w(X_j; z_{-\{i,j\}})$, so we let L^w_{ij} denote the rectangle with these dimensions: $\text{Area}(L^w_{ij}) = \text{Area}(L^w_{ji}) = w(X_i; z_{-\{i,j\}})w(X_j; z_{-\{i,j\}}) = [\hat{W} \odot \hat{W}^\top]_{ij}$. Moreover, note that we can interpret $\sum_{j'=1}^{n+1} \text{Area}(L^w_{ij'}) = \sum_{j'=1}^{n+1} \left[w(X_i; z_{-\{i,j'\}})w(X_{j'}; z_{-\{i,j'\}})\right] = \sum_{j'=1}^{n+1} [\hat{W} \odot \hat{W}^\top]_{ij'}$ as the total weighted importance of the games that point i plays in (including a "game" against itself), so thus the normalized weight $\tilde{w}_{i,j}(X_{n+1})$ defined in (19) could be interpreted as the relative importance of the game between i and j among all the games that point i plays in the tournament.

We now interpret our definition of strange points (18) with our tournament in mind to write an upper bound that will soon be useful to us. That is, we can interpret (18) as telling us that, if i is a strange point, then the total importance of the games that point i wins in is greater than or equal to a $1 - \alpha$ portion of the total importance of all the games that i plays in—that is, (18) tells us that if i is strange, then

$$\begin{array}{c}\text{Total weighted importance of games} \\ \text{where strange point } i \text{ wins}\end{array} = \sum_{j \text{ s.t. } i \text{ wins against } j} \text{Area}(L^w_{ij}) \quad \geq \quad (1-\alpha) \sum_{j'=1}^{n+1} \text{Area}(L^w_{ij'}).$$

Which implies that the total weighted importance of the games that strange point i *loses* in against *other* points ($j \neq i$) is at most an α portion of the total importance of i's games minus the weight of i's "game" against itself (our construction does not allow i to win against itself):

$$\begin{array}{c}\text{Total weighted importance of games against} \\ \textit{other} \text{ points where strange point } i \text{ loses}\end{array} = \sum_{j \text{ s.t. } j \neq i,\, i \text{ loses against } j} \text{Area}(L^w_{ij}) \quad \leq \quad \alpha \cdot \sum_{j'=1}^{n+1} \text{Area}(L^w_{ij'}) - \text{Area}(L^w_{ii}). \tag{21}$$

Now, as in Barber et al. (2021), consider that every pair of distinct strange points i and j ($i \neq j$) is also a pair of points where one point is strange and the other point is a point that the strange point loses to (since in our construction no game can be won by both points). In other words, the set of all pairs of strange points is a subset of the set of all pairs of points where at least one point is strange and the other point is a point that the strange point loses to. Accordingly, the total weighted importance of all the games played between two strange points is at most the total weighted importance of all the games between a strange point and a point that the strange point loses to:

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} \sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \frac{1}{2} \cdot \text{Area}(L^w_{ij}) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} \Big(\alpha \cdot \sum_{j'=1}^{n+1} \text{Area}(L^w_{ij'}) - \text{Area}(L^w_{ii})\Big), \tag{22}$$

where the $\frac{1}{2}$ on the left hand side is to avoid double-counting games (without the $\frac{1}{2}$, then the double summation would count $\text{Area}(L^w_{ij})$ and $\text{Area}(L^w_{ji})$ separately, while with the $\frac{1}{2}$ we appropriately count the weight of the game between i and j as $\frac{1}{2}\text{Area}(L^w_{ij}) + \frac{1}{2}\text{Area}(L^w_{ji}) = \text{Area}(L^w_{ij})$). Note that for exchangeable data where the weights are uniform

[4] Here we focus on a bound on the *expectation* of the total normalized weight of strange points roughly because in the current setting the weight normalization is done with respect to a given row of $\hat{A}^w$ (whereas in Prinster et al. (2022) the weight for each row cancels out in the strange point definition). This bound on the expectation of total strange point weights is still sufficient for use later in proof step 2. We also have some stylistic differences from the proof step 1 in Prinster et al. (2022) intended to improve clarity: in particular here we aim to simplify the argument for obtaining the initial bound (22) in part by initially using unnormalized rather than normalized weights to set up the argument.

$w(X_i; z_{-\{i,j\}}) = w(X_j; z_{-\{i,j\}}) = 1 \; \forall \; i \in \{1, ..., n+1\}$, then Area$(L_{ij}^w) = 1 \; \forall \; i, j \in \{1, ..., n+1\}$ and (22) reduces to the bound on the number of pairs of strange points in Barber et al. (2021). Moreover, in the standard covariate shift case, (22) is equivalent up to a normalization constant with the analogous bound in Prinster et al. (2022).

Writing out (22) explicitly with how we defined the rectangles L_{ij}^w, we have

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} \Big(\sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \frac{1}{2} w(X_i; z_{-\{i,j\}}) w(X_j; z_{-\{i,j\}}) \Big) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} \Big(\alpha \cdot \sum_{j'=1}^{n+1} \big[w(X_i; z_{-\{i,j'\}}) w(X_{j'}; z_{-\{i,j'\}}) \big] - w(X_i; z_{-\{i\}})^2 \Big).$$

Recall that $\sum_{j'=1}^{n+1} \big[w(X_i; z_{-\{i,j'\}}) w(X_{j'}; z_{-\{i,j'\}}) \big]$ is a positive normalization term that depends only on point i, so let us temporarily denote this term $C_i^w = \sum_{j'=1}^{n+1} \big[w(X_i; z_{-\{i,j'\}}) w(X_{j'}; z_{-\{i,j'\}}) \big]$. Substituting in this notation and multiplying each summand within the summation $\sum_{i \in \mathcal{S}(\hat{A}^w)}$ on both sides by $1 = \frac{C_i^w}{C_i^w}$, we obtain

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} \frac{C_i^w}{C_i^w} \cdot \Big(\sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \frac{1}{2} w(X_i; z_{-\{i,j\}}) w(X_j; z_{-\{i,j\}}) \Big) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} \frac{C_i^w}{C_i^w} \cdot \Big(\alpha \cdot C_i^w - w(X_i; z_{-\{i\}})^2 \Big)$$

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \frac{1}{2} \frac{w(X_i; z_{-\{i,j\}}) w(X_j; z_{-\{i,j\}})}{C_i^w} \Big) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\alpha - \frac{w(X_i; z_{-\{i\}})^2}{C_i^w} \Big)$$

and recalling our definition of normalized weights in (19), along with our notation for the normalizing term C_i^w, this is equivalent to

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\frac{1}{2} \sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \tilde{w}_{ij}(X_{n+1}) \Big) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\alpha - \tilde{w}_{ii}(X_{n+1}) \Big).$$

Adding $\frac{1}{2} \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \tilde{w}_{ii}(X_{n+1})$ to each side and simplifying we have

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\frac{1}{2} \sum_{j \in \mathcal{S}(\hat{A}^w) \setminus i} \tilde{w}_{ij}(X_{n+1}) \Big) + \frac{1}{2} \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \tilde{w}_{ii}(X_{n+1}) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\alpha - \tilde{w}_{ii}(X_{n+1}) \Big) + \frac{1}{2} \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \tilde{w}_{ii}(X_{n+1})$$

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\frac{1}{2} \sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{ij}(X_{n+1}) \Big) \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \Big(\alpha - \frac{1}{2} \tilde{w}_{ii}(X_{n+1}) \Big).$$

Note that we have written a form of the inequality where both sides take the form $\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot (\text{second term})$, that is the left and right side only differ in the second term in the summand of $\sum_{i \in \mathcal{S}(\hat{A}^w)}$. With this observation we can take the expectation of this second term on either side while maintaining the inequality, factor out the expectation terms, and simplify

$$\sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \mathbb{E}\Big[\frac{1}{2} \sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{ij}(X_{n+1}) \Big] \leq \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \cdot \mathbb{E}\Big[\alpha - \frac{1}{2} \tilde{w}_{ii}(X_{n+1}) \Big]$$

$$\mathbb{E}\Big[\frac{1}{2} \sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{ij}(X_{n+1}) \Big] \cdot \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w \leq \mathbb{E}\Big[\alpha - \frac{1}{2} \tilde{w}_{ii}(X_{n+1}) \Big] \cdot \sum_{i \in \mathcal{S}(\hat{A}^w)} C_i^w$$

$$\frac{1}{2} \mathbb{E}\Big[\sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{ij}(X_{n+1}) \Big] \leq \alpha - \frac{1}{2} \mathbb{E}\Big[\tilde{w}_{ii}(X_{n+1}) \Big]$$

where $\mathbb{E}[\tilde{w}_{ii}(X_{n+1})] > 0$, so we can write

$$\mathbb{E}\Big[\sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{ij}(X_{n+1})\Big] \leq 2\alpha, \tag{23}$$

which completes step 1 of the proof.

Step 2: Pseudo-exchangeability of the datapoints. We now leverage the pseudo-exchangeability of the data to show that, since the total weight of the strange points, normalized with respect to any point $i \in \{1, ..., n+1\}$, is in expectation at most 2α, that a test point has at most 2α probability of being strange. We organize this step into the following pieces:

- Step 2.1: Argue that $\hat{A}^w \overset{\text{d}}{=} P_\pi \hat{A}^w P_\pi^\top$ for any $(n+1) \times (n+1)$ permutation matrix P_π.
- Step 2.2: Argue that $\mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} = \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\}$ for all $j \in \{1, ..., n+1\}$.
- Step 2.3: Use the fact that the total expected weight of the strange points is at most 2α (from Step 1) to show that $\mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} \leq 2\alpha$.

Beginning with Step 2.1, for a permutation π of $\{1, ..., n+1\}$, let P_π denote the corresponding permutation matrix—that is, $\pi(i') = i \iff P_\pi(i', i) = 1$, which corresponds to the ith row in A becoming the i'th row in $P_\pi A$. Then, observe that $\hat{A}^w_{ii} = 0$ for all $i \in \{1, ..., n+1\}$, since $\mathbb{1}\{R_{ii} > R_{ii}\} = 0$ for all i. So, since an entry in the diagonal of $\hat{A}^w$ will always be mapped to another location in the diagonal of $P_\pi \hat{A}^w P_\pi^\top$, then, deterministically, the diagonal entries of both $\hat{A}^w$ and $P_\pi \hat{A}^w P_\pi^\top$ will be all zeros. So, to prove $\hat{A}^w \overset{\text{d}}{=} P_\pi \hat{A}^w P_\pi^\top$ it is sufficient to prove that $\hat{A}^w$ and $P_\pi \hat{A}^w P_\pi^\top$ are equivalent in distribution in their off-diagonal entries.

Recall that $\hat{W} \odot A$ has entries $(\hat{W} \odot A)_{ij} = w(X_i; z_{-\{i,j\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\}$ (equivalent to A with each ith row weighted by $w(X_i; z_{-\{i,j\}})$) and that $A \odot \hat{W}^\top$ has entries $(A \odot \hat{W}^\top)_{ij} = w(X_j; z_{-\{i,j\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\}$ (equivalent to A with each jth column weighted by $w(X_j; z_{-\{i,j\}})$). Moreover, note that $P_\pi(\hat{W} \odot A)$—which results from permuting the rows of $\hat{W} \odot A$—does not change the column membership of any entry in $\hat{W} \odot A$. In particular, $P_\pi(\hat{W} \odot A)$ has entries $(P_\pi(\hat{W} \odot A))_{ij} = (\hat{W} \odot A)_{\pi(i)j}$. Similarly, $(A \odot \hat{W}^\top)P_\pi^\top$ does not change the row membership of any entry in $A \odot \hat{W}$, such that $(A \odot \hat{W}^\top)P_\pi^\top$ has entries $((A \odot \hat{W}^\top)P_\pi^\top)_{ji} = (A \odot \hat{W}^\top)_{j\pi(i)}$. So, to show that $\hat{A}^w$ and $P_\pi \hat{A}^w P_\pi^\top$ are equivalent in distribution in their off diagonal entries, it is sufficient to show each jth column in $\hat{W} \odot A$ is equivalent in distribution to the jth column in $P_\pi(\hat{W} \odot A)$ aside from the initial diagonal entries in $\hat{W} \odot A$, and that each jth row in $A \odot \hat{W}^\top$ is equivalent in distribution to the corresponding jth row in $(A \odot \hat{W}^\top)P_\pi^\top$ aside from the initial diagonal entries in $A \odot \hat{W}^\top$.

To show $P_\pi(\hat{W} \odot A) \overset{\text{d}}{=} \hat{W} \odot A$ aside from the initial diagonal entries of $\hat{W} \odot A$, we draw on and adapt ideas from the proof for Lemma 3 in Tibshirani et al. (2019). Using condensed notation for the data as $\{Z_1, ..., Z_{n+1}\} = \{(X_1, Y_1), ..., (X_{n+1}, Y_{n+1})\}$, denote by E_z the event that $\{Z_1, ..., Z_{n+1}\} = \{z_1, ..., z_{n+1}\}$, and let f denote the density function of the joint sample $Z_1, ..., Z_{n+1}$. To do so, we begin by conditioning on E_z and then inspecting the probability of the joint event $R_{n+1,j} = r_{ij}, R_{j,n+1} = r_{ji}$ for each $i \in \{1, ..., n+1\}$ in each jth column, which occurs when $Z_{n+1} = z_i$:

$$\begin{aligned}\mathbb{P}\{R_{n+1,j} = r_{ij}, R_{j,n+1} = r_{ji} \mid E_z\} &= \mathbb{P}\{Z_{n+1} = z_i \mid E_z\} \\ &= \frac{\sum_{\pi:\pi(n+1)=i} f(z_{\pi(1)}, ..., z_{\pi(n+1)})}{\sum_\pi f(z_{\pi(1)}, ..., z_{\pi(n+1)})},\end{aligned}$$

where the second line above follows by the same reasoning as in the proof for Lemma 3 in Tibshirani et al. (2019). Then, recalling that data from feedback covariate shift (5) are pseudo-exchangeable with weight functions $w_1 = ... = w_n = 1$ and $w_{n+1} = w = \frac{d\widetilde{P}_{X;D}}{dP_X}$, this becomes

$$\mathbb{P}\{R_{n+1,j} = r_{i,j}, R_{j,n+1} = r_{j,i} \mid E_z\} = \frac{\sum_{\pi:\pi(n+1)=i} w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}}) h(z_{\pi(1)}, ..., z_{\pi(n+1)})}{\sum_\pi w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}}) h(z_{\pi(1)}, ..., z_{\pi(n+1)})}$$

where the core function h does not depend on the order of its inputs, so we have

$$\begin{aligned}\mathbb{P}\{R_{n+1,j} = r_{i,j}, R_{j,n+1} = r_{j,i} \mid E_z\} &= \frac{\sum_{\pi:\pi(n+1)=i} w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}})h(z_1, ..., z_{n+1})}{\sum_{\pi} w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}})h(z_1, ..., z_{n+1})} \\ &= \frac{\sum_{\pi:\pi(n+1)=i} w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}})}{\sum_{\pi} w(x_{\pi(n+1)}; z_{-\{\pi(n+1),j\}})} \\ &= \frac{w(x_i; z_{-\{i,j\}})}{\sum_{i'=1}^{n+1} w(x_{i'}; z_{-\{i',j\}})},\end{aligned}$$

which is equivalent to the jth column of $\hat{W}$ divided by a normalization constant. We can then rewrite this probability statement as

$$(R_{n+1,j}, R_{j,n+1}) \mid E_z \sim \sum_{i=1}^{n+1} \frac{w(x_i; z_{-\{i,j\}})}{\sum_{i'=1}^{n+1} w(x_{i'}; z_{-\{i',j\}})} \delta_{(r_{ij}, r_{ji})}.$$

Due to the conditioning on E_z, this is equivalent to

$$(R_{n+1,j}, R_{j,n+1}) \mid E_z \sim \sum_{i=1}^{n+1} \frac{w(X_i; Z_{-\{i,j\}})}{\sum_{i'=1}^{n+1} w(X_{i'}; Z_{-\{i',j\}})} \delta_{(R_{ij}, R_{ji})},$$

and since this statement holds for any $\{Z_1, ..., Z_{n+1}\} = \{z_1, ..., z_{n+1}\}$, marginalization yields

$$(R_{n+1,j}, R_{j,n+1}) \sim \sum_{i=1}^{n+1} \frac{w(X_i; Z_{-\{i,j\}})}{\sum_{i'=1}^{n+1} w(X_{i'}; Z_{-\{i',j\}})} \delta_{(R_{ij}, R_{ji})}.$$

More generally, substituting in any index $i' \in \{1, ..., n+1\}$ in for $n+1$ in the argument above yields

$$(R_{i',j}, R_{j,i'}) \sim \sum_{i=1}^{n+1} \frac{w(X_i; Z_{-\{i,j\}})}{\sum_{i'=1}^{n+1} w(X_{i'}; Z_{-\{i',j\}})} \delta_{(R_{ij}, R_{ji})}, \tag{24}$$

where the only difference is on the left-hand side.

Statement (24) tells us that within each jth column, draws of $(R_{i',j}, R_{j,i'})$ from this discrete distribution resemble the analogous draw $(R_{n+1,j}, R_{j,n+1})$ for the test point. That is, the distribution of $(R_{i',j}, R_{j,i'})$ in (24) is irrespective of the index i' and so these draws "look exchangeable". Thus, the distribution of the off diagonal entries in the jth column of $\hat{W} \odot A$ do not depend on the ordering of the elements. By a similar argument, the distribution of the off diagonal entries in the ith row of $A \odot \hat{W}$ do not depend on the ordering of the elements, and therefore $P_\pi \hat{A}^w P_\pi^\top \overset{\text{d}}{=} \hat{A}^w$ for any $(n+1) \times (n+1)$ permutation matrix P_π, the desired result for Step 2.1.

Because $P_\pi \hat{A}^w P_\pi^\top \overset{\text{d}}{=} \hat{A}^w$ from Step 2.1, this implies $\mathbb{P}\{j \in \mathcal{S}(P_\pi \hat{A}^w P_\pi^\top)\} = \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\}$. Now, let P_π denote a specific permutation matrix that maps $n+1$ to j, that is where $P_\pi(j, n+1) = 1$. Then, deterministically, $n+1 \in \mathcal{S}(\hat{A}^w) \iff j \in \mathcal{S}(\Pi \hat{A}^w \Pi^\top)$, so we have

$$\mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} = \mathbb{P}\{j \in \mathcal{S}(P_\pi \hat{A}^w P_\pi^\top)\} = \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\}$$

for all $j = 1, ..., n+1$. That is, an arbitrary training point j is equally likely to be strange as the test point $n+1$, which concludes Step 2.2.

Then, we begin Step 2.3 by multiplying the result from Step 2.2 by $\tilde{w}_{i,j}(X_{n+1})$ and summing over j to obtain

$$\begin{aligned}
\sum_{j=1}^{n+1} \tilde{w}_{i,j}(X_{n+1}) \cdot \mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} &= \sum_{j=1}^{n+1} \tilde{w}_{i,j}(X_{n+1}) \cdot \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\} \\
\mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} \cdot \sum_{j=1}^{n+1} \tilde{w}_{i,j}(X_{n+1}) &= \sum_{j=1}^{n+1} \tilde{w}_{i,j}(X_{n+1}) \cdot \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\} \\
\mathbb{P}\{n+1 \in \mathcal{S}(\hat{A}^w)\} &= \sum_{j=1}^{n+1} \tilde{w}_{i,j}(X_{n+1}) \cdot \mathbb{P}\{j \in \mathcal{S}(\hat{A}^w)\} \\
&= \mathbb{E}\Big[\sum_{j \in \mathcal{S}(\hat{A}^w)} \tilde{w}_{i,j}(X_{n+1}) \Big] \\
&\leq 2\alpha,
\end{aligned}$$

where the last line follows from Step 1.

Step 3: Connection to JAW-FCS: We would now like to connect our strange point result from Step 2 to coverage of the JAW-FCS prediction interval. Following the approach of Barber et al. (2021), suppose that $Y_{n+1} \notin \widehat{C}^{\text{JAW-FCS}}_{n,\alpha}(X_{n+1})$. Then, either

$$\begin{aligned}
Y_{n+1} > Q_{1-\alpha}\Big(\sum_{j=1}^{n} \big[\tilde{w}_{n+1,j}(X_{n+1}) \delta_{\widehat{\mu}_{-j}(X_{n+1}) + R_j^{LOO}} \big] + \tilde{w}_{(n+1)^2}(X_{n+1}) \delta_{\infty} \Big) \\
\implies \sum_{j=1}^{n} \tilde{w}_{n+1,j}(X_{n+1}) \cdot \mathbb{1}\big\{ Y_{n+1} > \widehat{\mu}_{-j}(X_{n+1}) + R_j^{LOO} \big\} \geq 1 - \alpha
\end{aligned}$$

or otherwise

$$\begin{aligned}
Y_{n+1} < Q_{\alpha}\Big(\sum_{j=1}^{n} \big[\tilde{w}_{n+1,j}(X_{n+1}) \delta_{\widehat{\mu}_{-j}(X_{n+1}) + R_j^{LOO}} \big] + \tilde{w}_{(n+1)^2}(X_{n+1}) \delta_{-\infty} \Big) \\
\implies \sum_{j=1}^{n} \tilde{w}_{n+1,j}(X_{n+1}) \cdot \mathbb{1}\big\{ Y_{n+1} < \widehat{\mu}_{-j}(X_{n+1}) - R_j^{LOO} \big\} \geq 1 - \alpha
\end{aligned}$$

And we can write the union of these two events as

$$\begin{aligned}
1 - \alpha &\leq \sum_{j=1}^{n} \tilde{w}_{n+1,j}(X_{n+1}) \cdot \mathbb{1}\big\{ Y_{n+1} \notin \widehat{\mu}_{-j}(X_{n+1}) \pm R_j^{LOO} \big\} \\
&= \sum_{j=1}^{n} \tilde{w}_{n+1,j}(X_{n+1}) \cdot \mathbb{1}\big\{ \big| Y_j - \widehat{\mu}_{-j}(X_j) \big| < \big| Y_{n+1} - \widehat{\mu}_{-j}(X_{n+1}) \big| \big\} \\
&= \sum_{j=1}^{n+1} \tilde{w}_{n+1,j}(X_{n+1}) \cdot \mathbb{1}\big\{ R_{j,n+1} < R_{n+1,j} \big\}
\end{aligned}$$

from which we see that $n+1 \in \mathcal{S}(\hat{A}^w)$—that is, $n+1$ is a strange point. This result together with the result from Step 2 gives us

$$\begin{aligned}
\mathbb{P}\big\{ Y_{n+1} \notin \widehat{C}^{\text{JAW-FCS}}_{n,\alpha}(X_{n+1}) \big\} \leq \mathbb{P}\big\{ n+1 \in \mathcal{S}(\hat{A}^w) \big\} \leq 2\alpha \\
\therefore\ \mathbb{P}\big\{ Y_{n+1} \in \widehat{C}^{\text{JAW-FCS}}_{n,\alpha}(X_{n+1}) \big\} \geq 1 - 2\alpha
\end{aligned}$$

A.3. Proof for JAW-KLOO coverage under feedback covariate shift

We first restate the theorem before proceeding with the proof.

Theorem 4.1 *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0,1)$, the JAW-KLOO predictive interval in* (10) *satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{JAW-}K\text{LOO}}(X_{n+1})\} \geq 1 - 2\alpha.$$

Theorem 4.1 follows from Lemma A.2 and Theorem 3.1. With training data $Z_1, ..., Z_n$ and test point Z_{n+1} generated under feedback covariate shift (5), let $S_{\text{LOO}} \subseteq \{1, ..., n\}$ denote a subset of the training data selected using a procedure that is invariant to the ordering of the data, where we retrain a leave-one-out model $\widehat{\mu}_{-j}$ for each $j \in S_{\text{LOO}}$. By Lemma A.2 the random variables $\{Z_j : j \in S_{\text{LOO}}\} \cup \{Z_{n+1}\}$ are pseudo-exchangeable, generated under FCS. Note that the training procedure for every leave-one-out model $\widehat{\mu}_{-j}$ for $j \in S_{\text{LOO}}$ includes the points in the subset $\{Z_{j'} : j' \notin S_{\text{LOO}}\}$ in its training data, so assuming that the model-fitting algorithm $\mathcal{A}$ treats the data symmetrically, in the fitting each $\widehat{\mu}_{-j}$, training on the data $\{Z_{j'} : j' \notin S_{\text{LOO}}\}$ can be considered a subroutine of the model-fitting algorithm that is invariant to the ordering of the remaining data. Thus, treating JAW-KLOO as an instance of JAW-FCS where the former is given a smaller dataset S_{LOO} and a different model-fitting algorithm $\mathcal{A}_{-S_{\text{LOO}}}$ that depends on the data $\{1, ..., n\} \backslash S_{\text{LOO}}$ but that still treats data symmetrically, the JAW-KLOO coverage guarantee follows from the guarantee for JAW-FCS given in Theorem 3.1.

We note that K points in S_{LOO} used for leave-one-out training can be selected in a range of ways, for example based on the K points with largest weight or using uniform random sampling. Different approaches to selecting S_{LOO} will result in different tradeoffs—for example, as we saw in the experimental results of our main paper, deterministically selecting S_{LOO} as the points with the largest weight can result in overly wide intervals, whereas uniform random sampling (without replacement) could result in higher coverage variance. A further alternative variant to JAW-KLOO would be to sample points with replacement using sampling probabilities proportional to the normalized weights, and then using the standard jackknife+ on the sampled points (which could include duplicates due to sampling with replacement, thus approximating the JAW-KLOO weighted quantiles based on the empirical sampling frequencies). This last approach of random sampling with probabilities proportional to normalized weights is the approach that we take for the randomized JAW-KLOO variant evaluated in Figure 5, which pays the price of higher coverage variance.

A.4. Proof of WCV+ coverage under feedback covariate shift

We first restate the theorem before proceeding with the proof.

Theorem A.3. *Suppose data are generated under feedback covariate shift* (5) *and assume $\widetilde{P}_{X;D}$ is absolutely continuous with respect to P_X for all possible values of D. Then, for any miscoverage level, $\alpha \in (0,1)$, the K-fold WCV+ predictive interval in* (13) *satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \widehat{C}_{n,K,\alpha}^{\textit{WCV+FCS}}(X_{n+1})\big\} \geq 1 - 2\alpha - \mathbb{E}\Big[\sum_{j \in S_{k(i)} \backslash i} \tilde{w}_{ij}^{CV}(X_{n+1})\Big].$$

The proof for weighted cross validation+ under feedback covariate shift (WCV+FCS) coverage follows a similar structure as the proof for JAW-FCS coverage presented in Appendix A.2, so here we try to focus on the key differences.

The first two setup steps are identical to the corresponding setup steps in the proof for CV+ coverage, or Theorem 4, in Barber et al. (2021):

(a) We now suppose the hypothetical scenario where we have access to $n/K - 1$ additional test points, for a total of $m = n/K$ test points $\{(X_{n+1}, Y_{n+1}), ..., (X_{n+m}, Y_{n+m})\}$. We then partition the training data into sets $S_1, ..., S_K$ with m datapoints each and define $S_{K+1} = \{n+1, ..., n+m\}$ as the set of test points. For any pair of distinct partition indices $k, k' \in \{1, ..., K+1\}$ such that $k \neq k'$ we define $\tilde{\mu}_{-(S_k, S_{k'})}$ as the regression model fit with training and test data except with S_k and $S_{k'}$ removed (i.e., fit with $\{1, ..., n+m\} \backslash \{S_k \cup S_{k'}\}$).

(b) We then define the matrix of residuals $R^{CV} \in \mathbb{R}^{(n+m)\times(n+m)}$ as follows, where $k(i)$ denotes the index of the partition that contains point i (so that for $i, j \in \{1, ..., n+m\}$ where $i \neq j$, $k(i) = k(j) \iff i, j \in S_k$):

$$R_{ij}^{CV} = \begin{cases} +\infty & k(i) = k(j) \\ |Y_i - \tilde{\mu}_{-(S_{k(i)}, S_{k(j)})}| & k(i) \neq k(j) \end{cases}$$

In the next two setup steps we introduce changes to the proof for Theorem 4 in Barber et al. (2021) that are analogous to setup steps (c) and (d) in our proof for JAW-FCS coverage.

(c) We define a weighted comparison matrix $\hat{A}^{wCV} \in \mathbb{R}^{(n+m)\times(n+m)}$ analogously as with the JAW-FCS proof, but with some modifications to appropriately account for whether or not points are in the same cross-validation fold. First, let A^{CV} be the unweighted comparison matrix with entries $A^{CV}_{ij} = \mathbb{1}\{R^{CV}_{ij} > R^{CV}_{ji}\}$ (where $A^{CV}_{ij} = 0$ if $k(i) = k(j)$); A^{CV} is thus a block off-diagonal matrix with square, size $m \times m$ zero matrices along the diagonal and block off-diagonal entries $A^{CV}_{ij} = \mathbb{1}\{R^{CV}_{ij} > R^{CV}_{ji}\}$. Next, we define $\hat{W}^{CV}$ as the weight matrix with entries $\hat{W}^{CV}_{ii} = w(X_i; z_{-\{S_{k(i)}\}})$ along the diagonal; with entries $\hat{W}^{CV}_{ij} = w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}})$ in the block off-diagonal which corresponds to pairs of points not in the same cross-validation fold ($k(i) \neq k(j)$); and with zero otherwise for distinct points in the same cross-validation fold ($k(i) = k(j), i \neq j$). Then, define $\hat{A}^{wCV} = \hat{W}^{CV} \odot A \odot \hat{W}^{CV\top}$ to be the block off-diagonal matrix with entries $\hat{A}^{wCV}_{ij} = w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}})w(X_j; z_{-\{S_{k(i)},S_{k(j)}\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\}$ in the block off-diagonal entries, and square, size $m \times m$ zero matrices along the diagonal.

(d) Next, as in the JAW-FCS proof we define a set of strange points as points with unusually large residuals. By "unusually large" we again mean points i where $\mathbb{1}\{R^{CV}_{ij} > R^{CV}_{ji}\}$ for a sufficiently large *weighted portion* of other points j that we might compare point i to, and we can thus similarly reference the ith row of our weighted comparison matrix $\hat{A}^{wCV}$ for this information. In particular, we define the set of "strange" points $\mathcal{S}(\hat{A}^{wCV}) \subseteq \{1, ..., n+m\}$ as the set of points i where the sum of the ith row in $\hat{A}^{wCV}$ is at least a $1-\alpha$ portion of the sum of the ith row in $\hat{W}^{CV} \odot \hat{W}^{CV\top}$:

$$\begin{aligned}
\mathcal{S}(\hat{A}^{wCV}) &= \Big\{ i \in [n+m] \ : \ \sum_{j=1}^{n+m} \hat{A}^{wCV}_{ij} \geq (1-\alpha) \sum_{j'=1}^{n+m} [\hat{W}^{CV} \odot \hat{W}^{CV\top}]_{ij'} \Big\} \\
&= \Big\{ i \in [n+m] \ : \ \sum_{j \in \{1,...,n+m\} \setminus S_{k(i)}} \big[w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}}) w(X_j; z_{-\{S_{k(i)},S_{k(j)}\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\} \big] \\
&\qquad \geq (1-\alpha) \sum_{j' \in \{i\} \cup \{1,...,n+m\} \setminus S_{k(i)}} \big[w(X_i; z_{-\{S_{k(i)},S_{k(j')}\}}) w(X_{j'}; z_{-\{S_{k(i)},S_{k(j')}\}}) \big] \Big\}.
\end{aligned} \tag{25}$$

Dividing both sides of the inequality in our definition of $\mathcal{S}(\hat{A}^{wCV})$ by the normalization term $\sum_{j' \in \{i\} \cup \{1,...,n+m\} \setminus S_{k(i)}} \big[w(X_i; z_{-\{S_{k(i)},S_{k(j')}\}}) w(X_{j'}; z_{-\{S_{k(i)},S_{k(j')}\}}) \big]$ yields

$$\begin{aligned}
\mathcal{S}(\hat{A}^{wCV}) = \Big\{ & i \in [n+m] \ : \\
& \sum_{j \in \{1,...,n+m\} \setminus S_{k(i)}} \frac{w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}}) w(X_j; z_{-\{S_{k(i)},S_{k(j)}\}}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\}}{\sum_{j' \in \{i\} \cup \{1,...,n+m\} \setminus S_{k(i)}} \big[w(X_i; z_{-\{S_{k(i)},S_{k(j')}\}}) w(X_{j'}; z_{-\{S_{k(i)},S_{k(j')}\}}) \big]} \geq 1-\alpha \Big\},
\end{aligned}$$

and to further simplify notation we denote the resulting fraction on the left side of the inequality as $\tilde{w}^{CV}_{i,j}(X_{n+1})$, that is

$$\tilde{w}^{CV}_{i,j}(X_{n+1}) = \frac{w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}}) w(X_j; z_{-\{S_{k(i)},S_{k(j)}\}})}{\sum_{j' \in \{i\} \cup \{1,...,n+m\} \setminus S_{k(i)}} \big[w(X_i; z_{-\{S_{k(i)},S_{k(j')}\}}) w(X_{j'}; z_{-\{S_{k(i)},S_{k(j')}\}}) \big]}, \tag{26}$$

so our strange point definition becomes

$$\mathcal{S}(\hat{A}^{wCV}) = \Big\{ i \in [n+1] \ : \ \sum_{j \in \{1,...,n+m\} \setminus S_{k(i)}} \Big(\tilde{w}^{CV}_{i,j}(X_{n+1}) \cdot \mathbb{1}\{R_{ij} > R_{ji}\} \Big) \geq 1-\alpha \Big\}. \tag{27}$$

As in the proof for JAW-FCS, we will call the quantity $\tilde{w}^{CV}_{i,j}(X_{n+1})$ defined in (26) a "normalized cross-validation weight for point j, normalized with respect to i" since it is normalized with respect to the sum over the ith row in the matrix $\hat{W}^{CV} \odot \hat{W}^{CV\top}$. Note that letting $i = n+1$ in (26) yields the weights $\tilde{w}^{CV}_{n+1,j}(X_{n+1})$ defined in the main paper (12) for the WCV+ predictive interval, and moreover for exchangeable data where $w(X_i; z_{-\{S_{k(i)},S_{k(j)}\}}) = 1$, the weights reduce to uniform weights and the set of strange points reduces to that in the CV+ coverage proof in Barber et al. (2021).

Step 1: Bounding the expected normalized weight of strange points

We begin similarly as Step 1 in our JAW-FCS coverage proof, except we need to make several adjustments to account for the fact that points in the same fold do not play against each other in the "tournament". This proof step extends the analogous exchangeable cross validation+ coverage proof in Barber et al. (2021) to feedback covariate shift.

The tournament setup for WCV+ is similar as that for JAW-FCS, except with the key difference that distinct points in the same cross-validation fold (i and j s.t. $i \neq j, k(i) = k(j)$) do not play a game against each other in the tournament. For the remaining pairs of points that do play games against each other in the tournament, as before we say that i "wins" its game against point j if $\hat{A}^{wCV}_{ij} > 0$ and that i loses against j otherwise, and we similarly "weight" the importance of a game between points i and j by the product $w(X_i; z_{-\{S_{k(i)}, S_{k(j)}\}})w(X_j; z_{-\{S_{k(i)}, S_{k(j)}\}}) = [\hat{W}^{CV} \odot \hat{W}^{CV\top}]_{ij}$. We moreover define L^{wCV}_{ij} as the rectangle with width $w(X_i; z_{-\{S_{k(i)}, S_{k(j)}\}})$ and height $w(X_j; z_{-\{S_{k(i)}, S_{k(j)}\}})$, so that $\text{Area}(L^{wCV}_{ij}) = \text{Area}(L^{wCV}_{ji}) = w(X_i; z_{-\{S_{k(i)}, S_{k(j)}\}})w(X_j; z_{-\{S_{k(i)}, S_{k(j)}\}}) = [\hat{W}^{CV} \odot \hat{W}^{CV\top}]_{ij}$. We can thus again interpret $\sum_{j' \in \{i\} \cup \{1,\ldots,n+m\} \setminus S_{k(i)}} \text{Area}(L^{wCV}_{ij'}) = \sum_{j' \in \{i\} \cup \{1,\ldots,n+m\} \setminus S_{k(i)}} \left[w(X_i; z_{-\{S_{k(i)}, S_{k(j)}\}})w(X_j; z_{-\{S_{k(i)}, S_{k(j)}\}})\right] = \sum_{j' \in \{1,\ldots,n+m\}} [\hat{W}^{CV} \odot \hat{W}^{CV\top}]_{ij'}$ as the total weighted importance of the games that point i plays in (including a "game" against itself), and $\tilde{w}^{CV}_{i,j}(X_{n+1})$ in (26) as the relative importance of i's game against j, with respect to all the games that i plays.

With a similar argument as in the corresponding JAW-FCS proof step, from our definition of strange points we can again obtain a bound on the total weighted importance of the games where strange point i plays against and loses to as

$$\begin{array}{l}\text{Total weighted importance of} \\ \text{games against } \textit{other} \text{ points} \\ \text{where strange point } i \text{ loses}\end{array} = \sum_{j \text{ s.t. } j \neq i,\, i \text{ loses against } j} \text{Area}(L^{w}_{ij}) \quad \leq \quad \alpha \cdot \sum_{j' \in \{i\} \cup \{1,\ldots,n+m\} \setminus S_{k(i)}} \text{Area}(L^{w}_{ij'}) - \text{Area}(L^{w}_{ii}). \tag{28}$$

Now, whereas in the JAW-FCS coverage proof we derive the inequality (22) by leveraging the observation that (in JAW-FCS's leave-one-out construction) a pair of strange points is also a pair of points where one point is strange and the other is a point that loses to the strange point, the same statement is not true for the WCV+. In particular, in WCV+ a pair of strange points $\{i, j\}$ s.t. $i, j \in \mathcal{S}(A^w)$ might consist of points in the same fold, that is where $k(i) = k(j)$, which implies that i and j do not play against each other in the tournament, and thus there is no loser. We thus need to separately account for two types of pairs of strange points $\{i, j\}$ s.t. $i, j \in \mathcal{S}(A^w), i \neq j$: one type where the strange points in the pair play against each other in the tournament (i.e., where $k(i) \neq k(j)$), and another type where the strange points in a pair do not play against each other in the tournament (i.e., where $k(i) = k(j)$).

For the first type of strange point pair, by essentially the same arguments as in the JAW-FCS coverage proof to obtain (22), we can obtain an inequality that appears similar to (22) but with summations restricted only to pairs of strange points that play each other in the tournament:

$$\sum_{i \in \mathcal{S}(\hat{A}^{wCV})} \sum_{j \in \mathcal{S}(\hat{A}^{wCV}) \setminus S_{k(i)}} \frac{1}{2} \cdot \text{Area}(L^{wCV}_{ij}) \leq \sum_{i \in \mathcal{S}(\hat{A}^{wCV})} \Big(\alpha \cdot \sum_{j' \in \{i\} \cup \{1,\ldots,n+m\} \setminus S_{k(i)}} \text{Area}(L^{wCV}_{ij'}) - \text{Area}(L^{wCV}_{ii})\Big). \tag{29}$$

Separately, we also account for the weights of strange point pairs in the same fold. We can write the (adjusted by $\frac{1}{2}$) sum of $\text{Area}(L^{wCV}_{ij}) = w(X_i; z_{-\{S_{k(i)}, S_{k(j)}\}})w(X_j; z_{-\{S_{k(i)}, S_{k(j)}\}})$ (which we can interpret as the weight of a hypothetical game between points i and j that do not play each other in the tournament) over all pairs of distinct strange points in the same fold ($i, j \in \mathcal{S}(\hat{A}^{wCV}), k(i) = k(j), i \neq j$) as $\frac{1}{2} \sum_{i \in \mathcal{S}(A^{wCV})} \sum_{j \in S_{k(i)} \cap \mathcal{S}(A^{wCV}) \setminus i} \text{Area}(L^{wCV}_{ij})$, and adding this quantity

to both sides of (29) we obtain

$$\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}\sum_{j\in\mathcal{S}(\hat{A}^{wCV})\setminus i}\frac{1}{2}\cdot\text{Area}(L_{ij}^{wCV}) \leq \sum_{i\in\mathcal{S}(\hat{A}^{wCV})}\Big(\alpha\cdot\sum_{j'\in\{i\}\cup\{1,\ldots,n+m\}\setminus S_{k(i)}}\text{Area}(L_{ij'}^{wCV})-\text{Area}(L_{ii}^{wCV})+\frac{1}{2}\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\text{Area}(L_{ij}^{wCV})\Big) \tag{30}$$

Then, as in the JAW-FCS proof, we can denote the normalization term $C_i^{wCV} = \sum_{j'\in\{i\}\cup\{1,\ldots,n+m\}\setminus S_{k(i)}}\left[w(X_i;z_{-\{S_{k(i)},S_{k(j')}\}})w(X_{j'};z_{-\{S_{k(i)},S_{k(j')}\}})\right]$, and multiplying each summand inside $\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}$ by $\frac{C_i^{wCV}}{C_i^{wCV}}$ and simplifying using our definitions of L_{ij}^{wCV}, C_i^{wCV}, and $\tilde{w}_{ij}^{CV}(X_{n+1})$ to obtain

$$\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\sum_{j\in\mathcal{S}(\hat{A}^{wCV})\setminus i}\frac{1}{2}\cdot\tilde{w}_{ij}^{CV}(X_{n+1}) \leq \sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\Big(\alpha-\tilde{w}_{ii}^{CV}(X_{n+1})+\frac{1}{2}\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big).$$

Then, adding $\frac{1}{2}\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\tilde{w}_{ii}^{CV}(X_{n+1})$ and simplifying, we have

$$\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\frac{1}{2}\cdot\tilde{w}_{ij}^{CV}(X_{n+1}) \leq \sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\Big(\alpha-\frac{1}{2}\tilde{w}_{ii}^{CV}(X_{n+1})+\frac{1}{2}\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big)$$

$$\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\tilde{w}_{ij}^{CV}(X_{n+1}) \leq \sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\Big(2\alpha-\tilde{w}_{ii}^{CV}(X_{n+1})+\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big)$$

As in the JAW-FCS proof, the inequality is of the form $\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot(\text{second term})$ for a second term on each side. We can thus take the expectation of the second term on each side while maintaining the inequality, factor out the expectation term, and simplify to obtain

$$\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\mathbb{E}\Big[\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\tilde{w}_{ij}^{CV}(X_{n+1})\Big] \leq \sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}\cdot\mathbb{E}\Big[2\alpha-\tilde{w}_{ii}^{CV}(X_{n+1})+\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big]$$

$$\mathbb{E}\Big[\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\tilde{w}_{ij}^{CV}(X_{n+1})\Big]\cdot\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV} \leq \mathbb{E}\Big[2\alpha-\tilde{w}_{ii}^{CV}(X_{n+1})+\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big]\cdot\sum_{i\in\mathcal{S}(\hat{A}^{wCV})}C_i^{wCV}$$

$$\mathbb{E}\Big[\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\tilde{w}_{ij}^{CV}(X_{n+1})\Big] \leq 2\alpha-\mathbb{E}[\tilde{w}_{ii}^{CV}(X_{n+1})]+\mathbb{E}\Big[\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big].$$

Note that the added expectation term on the right hand side $\mathbb{E}[\sum_{j\in S_{k(i)}\cap\mathcal{S}(A^{wCV})\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})]$ is the expected normalized weight of strange points $j\neq i$ in cross-validation a fold $k(i)$, so as an upper bound for this term we can use the expected normalized weight of *all* (strange and not strange) points $j\neq i$ in fold $k(i)$, that is $\mathbb{E}[\sum_{j\in S_{k(i)}\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})]$. With this

observation and noting that $\mathbb{E}[\tilde{w}_{ii}^{CV}(X_{n+1})]$ is positive, we obtain the bound

$$\mathbb{E}\Big[\sum_{j\in\mathcal{S}(\hat{A}^{wCV})}\tilde{w}_{ij}^{CV}(X_{n+1})\Big] \leq 2\alpha + \mathbb{E}\Big[\sum_{j\in S_{k(i)}\setminus i}\tilde{w}_{ij}^{CV}(X_{n+1})\Big], \tag{31}$$

which completes step 1 for our WCV+ proof.

Step 2: Pseudo exchangeability of the datapoints. We now leverage the pseudo exchangeability of the data to show that, since the expected total weight of the strange points is at most $2\alpha + \mathbb{E}\big[\sum_{j\in S_{k(i)}\setminus i}\big(\tilde{w}_{ij}^{CV}(X_{n+1})\big)\big]$, that a test point has at most $2\alpha + \mathbb{E}\big[\sum_{j\in S_{k(i)}\setminus i}\big(\tilde{w}_{ij}^{CV}(X_{n+1})\big)\big]$ probability of being strange.

This step proceeds similarly as the analogous step 2 in the JAW-FCS proof, except with a restriction on the permutation matrix P_π to maintain the fold structure of the data. That is, for any $(n+m)\times(n+m)$ permutation matrix P_π where $i \sim j$ if $k(i) = k(j)$, through a similar argument as in the JAW-FCS proof we can show that $\overset{\text{d}}{=} P_\pi \hat{A}^w P_\pi^\top$ for any such permutation matrix P_π. When combined with (31), the result for this step then follows.

Step 3: Connection to weighted CV+: The last proof step proceeds similarly as the third main step in the JAW-FCS proof. Through the same procedure, we establish that

$$\therefore\ \mathbb{P}\big\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{JAW}}(X_{n+1})\big\} \geq 1 - 2\alpha - \mathbb{E}\Big[\sum_{j\in S_{k(i)}\setminus i}\Big(\tilde{w}_{ij}^{CV}(X_{n+1})\Big)\Big] \tag{32}$$