Training-Conditional Coverage Bounds for Uniformly Stable Learning Algorithms

Mehrdad Pournaderi and Yu Xiang
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT 84112, USA
{m.pournaderi, yu.xiang}@utah.edu

***Abstract*—The training-conditional coverage performance of the conformal prediction is known to be empirically sound. Recently, there have been efforts to support this observation with theoretical guarantees. The training-conditional coverage bounds for jackknife+ and full-conformal prediction regions have been established via the notion of (m, n)-stability by Liang and Barber [2023]. Although this notion is weaker than uniform stability, it is not clear how to evaluate it for practical models. In this paper, we study the training-conditional coverage bounds of full-conformal, jackknife+, and CV+ prediction regions from a uniform stability perspective which is known to hold for empirical risk minimization over reproducing kernel Hilbert spaces with convex regularization. We derive coverage bounds for finite-dimensional models by a concentration argument for the (estimated) predictor function, and compare the bounds with existing ones under ridge regression.**

I. Introduction and Problem Formulation

Conformal prediction is a framework for constructing *distribution-free* predictive confidence regions as long as the training and test data are exchangeable [1] (also see [2]–[4]). Specifically, let $\mathcal{D}_n \cup (X_{\text{test}}, Y_{\text{test}})$ denote a dataset with exchangeable data points, consisting of a training set of n samples $\mathcal{D}_n := \{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ and one test sample $(X_{\text{test}}, Y_{\text{test}})$, where $[n] := \{1, 2, ..., n\}$. The conformal prediction provides a coverage of Y_{test} in the sense of

$$\mathbb{P}(Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})) \geq 1 - \alpha, \tag{1}$$

where $\hat{C}_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ is a data-dependent map. This type of guarantee is referred to as *marginal* coverage, as it is averaged over all the training and test data. One natural direction to stronger results is to devise *conditional* coverage guarantee

$$\mathbb{P}(Y_{\text{test}} \in \hat{\mathcal{C}}_\alpha(X_{\text{test}}) | X_{\text{test}}) \geq 1 - \alpha.$$

However, it has been shown in [4]–[6] that it is *impossible* to obtain (non-trivial) distribution-free prediction regions $\hat{\mathcal{C}}(x)$ in the finite-sample regime; relaxed versions of this type of guarantee have been extensively studied (see [7]–[9] and references therein). As an alternative approach, several results (e.g., [4], [10]) have been reported on the *training-conditional* guarantee by conditioning on $\mathcal{D}_n$, which is also more appealing than the marginal guarantee as can be seen below. Define the following miscoverage rate as a function of the training data,

$$P_e(\mathcal{D}_n) := \mathbb{P}(Y_{\text{test}} \notin \hat{\mathcal{C}}(X_{\text{test}}) | \mathcal{D}_n).$$

Note that the marginal coverage in (1) is equivalent to $\mathbb{E}[P_e(\mathcal{D}_n)] \leq \alpha$. The training-conditional guarantees are of the following form, for some small δ,

$$\mathbb{P}(P_e(\mathcal{D}_n) \geq \alpha) \leq \delta$$

or its asymptotic variants. Roughly speaking, this guarantee means that the $(1 - \alpha)$-level coverage lower bounds hold for a *generic* dataset.

In this line of research, samples are assumed to be i.i.d., which is not only exchangeable but also ergodic and admits some nice concentration properties. For the K*-fold CV+* with m samples in each fold, the conditional coverage bound

$$\mathbb{P}\Big(P_e(\mathcal{D}_n) \geq 2\alpha + \sqrt{2 \log(K/\delta)/m}\Big) \leq \delta \tag{2}$$

is established in [10]. They have also shown that distribution-free training-conditional guarantees for full-conformal and *jackknife+* methods are impossible without further assumptions; in particular, they conjectured that a certain form of algorithmic stability is needed for full-conformal and jackknife+. Recently, [11] proposed (asymptotic) conditional coverage bounds for jackknife+ and full-conformal prediction sets under the assumption that the training algorithm is symmetric. The bound, however, depends on the distribution of the data through the so-called (m, n)-stability parameters, where the convergence rate can be slow (see Section IV).

This work is motivated by a large class of regression models that can be written as finite-dimensional empirical risk minimization over a reproducing kernel Hilbert space with regularization, i.e., $\hat{\mu}_{\mathcal{D}_n} = g_{\hat{\theta}_n}$ with

$$\hat{\theta}_n = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i \in [n]} \ell(g_\theta(X_i), Y_i) + \lambda \|g_\theta\|^2,$$

where g_θ, $\theta \in \mathbb{R}^p$ is a family of predictor functions parametrized by θ and ℓ is some suitable loss function. These models are known to be *uniformly stable* [12] in the sense that

$$\|\hat{\mu}_{\mathcal{D}_n} - \hat{\mu}_{\mathcal{D}'_n}\|_\infty \leq \beta \tag{3}$$

with $\beta = O(1/n)$ for any two datasets $(\mathcal{D}_n, \mathcal{D}'_n)$ that differ in one data sample, which is a stronger notion than the stability assumed in [11]. We aim to improve the training-conditional coverage guarantees of these learning models by establishing better rates of convergence for full-conformal and jackknife+.

II. Background and Related Work

A. Full-Conformal and Split-Conformal

Let T denote a symmetric training algorithm, i.e., the predictor function $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ is invariant under permutations of the training data points, and $\hat{\mu}_{(x,y)} := T(\mathcal{D}_n \cup (x,y))$ is a regression fucntion by running T on $\mathcal{D}_n \cup (x,y)$. Define the score function $s(x', y'; \hat{\mu}_{(x,y)}) := f(\hat{\mu}_{(x,y)}(x'), y')$ via some arbitrary (measurable) cost function f. For instance, $s(x', y'; \hat{\mu}_{(x,y)}) = |y' - \hat{\mu}_{(x,y)}(x')|$ when $f(y, y') = |y - y'|$. Let

$$\mathcal{S}(x, y; \mathcal{D}_n) := \{s(x', y'; \hat{\mu}_{(x,y)}) : (x', y') \in \mathcal{D}_n \cup (x, y)\}$$

and observe that the elements of $\mathcal{S}(X_{\text{test}}, Y_{\text{test}}; \mathcal{D}_n)$ are exchangeable. Therefore,

$$\mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}_{(X_{\text{test}}, Y_{\text{test}})}) \leq \hat{F}^{-1}_{\mathcal{S}(X_{\text{test}}, Y_{\text{test}};\, \mathcal{D}_n)}(1-\alpha)\Big) \geq 1 - \alpha,$$

where $\hat{F}^{-1}_{\mathcal{S}(X_{\text{test}}, Y_{\text{test}};\, \mathcal{D}_n)}(1-\alpha)$ denotes the *empirical* quantile function with respect to the set of values $\{\mathcal{S}(X_{\text{test}}, Y_{\text{test}};\, \mathcal{D}_n)\}$. Thus,

$$\mathbb{P}(Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})) \geq 1 - \alpha,$$

where the following confidence region is referred to as *full-conformal* in the literature

$$\hat{C}_\alpha(x) = \{y : s(x, y; \hat{\mu}_{(x,y)}) \leq \hat{F}^{-1}_{\mathcal{S}(x,y;\mathcal{D}_n)}(1-\alpha)\}.$$

It is well-known that this approach can be computationally intensive when $\mathcal{Y} = \mathbb{R}$ since to find out whether $y \in \hat{C}_\alpha(x)$ one needs to train the model with the dataset including (x, y) with $y \in \mathbb{R}$. One simple way to alleviate this issue is to split the data into training and calibration datasets, namely $\mathcal{D}_n = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}$. First one finds the regression $\hat{\mu} := T(\mathcal{D}^{\text{train}})$ and treats $\hat{\mu}$ as fixed. Let $\tilde{\mathcal{S}}(\mathcal{D}_n) := \{s(x, y; \hat{\mu}) : (x, y) \in \mathcal{D}_n\}$, and note that the elements of $\tilde{\mathcal{S}}((X_{\text{test}}, Y_{\text{test}}) \cup \mathcal{D}^{\text{cal}})$ are exchangeable. Hence, we get

$$\mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}) \leq \hat{F}^{-1}_{\tilde{\mathcal{S}}((X_{\text{test}}, Y_{\text{test}}) \cup \mathcal{D}^{\text{cal}})}(1-\alpha)\Big) \geq 1 - \alpha.$$

Hence,

$$\mathbb{P}(Y_{\text{test}} \in \hat{C}^{\text{split}}_\alpha(X_{\text{test}})) \geq 1 - \alpha,$$

for

$$\begin{aligned}\hat{C}^{\text{split}}_\alpha(x) &= \Big\{y : s(x, y; \hat{\mu}) \leq \hat{F}^{-1}_{\tilde{\mathcal{S}}(\mathcal{D}^{\text{cal}}) \cup \{\infty\}}(1-\alpha)\Big\} \\ &\supseteq \Big\{y : s(x, y; \hat{\mu}) \leq \hat{F}^{-1}_{\tilde{\mathcal{S}}((x,y) \cup \mathcal{D}^{\text{cal}})}(1-\alpha)\Big\}.\end{aligned}$$

B. Jackknife+

Although the split-conformal approach resolves the computational efficiency problem of the full-conformal method, it is somewhat inefficient in using the data and may not be useful in situations where the number of samples is limited. A heuristic alternative has long been known in the literature, namely, jackknife or leave-one-out cross-validation that can provide a compromise between the full conformal and split conformal methods. In particular,

$$\hat{C}^{\text{J}}_\alpha(x) = \{y : s(x, y; \hat{\mu}) \leq \hat{F}^{-1}_{\mathcal{S}^{\text{cal}}}(1-\alpha)\}$$

where $\mathcal{S}^{\text{cal}} := \{s(X_i, Y_i; \hat{\mu}^{-i}) : 1 \leq i \leq |\mathcal{D}^{\text{train}}|\}$ and $\hat{\mu}^{-i} := T(\mathcal{D}^{\text{train}} \setminus \{(X_i, Y_i)\})$. Despite its effectiveness, no general finite-sample guarantees are known for jackknife. Recently, [13] proposed jackknife+, a modified version of the jackknife for $\mathcal{Y} = \mathbb{R}$ and $f(y, y') = |y - y'|$, and established $(1 - 2\alpha)$ coverage lower bound for it. Let $\hat{q}^+_\alpha(A)$ and $\hat{q}^-_\alpha(A)$ denote the $\lceil (1-\alpha)(|A|+1) \rceil$-th and $\lfloor \alpha(|A|+1) \rfloor$-th smallest values of the set A, respectively, with the convention $\hat{q}^+_\alpha(A) = \infty$ if $\alpha < 1/(n+1)$. Let

$$\begin{aligned}\mathcal{S}^-(x) &= \{\hat{\mu}^{-i}(x) - |Y_i - \hat{\mu}^{-i}(X_i)| : i \in [n]\}, \\ \mathcal{S}^+(x) &= \{\hat{\mu}^{-i}(x) + |Y_i - \hat{\mu}^{-i}(X_i)| : i \in [n]\},\end{aligned}$$

and the jackknife+ prediction interval is defined as

$$\hat{C}^{\text{J+}}_\alpha(x) = [\hat{q}^-_\alpha(\mathcal{S}^-(x)),\ \hat{q}^+_\alpha(\mathcal{S}^+(x))].$$

In the same paper, an ϵ-inflated version of the jackknife+

$$\hat{C}^{\text{J+},\epsilon}_\alpha(x) = [\hat{q}^-_\alpha(\mathcal{S}^-(x)) - \epsilon,\ \hat{q}^+_\alpha(\mathcal{S}^+(x)) + \epsilon] \tag{4}$$

is proposed which has $1 - \alpha - 4\sqrt{\nu}$ coverage lower bound, instead of $1 - 2\alpha$, if the training procedure satisfies

$$\max_{i \in [n]} \mathbb{P}(|\hat{\mu}(X_{\text{test}}) - \hat{\mu}^{-i}(X_{\text{test}})| > \epsilon) < \nu.$$

Also, the jackknife+ has been generalized to CV+ for K-fold cross-validation, and $(1 - 2\alpha - \sqrt{2/|\mathcal{D}^{\text{train}}|})$ coverage lower bound is established.

C. Asymptotic Training Conditional Coverage [11]

The bounds established in [11] depend on the distribution of the data through the (m, n)-stability parameters,

$$\psi^{\text{out}}_{m,n} = \mathbb{E}_{\mathcal{D}_{n+m}} |\hat{\mu}_{\mathcal{D}_n}(X_{\text{test}}) - \hat{\mu}_{\mathcal{D}_{n+m}}(X_{\text{test}})|, \tag{5}$$

$$\psi^{\text{in}}_{m,n} = \mathbb{E}_{\mathcal{D}_{n+m}} |\hat{\mu}_{\mathcal{D}_n}(X_1) - \hat{\mu}_{\mathcal{D}_{n+m}}(X_1)|. \tag{6}$$

where $\hat{\mu}_{\mathcal{D}_n} = T(\mathcal{D}_n)$ and $X_{\text{test}} \perp\!\!\!\perp \mathcal{D}_{n+m}$ with $\mathcal{D}_{n+m} = \{(X_1, Y_1), ..., (X_{n+m}, Y_{n+m})\}$. Tight bounds for this parameter are not known yet. Therefore, the current convergence rates appear to be slow in sample size — see in Section IV for details. Furthermore, in this analysis, a γ-inflated version (as in (4)) of the method is considered, hence, one needs to deal with terms of the form $(\psi_{m,n}/\gamma)^{1/3}$ in the bound which can make the rates even slower if one let $\gamma \to 0$. We aim to improve the training-conditional coverage guarantees of these learning models in the following ways: (1) establishing $n^{-1/2}$ rates with explicit dependence on the dimension of the problem and (2) removing the interval inflation.

III. Conditional Coverage Guarantees

Let $\mu_\beta \in L^\infty(\mathcal{X})$ denote a predictor function parameterized by $\beta \in \mathbb{R}^p$. By a slight abuse of notation, let the map $T : \cup_{n\geq 1}(\mathcal{X}\times\mathcal{Y})^n \to \mathbb{R}^p$ denote a training algorithm for estimating β, hence, $\hat\beta_n = T(\mathbf{D}_n)$ where $\mathbf{D}_n := ((X_1,Y_1),\ldots,(X_n,Y_n)) \in (\mathcal{X}\times\mathcal{Y})^n$ denotes the i.i.d. training tuple of data points. In this case, we have $\hat\mu_{\mathbf{D}_n} = \mu_{\hat\beta_n}$.

Assumption 1 (Uniform stability): For all $i\in[n]$, we have

$$\sup_{z_1,\ldots,z_n} \|\mu_{T(z_1,\ldots,z_{i-1},z_{i+1}\ldots,z_n)} - \mu_{T(z_1,\ldots,z_i,\ldots,z_n)}\|_\infty \leq \frac{c_n}{2}.$$

In the case of the ridge regression [14] with $\mathcal{Y}=[-B,B]$ and $\mathcal{X} = \{x : \|x\|_2 \leq b\}$, this assumption holds with $c_n = 16\, b^2B^2/(\lambda\, n)$ where λ denotes the regularization parameter [12].

Assumption 2: The model is bi-Lipschitz (Lipeomorphism) in parameters,

$$\kappa_1\big\|\beta-\beta'\big\|_\infty \leq \big\|\mu_\beta - \mu_{\beta'}\big\|_\infty \leq \kappa_2\big\|\beta-\beta'\big\|_\infty,$$

with $\kappa_1 > 0$ and $\kappa_2 < \infty$.

Remark 1: It is worth noting that if the parameter space Θ is compact, $\Phi : U \to L^\infty(\mathcal{X})$ given by $\beta \mapsto \mu_\beta$ is continuously differentiable for some open $U \supseteq \Theta$, then $\kappa_2 < \infty$. Moreover, the inverse function theorem (for Banach spaces), gives the sufficient condition under which the inverse is continuously differentiable over $\Phi(U)$ and hence $\kappa_1 > 0$.

In the case of linear regression with $\mathcal{X} = \{x : \|x\|_2 \leq b\}$, one can verify that Assumption 2 holds with $\kappa_1 = b$ and $\kappa_2 = \sqrt{p}b$.

Let $\overline{\beta}_n = \mathbb{E}\,\hat\beta_n$, $\hat\beta_{-i} = T(Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)$ where $Z_i=(X_i,Y_i)$, and $\overline{\beta}_{-i} = \mathbb{E}\,\hat\beta_{-i}$. Define

$$F_{n-1}(t) := \mathbb{P}\Big(\Big|Y_1 - \mu_{\overline{\beta}_{-1}}(X_1)\Big| \leq t\Big),$$
$$\hat F_{n-1}(t) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\Big\{\Big|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\Big| \leq t\Big\}.$$

Assumption 3 (Bounded density): $F'_n < L_n$.

Theorem 1 (Jackknife+): Under Assumptions 1—3, for all $\epsilon,\delta>0$, it holds that

$$\mathbb{P}\left(P_e^{\text{J+}}(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + 2\,L_{n-1}\,\kappa_2\,c_{n-1}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right)\right) \leq \epsilon+\delta.$$

Using the same arguments as in the proof of this theorem, one can get a coverage bound for the CV+ as well. Unlike (2) which is meaningful only if the number of samples in each fold m is large, the bound we present in the following corollary is suitable for cases where $m/n \to 0$.

Corollary 1 (CV+): Under Assumptions 1—3, for all $\epsilon,\delta>0$, it holds that

$$\mathbb{P}\left(P_e^{\text{CV+}}(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + 2\,m\,L_{n-m}\,\kappa_2\,c_{n-m}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right)\right) \leq \epsilon+\delta.$$

The following theorem concerns the training-conditional guarantees for the full-conformal prediction regions.

Theorem 2 (Full-conformal): Under Assumptions 1—3, for all $\epsilon,\delta>0$, it holds that

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + L_n\left(c_{n+1} + \sqrt{2n\log\frac{2p}{\epsilon}}\,\frac{\kappa_2\,c_n}{\kappa_1}\right)\right) \leq \epsilon+\delta.$$

IV. Coverage Bounds for Ridge Regression

In this section, we wish to evaluate the bounds for the ridge regression with $\mathcal{X} = \{x : \|x\| \leq b\}$ and $\mathcal{Y} = [-B,B]$. As stated in the previous section, this regression model satisfies $c_n = 16\,b^2B^2/(\lambda\,n)$, $\kappa_1 = b$ and $\kappa_2 = \sqrt{p}\,b$. Hence, we get the following bound for both full-conformal and jackknife+ methods,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + O\left(n^{-1/2}\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{p\log(\frac{2p}{\epsilon})}\right)\right)\right) \leq \epsilon+\delta.$$

On the other hand, the following bound is proposed for the γ-inflated jackknife in [11],

$$\mathbb{P}\left(P_e^{\text{J+},\gamma}(\mathbf{D}_n) > \alpha + 3\sqrt{\frac{\log(1/\delta)}{\min(m,n)}} + 2\sqrt[3]{\frac{\psi^{\text{out}}_{m,n-1}}{\gamma}}\right) \leq 3\delta + \sqrt[3]{\frac{\psi^{\text{out}}_{m,n-1}}{\gamma}}. \tag{7}$$

for all $m \geq 1$. We get $\psi^{\text{out}}_{m,n} = O(mc_n)$ since $\psi^{\text{out}}_{1,n} \leq c_{n+1}/2$ by definition (5) and Assumption 1, and $\psi^{\text{out}}_{m,n} \leq \sum_{k=n}^{n+m-1}\psi^{\text{out}}_{1,k}$ holds according to in [11, Lemma 5.2] . Substituting for $\psi^{\text{out}}_{m,n-1}$ in bound (7), we obtain

$$\mathbb{P}\left(P_e^{\text{J+},\gamma}(\mathbf{D}_n) > \alpha + O\left(\sqrt{\frac{\log(1/\delta)}{\min(m,n)}} + \sqrt[3]{\frac{m\,c_{n-1}}{\gamma}}\right)\right) \leq 3\delta + O\left(\sqrt[3]{\frac{mc_{n-1}}{\gamma}}\right).$$

Letting $m^{-1/2} = (m/n)^{1/3}$ to balance the two terms $\sqrt{\frac{\log(1/\delta)}{\min(m,n)}}$ and $\sqrt[3]{mc_{n-1}/\gamma}$, we get $m = n^{2/5}$. By plugging $m = n^{2/5}$ in, we get

$$\mathbb{P}\left(P_e^{\text{J+},\gamma}(\mathbf{D}_n) > \alpha + O\left(n^{-1/5}\left(\sqrt{\log(1/\delta)} + \gamma^{-1/3}\right)\right)\right) \leq 3\delta + O\left(n^{-1/5}\gamma^{-1/3}\right). \tag{8}$$

This bound, although dimension-free, is very slow in the sample size. In [11], the same bound as (7) is established for γ-inflated full-conformal method except with $\psi^{\text{in}}_{m-1,n+1}$ instead of $\psi^{\text{out}}_{m,n-1}$. Hence, the same bound as (8) can be obtained for the γ-inflated full-conformal method via $\psi^{\text{in}}_{m,n} = O(mc_n)$.

V. Conclusion

The (m, n)-stability is a new measure of the stability of a regression model. It was recently introduced in [11] and used to compute training-conditional coverage bounds for full-conformal and jackknife+ prediction intervals. Unlike uniform stability which is a distribution-free property of a training process, (m, n)-stability depends on both the training algorithm and the distributions of the data. Although weaker than uniform stability, the parameter is not well-understood in a practical sense yet. In this work, we have studied the training-conditional coverage bounds of full-conformal, jackknife+, and CV+ prediction regions from a uniform stability perspective which is well understood for convexly regularized empirical risk minimization over reproducing kernel Hilbert spaces. We have derived new bounds via a concentration argument for the (estimated) predictor function. In the case of ridge regression, we have used the uniform stability parameter to derive a bound for the (m, n)-stability and compare the resulting bounds from [11] to the bounds established in this paper. We have observed that our rates are faster in sample size but dependent to the dimension of the problem.

References

[1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[2] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[3] V. Vovk, I. Nouretdinov, and A. Gammerman, "On-line predictive linear regression," *The Annals of Statistics*, pp. 1566–1590, 2009.

[4] V. Vovk, "Conditional validity of inductive conformal predictors," in *Asian conference on machine learning*. PMLR, 2012, pp. 475–490.

[5] R. Foygel Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Information and Inference: A Journal of the IMA*, vol. 10, no. 2, pp. 455–482, 2021.

[6] J. Lei and L. Wasserman, "Distribution-free prediction bands for non-parametric regression," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 76, no. 1, pp. 71–96, 2014.

[7] C. Jung, G. Noarov, R. Ramalingam, and A. Roth, "Batch multivalid conformal prediction," *arXiv preprint arXiv:2209.15145*, 2022.

[8] I. Gibbs, J. J. Cherian, and E. J. Candès, "Conformal prediction with conditional guarantees," *arXiv preprint arXiv:2305.12616*, 2023.

[9] V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman, "Mondrian confidence machine," *Technical Report*, 2003.

[10] M. Bian and R. F. Barber, "Training-conditional coverage for distribution-free predictive inference," *Electronic Journal of Statistics*, vol. 17, no. 2, pp. 2044–2066, 2023.

[11] R. Liang and R. F. Barber, "Algorithmic stability implies training-conditional coverage for distribution-free prediction methods," *arXiv preprint arXiv:2311.04295*, 2023.

[12] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[13] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," 2021.

[14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[15] C. McDiarmid *et al.*, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[16] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

APPENDIX A
PROOF FOR JACKKNIFE+

Lemma 1: If Assumption 1 and 2 hold, then

$$\mathbb{P}\left(\left\|\hat{\beta}_n - \mathbb{E}\,\hat{\beta}_n\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right).$$

Proof: Assumption 1 and 2 imply that

$$\sup_{z_1,\ldots,z_n,z_i'} \|T(z_1,\ldots,z_i\ldots,z_n) - T(z_1,\ldots,z_i',\ldots,z_n)\|_\infty \leq \frac{c_n}{\kappa_1}.$$

By McDiarmid's inequality [15] we get

$$\mathbb{P}\left(\|\hat{\beta}_n - \mathbb{E}\,\hat{\beta}_n\|_\infty \geq \epsilon\right) = \mathbb{P}\left(\left\|T(Z_1,\ldots,Z_n) - \mathbb{E}\;T(Z_1,\ldots,Z_n)\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right) \tag{9}$$

for independent Z_i and all $\epsilon > 0$. ■

Lemma 2: Under Assumptions 1 and 2 we have

$$\mathbb{P}\left(\max_i\left\|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right).$$

Proof: From Assumption 1 and 2, it follows that

$$\max_{i,j}\|\hat{\beta}_{-i} - \hat{\beta}_{-j}\|_\infty \leq \frac{c_{n-1}}{\kappa_1}. \tag{10}$$

Also, according to (1), we have $\|\hat{\beta}_{-1} - \overline{\beta}_{-1}\|_\infty < \epsilon$ with probability at least $1 - 2p\exp(-2\kappa_1^2\epsilon^2/(nc_{n-1}^2))$. We note that,

$$\begin{aligned}
\mathbb{P}\left(\max_i\left\|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\right\|_\infty \geq \epsilon\right) &\overset{(*)}{\leq} \mathbb{P}\left(\kappa_2\max_i\left\|\hat{\beta}_{-i} - \overline{\beta}_{-1}\right\|_\infty \geq \epsilon\right)\\
&\overset{(**)}{\leq} \mathbb{P}\left(\kappa_2\left(\frac{c_{n-1}}{\kappa_1} + \left\|\hat{\beta}_{-1} - \overline{\beta}_{-1}\right\|_\infty\right) \geq \epsilon\right)\\
&\leq 2p\exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right).
\end{aligned}$$

where (*) and (**) hold according to Assumption 2 and (10), respectively. ■

Let $\hat{\mathcal{C}}_\alpha(X_{n+1})$ denote the Jackknife+ α-level interval for test data-point X_{n+1} and define $P_e(\mathbf{D}_n) := \mathbb{P}(Y_{n+1} \notin \hat{\mathcal{C}}(X_{n+1})|\mathbf{D}_n)$.

Proof: We note,

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{-i}}(X_{n+1})\right|\right\} > \alpha\right\}\\
&\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| - \left|\mu_{\hat{\beta}_{-i}}(X_i) - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \right.\right.\\
&\qquad\qquad \left.\left.\left|y - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right| + \left|\mu_{\hat{\beta}_{-i}}(X_{n+1}) - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right|\right\} > \alpha\right\},
\end{aligned}$$

where the first relation holds according to [10]. Assuming $\max_i\|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon$, we obtain

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right| + 2\epsilon\right\} > \alpha\right\}\\
&\supseteq \left\{y \in \mathbb{R} : 1 - \hat{F}_{n-1}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right| + 2\epsilon\right) > \alpha\right\}.
\end{aligned}$$

Assuming $\left\|\hat{F}_{n-1} - F_{n-1}\right\|_\infty < \delta$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : 1 - F_{n-1}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right| + 2\epsilon\right) > \alpha + \delta\right\} \\ &\supseteq \left\{y \in \mathbb{R} : 1 - F_{n-1}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{n+1})\right|\right) > \alpha + \delta + 2\epsilon L\right\}\end{aligned}$$

Therefore,

$$\begin{aligned}P_e(\mathbf{D}_n) = \mathbb{P}(Y_{n+1} \notin \hat{\mathcal{C}}(X_{n+1})|\mathbf{D}_n) &\leq \mathbb{P}\Big(1 - F_{n-1}\Big(\Big|Y_{n+1} - \mu_{\overline{\beta}_{-1}}(X_{n+1})\Big|\Big) \leq \alpha + \delta + 2\epsilon L\Big) \\ &= \alpha + \delta + 2\epsilon L\end{aligned}$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}$ where $\mathcal{A} := \left\{D : \max_i \|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon\right\}$ and $\mathcal{B} := \left\{D : \left\|\hat{F}_{n-1} - F_{n-1}\right\|_\infty < \delta\right\}$. From Lemma 2, we know $\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p \exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right)$. Also, according to Dvoretzky–Kiefer–Wolfowitz inequality [16], we have $\mathbb{P}(\mathbf{D}_n \notin \mathcal{B}) \leq 2e^{-2n\delta^2}$. Thus,

$$\mathbb{P}(P_e(\mathbf{D}_n) > \alpha + \delta + \epsilon) \leq \mathbb{P}((\mathcal{A} \cap \mathcal{B})^c) \leq 2e^{-2n\delta^2} + 2p \exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{2L_{n-1}\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + 2\, L_{n-1}\, \kappa_2\, c_{n-1}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right)\right) \leq \epsilon + \delta.$$

■

Appendix B
Proof for Full-Conformal

Lemma 3: Under Assumptions 1 and 2, we have

$$\mathbb{P}\left(\left\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\right\|_\infty \geq \epsilon\right) \leq 2p \exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right).$$

Proof: According to Lemma 1, we have $\|\hat{\beta}_n - \overline{\beta}_n\|_\infty < \epsilon$ with probability at least $1 - 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right)$. It follows from Assumption 2 that,

$$\mathbb{P}\left(\left\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\right\|_\infty \geq \epsilon\right) \leq \mathbb{P}\left(\kappa_2\left\|\hat{\beta}_n - \overline{\beta}_n\right\|_\infty \geq \epsilon\right) \leq 2p \exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right).$$

■

Let $\hat{\mathcal{C}}_\alpha(X_{n+1})$ denote the full-conformal α-level interval for test data-point X_{n+1} and define $P_e(\mathbf{D}_n) := \mathbb{P}(Y_{n+1} \notin \hat{\mathcal{C}}(X_{n+1})|\mathbf{D}_n)$. Define $\hat{\beta}_{X_{n+1},y} := T((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$.

Proof: We note,

$$\begin{aligned}\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{X_{n+1},y}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{X_{n+1},y}}(X_{n+1})\right|\right\} > \alpha\right\} \\ &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_n}(X_i)\right| - \left|\mu_{\hat{\beta}_n}(X_i) - \mu_{\hat{\beta}_{X_{n+1},y}}(X_i)\right| \geq \right.\right. \\ &\qquad\qquad \left.\left.\left|y - \mu_{\hat{\beta}_n}(X_{n+1})\right| + \left|\mu_{\hat{\beta}_n}(X_{n+1}) - \mu_{\hat{\beta}_{X_{n+1},y}}(X_{n+1})\right|\right\} > \alpha\right\} \\ &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_n}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_n}(X_{n+1})\right| + c_{n+1}\right\} > \alpha\right\},\end{aligned}$$

where the first and last relations hold according to the definition of $\hat{\mathcal{C}}_\alpha(X_{n+1})$ and Assumption 1. Assuming $\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_n}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_n}(X_{n+1})\right| + c_{n+1} + 2\epsilon\right\} > \alpha\right\} \\ &\supseteq \left\{y \in \mathbb{R} : 1 - \hat{F}_n\left(\left|y - \mu_{\overline{\beta}_n}(X_{n+1})\right| + c_{n+1} + 2\epsilon\right) > \alpha\right\}.\end{aligned}$$

Assuming $\left\|\hat{F}_n - F_n\right\|_\infty < \delta$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha(X_{n+1}) &\supseteq \left\{y \in \mathbb{R} : 1 - F_n\left(\left|y - \mu_{\overline{\beta}_n}(X_{n+1})\right| + c_{n+1} + 2\epsilon\right) > \alpha + \delta\right\} \\ &\supseteq \left\{y \in \mathbb{R} : 1 - F_n\left(\left|y - \mu_{\overline{\beta}_n}(X_{n+1})\right|\right) > \alpha + \delta + (2\epsilon + c_{n+1})L_n\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}P_e(\mathbf{D}_n) &= \mathbb{P}(Y_{n+1} \notin \hat{\mathcal{C}}(X_{n+1})|\mathbf{D}_n) \\ &\leq \mathbb{P}\left(1 - F_n\left(\left|Y_{n+1} - \mu_{\overline{\beta}_n}(X_{n+1})\right|\right) \leq \alpha + \delta + (2\epsilon + c_{n+1})L_n\right) \\ &= \alpha + \delta + (2\epsilon + c_{n+1})L_n\end{aligned}$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}$ where $\mathcal{A} := \left\{D : \|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon\right\}$ and $\mathcal{B} := \left\{D : \left\|\hat{F}_n - F\right\|_\infty < \delta\right\}$. From Lemma 2, we know $\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right)$. Also, according to Dvoretzky–Kiefer–Wolfowitz inequality, we have $\mathbb{P}(\mathbf{D}_n \notin \mathcal{B}) \leq 2e^{-2n\delta^2}$. Thus,

$$\mathbb{P}(P_e(\mathbf{D}_n) > \alpha + \delta + \epsilon) \leq \mathbb{P}((\mathcal{A} \cap \mathcal{B})^c) \leq 2e^{-2n\delta^2} + 2p\exp\left(-\left(\frac{\kappa_1(\epsilon/L_n - c_{n+1})}{\sqrt{2n}\kappa_2 c_n}\right)^2\right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + L_n\left(c_{n+1} + \sqrt{2n\log\frac{2p}{\epsilon}}\,\frac{\kappa_2\, c_n}{\kappa_1}\right)\right) \leq \epsilon + \delta.$$

■