2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Single Trajectory Conformal Prediction

Brian Lee and Nikolai Matni

Abstract— **We study the performance of risk-controlling prediction sets (RCPS), an empirical risk minimization-based formulation of conformal prediction, on a single trajectory of data from an unknown stochastic process. Our analysis characterizes the graceful degradation in RCPS performance as data becomes nearly arbitrarily dependent and nonstationary, subject only to a mild requirement that the underlying process is causal. By specializing this analysis, we find that RCPS attains guarantees comparable to those enjoyed on independent and identically distributed data whenever data is generated by an asymptotically stationary and mixing process. We then relate these conditions to system-theoretic properties like contractivity.**

I. Introduction

Characterizing uncertainty in the predictions of machine learning models is a key step in safely bridging learning and control. However, traditional uncertainty quantification methods, like confidence sets around maximum likelihood estimates, rely on strong assumptions about the data generating process and the performance of the training algorithm that are incompatible with the often black-box nature of machine learning practice.

A simple alternative approach is treating a given machine learning model as arbitrary but fixed, directly estimating metrics of its predictive uncertainty over labeled holdout data, and relating the estimate to the true model uncertainty at test time under weak assumptions about the data generating process. An example of this approach is conformal prediction (CP), a family of assumption-light methods that post-process an arbitrary point predictor, like a black-box neural network, into a *set predictor* that provably covers the true label at test time with high probability [1]–[3].[1] However, most existing proofs of CP coverage rely on the holdout and test data being independent and identically distributed (iid), or at least exchangeable, limiting their use on non-iid, non-exchangeable dependent data, as is typical in control applications.

Existing works that apply CP to control settings [4]–[6] overcome this issue by sampling N iid-initialized trajectories, each of length T, from a process of interest, $\left\{\left(Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_T^{(1)}\right), \ldots, \left(Z_1^{(N)}, Z_2^{(N)}, \ldots, Z_T^{(N)}\right)\right\}$, and then leveraging the fact that under mild regularity conditions, for any fixed timestep $t \in [T]$, the set of observations aggregated across trajectories, $\left\{Z_t^{(1)}, \ldots, Z_t^{(N)}\right\}$, is iid. While intuitive, this multiple trajectories approach falls short of characterizing what makes CP with dependent data easy or hard, and does not cover settings, e.g., without simulator access, where only a single trajectory of data, or multiple trajectories with varying lengths and non-iid initializations, can be collected.

In this work, we generalize the risk-controlling prediction sets (RCPS) algorithm, which was first designed for iid data, and theoretically characterize its performance on a single trajectory of data from an unknown stochastic process. By approximating sequences of dependent data with more structured ones, we prove algorithm-dependent test-time risk upper bounds of the form $\varepsilon + \gamma(w)$, where ε is the user-specified risk tolerance and $\gamma(w)$ is the excess risk incurred due to dependence, which can be minimized with the choice of weight parameters w. Specifically,

- In Section III, we use the decoupling technique [7] toward a general analysis of the excess risk $\gamma(w)$ that characterizes the graceful degradation in RCPS performance as data becomes nearly arbitrarily dependent and nonstationary. The only required assumption is that the underlying process is adapted, or causal. See Theorem 1.
- In Section IV, we use the blocking technique [8] to show that even with the simplest weights w, excess risk can be made vanishingly small when data is generated by asymptotically stationary and mixing processes, which roughly correspond to contractive dynamical systems. See Theorem 2 and Corollary 1.

A. Related Works

1) Conformal prediction with dependent data: Recent works [9]–[11] provide a general analysis of the performance of split CP and conformal risk control over non-exchangeable data. While these results are similar in spirit to our general analysis of RCPS, the underlying techniques are different, and split CP and conformal risk control guarantees are weaker than those of RCPS. Prior work [12] also uses the blocking technique, but to analyze split CP over strictly stationary and mixing data. Like us, concurrent work [13] relaxes the strict stationarity requirement, but to analyze split CP over geometrically ergodic Markovian data. In contrast, we also study RCPS in more general dependent data settings.

2) Statistical learning with dependent data: Our techniques are adapted from statistical learning works [14], [15] that use the decoupling and blocking techniques to prove generalization bounds over adapted and mixing data. While these works seek uniform convergence guarantees over

This work was supported in part by NSF Awards SLES-2331880, ECCS-2045834, ECCS-2231349, and AFOSR Award FA9550-24-1-0102. B.L. is with the University of California, Berkeley, Berkeley, CA 94720, USA. N.M. is with the University of Pennsylvania, Philadelphia, PA 19104, USA. bl.ee@berkeley.edu, nmatni@seas.upenn.edu.

[1]We abuse terminology and refer to any post-hoc uncertainty quantification method that builds set predictors with finite-sample coverage guarantees as *conformal prediction*. In the literature, these methods are varyingly called split conformal prediction, conformal risk control, and risk-controlling prediction sets. These methods moderately differ from one another in the algorithm that is implemented and the guarantee attained.

dependent data, RCPS requires only pointwise convergence, leading to simpler proofs that avoid potential looseness.

B. Notation

A filtered probability space $(\Omega, \mathscr{F}, \mathbf{P}; \mathscr{F}_{1:\infty})$ has a sequence of σ-algebras, or filtration, $\mathscr{F}_{1:\infty} \triangleq \{\mathscr{F}_t\}_{t=1}^{\infty}$ that satisfies $\mathscr{F}_t \subseteq \mathscr{F}_{t+1}$ and $\mathscr{F}_t \subset \mathscr{F}$ for all $t \in \mathbb{N}$. Given a stochastic process $Z_{1:\infty}$, $\mathbf{P}_i, \mathbf{P}_j$ are the marginal distributions of Z_i, Z_j. The total variation distance between two probability measures $\mathbf{P}, \mathbf{Q}$ is defined as $\|\mathbf{P}-\mathbf{Q}\|_{\mathrm{TV}} \triangleq \sup_A |\mathbf{P}(A)-\mathbf{Q}(A)| \in [0,1]$, where the supremum is taken over sets A in a common σ-algebra. The spectral radius of matrix G is denoted by $\rho(G)$, the resolvent by $R_G(z) = (zI-G)^{-1}$, and the H_∞-norm by $\|G\|_{H_\infty} = \sup_{z\in\mathbb{T}} \|G(z)\|$, where $\mathbb{T}$ is the complex unit circle. $O(\cdot), \Omega(\cdot)$ denote orderwise upper and lower bounds, and $\tilde{O}(\cdot)$ elides polylogarithmic factors.

II. Problem Formulation

Given $\mathscr{X} \subseteq \mathbb{R}^{d_X}, \mathscr{Y} \subseteq \mathbb{R}^{d_Y}$, consider the model

$$Y_t = f_\star(X_t) + W_t \tag{1}$$

with unknown $f_\star : \mathscr{X} \to \mathscr{Y}$ and noise W_t. When $Y_t = X_{t+1}$, we recover an autonomous dynamical system. Otherwise, we assume there exists a marginal process for X_t that is consistent with the conditions that Section II-A imposes on the joint process for (X_t, Y_t).

Suppose we have access to a learned model $\hat{f}$ of $f_\star$, which comes with no *a priori* guarantees, and a single trajectory $Z_{1:\infty} \triangleq \{(X_t, Y_t)\}_{t=1}^{\infty}$ drawn from (1) that, for some $T, k \in \mathbb{N}$, we split into training trajectory $Z_{1:T} \triangleq \{(X_t, Y_t)\}_{t=1}^{T}$ and test point $Z_{T+k} \triangleq (X_{T+k}, Y_{T+k})$. When appropriate, we draw test point $Z' = (X', Y')$ from the stationary distribution Π of $Z_{1:\infty}$.

The RCPS problem setting is as follows. We use black-box model $\hat{f}$ and training trajectory $Z_{1:T}$ to build a set predictor $C_\lambda : \mathscr{X} \times \mathbb{R} \to 2^{\mathscr{Y}}$, which is centered at $\hat{f}(\cdot)$ and parameterized by $\lambda \in \Lambda \subseteq \mathbb{R}$ that controls the radius of the prediction sets, i.e., $\lambda < \lambda' \implies C_\lambda \subset C_{\lambda'}$. See prior work [16] for connections between this "nested sets" approach to CP and the more standard non-conformity score approach. Specifically, given a loss $\ell : \mathscr{Y} \times 2^{\mathscr{Y}} \to [0,1]$ that decreases monotonically as λ increases (usually chosen as the indicator loss $\mathbb{1}(Y \notin C_\lambda(X))$, but other choices are admissible, e.g., false negative rate), risk tolerance $\varepsilon \in (0,1)$, and failure probability $\delta \in (0,1)$, we aim to ensure that

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{Z_{T+k}}[\ell(Y_{T+k}, C_\lambda(X_{T+k}))|Z_{1:T}] \le \varepsilon\right) \ge 1-\delta, \tag{2}$$

i.e., with probability at least $(1-\delta)$ over the draw of the training trajectory, the expectation of $\ell(\cdot,\cdot)$ taken over the draw of Z_{T+k} *after conditioning on* $Z_{1:T}$ is at most ε. We also consider a weaker guarantee,

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{Z_{T+k}}[\ell(Y_{T+k}, C_\lambda(X_{T+k}))] \le \varepsilon\right) \ge 1-\delta, \tag{3}$$

i.e., the expectation is taken over the *marginal draw* of Z_{T+k}. The two expectations are identical if $Z_{1:T} \cup Z_{T+k}$ are iid, but is not necessarily so otherwise. This distinction highlights how the "conditional validity" of CP, first discussed in the iid data setting [17], becomes more nuanced when data is dependent. When appropriate, we take the expectation of $\ell(\cdot,\cdot)$ over the draw of Z' from stationary distribution Π of $Z_{1:\infty}$. For convenience, we often denote $\ell_\lambda(Z) \triangleq \ell(Y, C_\lambda(X))$.

The RCPS algorithm works as follows. Given weights $w = (w_1, w_2, \ldots, w_T)^\top$ from the probability simplex $\Delta(\mathbb{R}^T)$ and an upper bound $U : (0,1) \times \Delta(\mathbb{R}^T) \to \mathbb{R}_+$, we select radius

$$\hat{\lambda} \triangleq \inf\left\{\lambda \in \Lambda : \sum_{t=1}^{T} w_t \ell_\lambda(Z_t) + U(\delta, w) < \varepsilon,\ \forall \lambda' \ge \lambda\right\}. \tag{4}$$

When $Z_{1:T} \cup Z_{T+k}$ are iid, we choose $w_t = 1/T$ for all $t \in [T]$ and $U(\delta, w) = \sqrt{\log(1/\delta)/T}$, and use Hoeffding's inequality to show that (2), (3) hold. This idea of leveraging a pointwise concentration inequality carries over to the dependent data setting, although we sometimes set $U(\delta, w) = 0$ because analogs of the Hoeffding bandwidth can depend on unknown parameters that quantify the degree of dependence in data, and instead suffer that term as excess risk.

A. Data Generating Processes

Now we define data generating processes for $Z_{1:\infty}$ that we consider in each section.

Definition 1 (Adapted process [7])**.** *A stochastic process $Z_{1:\infty}$ is adapted to filtration $\mathscr{F}_{1:\infty}$ if Z_t is $\mathscr{F}_t$-measurable $\forall t \in \mathbb{N}$.*

In Section III, we require that $Z_{1:\infty}$ is adapted to some filtration $\mathscr{F}_{1:\infty}$. This simply prevents Z_t from depending on the future, i.e., $\mathscr{F}_{t+1:\infty}$, and is a very mild assumption that all causal systems satisfy.

Definition 2 ($\phi^\star$-mixing processes [8], [15])**.** *A stochastic process $Z_{1:\infty}$ that is adapted to filtration $\mathscr{F}_{1:\infty}$ and has stationary distribution Π has ϕ^*-mixing coefficient*

$$\phi^*(k) \triangleq \phi_Z^\star(k) \triangleq \sup_{t\in\mathbb{N}} \sup_{A\in\mathscr{F}_t} \left[\left\|\mathbf{P}_{t+k}(\cdot|A) - \Pi\right\|_{TV}\right].$$

$Z_{1:\infty}$ is said to be $\phi^\star$-mixing if $\phi^\star(k) \to 0$ as $k \to \infty$.

In Section III, we also consider ϕ^*-mixing $Z_{1:\infty}$, i.e., the data generating process converges to a stationary distribution and samples that are sufficiently separated in time are approximately independent, even after conditioning on the realization of a "bad" trajectory $Z_{1:t}$. This imposes a very weak dependence structure on $Z_{1:\infty}$ and is a much more stringent assumption than adaptivity.

Definition 3 ($\beta^\star$-mixing processes [8], [15])**.** *A stochastic process $Z_{1:\infty}$ that is adapted to filtration $\mathscr{F}_{1:\infty}$ and has stationary distribution Π has β^*-mixing coefficient*

$$\beta^\star(k) \triangleq \beta_Z^\star(k) \triangleq \sup_{t\in\mathbb{N}} \mathbb{E}_{Z_{1:t}}\left[\left\|\mathbf{P}_{t+k}(\cdot|Z_{1:t}) - \Pi\right\|_{TV}\right].$$

$Z_{1:\infty}$ is said to be $\beta^\star$-mixing if $\beta^\star(k) \to 0$ as $k \to \infty$.

In Section IV, we require that $Z_{1:\infty}$ is $\beta^\star$-mixing, which lies between being adapted and $\phi^\star$-mixing. This is because β^*-mixing only requires that convergence to the stationary distribution and approximate independence with time separation occur over "average" trajectories $Z_{1:t}$. In particular, $\beta^\star(k) \le \phi^\star(k)$, so all ϕ^*-mixing processes are β^*-mixing.

A simple $\beta^\star$-mixing dynamical system is the stable linear time-invariant (LTI) system with iid Gaussian noise, which we use to build system-theoretic intuition about results.

III. Decoupling Technique

In this section, we assume that $Z_{1:\infty}$ is adapted to some filtration $\mathscr{F}_{1:\infty}$, per Definition 1, and use the decoupling technique to study RCPS guarantees of the form (2). This yields a general analysis of RCPS that remains valid even as data becomes nearly arbitrarily dependent and nonstationary.

A sequence being adapted is a sufficient condition for there to exist a *decoupled tangent sequence* of conditionally independent random variables with the following properties.

Proposition 1 (Decoupling technique [7])**.** *Let $Z_{1:\infty}$ be a stochastic process adapted to some filtration $\mathscr{F}_{1:\infty}$ that is contained in the σ-algebra $\mathscr{F}$. Then there exists a decoupled tangent sequence $Z'_{1:\infty} = \{Z'_t\}_{t=1}^{\infty}$ that satisfies*

1) *Z_t, Z'_t are iid with respect to $\mathbf{P}(\cdot|\mathscr{F}_{t-1})\ \forall t \in \mathbb{N}$, and*
2) *$Z'_{1:\infty}$ is conditionally independent given some σ-algebra $\mathscr{G} \subset \mathscr{F}$ for which $\mathbf{P}(Z'_t|\mathscr{F}_{t-1}) = \mathbf{P}(Z'_t|\mathscr{G})\ \forall t \in \mathbb{N}$. $\mathscr{G}$ is often $\sigma(Z_{1:\infty})$, the σ-algebra induced by $Z_{1:\infty}$.*

Then for any function ℓ that maps Z_t to $\mathbb{R}$, by the linearity of expectation and the law of iterated expectations,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(Z_t)\right] = \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\ell(Z_t)|\mathscr{F}_{t-1}\right]\right] = \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\ell(Z'_t)|\mathscr{F}_{t-1}\right]\right]$$
$$= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\ell(Z'_t)|\mathscr{G}\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T}\ell(Z'_t)\right].$$

Instead of directly evaluating RCPS performance on an arbitrarily dependent and nonstationary sequence, we use Proposition 1 to do so on its tangent sequence, whose conditional independence structure can be exploited in a sequential symmetrization scheme.

Theorem 1 (Decoupled RCPS)**.** *Fix risk tolerance $\varepsilon \in (0,1)$, failure probability $\delta \in (0,1)$, weights $w \in \Delta(\mathbb{R}^T)$, and upper bound $U(\delta, w) = \|w\|_2\sqrt{8\log(1/\delta)}$. Suppose $Z_{1:\infty}$ is adapted to filtration $\mathscr{F}_{1:\infty}$. Then setting $\hat{\lambda}$ according to* (4) *attains*

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{Z_{T+k}}\left[\ell_{\hat{\lambda}}(Z_{T+k})|Z_{1:T}\right] \le \varepsilon + \gamma(w)\right) \ge 1-\delta,$$

where the expectation is taken over the draw of the test point after conditioning on realized training trajectory $Z_{1:T}$, and

$$\gamma(w) \triangleq \mathbb{E}\left[\ell_{\hat{\lambda}}\left(Z_{T+k}\right)\right)|Z_{1:T}\Big] - \sum_{t=1}^{T} w_t \cdot \mathbb{E}\left[\ell_{\hat{\lambda}}\left(Z_t\right)|Z_{1:t-1}\right].$$

Theorem 1 states that when data is causal, the test-time risk upper bound is the sum of risk tolerance ε and an excess risk term $\gamma(w)$ that captures the drift between the test distribution $\mathbf{P}_{Z_{T+k}}(\cdot|Z_{1:T})$ and the mixture of past distributions $\sum_{t=1}^{T} w_t \mathbf{P}_{Z_t}(\cdot|Z_{1:t-1})$. The weights w can be selected to minimize the latter term; prior works [9], [14] discuss strategies for selecting these weights. In Section IV, we show that even the simplest weights $w_t = 1/T$ for all $t \in [T]$ attains desirable guarantees when data is appropriately mixing.

Proof of Theorem 1. We adapt the proof of Theorem 1 in [14]. Suppose $\mathop{\mathbb{E}}_{Z_{T+k}}\left[\ell_{\hat{\lambda}}(Z_{T+k})|Z_{1:T}\right] > \varepsilon + \gamma(w)$, i.e.,

$$\sum_{t=1}^{T} w_t \mathbb{E}\left[\ell_{\hat{\lambda}}(Z_t)|Z_{1:t-1}\right] > \varepsilon.$$

Then $\hat{\lambda} < \lambda^* \triangleq \inf\left\{\lambda \in \Lambda : \sum_{t=1}^{T} w_t \mathbb{E}\left[\ell_\lambda(Z_t)|Z_{1:t-1}\right] \le \varepsilon\right\}$. However, by (4), $\sum_{t=1}^{T} w_t \ell_{\lambda^*}(Z_t) + U(\delta, w) < \varepsilon$ must have held. We show this occurs with probability at most δ, i.e.,

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\sum_{t=1}^{T} w_t\left(\mathop{\mathbb{E}}_{Z_t}\left[\ell_{\lambda^*}(Z_t)|Z_{1:t-1}\right] - \ell_{\lambda^*}(Z_t)\right) > U(\delta, w)\right) \le \delta.$$

Denote $U \triangleq U(\delta, w)$ for convenience. Then for any $a > 0$,

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\sum_{t=1}^{T} w_t\left(\mathop{\mathbb{E}}_{Z_t}\left[\ell_{\lambda^*}(Z_t)|Z_{1:t-1}\right] - \ell_{\lambda^*}(Z_t)\right) > U\right)\exp(aU)$$
$$\le \mathop{\mathbb{E}}_{Z_{1:T}}\left[\exp\left(a\sum_{t=1}^{T} w_t\left(\mathop{\mathbb{E}}_{Z_t}\left[\ell_{\lambda^*}(Z_t)|Z_{1:t-1}\right] - \ell_{\lambda^*}(Z_t)\right)\right)\right]$$
$$= \mathop{\mathbb{E}}_{Z_{1:T}}\left[\exp\left(a\sum_{t=1}^{T} w_t\left(\mathop{\mathbb{E}}_{Z'_t}\left[\ell_{\lambda^*}\left(Z'_t\right)|Z_{1:T}\right] - \ell_{\lambda^*}(Z_t)\right)\right)\right]$$
$$\le \mathop{\mathbb{E}}_{Z_{1:T}}\left[\mathop{\mathbb{E}}_{Z'_{1:T}}\left[\exp\left(a\sum_{t=1}^{T} w_t\left(\ell_{\lambda^*}\left(Z'_t\right) - \ell_{\lambda^*}(Z_t)\right)\right)\Big|Z_{1:T}\right]\right]$$
$$= \mathop{\mathbb{E}}_{Z_{1:T}, Z'_{1:T}}\left[\exp\left(a\sum_{t=1}^{T} w_t\left(\ell_{\lambda^*}\left(Z'_t\right) - \ell_{\lambda^*}(Z_t)\right)\right)\right],$$

where the first line holds by the Chernoff bound, the second line holds because $\mathbb{E}\left[\ell_{\lambda^*}(Z_t)|Z_{1:t-1}\right] = \mathbb{E}\left[\ell_{\lambda^*}(Z'_t)|Z_{1:t-1}\right] = \mathbb{E}\left[\ell_{\lambda^*}(Z'_t)|Z_{1:T}\right]$ by Proposition 1, the third line holds by Jensen's inequality, and the final line holds by the law of iterated expectations.

Now we equate the expectation over $Z_{1:T}, Z'_{1:T}$ to an expectation over auxiliary random variables through a sequential symmetrization argument, following [14], [18]. This involves introducing $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_T)$, where the $\{\sigma_t\}$ are pairwise independent, distributed uniformly over $\{-1, 1\}$, and each σ_t drawn independently of Z_t, Z'_t, so that the expression inside the expectation becomes $\exp\left(a\sum_{t=1}^{T}\sigma_t w_t\left(\ell_{\lambda^*}(Z'_t) - \ell_{\lambda^*}(Z_t)\right)\right)$. Consider $t = 1$. If $\sigma_1 = 1$, $w_1\left(\ell_{\lambda^*}(Z'_1) - \ell_{\lambda^*}(Z_1)\right)$ is unchanged. If $\sigma_1 = -1$, the order of subtraction is flipped, i.e., $w_1\left(\ell_{\lambda^*}(Z_1) - \ell_{\lambda^*}(Z'_1)\right)$. Then Z_1 becomes part of the tangent sequence and Z'_1, the original sequence. Iteratively applying this argument to all indices $t \in [T]$ motivates using binary tree structures to keep track of "who is tangent to who." Define binary tree $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ as a T-tuple of mappings $\mathbf{z}_t : \{-1, 1\}^{t-1} \to \mathscr{X} \times \mathscr{Y}$, i.e., from $\sigma_1, \ldots, \sigma_{t-1}$ to a value of $Z_t = (X_t, Y_t)$. Define the tangent binary tree $\mathbf{z}' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_T)$ as the T-tuple of maps $\mathbf{z}'_t$ from $\sigma_1, \ldots, \sigma_{t-1}$ to $Z'_t = (X'_t, Y'_t)$. Lemma 2 of [14] and Theorem 3 of [18] prove that the expectation over the draw of $Z_{1:T}, Z'_{1:T}$ is equivalent to the expectation over the independent draw of the trees $\mathbf{z}, \mathbf{z}'$ and the "path" σ. Then

$$\mathop{\mathbb{E}}_{Z_{1:T}, Z'_{1:T}}\left[\exp\left(a\sum_{t=1}^{T} w_t\left(\ell_{\lambda^*}\left(Z'_t\right) - \ell_{\lambda^*}(Z_t)\right)\right)\right]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\mathbf{z}',\mathbf{z}}\left[\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(a\sum_{t=1}^{T}\sigma_t w_t\left(\ell_{\lambda^*}(\mathbf{z}'_t)-\ell_{\lambda^*}(\mathbf{z}_t)\right)\right)\right]\right]\\
&= \mathop{\mathbb{E}}_{\mathbf{z}',\mathbf{z}}\left[\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(a\sum_{t=1}^{T}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}'_t)+\sum_{t=1}^{T}-\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\right]\right]\\
&\le \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{z}'}\left[\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2a\sum_{t=1}^{T}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}'_t)\right)\right]\right]\\
&\qquad + \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{z}}\left[\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2a\sum_{t=1}^{T}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\right]\right]\\
&= \mathop{\mathbb{E}}_{\mathbf{z}}\left[\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2a\sum_{t=1}^{T}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\right]\right],
\end{aligned}
$$

where the fourth line holds by Young's inequality and the final line holds by symmetry. Now observe that

$$
\mathop{\mathbb{E}}_{\sigma_T,\mathbf{z}_T}\left[\exp\left((2aw_T)\sigma_T\ell_{\lambda^*}(\mathbf{z}_T)\right)\right] \le \exp\left(2a^2w_T^2\right)
$$

by Hoeffding's lemma and the independence of $\sigma_T, \mathbf{z}_T$. So,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{z},\sigma}\left[\exp\left(2a\sum_{t=1}^{T}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\right]\\
&\quad= \mathop{\mathbb{E}}_{\mathbf{z},\sigma}\left[\exp\left(2a\sum_{t=1}^{T-1}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\right.\\
&\qquad\qquad\left.\cdot \mathop{\mathbb{E}}_{\mathbf{z}_T,\sigma_T}\left[\exp\left(2a\sigma_T w_T \ell_{\lambda^*}(\mathbf{z}_T)\right)|\mathbf{z}_{1:T-1},\sigma_{1:T-1}\right]\right]\\
&\quad\le \mathop{\mathbb{E}}_{\mathbf{z},\sigma}\left[\exp\left(2a\sum_{t=1}^{T-1}\sigma_t w_t \ell_{\lambda^*}(\mathbf{z}_t)\right)\exp\left(2a^2w_t^2\right)\right]\\
&\quad\le \exp\left(2a^2\|w\|_2^2\right),
\end{aligned}
$$

where the second line holds by the law of iterated expectations and the final line holds by iteratively applying the previous one. Collecting the $\exp(-aU)$ term from the beginning of the argument and optimizing over a completes the proof. □

The key difference between the general RCPS guarantee in Theorem 1 and the guarantee in the iid data setting is the excess risk term $\gamma(w)$. When $Z_{1:T} \cup Z_{T+k}$ are iid, $\gamma(w) = 0$ for any w. Now we show a bound on $\gamma(w)$ for when data is only very weakly dependent, i.e., ϕ^*-mixing per Definition 2.

Example 1 ($\phi^\star$-mixing process)**.** *If $Z_{1:\infty}$ is $\phi^\star$-mixing with stationary distribution Π, then implementing RCPS with $w_t = 1/T$ for all $t \in [T]$, $U(\delta, w) = \sqrt{8\log(1/\delta)/T}$, and test point $Z' \sim \Pi$ ensures that*

$$
\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\hat{\lambda}}(Z')\right] \le \varepsilon + \gamma(w)\right) \ge 1-\delta,
$$

where

$$
\begin{aligned}
\gamma(w) &\le \left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{P}_{Z_t}(\cdot|Z_{1:t-1}) - \Pi\right\|_{TV}\\
&\le \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{P}_{Z_t}(\cdot|Z_{1:t-1}) - \Pi\|_{TV} \le \phi^*(1)
\end{aligned}
$$

The first inequality captures the intuition that $\gamma(w)$ is small when w is chosen so that the mixture of past distributions closely approximates the test distribution Π. Naturally, we hope to relate $\gamma(w)$ to the ϕ^*-mixing coefficient, which measures distance to Π. However, this is complicated by the fact that each past distribution at time t is conditioned on $Z_{1:t-1}$, which interacts poorly with the definition of the ϕ^*-mixing coefficient. In the second inequality, we resort to the triangle inequality and recover the constant upper bound $\phi^*(1)$, which cuts against intuition that uniform weights over past distributions should be sufficient to capture the stationary distribution of an ergodic process. Unfortunately, conditioning on $Z_{1:t-1}$ at each time t is a key part of the decoupling technique and cannot be easily removed. This suggests that because the decoupling technique is designed for incredibly general settings, it can refrain from exploiting additional structures available in weakly dependent data.

This is not a fundamental shortcoming of the RCPS algorithm: next, we fix uniform weights and modify our analysis to prove RCPS performs well on mixing data.

IV. Blocking Technique

In this section, we fix $w_t = 1/T$ for all $t \in [T]$, assume $Z_{1:\infty}$ is $\beta^\star$-mixing per Definition 3, and use the blocking technique to study RCPS guarantees of the form (3). We also study these guarantees in terms of system-theoretic quantities when the underlying process is a stable LTI system with iid Gaussian noise—a β^*-mixing process that does not satisfy the strict stationarity assumption required in prior work [12].

The blocking technique can be viewed as a specialization of the decoupling technique that goes beyond exploiting conditional independence structures to *outright iid-like structures* in mixing data. Specifically, this approach transforms the analysis of a subsample of β^*-mixing random variables into one of a subsample of iid random variables from stationary distribution Π, at the expense of an additive error term that is minimized with our choice of subsample size.

Proposition 2 (Blocking technique [8], [15], [19])**.** *Suppose $Z_{1:\infty}$ is a $\beta^\star$-mixing sequence with stationary distribution Π. Fix $m, n \in \mathbb{N}$ and suppose without loss of generality that $T = mn$. Construct n subsampled blocks of size m each, where*

$$
Z_{(j)} = \{Z_t : (t-1 \mod n) = j-1\} \quad \text{for} \quad j = 1, 2, \ldots, n.
$$

Let $\tilde{Z}_\Pi$ be a block of m iid draws from Π. Then for any measurable function ℓ that takes values in $[0,1]$,

$$
\left|\mathbb{E}\left[\ell\left(\tilde{Z}_\Pi\right)\right] - \mathbb{E}\left[\ell\left(Z_{(j)}\right)\right]\right| \le m\beta^\star(n).
$$

Using the blocking technique, we first study RCPS performance on the stationary distribution Π, then on the marginal distribution $\mathbf{P}_{T+k}$ of the test point $Z_{T+k} = (X_{T+k}, Y_{T+k})$.

Theorem 2 (Blocked RCPS, Π.)**.** *Suppose $Z_{1:\infty}$ is β^*-mixing with $\beta^*(k) = O(1/k)$. Fix block size m and number of blocks n, and assume without loss of generality that $T = mn$. Also fix risk tolerance $\varepsilon \in (0,1)$, failure probability $(T\beta^*(n), 1)$, and trivial upper bound $U(\delta, w) = 0$. Then selecting $\hat{\lambda}$ according to* (4) *attains*

$$
\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\hat{\lambda}}(Z')\right] \le \varepsilon + \eta\right) \ge 1-\delta,
$$

where the expectation is over the draw of $Z' = (X', Y')$ from stationary distribution Π and $\eta = \sqrt{\log\left(\frac{n}{\delta - T\beta^\star(n)}\right)/m}$.

Theorem 2 states that for a sufficiently long training trajectory, the upper bound on true risk $\mathop{\mathbb{E}}_{\Pi}\left[\ell(Y', C_{\hat{\lambda}}(X'))\right]$ can be made arbitrarily close to ε at the blocking-deflated rate $\tilde{O}(1/\sqrt{m})$. The assumption that $\beta^*(k) = O(1/k)$ is mild, as many processes of interest mix faster. With this assumption, the requirement that $\delta > T\beta^\star(n)$ is satisfied for any $\delta \in (0,1)$ as long as T exceeds a burn-in time that is at most polynomial in $1/\delta$. Importantly, because this result does not rely on any particular values of m, n, they can be selected to balance the required burn-in time for T and the η term (as part of the analysis, not the implementation, of RCPS). We discuss these details in Example 2.

Proof of Theorem 2. We adapt the proof of Theorem 2.2 in [19]. Suppose $\mathop{\mathbb{E}}_{Z'}\left[\ell_{\hat{\lambda}}(Z')\right] > \varepsilon + \eta$. Then

$$\hat{\lambda} < \lambda^* \triangleq \inf\left\{\lambda \in \Lambda : \mathbb{E}_{\Pi}\left[\ell_\lambda(Z')\right] \le \varepsilon\right\}.$$

However, by (4), $\sum_{t=1}^{T} w_t \ell_{\lambda^*}(Z_t) < \varepsilon$ must have held. We show that this occurs with probability at most δ,

For $j \in \{1, 2, \ldots, n\}$, let I_j be the set of indices included in the jth block. Then

$$\begin{aligned}
&\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \frac{1}{T}\sum_{t=1}^{T}\ell_{\lambda^*}(Z_t) > \eta\right) \\
&= \mathop{\mathbf{P}}_{Z_{1:T}}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{m}\sum_{t\in I_j}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \ell_{\lambda^*}(Z_t)\right) > \eta\right) \\
&\le \sum_{j=1}^{n}\mathop{\mathbf{P}}_{Z_{(j)}}\left(\frac{1}{m}\sum_{t\in I_j}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \ell_{\lambda^*}(Z_t)\right) > \eta\right) \\
&\le T\beta^\star(n) + n\mathop{\mathbf{P}}_{\tilde{Z}_\Pi}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \frac{1}{m}\sum_{t\in I_j}\left(\ell_{\lambda^*}(Z_t)\right) > \eta\right) \\
&\le T\beta^\star(n) + (\delta - T\beta^\star(n)) = \delta.
\end{aligned}$$

The second line holds by the union bound: if

$$\frac{1}{n}\sum_{j=1}^{n}\frac{1}{m}\sum_{t\in I_j}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \ell_{\lambda^*}(Z_t)\right) > \eta$$

holds, then there must exist an index j for which

$$\frac{1}{m}\sum_{t\in I_j}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \ell_{\lambda^*}(Z_t)\right) > \eta$$

holds. The third line follows from Proposition 2, and the final line follows from setting η as in the theorem statement and applying Hoeffding's inequality. □

Theorem 2 shows that when data is β^*-mixing, RCPS with respect to the stationary distribution is nearly as easy as RCPS in the iid setting: there is no excess risk term analogous to $\gamma(w)$ in Theorem 1. Instead, excess risk is due to the η term, which is simply the Hoeffding bandwidth that scales with block size, rather than training trajectory length. We set $U(\delta, w) = 0$ because the desired block size is not known *a priori*, but if it is, η can be offset.

Corollary 1 (Blocked RCPS, $\mathbf{P}_{T+k}$)**.** *Under the assumptions of Theorem 2, we have that*

$$\mathop{\mathbf{P}}_{Z_{1:T}}\left(\mathop{\mathbb{E}}_{\mathbf{P}_{T+k}}\left[\ell_{\hat{\lambda}}(Z_{T+k})\right] \le \varepsilon + \eta + \gamma\right) \ge 1 - \delta,$$

where the expectation is taken over the marginal draw of test point Z_{T+k}, $\eta = \tilde{O}(1/\sqrt{m})$ as in Theorem 2, and $\gamma \le \beta^\star(k)$ is an excess risk term.

Proof of Corollary 1. Observe that

$$\begin{aligned}
\gamma &\triangleq \mathop{\mathbb{E}}_{\mathbf{P}_{T+k}}\left[\ell(Y_{T+k}, C_{\hat{\lambda}}(X_{T+k}))\right] - \mathop{\mathbb{E}}_{\Pi}\left[\ell(Y', C_{\hat{\lambda}}(X'))\right] \\
&\le \|\mathbf{P}_{T+k} - \Pi\|_{\mathrm{TV}} \le \mathop{\mathbb{E}}_{Z_{1:T}}\left[\|\mathbf{P}_{T+k}(\cdot|Z_{1:T}) - \Pi\|_{\mathrm{TV}}\right] \le \beta^\star(k),
\end{aligned}$$

where the second line holds by the law of iterated expectations and Jensen's inequality. Introducing γ to the result of Theorem 2 completes the proof. □

An excess risk term analogous to $\gamma(w)$ in Theorem 1 is found in Corollary 1, but this term is at most $\beta^*(k)$, which captures the intuition that RCPS with respect to a marginal distribution that is sufficiently separated from the training trajectory should be as easy as RCPS with respect to the stationary distribution. This is a significant improvement over the constant $\phi^*(1)$ bound shown in Example 1.

Next, we study these results in the specific case where data is generated by a stable LTI system with iid Gaussian process noise.

Example 2 (Stable LTI [19])**.** *Consider the stable and autonomous LTI system*

$$X_{t+1} = AX_t + W_t$$

with spectral radius $\rho = \rho(A) < 1$ and $W_t \sim N(0, I)$. The associated stochastic process $X_{1:\infty} = \{X_t\}_{t=1}^{\infty}$ has marginal distribution $X_t \sim N(0, \Sigma_t)$, where $\Sigma_t = \sum_{j=1}^{t-1}(A^j)(A^j)^\top$, and stationary distribution $N(0, \Sigma_\infty)$, where Σ_∞ is the unique solution to the discrete-time Lyapunov equation

$$A\Sigma_\infty A^\top - \Sigma_\infty + I = 0.$$

For any initial $\Sigma_0 \neq \Sigma_\infty$, this system is not strictly stationary, but is asymptotically so. Suppose $X_0 = 0$. Then

$$\begin{aligned}
\beta^\star_X(k) &= \sup_{t\in\mathbb{N}}\mathop{\mathbb{E}}_{X_{1:t}}\left[\left\|\mathbf{P}_{t+k}(\cdot|X_{1:t}) - \Pi\right\|_{TV}\right] \\
&\le \left(\frac{\|R_{\rho^{-1}A}\|_{H_\infty}}{2}\sqrt{\mathbf{Tr}(\Sigma_\infty) + \frac{d_X}{1-\rho^2}}\right)\rho^k \triangleq \Gamma\rho^k,
\end{aligned}$$

which tends to 0 as $k \to \infty$. By an application of Pinsker's inequality and the chain rule for the Kullback-Leibler divergence, the process $Z_{1:\infty} = \{(X_t, X_{t+1})\}_{t=1}^{\infty}$, has mixing coefficient $\beta^*_Z(k) \le 2\beta^*_X(k)$. Hence $Z_{1:\infty}$ is β^*-mixing.

When implementing RCPS on $Z_{1:\infty}$, we can select

$$n = \left\lceil \frac{1}{1-\rho}\log\left(\frac{4\Gamma T}{\delta}\right)\right\rceil \quad \text{and} \quad \gamma = \sqrt{\frac{\log\left(\frac{2n}{\delta}\right)}{m}},$$

to enforce, in the fourth line of the proof of Theorem 2,

$$\begin{aligned}
&T\beta^\star(n) + n\mathop{\mathbf{P}}_{\tilde{Z}_\Pi}\left(\mathop{\mathbb{E}}_{\Pi}\left[\ell_{\lambda^*}(Z')\right] - \frac{1}{m}\sum_{t\in I_j}\left(\ell_{\lambda^*}(Z_t)\right) > \eta\right) \\
&\qquad \le 2T\Gamma\rho^n + \frac{\delta}{2} = \delta.
\end{aligned}$$

Notably, $\beta^*_Z(k)$ decays geometrically in k, which is much faster than the $O(1/k)$ decay required in Theorem 2. This means the required burn-in for T, implicitly given by n, is logarithmic in $1/\delta$, far better than the worst-case polynomial dependence discussed previously. Using techniques from learning for control [20], similar decay results can be shown for contractive dynamical systems with non-iid Gaussian noise, which also lead to efficient burn-in requirements. This shows that intuition from Example 2 can inform the application of RCPS to a larger class of systems of interest.

V. Discussion

We theoretically characterize the performance of RCPS on a single trajectory of non-iid, non-exchangeable data using the decoupling and blocking techniques, which allow us to approximate dependent sequences with more structured ones. Notably, neither technique strictly dominates the other in usefulness. The decoupling technique handles very general data and highlights the possibility of re-weighing training samples to reduce the $\gamma(w)$ excess risk term, echoing algorithmic modifications that prior works have suggested for split CP and conformal risk control [9], [10]. However, it can be difficult to directly relate $\gamma(w)$ to well-studied metrics of weak dependence. This motivates the use of the blocking technique with mixing data, albeit at the cost of introducing burn-in requirements for training trajectory length T and suffering the blocking-deflated excess risk term $\tilde{O}(1/\sqrt{m})$. It is worth noting that this deflated rate may be an artifact of our analysis and could possibly be removed using techniques from [21]. Importantly, these remaining questions about the analysis of RCPS do not affect its implementation.

An interesting direction for future work is better understanding the interplay between the decoupling and blocking techniques so that CP algorithms' performance on data generating processes that lie between the adapted, or adversarial, and iid extremes can be more sharply analyzed. This question relates to recent works on "best-of-both-worlds" CP [22], [23], which study online CP algorithms that simultaneously attain desirable guarantees over adversarial and iid data.

Finally, a natural question is whether CP guarantees that remain valid after (approximately) conditioning on the test input X_{T+k}, possibly in addition to training trajectory $Z_{1:T}$, is possible when data is non-iid. Exact conditioning is known to be impossible even when data is iid [17], [24], but various notions of approximate conditioning have been attained by recent works whose methods seek robustness to specific types of distribution shifts [25], [26]. Test input-conditional CP guarantees, if attained over dependent data, would naturally be useful for safety-critical control applications.

VI. Acknowledgements

B.L. thanks Ingvar Ziemann, Lars Lindemann, and Rahul Ramesh for helpful discussions.

References

[1] A. Gammerman, G. Shafer, and V. Vovk, *Algorithmic Learning in a Random World*. New York: Springer-Verlag, 2005.

[2] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal Risk Control," Apr. 2023, arXiv:2208.02814.

[3] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan, "Distribution-free, Risk-controlling Prediction Sets," *Journal of the ACM*, vol. 68, no. 6, pp. 1–34, Dec. 2021.

[4] N. Hashemi, X. Qin, L. Lindemann, and J. V. Deshmukh, "Data-Driven Reachability Analysis of Stochastic Dynamical Systems with Conformal Inference," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, Dec. 2023, pp. 3102–3109.

[5] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe Planning in Dynamic Environments Using Conformal Prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5116–5123, Aug. 2023.

[6] R. Tumu, L. Lindemann, T. Nghiem, and R. Mangharam, "Physics Constrained Motion Prediction with Uncertainty Quantification," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2023, pp. 1–8.

[7] V. H. De La Peña and E. Giné, *Decoupling*, ser. Probability and its Applications, J. Gani, C. C. Heyde, and T. G. Kurtz, Eds. New York, NY: Springer, 1999.

[8] B. Yu, "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences," *The Annals of Probability*, vol. 22, no. 1, pp. 94–116, Jan. 1994.

[9] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, Apr. 2023.

[10] A. Farinhas, C. Zerva, D. Ulmer, and A. F. T. Martins, "Non-Exchangeable Conformal Risk Control," Jan. 2024, arXiv:2310.01262.

[11] D. Prinster, S. Stanton, A. Liu, and S. Saria, "Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them)," May 2024, arXiv:2405.06627.

[12] R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano, "Split Conformal Prediction for Dependent Data," Mar. 2022.

[13] F. Zheng and A. Proutiere, "Conformal Predictions under Markovian Data," Jul. 2024, arXiv:2407.15277.

[14] V. Kuznetsov and M. Mohri, "Discrepancy-Based Theory and Algorithms for Forecasting Non-Stationary Time Series," *Annals of Mathematics and Artificial Intelligence*, vol. 88, no. 4, pp. 367–399, Apr. 2020.

[15] ——, "Generalization bounds for non-stationary mixing processes," *Machine Learning*, vol. 106, no. 1, pp. 93–117, Jan. 2017.

[16] C. Gupta, A. K. Kuchibhotla, and A. Ramdas, "Nested conformal prediction and quantile out-of-bag ensemble methods," *Pattern Recognition*, vol. 127, p. 108496, Jul. 2022.

[17] V. Vovk, "Conditional Validity of Inductive Conformal Predictors," in *Proceedings of the Asian Conference on Machine Learning*. PMLR, Nov. 2012, pp. 475–490.

[18] A. Rakhlin, K. Sridharan, and A. Tewari, "Online Learning: Stochastic, Constrained, and Smoothed Adversaries," in *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011.

[19] S. Tu and B. Recht, "Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 5005–5014.

[20] I. M. Ziemann, H. Sandberg, and N. Matni, "Single Trajectory Nonparametric Learning of Nonlinear Dynamics," in *Proceedings of Thirty Fifth Conference on Learning Theory*. PMLR, Jun. 2022, pp. 3333–3364.

[21] I. Ziemann, S. Tu, G. J. Pappas, and N. Matni, "The noise level in linear regression with dependent data," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74 903–74 920, Dec. 2023.

[22] I. Gibbs and E. Candes, "Adaptive Conformal Inference Under Distribution Shift," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 1660–1672.

[23] A. N. Angelopoulos, R. Barber, and S. Bates, "Online conformal prediction with decaying step sizes," in *Proceedings of the 41st International Conference on Machine Learning*. PMLR, Jul. 2024, pp. 1616–1630.

[24] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," Apr. 2020, arXiv:1903.04684.

[25] I. Gibbs, J. J. Cherian, and E. J. Candès, "Conformal Prediction With Conditional Guarantees," Dec. 2023, arXiv:2305.12616.

[26] M. Zecchin and O. Simeone, "Localized Adaptive Risk Control," Jun. 2024, arXiv:2405.07976.