Localized Adaptive Risk Control

Matteo Zecchin   Osvaldo Simeone
Centre for Intelligent Information Processing Systems
Department of Engineering
King's College London
London, United Kingdom
{matteo.1.zecchin,osvaldo.simeone}@kcl.ac.uk

Abstract

Adaptive Risk Control (ARC) is an online calibration strategy based on set prediction that offers worst-case deterministic long-term risk control, as well as statistical marginal coverage guarantees. ARC adjusts the size of the prediction set by varying a single scalar threshold based on feedback from past decisions. In this work, we introduce Localized Adaptive Risk Control (L-ARC), an online calibration scheme that targets statistical localized risk guarantees ranging from conditional risk to marginal risk, while preserving the worst-case performance of ARC. L-ARC updates a threshold function within a reproducing kernel Hilbert space (RKHS), with the kernel determining the level of localization of the statistical risk guarantee. The theoretical results highlight a trade-off between localization of the statistical risk and convergence speed to the long-term risk target. Thanks to localization, L-ARC is demonstrated via experiments to produce prediction sets with risk guarantees across different data subpopulations, significantly improving the fairness of the calibrated model for tasks such as image segmentation and beam selection in wireless networks.

1 Introduction

Adaptive risk control (ARC), also known as online risk control, is a powerful tool for reliable decision-making in online settings where feedback is obtained after each decision [Gibbs and Candes, 2021, Feldman et al., 2022]. ARC finds applications in domains, such as finance, robotics, and health, in which it is important to ensure reliability in forecasting, optimization, or control of complex systems [Wisniewski et al., 2020, Lekeufack et al., 2023, Zhang et al., 2023, Zecchin et al., 2024]. While providing worst-case deterministic guarantees of reliability, ARC may distribute such guarantees *unevenly* in the input space, favoring a subpopulation of inputs at the detriment of another subpopulation.

As an example, consider the tumor segmentation task illustrated in Figure 1. In this setting, the objective is to calibrate a pre-trained segmentation model to generate masks that accurately identify tumor areas according to a user-defined reliability level [Yu et al., 2016]. The calibration process typically involves combining data from various datasets, such as those collected from different hospitals. For an online setting, as visualized in the figure, ARC achieves the desired long-term reliability in terms of false negative ratio. However, it does so by prioritizing certain datasets, resulting in unsatisfactory performance on other data sources. Such behavior is particularly dangerous, as it may result in some subpopulations being poorly diagnosed. This paper addresses this shortcoming of ARC by proposing a novel *localized* variant of ARC.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

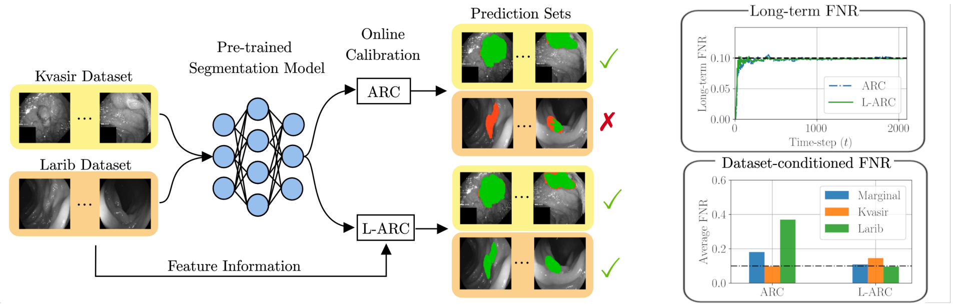

Figure 1: Calibration of a tumor segmentation model via ARC [Angelopoulos et al., 2024a] and the proposed localized ARC, L-ARC. Calibration data comprises images from multiple sources, namely, the Kvasir data set [Jha et al., 2020] and the ETIS-LaribPolypDB data set [Silva et al., 2014]. Both ARC and L-ARC achieve worst-case deterministic long-term risk control in terms of false negative rate (FNR). However, ARC does so by prioritizing Kvasir samples at the detriment of the Larib data source, for which the model has poor FNR performance. In contrast, L-ARC can yield uniformly satisfactory performance for both data subpopulations.

1.1 Adaptive Risk Control

To elaborate, consider an online decision-making scenario in which inputs are provided sequentially to a pre-trained model. At each time step $t \geq 1$, the model observes a feature vector X_t, and based on a bounded *non-conformity scoring function* $s : \mathcal{X} \times \mathcal{Y} \to [0, S_{\text{max}}]$ and a threshold $\lambda_t \in \mathbb{R}$, it outputs a prediction set

$$C_t = C(X_t, \lambda_t) = \{y \in \mathcal{Y} : s(X_t, y) \leq \lambda_t\}, \tag{1}$$

where $\mathcal{Y}$ is the domain of the target variable Y. After each time step t, the model receives feedback in the form of a loss function

$$L_t = \mathcal{L}(C_t, Y_t) \tag{2}$$

that is assumed to be non-negative, upper bounded by $B < \infty$ and non-increasing in the predicted set size $|C_t|$. A notable example is the miscoverage loss

$$\mathcal{L}(C, y) = \mathbb{1}\{y \notin C\}. \tag{3}$$

Accordingly, for an input-output sequence $\{(X_t, Y_t)\}_{t=1}^{T}$ the performance of the set predictions $\{C_t\}_{t=1}^{T}$ in (1) can be gauged via the cumulative risk

$$\bar{L}(T) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(C_t, Y_t) = \frac{1}{T} \sum_{t=1}^{T} L_t. \tag{4}$$

For a user-specified loss level α and a learning rate sequence $\{\eta_t\}_{t=1}^{T}$, ARC updates the threshold λ_t in (1) as [Feldman et al., 2022]

$$\lambda_{t+1} = \lambda_t + \eta_t (L_t - \alpha), \tag{5}$$

where $L_t - \alpha$ measures the discrepancy between the current loss (2) and the target α. For step size decreasing as $\eta_t = \eta_1 t^{-1/2}$ for $a \in (0, 1)$ and an arbitrary $\eta_1 > 0$, the results in [Angelopoulos et al., 2024b] imply that the update rule (5) guarantees that the cumulative risk (4) for the miscoverage loss (3) converges to target level α for *any* data sequence $\{(X_t, Y_t)\}_{t \geq 1}$ as

$$\left|\bar{L}(T) - \alpha\right| \leq \frac{S_{\text{max}} + \eta_1 B}{\sqrt{T}}, \tag{6}$$

thus offering a worst-case deterministic long-term guarantee. Furthermore when data are generated i.i.d. as $(X_t, Y_t) \sim P_{XY}$ for all $t \geq 1$, in the special case of the miscoverage loss (3), the set predictor produced by (5) enjoys the asymptotic *marginal* coverage guarantee

$$\lim_{T \to \infty} \Pr\left[Y \notin C_T\right] \overset{p}{=} \alpha, \tag{7}$$

where the probability is computed with respect to the test sample $(X, Y) \sim P_{XY}$, which is independent of the sequence of samples $\{(X_t, Y_t)\}_{t=1}^{T}$, and the convergence is in probability with respect to the sequence $\{(X_t, Y_t)\}_{t \geq 1}$. Note that in [Angelopoulos et al., 2024b], a stronger version of (7) is provided, in which the limit holds almost surely.

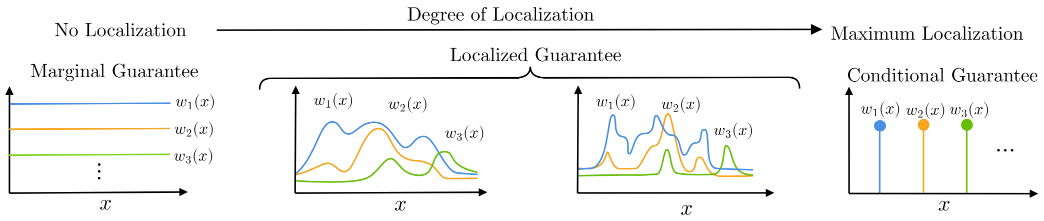

Figure 2: The degree of localization in L-ARC is dictated by the choice of the reweighting function class $\mathcal{W}$ via the marginal-to-conditional guarantee (9). At the leftmost extreme, we illustrate constant reweighting functions, for which marginal guarantees are recovered. At the rightmost extreme, reweighting with maximal localization given by Dirac delta functions for which the criterion (9) corresponds to a conditional guarantee. In between the two extremes lie function sets $\mathcal{W}$ with an intermediate level of localization yielding localized guarantees.

1.2 Conditional and Localized Risk

The convergence guarantee (7) for ARC is marginalized over the covariate X. Therefore, there is no guarantee that the conditional miscoverage $\Pr[Y \notin C_T|X=x]$ is smaller than the target α. This problem is particularly relevant for high-stakes applications in which it is important to ensure a homogeneous level of reliability across different regions of the input space, such as across subpopulations. That said, even when the set predictor $C(X|\mathcal{D}_{\text{cal}})$ is obtained based on an offline calibration data set $\mathcal{D}_{\text{cal}}$ with i.i.d. data $(X,Y) \sim P_{XY}$, it is generally impossible to control the conditional miscoverage probability as

$$\Pr[Y \notin C(X|\mathcal{D}_{\text{cal}})|X=x] \leq \alpha \text{ for all } x \in \mathcal{X} \tag{8}$$

without making further assumptions about the distribution P_{XY} or producing uninformative prediction sets [Vovk, 2012, Foygel Barber et al., 2021].

A relaxed marginal-to-conditional guarantee was considered by Gibbs et al. [2023], which relaxed the marginal miscoverage requirement (8) as

$$\mathbb{E}_{X,Y,\mathcal{D}_{\text{cal}}}\left[\frac{w(X)}{\mathbb{E}_X[w(X)]}\mathbb{1}\{Y \notin C(X|\mathcal{D}_{\text{cal}})\}\right] \leq \alpha \text{ for all } w(\cdot) \in \mathcal{W}, \tag{9}$$

where $\mathcal{W}$ is a set of non-negative reweighting functions, and the expectation is taken over the joint distribution of the calibration data $\mathcal{D}_{\text{cal}}$ and the test pair (X,Y). Note that with a singleton set $\mathcal{W}$ encompassing a single constant function, e.g., $w(x)=1$, the criterion (9) reduces to marginal coverage. Furthermore, as illustrated in Figure 2, depending on the degree of localization of the functions in set $\mathcal{W}$, the criterion (9) interpolates between marginal and conditional guarantees.

At the one extreme, a marginal guarantee like (7) is recovered when the reweighting functions are constant. Conversely, at the other extreme, conditional guarantees as in (8) emerge when the reweighting functions are maximally localized, i.e., when $\mathcal{W}=\{w(x)=\delta(x-\mu):\mu\in\mathcal{X}\}$, where $\delta(x)$ denotes the Dirac delta function. In between these two extremes, one obtains an intermediate degree of localization. For example, this can be done by considering reweighting functions such as

$$\mathcal{W}=\left\{w(x)=\sum_{i=1}^{\infty}\beta_i\left(\kappa\exp\left(-\frac{\|x-\mu_i\|^2}{l}\right)+1\right): \mathbb{E}_X[w(X)]>0,\ \text{and } w(x)\geq 0\ \forall x\in\mathcal{X}\right\}, \tag{10}$$

where $l \geq 0$ is a fixed length scale, $\kappa \geq 0$ is a fixed scaling parameter, and $\|\cdot\|$ denotes the Euclidean norm. Furthermore, function $w(x)$ may also depend on the output of the pre-trained model, supporting calibration requirements via constraints of the form (9) [Zhang et al., 2024].

In Gibbs et al. [2023], the authors demonstrated that it is possible to design *offline* set predictors $C(X|\mathcal{D}_{\text{cal}})$ that *approximately* control risk (9), with an approximation gap that depends on the degree of localization of the family $\mathcal{W}$ of weighting functions.

1.3 Localized Risk Control

Motivated by the importance of conditional risk guarantees, we propose Localized ARC (L-ARC), a novel online calibration algorithm that produces prediction sets with localized statistical risk control guarantees as in (9), while also retaining the worst-case deterministic long-term guarantees (6) of ARC. Unlike Gibbs et al. [2023], our work focuses on *online* settings in which calibration is carried out sequentially based on feedback received on past decisions.

The key technical innovation of L-ARC lies in the way set predictions are constructed. As detailed in Section 2, L-ARC prediction sets replace the single threshold in (1) with a threshold function $g(\cdot)$ mapping covariate X to a localized threshold value $g(X)$. The threshold function is adapted in an online fashion within a reproducing kernel Hilbert space (RKHS) family $\mathcal{G}$ based on an input data stream and loss feedback. The choice of the RKHS family determines the family $\mathcal{W}$ of weighting functions in the statistical guarantee of the form (9), thus dictating the desired level of localization.

The main technical results, presented in Section 2.3, are as follows.

- In the case of i.i.d. sequences, $(X_t, Y_t) \sim P_{XY}$ for all $t \geq 1$, L-ARC provides localized statistical risk guarantees where the reweighting class $\mathcal{W}$ corresponds to all non-negative functions $w \in \mathcal{G}$ with a positive mean under distribution P_{XY}. More precisely, given a target loss value α, the time-averaged threshold function

  $$\bar{g}_T(\cdot) = \frac{1}{T}\sum_{t=1}^{T} g_t(\cdot), \tag{11}$$

  ensures that for any function $w \in \mathcal{W}$, the limit

  $$\limsup_{T\to\infty} \mathbb{E}_{X,Y}\left[\frac{w(X)}{\mathbb{E}_X[w(X)]}\mathcal{L}(C(X,\bar{g}_T),Y)\right] \overset{p}{\leq} \alpha + A(\mathcal{G},w) \tag{12}$$

  holds, where convergence is in probability with respect to the sequence $\{(X_t,Y_t)\}_{t\geq 1}$ and the average is over the test pair (X,Y). The gap $A(\mathcal{G},w)$ depends on both the RKHS $\mathcal{G}$ and function w; it increases with the level of localization of the functions in the RKHS $\mathcal{G}$; and it equals zero in the case of constant threshold functions, recovering (7) for the special case of the miscoverage loss.
- Furthermore, for an arbitrary sequence $\{(X_t,Y_t)\}_{t\geq 1}$ L-ARC has a cumulative loss that converges to a neighborhood of the nominal reliability level α as

  $$\left|\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(C(X_t,g_t),Y_t) - \alpha\right| \leq \frac{B(\mathcal{G})}{\sqrt{T}} + C(\mathcal{G}), \tag{13}$$

  where $B(\mathcal{G})$ and $C(\mathcal{G})$ are terms that increase with the level of localization of the function in the RKHS $\mathcal{G}$. The quantity $C(\mathcal{G})$ equals zero in the case of constant threshold functions, recovering the guarantee (6) of ARC.

In Section 3 we showcase the superior conditional risk control properties of L-ARC as compared to ARC for the task of electricity demand forecasting, tumor segmentation, and beam selection in wireless networks.

2 Localized Adaptive Risk Control

2.1 Setting

Unlike the ARC prediction set (1), L-ARC adopts prediction sets that are defined based on a threshold function $g_t : \mathcal{X} \to \mathbb{R}$. Specifically, at each time $t \geq 1$ the L-ARC prediction set is obtained based on a non-conformity scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as

$$C_t = C(X_t, g_t) := \{y \in \mathcal{Y} : s(X_t, y) \leq g_t(X_t)\}\,. \tag{14}$$

By (14), the threshold $g_t(X_t)$ is localized, i.e., it is selected as a function of the current input X_t. In this paper, we consider threshold functions of the form

$$g_t(\cdot) = f_t(\cdot) + c_t, \tag{15}$$

where $c_t \in \mathbb{R}$ is a constant and function $f_t(\cdot)$ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated to a kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Note that the threshold function $g_t(\cdot)$ belongs to the RKHS $\mathcal{G}$ determined by the kernel $k'(\cdot,\cdot) = k(\cdot,\cdot) + 1$.

We focus on the online learning setting, in which at every time set $t \geq 1$, the model observes an input feature X_t, produces a set C_t, and receives as feedback the loss $L_t = \mathcal{L}(C_t, Y_t)$. Note that label Y_t may not be directly observed, and only the loss $\mathcal{L}(C_t, Y_t)$ may be recorded. Based on the observed sequence of features X_t and feedback L_t, we are interested in producing prediction sets as in (14) that satisfy the reliability guarantees (12) and (13), with a reweighting function set $\mathcal{W}$ encompassing all non-negative functions $w(\cdot) \in \mathcal{G}$ with a positive mean $\mathbb{E}_X[w(X)]$ under distribution P_X, i.e.,

$$\mathcal{W} = \{w(\cdot) \in \mathcal{G} : \ \mathbb{E}_X[w(X)] > 0, \text{ and } w(x) \geq 0 \text{ for all } x \in \mathcal{X}\}. \tag{16}$$

Importantly, as detailed below, the level of localization in guarantee (12) depends on the choice of the kernel $k(\cdot,\cdot)$.

2.2 L-ARC

Given a regularization parameter $\lambda > 0$ and a learning rate $\eta_t \leq 1/\lambda$, L-ARC updates the threshold function $g_t(\cdot) = f_t(\cdot) + c_t$ in (14) based on the recursive formulas

$$c_{t+1} = c_t - \eta_t(\alpha - L_t) \tag{17}$$

$$f_{t+1}(\cdot) = (1 - \lambda\eta_t) f_t(\cdot) - \eta_t(\alpha - L_t) k(X_t, \cdot), \tag{18}$$

with $f_1(\cdot) = 0$ and $c_1 = 0$. In order to implement the update (17)-(18), it is useful to rewrite the function $g_{t+1}(\cdot)$ as

$$g_{t+1}(\cdot) = \sum_{i=1}^{t} a_{t+1}^{i} k(X_i, \cdot) + c_{t+1}, \tag{19}$$

where the coefficients $\{a_{t+1}^i\}_{i=1}^t$ are recursively defined as

$$a_{t+1}^t = -\eta_t(\alpha - L_t) \tag{20}$$

$$a_{t+1}^i = (1 - \eta_t\lambda) a_t^i, \quad \text{ for } i = 1, 2, \ldots, t-1. \tag{21}$$

Accordingly, if the loss L_t is larger than the long-term target α, the update rule (20)-(21) increases the function $g_{t+1}(\cdot)$ around the current input X_t, while decreasing it around the previous inputs $X_1, \ldots, X_{t-1}$. Intuitively, this change enhances the reliability for inputs in the neighborhood of X_t.

It is important to note that, at any time t, computing the threshold function (19) requires storing the coefficients $\{a_t^i\}_{i=1}^{t-1}$ and c_t, as well as the input data $\{X_t\}_{i=1}^t$. Consequently, L-ARC has a linear memory requirement in t, which is a known limitation of non-parametric learning in online settings [Koppel et al., 2020]. Previous research has explored methods that trade memory efficiency for accuracy [Kivinen et al., 2004]. In Appendix C.3, we build on these approaches to present a memory-efficient variant of L-ARC that allows for a trade-off between localized risk control and memory requirements.

2.3 Theoretical Guarantees

In this section, we formalize the theoretical guarantees of L-ARC, which were informally stated in Section 1.3 as (12) and (13).

Assumption 1 (Stationary and bounded kernel)**.** *The kernel function is stationary, i.e.,* $k(x, x') = \tilde{k}(\|x - x'\|)$*, for some non-negative function* $\tilde{k}(\cdot)$*, which is* ρ*-Lipschitz for some* $\rho > 0$*, upper bounded by* $\kappa < \infty$*, and coercive, i.e.,* $\lim_{z\to\infty} \tilde{k}(z) = 0$*.*

Many well-known stationary kernels, such as the radial basis function (RBF), Cauchy, and triangular kernels, satisfy Assumption 1. The smoothness parameter ρ and the maximum value of the kernel function κ determine the localization of the threshold function $g_t(\cdot) \in \mathcal{G}$. For example, the set of functions $\mathcal{W}$ defined in (10) corresponds to the function class (16) associated with the RKHS defined by the raised RBF kernel $k(x, x') = \kappa \exp(-\|x - x'\|^2 / l) + 1$, with length scale $l = 2e(\kappa/\rho)^2$. As illustrated in Figure 2, by increasing κ and ρ, we obtain functions with an increasing level of localization, ranging from constant functions to maximally localized functions.

Assumption 2 (Bounded non-conformity scores)**.** *The non-conformity scoring function is non-negative and bounded, i.e.,* $s(x, y) \leq S_{max} < \infty$ *for any pair* $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Assumption 3 (Bounded and monotone loss)**.** *The loss function is non-negative; bounded, i.e.,* $\mathcal{L}(C, Y) \leq B < \infty$ *for any* $C \subseteq \mathcal{Y}$ *and* $Y \in \mathcal{Y}$*; and monotonic, in the sense that for prediction sets* C' *and* C *such that* $C' \subseteq C$*, the inequality* $\mathcal{L}(C, Y) \leq \mathcal{L}(C', Y)$ *holds for any* $Y \in \mathcal{Y}$.

2.3.1 Statistical Localized Risk Control

To prove the localized statistical guarantee (12) we will make the following assumption.

Assumption 4 (Strictly decreasing loss)**.** *For any fixed threshold function* $g(\cdot) \in \mathcal{G}$*, the loss* $\mathbb{E}_Y[\mathcal{L}(C(X, g), Y)|X = x]$ *is strictly decreasing in the threshold* $g(x)$ *for any* $x \in \mathcal{X}$.

Assumption 5 (Left-continuous loss)**.** *For any fixed threshold function* $g(\cdot) \in \mathcal{G}$*, the loss* $\mathcal{L}(C(x, g + h), y)$ *is left-continuous in* $h \in \mathbb{R}$ *for any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Theorem 1. *Fix a user-defined target reliability* α*. For any regularization parameter* $\lambda > 0$ *and any learning rate sequence* $\eta_t = \eta_1 t^{-1/2} < 1/\lambda$*, for some* $\eta_1 > 0$*, given a sequence* $\{(X_t, Y_t)\}_{t=1}^T$ *of i.i.d. samples from* P_{XY}*, the time-averaged threshold function (11) satisfies the limit*

$$\limsup_{T\to\infty} \mathbb{E}_{X,Y} \left[\frac{w(X)}{\mathbb{E}_X[w(X)]} \mathcal{L}(C(X, \bar{g}_T), Y) \right] \overset{p}{\leq} \alpha + \kappa B \frac{\|f_w\|_{\mathcal{H}}}{\mathbb{E}_X[w(X)]}, \tag{22}$$

for any weighting function $w(\cdot) = f_w(\cdot) + c_w \in \mathcal{W}$ *where the expectation is with respect to the test sample* (X, Y).

Proof. See Appendix A. □

By (22), the average localized loss converges in probability to a quantity that can be bounded by the target α with a gap $A(\mathcal{G}, w)$ that increases with the level of localization κ.

2.3.2 Worst-Case Deterministic Long-Term Risk Control

Theorem 2. *Fix a user-defined target reliability* α*. For any regularization parameter* $\lambda > 0$ *and any learning rate sequence* $\eta_t = \eta_1 t^{-1/2} < 1/\lambda$ *with* $\eta_1 > 0$*, given any sequence* $\{(X_t, Y_t)\}_{t=1}^T$ *with bounded input* $\|X_t\| \leq D < \infty$*, L-ARC produces a sequence of threshold functions* $\{g_t(\cdot)\}_{t=1}^T$ *in (19) that satisfy the inequality*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(C(X_t, g_t), Y_t) - \alpha \right| \leq \frac{1}{\sqrt{T}} \left(\frac{S_{max}}{\eta_1} + \frac{4B\sqrt{\rho\kappa D}}{\eta_1 \lambda} + 2B(2\kappa + 1) \right) + \kappa B. \tag{23}$$

Proof. We defer the proof to Appendix B. □

Formalizing the upper bound in (13), Theorem 2 states that the difference between the long-term cumulative risk and the target reliability level α decreases with a rate $B(\mathcal{G})T^{-1/2}$ to a value $C(\mathcal{G}) = \kappa B$ that is increasing with the maximum value of the kernel κ. In the special case, $\kappa = 0$ which corresponds to no localization, the right-hand side of (23) vanishes in T, recovering ARC long-term guarantee (6).

3 Experiments

In this section, we explore the worst-case long-term and statistical localized risk control performance of L-ARC as compared to ARC. Firstly, we address the task of electricity demand forecasting, utilizing data from the Elec2 dataset [Harries et al., 1999]. Next, we present an experiment focusing on tumor segmentation, where the data comprises i.i.d. samples drawn from various image datasets [Jha et al., 2020, Bernal et al., 2015, 2012, Silva et al., 2014, Vázquez et al., 2017]. Finally, we study a problem in the domain of communication engineering by focusing on beam selection, a key task in wireless systems [Ali et al., 2017]. A further example concerning applications with calibration constraints can be found in Appendix C.2. Unless stated otherwise, we instantiate L-ARC with the RBF kernel

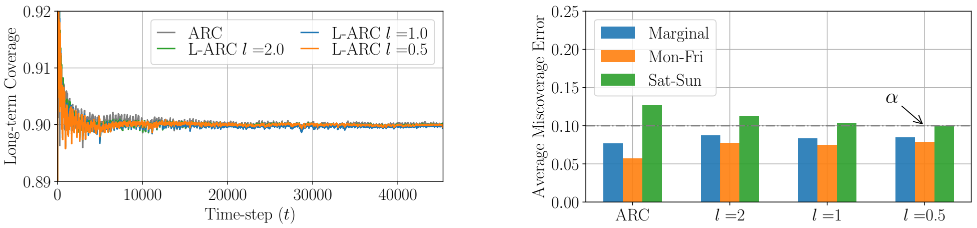

Figure 3: Long-term coverage (left) and average miscoverage error (right), marginalized and conditioned on weekdays and weekends. for ARC and L-ARC with varying values of the localization parameter l on the Elec2 dataset.

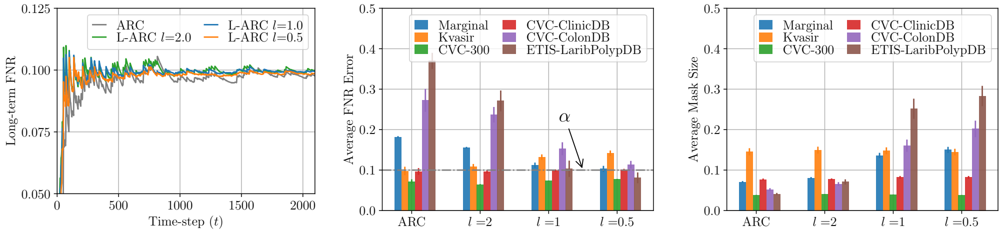

Figure 4: Long-term FNR (left), average FNR across different data sources (center), and average mask size across different data sources (right) for ARC and L-ARC with varying values of the localization parameter l for the task of tumor segmentation [Fan et al., 2020].

$k(x, x') = \kappa \exp(-\|x - x'\|^2 / l)$ with $\kappa = 1$, length scale $l = 1$ and regularization parameter $\lambda = 10^{-4}$. With a smaller length scale l, we obtain increasingly localized weighting functions. All the experiments are conducted on a consumer-grade Mac Mini with an M1 chip. The simulation code is available at `https://github.com/kclip/localized-adaptive-risk-control.git`.

3.1 Electricity Demand

The Elec2 dataset comprises $T = 45312$ hourly recordings of electricity demands in New South Wales, Australia. The data sequence $\{Y_t\}_{t=1}^T$ is subject to distribution shifts due to fluctuations in demand over time, such as between day and night or between weekdays and weekends. We adopt a setup akin to that of Angelopoulos et al. [2024b], wherein the even-time data samples are used for online calibration while odd-time data samples are used to evaluate coverage after calibration. At time t, the observed covariate X_t corresponds to the past time series $Y_{1:t-1}$, and the forecasted electricity demand $\hat{Y}_t$ is obtained based on a moving average computed from demand data collected within the preceding 24 to 48 hours. We produce prediction sets C_t based on the non-conformity score $s(X_t, Y_t) = |\hat{Y}_t - Y_t|$ and we target a miscoverage rate $\alpha = 0.1$ using the miscoverage loss (3). Both ARC and L-ARC use the learning rate $\eta_t = t^{-1/2}$. L-ARC is instantiated with the RBF kernel $k(x, x') = \kappa \exp(-\|\phi(x) - \phi(x')\|^2 / l)$, where $\phi(x)$ is a 7-dimensional feature vector corresponding to the daily average electricity demand during the past 7 days.

In the left panel of Figure 3, we report the cumulative miscoverage error of ARC and L-ARC for different values of the localization parameter l. All algorithms converge to the desired coverage level of 0.9 in the long-term. The right panel of Figure 3, displays the average miscoverage error on the hold-out dataset at convergence. We specifically evaluate both the marginalized miscoverage rate and the conditional miscoverage rate separately over weekdays and weekends. L-ARC is shown to reduce the weekend coverage error rate as compared to ARC providing balanced coverage as the length scale l decreases.

3.2 Tumor Image Segmentation

In this section, we focus on the task of calibrating a predictive model for tumor segmentation. Here, the feature vector X_t represents a $d_{\mathrm{H}} \times d_{\mathrm{W}}$ image, while the label $Y_t \subseteq \mathcal{P}$ identifies a subset of the image pixels $\mathcal{P} = \{(1,1), \ldots, (d_{\mathrm{H}}, d_{\mathrm{W}})\}$ that encompasses the tumor region. As in Angelopoulos et al. [2022], the dataset is a compilation of samples from several open-source online repositories: Kvasir, CVC-300, CVC-ColonDB, CVC-ClinicDB, and ETIS-LaribDB. We reserve 50 samples from each repository for testing the performance post-calibration, while the remaining $T = 2098$ samples are used for online calibration. Predicted sets are obtained by applying a threshold $g(X_t)$ to the pixel-wise logits $f(p_{\mathrm{H}}, p_{\mathrm{W}})$ generated by the PraNet segmentation model [Fan et al., 2020], with the objective of controlling the false negative ratio (FNR) $\mathcal{L}(C_t, Y_t) = 1 - |C_t \cap Y_t|/|Y_t|$. Both ARC and L-ARC are run using the same decaying learning rate $\eta_t = 0.1t^{-1/2}$. L-ARC is instantiated with the RBF kernel $k(x, x') = \kappa \exp(-\|\phi(x) - \phi(x')\|^2 /l)$, where $\phi(x)$ is a 5-dimensional feature vector obtained via the principal component analysis (PCA) from the last hidden layer of the ResNet model used in PraNet.

In the leftmost panel of Figure 4, we report the long-term FNR for varying values of the localization parameter l, targeting an FNR level $\alpha = 0.1$. All methods converge rapidly to the desired FNR level, ensuring long-term risk control. The calibrated models are then tested on the hold-out data, and the FNR and average predicted set size are separately evaluated across different repositories. In the middle and right panels of Figure 4, we report the average FNR and average prediction set size averaged over 10 trials.

The model calibrated via ARC has a marginalized FNR error larger than the target value α. Moreover, the FNR error is unevenly distributed across the different data repositories, ranging from FNR $= 0.08$ for CVC-300 to FNR $= 0.32$ for ETIS-LaribPolypDB. In contrast, L-ARC can equalize performance across repositories, while also achieving a test FNR closer to the target level. In particular, as illustrated in the rightmost panel, L-ARC improves the FNR for the most challenging subpopulation in the data by increasing the associated prediction set size, while maintaining a similar size for subpopulations that already have satisfactory performance.

3.3 Beam Selection

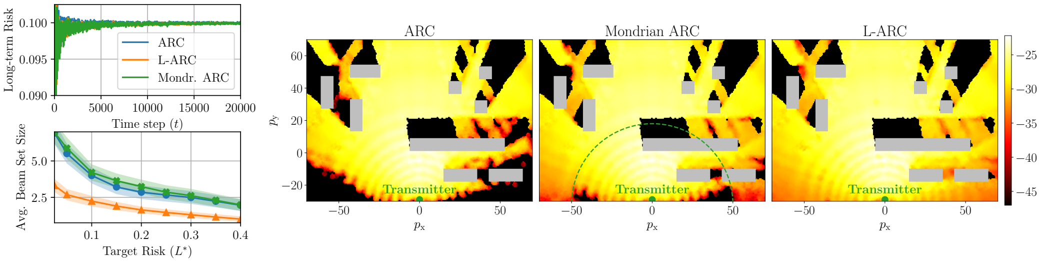

Figure 5: Long-term risk (left-top), average beam set size (left-bottom), and SNR level across the deployment area (right) for ARC, Mondrian ARC, and L-ARC. The transmitter is denoted as a green circle and obstacles to propagation are shown as grey rectangles.

Motivated by the importance of reliable uncertainty quantification in engineering applications, we address the task of selecting location-specific beams for the initial access procedure in sixth-generation wireless networks [Ali et al., 2017]. Further details regarding the engineering aspects of the problem and the simulation scenario are provided in Appendix C.1.1. In the beam selection task, at each time t, the observed covariate corresponds to the location $X_t = [p_{\mathrm{x}}, p_{\mathrm{y}}]$ of a receiver within the network deployment, where p_{x} and p_{y} represent the geographical coordinates. Based on the observed covariate, the transmitter chooses a set, denoted as $C_t \subseteq [1, \cdots, B_{\mathrm{max}}]$, consisting of a subset of the B_{max} available communication beams.

Each communication beam i is associated with a wireless link characterized by a signal-to-noise ratio $Y_{t,i}$, which follows an unknown distribution depending on the user's location X_t. We represent the vector of signal-to-noise ratios as $Y_t = [Y_{t,1}, \ldots, Y_{t,B_{\mathrm{max}}}] \in \mathbb{R}^{B_{\mathrm{max}}}$. For a set C_t, the transmitter sweeps over the beam set C_t, and the performance is measured by the ratio between the SNR obtained on the best beam in set C_t and the best SNR on all the beams, i.e.,

$$\mathcal{L}(C_t, Y_t) = L_t = 1 - \frac{\max_{i \in \mathcal{C}_t} Y_{t,i}}{\max_{i \in \{1,\ldots,B_{\max}\}} Y_{t,i}}. \tag{24}$$

Given an SNR predictor $\hat{Y}_t = f_{\mathrm{SNR}}(X_t)$ for all beams at location X_t, we consider sets that include only beams with a predicted SNR exceeding a threshold $g_t(X_t)$ as

$$C(X_t, g_t) = \{i \in [1, \ldots, B_{\text{max}}] : \hat{Y}_{t,i} > g_t(X_t)\}. \tag{25}$$

In this setting, localization refers to the fair provision of service across the entire deployment area. As a benchmark, we thus also consider an additional calibration strategy that divides the deployment area into two regions: one encompassing all locations near the transmitter, which are more likely to experience high SNR levels, and the other including locations far from the transmitter. For each of these regions, we run two separate instances of ARC algorithms. Inspired by the method introduced in Boström et al. [2021] for offline settings, we refer to this baseline approach as Mondrian ARC.

In the left panels of Figure 5, we compare the performance of ARC, Mondrian ARC, and L-ARC with an RBF kernel with $l = 10$, using a calibration data sequence of length $T = 25000$. All methods achieve the target long-term SNR regret, but L-ARC achieves this result while selecting sets with smaller sizes, thus requiring less time for beam sweeping. Additionally, as illustrated on the right panel, thanks to the localization of the threshold function, L-ARC ensures a satisfactory communication SNR level across the entire deployment area. In contrast, both ARC and Mondrian ARC produce beam-sweep sets with uneven guarantees over the network deployment area.

4 Related Work

Our work contributes to the field of adaptive conformal prediction (CP), originally introduced by Gibbs and Candes [2021]. Adaptive CP extends traditional CP [Vovk et al., 2005] to online settings, where data is non-exchangeable and may be affected by distribution shifts. This extension has found applications in reliable time-series forecasting [Xu and Xie, 2021, Zaffran et al., 2022], control [Lekeufack et al., 2023, Angelopoulos et al., 2024a], and optimization [Zhang et al., 2023, Deshpande et al., 2024]. Adaptive CP ensures that prediction sets generated by the algorithm contain the response variable with a user-defined coverage level on average across the entire time horizon. Recently, Bhatnagar et al. [2023] proposed a variant of adaptive CP based on strongly adaptive online learning, providing coverage guarantees for any subsequence of the data stream. While their approach offers localized guarantees in time, L-ARC provides localized guarantees in the covariate space. More similar to our work is [Bastani et al., 2022], which studies group-conditional coverage. Our work extends beyond coverage guarantees to a more general risk definition, akin to Feldman et al. [2022]. Angelopoulos et al. [2024a] studied the asymptotic coverage properties of adaptive conformal predictions in the i.i.d. setting; and our work extends these results to encompass covariate shifts. Finally, the guarantee provided by L-ARC is similar to that of Gibbs et al. [2023], albeit for an offline conformal prediction setting.

5 Conclusion and Limitations

We have presented and analyzed L-ARC, a variant of adaptive risk control that produces prediction sets based on a threshold function mapping covariate information to localized threshold values. L-ARC can guarantee both worst-case deterministic long-term risk control and statistical localized risk control. Empirical analysis demonstrates L-ARC's ability to effectively control risk for different tasks while providing prediction sets that exhibit consistent performance across various data subpopulations. The effectiveness of L-ARC is contingent upon selecting an appropriate kernel function. Furthermore, L-ARC has memory requirements that grow with time due to the need to store the input data $\{X_t\}_{t>1}$ and coefficients (20)-(21). These limitations of L-ARC motivate future work aimed at optimizing online the kernel function based on hold-out data [Kiyani et al., 2024] or in an online manner [Angelopoulos et al., 2024a], and at studying the statistical guarantees of memory-efficient variants of L-ARC [Kivinen et al., 2004].

6 Acknowledgments

This work was supported by the European Union's Horizon Europe project CENTRIC (101096379). The work of Osvaldo Simeone was also supported by the Open Fellowships of the EPSRC (EP/W024101/1) by the EPSRC project (EP/X011852/1), and by Project REASON, a UK Government funded project under the Future Open Networks Research Challenge (FONRC) sponsored by the Department of Science Innovation and Technology (DSIT). We would also like to express our gratitude to Anastasios Angelopoulos for valuable insights on the technical content of the paper.

References

Anum Ali, Nuria González-Prelcic, and Robert W Heath. Millimeter wave beam-selection using out-of-band spatial information. *IEEE Transactions on Wireless Communications*, 17(2):1038–1052, 2017.

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal PID control for time series prediction. *Advances in Neural Information Processing Systems*, 36, 2024a.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*, 2024b.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR, 2023.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 24–38. PMLR, 2021.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Shachi Deshpande, Charles Marx, and Volodymyr Kuleshov. Online calibrated and conformal prediction improves bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1450–1458. PMLR, 2024.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *arXiv preprint arXiv:2205.09095*, 2022.

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

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Andrea Goldsmith. *Wireless communications*. Cambridge university press, 2005.

Michael Harries, New South Wales, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jakob Hoydis, Fayçal Aït Aoudia, Sebastian Cammerer, Merlin Nimier-David, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna RT: Differentiable ray tracing for radio propagation modeling. *arXiv preprint arXiv:2303.11103*, 2023.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

Shayan Kiyani, George Pappas, and Hamed Hassani. Conformal prediction with learned features. *arXiv preprint arXiv:2404.17487*, 2024.

Alec Koppel, Amrit Singh Bedi, Ketan Rajawat, and Brian M Sadler. Optimally compressed nonparametric online learning: Tradeoffs between memory and consistency. *IEEE Signal Processing Magazine*, 37(3):61–70, 2020.

Jordan Lekeufack, Anastasios A Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. *arXiv preprint arXiv:2310.05921*, 2023.

Horea Muresan and Mihai Oltean. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 10(1):26–42, 2018.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wojciech Wisniewski, David Lindsay, and Sian Lindsay. Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and probabilistic prediction and applications*, pages 285–301. PMLR, 2020.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR, 2021.

Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2016.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Forking uncertainties: Reliable prediction and model predictive control with sequence models via conformal risk control. *IEEE Journal on Selected Areas in Information Theory*, 2024.

Lujing Zhang, Aaron Roth, and Linjun Zhang. Fair risk control: A generalized framework for calibrating multi-group fairness risks. *arXiv preprint arXiv:2405.02225*, 2024.

Yunchuan Zhang, Sangwoo Park, and Osvaldo Simeone. Bayesian optimization with formal safety guarantees via online conformal prediction. *arXiv preprint arXiv:2306.17815*, 2023.

A Proof of Theorem 1

We are interested in bounding the localized risk in (22) of the threshold function (11) for all weighting functions in set $\mathcal{W}$ defined in (16). To study the limit in (22) we first note that L-ARC update rule (18) corresponds to an online gradient descent step for a loss function $\ell(g, x, y)$, with respect to function $f(\cdot)$ and constant c in function $g(\cdot) = f(\cdot) + c$ as in (15). In particular, interpreting the update rule (17)-(18) as a gradient descent step, we obtain that the partial derivatives of the loss function $\ell(g, x, y)$ evaluated at $g(\cdot) = f(\cdot) + c$ are

$$\nabla_f \ell(g) = \frac{\partial \ell(g, x, y)}{\partial f}(\cdot) = (\alpha - \mathcal{L}(C(x, g), y))k(x, \cdot) + \lambda f(\cdot) \in \mathcal{H}, \tag{26}$$

$$\nabla_c \ell(g) = \frac{\partial \ell(g, x, y)}{\partial c} = (\alpha - \mathcal{L}(C(x, g), y)) \in \mathbb{R}, \tag{27}$$

so that the first order approximation of the loss $\ell(g, x, y)$ around $g(\cdot)$ is given by

$$\ell(g + \epsilon\delta_f, x, y) \approx \ell(g, x, y) + \epsilon(\alpha - \mathcal{L}(C(x, g), y))\langle K_x, \delta_f\rangle + \epsilon\langle f, \delta_f\rangle \tag{28}$$

$$\ell(g + \epsilon\delta_c, x, y) \approx \ell(g, x, y) + \epsilon(\alpha - \mathcal{L}(C(x, g), y))\delta_c. \tag{29}$$

In order to study the convexity of the loss $\ell(g, x, y)$ in $g(\cdot)$, we compute the the derivatives of (26)-(27) with respect to $f(\cdot)$ and c. The derivative of (26) with respect to f is the operator $A : \mathcal{H} \to \mathcal{H}$ satisfying

$$\begin{aligned} A\delta_f &= \lim_{\epsilon\to 0} \frac{\nabla_f \ell(g + \epsilon\delta_f) - \nabla_f \ell(g)}{\epsilon} \\ &= -\lim_{\epsilon\to 0} \frac{\mathcal{L}(C(x, g), y) - \mathcal{L}(C(x, g + \epsilon\delta_f), y)}{\epsilon} K_x + \lambda\delta_f \\ &= -\Big\langle \frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} \frac{\partial g(x)}{\partial f}, \delta_f \Big\rangle K_x + \lambda\delta_f \\ &= -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} \langle K_x, \delta_f\rangle K_x + \lambda\delta_f. \end{aligned} \tag{30}$$

It follows that

$$\langle f, Af\rangle = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} f(x)^2 + \lambda \|f\|^2_{\mathcal{H}}. \tag{31}$$

Similarly, the derivative of (26) with respect to c is the operator $B : \mathbb{R} \to \mathcal{H}$ is given by

$$Bc = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} K_x c, \tag{32}$$

which satisfies

$$\langle f, Bc\rangle = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} f(x)c. \tag{33}$$

The derivative of (27) with respect to c is given by

$$Dc = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} c, \tag{34}$$

and the derivative with respect to to f is the operator $C : \mathcal{H} \to \mathcal{R}$ given by

$$C\delta_f = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} \langle K_x, \delta_f\rangle, \tag{35}$$

so that

$$\langle c, Cf\rangle = -\frac{\partial \mathcal{L}(C(x, g), y)}{\partial g(x)} f(x)c. \tag{36}$$

From Assumption 4, the inequality $L' = -\mathbb{E}_Y\left[\frac{\partial \mathcal{L}(C(x,g),Y)}{\partial g(x)}|X=x\right] \geq \gamma > 0$ holds. Thus, the second-order term of the approximation of $\mathbb{E}_Y\left[\ell(g,X,Y)|X=x\right]$ around $g(\cdot) \neq 0$ satisfies

$$\begin{aligned}
\mathbb{E}_Y\left[\begin{bmatrix} f & c \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} f \\ c \end{bmatrix}\middle| X=x\right] =& \mathbb{E}_Y\left[\langle f, Af\rangle + \langle f, Bc\rangle + \langle c, Cf\rangle + \langle c, Dc\rangle | X=x\right] \\
=& L'f(x)^2 + 2L'f(x)c + L'c^2 + \lambda\left\|f\right\|^2_{\mathcal{H}} \\
=& L'(f(x)+c)^2 + \lambda\left\|f\right\|^2_{\mathcal{H}} > 0. \qquad (37)
\end{aligned}$$

We then conclude that the loss function $\mathbb{E}_Y\left[\ell(g,X,Y)|X=x\right]$ is strongly convex in $g(\cdot)$, and that the population loss minimizer

$$g^*(\cdot) = f^*(\cdot) + c^* = \underset{g\in\mathcal{G}}{\arg\min}\, \mathbb{E}_{X,Y}\left[\ell(g,X,Y)\right] \qquad (38)$$

is unique. For any covariate shift $w(\cdot) \in \mathcal{W}$ denote its components $f_w(\cdot) \in \mathcal{H}$ and $c_w \in \mathbb{R}$ such that $w(\cdot) = f_w(\cdot) + c_w$. From the first order optimality conditions, it holds that the directional derivatives with respect to $f_w(\cdot)$ and c_w must satisfy

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\nabla_\epsilon \ell(g^* + \epsilon f_w, X, Y)|_{\epsilon=0}\right] =& \mathbb{E}_{X,Y}\left[(\alpha - \mathcal{L}(C_t(X,g^*),Y))f_w(X) + \lambda\langle f_w, f^*\rangle_{\mathcal{H}}\right] = 0, \qquad (39) \\
\mathbb{E}_{X,Y}\left[\nabla_\epsilon \ell(g^* + \epsilon c_w, X, Y)|_{\epsilon=0}\right] =& \mathbb{E}_{X,Y}\left[(\alpha - \mathcal{L}(C_t(X,g^*),Y))c_w\right] = 0, \qquad (40)
\end{aligned}$$

which implies that for the optimal solution $g^*(\cdot)$

$$\mathbb{E}_{X,Y}\left[\frac{w(X)}{\mathbb{E}_X[w(X)]}\mathcal{L}(C(X,g^*),Y)\right] = \alpha + \lambda\left\langle f^*, \frac{f_w}{\mathbb{E}_X[w(X)]}\right\rangle_{\mathcal{H}}. \qquad (41)$$

Equality (41) amounts to a localized risk control guarantee for the threshold $g^*(\cdot)$ for covariate shift in $w(\cdot) \in \mathcal{W}$. The following lemma states that the time-average L-ARC threshold function $\bar{g}_T(\cdot)$ defined in (11) converges to the population risk minimizer $g^*(\cdot)$.

Lemma 1. *For any regularization parameter $\lambda > 0$ and any learning rate sequence $\eta_t = \eta_1 t^{-1/2} < 1/\lambda$, for some $\eta_1 > 0$, given a sequence $\{(X_t, Y_t)\}_{t=1}^T$ of i.i.d. samples from P_{XY}, the time-averaged threshold function (11) satisfies for any $\epsilon > 0$*

$$\lim_{T\to\infty} \Pr[\left\|g^* - \bar{g}_T\right\|_\infty \geq \epsilon] = 0 \qquad (42)$$

Proof. To prove convergence in probability, we need to show that the loss function $\ell(g,X,Y)$ is bounded. To this end, we first show that $\ell(g,X,Y)$ is Lipschitz in $g(\cdot)$ by studying the norm of the derivatives (26)-(27). For $g_t(\cdot) = f_t(\cdot) + c_t$ returned by the update rule (18), the gradient with respect to $f_t(\cdot)$ satisfies

$$\begin{aligned}
\left\|\frac{\partial \ell(g_t,x,y)}{\partial f}(\cdot)\right\|_{\mathcal{H}} &= \left\|(\alpha - \mathcal{L}(C(x,g_t),y)k(x,\cdot) + \lambda f_t(\cdot)\right\|_{\mathcal{H}} \\
&\leq B\sqrt{\kappa} + \lambda\left\|f_t(\cdot)\right\|_{\mathcal{H}} \leq 2B\sqrt{\kappa}, \qquad (43)
\end{aligned}$$

where the first inequality follows from the boundedness on the kernel (Assumption 1) and the boundedness on the loss (Assumption 3), while the last follows from Proposition 1. The gradient with respect to c can be similarly bounded as

$$\left|\frac{\partial \ell(g_t,x,y)}{\partial c}(\cdot)\right| = |(\alpha - \mathcal{L}(C(x,g_t),y)| \leq B. \qquad (44)$$

From the mean value theorem it follows that for $g(\cdot) = f(\cdot) + c$ and $g'(\cdot) = f'(\cdot) + c'$

$$|\ell(g,X,Y) - \ell(g',X,Y)| \leq \left(2B\sqrt{k} + B\right)\left\|f - f'\right\|_{\mathcal{H}} + B|(c - c')|. \qquad (45)$$

Since L-ARC returns functions $f_t(\cdot)$ with bounded RKHS norm and infinity norm (Proposition 1), and thresholds function $g_t(\cdot)$ with bounded in infinity norm (Proposition 3), we conclude that there exists a finite $\ell_{max} < \infty$ such that $|\ell(g,X,Y)| \leq \ell_{max}$. Given that the loss is bounded we can apply

[Kivinen et al., 2004, Theorem 4] and obtain that for the threshold $\{g_t(\cdot)\}_{t\geq 1}$ returned by L-ARC and the population loss minimizer $g^*(\cdot)$ it holds

$$\frac{1}{T}\sum_{t=1}^{T}\ell(g_t, X_t, Y_t) \leq \frac{1}{T}\sum_{t=1}^{T}\ell(g^*, X_t, Y_t) + B^2\kappa^2(2\kappa^2+1)^2\left(\frac{2}{\sqrt{T}}\left(2\eta_1 + \frac{1}{\eta_1\lambda^2}\right) + \frac{1}{2\eta_1\lambda^2 T}\right). \tag{46}$$

By Hoeffding's inequality the empirical average on the right-hand side of (46) converges to its expected value. Formally, we have that with probability at least $1-\delta$ with respect to the sequence $\{(X_t, Y_t)\}_{t=1}^{T}$

$$\left|\frac{1}{T}\sum_{t=1}^{T}\ell(g^*, X_t, Y_t) - \mathbb{E}_{X,Y}[\ell(g^*, X, Y)]\right| \leq \ell_{max}\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)}. \tag{47}$$

Similarly, by [Cesa-Bianchi et al., 2004, Theorem 2] the empirical risk on the left-hand side of (46) converges to the population risk of the time-averaged solution (11). With probability at least $1-\delta$ with respect to the sequence of samples $\{(X_t, Y_t)\}_{t=1}^{T}$, it holds

$$\mathbb{E}_{X,Y}\left[\ell(\bar{g}_T, X, Y)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\ell(g_t, X_t, Y_t) + \ell_{max}\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)} \tag{48}$$

Combining the two inequalities, with probability at least $1-2\delta$ with respect to $\{(X_t, Y_t)\}_{t=1}^{T}$,

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\ell(\bar{g}_T, X, Y) - \ell(g^*, X, Y)\right] \leq & B^2\kappa^2(2\kappa^2+1)^2\left(\frac{2}{\sqrt{T}}\left(2\eta_1 + \frac{1}{\eta_1\lambda^2}\right) + \frac{1}{2\eta_1\lambda^2 T}\right) \\ & + 2\ell_{max}\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)}\end{aligned} \tag{49}$$

Since the $\mathbb{E}_{X,Y}[\ell(g, X, Y)]$ is strongly convex there exists a value $\gamma > 0$ such that the second order approximation of $\mathbb{E}_{X,Y}[\ell(g, X, Y)]$ at $g^*(\cdot)$ satisfies

$$\frac{\gamma}{2}\left(\left\|f^* - \bar{f}_T\right\|_{\mathcal{H}} + (c^* - \bar{c}_T)\right)^2 \leq \mathbb{E}_{X,Y}[\ell(\bar{g}_T, X, Y) - \ell(g^*, X, Y)] \tag{50}$$

Combining (50) and (49), and leveraging $\|f\|_\infty \leq \sqrt{\kappa}\,\|f\|_{\mathcal{H}}$, which follows from the Assumption 1, we conclude that with probability $1-2\delta$

$$\|g^* - \bar{g}_T\|_\infty \leq \sqrt{\frac{2B^2\kappa^2(2\kappa^2+1)^2}{\gamma(\kappa+1)}\left(\frac{2}{\sqrt{T}}\left(2\eta_1 + \frac{1}{\eta_1\lambda^2}\right) + \frac{1}{2\eta_1\lambda^2 T}\right) + \frac{4\ell_{max}}{\gamma(\kappa+1)}\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)}}. \tag{51}$$

Choosing $\delta = \frac{1}{T}$, for any $\epsilon > 0$, it holds

$$\lim_{T\to\infty}\Pr[\|g^* - \bar{g}_T\|_\infty \geq \epsilon] = 0. \tag{52}$$

□

By itself, the convergence of the threshold function $\bar{g}_T(\cdot)$ to the population risk minimizer $g^*(\cdot)$ is not sufficient to provide localized risk control guarantees for L-ARC time-averaged solution. However, under the additional loss regularity assumption in Assumption 5, we can show that set predictor $C(X, \bar{g}_T)$ enjoys conditional risk control for $T \to \infty$.

Having assumed that the loss $\mathcal{L}(C(x, g), y)$ is left-continuous and decreasing for larger prediction sets (Assumption 3 and 5), for any $\delta' > 0$ there exists $\epsilon > 0$ such that for $g(\cdot)$ such that $\|g^* - g\|_\infty \leq \epsilon$ it holds

$$\mathcal{L}\left(C\left(X, g\right), Y\right) \leq \mathcal{L}\left(C\left(X, g^*\right), Y\right) + \delta'. \tag{53}$$

For such $g(\cdot)$ the following inequality holds

$$\max_{w\in\mathcal{W}}\mathbb{E}\left[\frac{w(X)}{\mathbb{E}[w(X)]}\mathcal{L}\left(C\left(X, g\right), Y\right)\right] \leq \max_{w\in\mathcal{W}}\mathbb{E}\left[\frac{w(X)}{\mathbb{E}[w(X)]}\mathcal{L}\left(C\left(X, g^*\right), Y\right)\right] + \delta'. \tag{54}$$

As stated in Lemma 1, we can always find T large enough, such that $\|g^* - \bar{g}_T\|_\infty \leq \epsilon$ with arbitrary large probability. This implies, that for any $\delta' > 0$ and $w \in \mathcal{W}$,

$$\lim_{T\to\infty} \mathbb{E}\left[\frac{w(X)}{\mathbb{E}[w(X)]}\mathcal{L}\left(C\left(X, \bar{g}_T\right), Y\right)\right] \leq \mathbb{E}\left[\frac{w(X)}{\mathbb{E}[w(X)]}\mathcal{L}\left(C\left(X, g^*\right), Y\right)\right] + \delta' \tag{55}$$

$$\leq \alpha + \lambda \left\langle f^*, \frac{f_w}{\mathbb{E}_X[w(X)]}\right\rangle_{\mathcal{H}} + \delta' \tag{56}$$

$$\leq \alpha + \kappa B \frac{\|f_w\|_{\mathcal{H}}}{\mathbb{E}_X[w(X)]} + \delta', \tag{57}$$

where the inequality (55) follows from (54), the inequality (56) follows from (41) and the inequality (57) from Proposition 1.

B Proof of Theorem 2

We are interested in bounding the absolute difference between the cumulative loss value incurred by the set predictors $\{C(g_t, X_t)\}_{t=1}^T$ produced by L-ARC (18) and the target reliability level α, i.e.,

$$\left|\frac{1}{T}\sum_{t=1}^{T}(\underbrace{\mathcal{L}(C_t, Y_t)}_{L_t} - \alpha)\right|. \tag{58}$$

From Assumption 1 and having assumed $\|X_t\| \leq D$ for $t \geq 1$, it follows that

$$\lim_{\|x\|\to\infty} k(X_t, x) = 0. \tag{59}$$

A bound on the cumulative risk can then be obtained by bounding

$$\left\|\frac{1}{T}\sum_{t=1}^{T}(L_t - \alpha)(k(X_t, \cdot) + 1)\right\|_\infty, \tag{60}$$

where for a function $f : \mathcal{X} \to \mathbb{R}$, the infinity norm $\|f\|_\infty$ is defined as $\max_{x\in\mathcal{X}} |f(x)|$. In fact, from (60) we directly obtain a bound on the cumulative risk

$$\left|\frac{1}{T}\sum_{t=1}^{T}(L_t - \alpha)\right| = \lim_{\|x\|\to\infty}\left|\frac{1}{T}\sum_{t=1}^{T}(L_t - \alpha)(k(X_t, x) + 1)\right| \leq \left\|\frac{1}{T}\sum_{t=1}^{T}(L_t - \alpha)(k(X_t, \cdot) + 1)\right\|_\infty. \tag{61}$$

To this end, we first note that functions $\{f_t(\cdot)\}_{t\in\mathbb{N}}$ generated by (18) have bounded RKHS norm and are smooth.

Proposition 1. *For every $t \geq 1$, we have the inequalities $\|f_t\|_{\mathcal{H}} \leq \frac{B\sqrt{\kappa}}{\lambda}$ and $\|f_t\|_\infty \leq \frac{\kappa B}{\lambda}$.*

Proof. The proof is by induction, with the base case $\|f_1(\cdot)\|_{\mathcal{H}} \leq B\sqrt{\kappa}/\lambda$ being satisfied as $f_1(\cdot) = 0$. The induction step is given as

$$\|f_{t+1}\|_{\mathcal{H}} = \|(1 - \lambda\eta_t)f_t - \eta_t(\alpha - L_t)k(x_t, \cdot)\|_{\mathcal{H}} \tag{62}$$

$$\leq \|(1 - \lambda\eta_t)f_t\|_{\mathcal{H}} + \|\eta_t(\alpha - L_t)k(x_t, \cdot)\|_{\mathcal{H}} \tag{63}$$

$$\leq (1 - \lambda\eta_t)\,\|f_t\|_{\mathcal{H}} + \eta_t B\sqrt{\kappa} \tag{64}$$

$$\leq \frac{B\sqrt{\kappa}}{\lambda}, \tag{65}$$

where the equality (62) follows from the update rule (18); the inequality (63) from the properties of the norm, the inequality (64) from Assumption 1 and 3, and the inequality (65) from the induction hypothesis $\|f_t\|_{\mathcal{H}} \leq \frac{B\sqrt{\kappa}}{\lambda}$. □

Proposition 2. *For $t \geq 1$ and any $(x, x') \in \mathcal{X} \times \mathcal{X}$ we have*

$$|f(x) - f(x')| \leq \frac{B\sqrt{2\rho\kappa D}}{\lambda}. \tag{66}$$

Proof. Denote the evaluation function at x as $K_x = k(x, \cdot)$. From Proposition 1 and the Lipschitz continuity assumed in Assumption 1, it follows that

$$\begin{aligned}|f(x) - f(y)| &= |\langle f, K_x\rangle_{\mathcal{H}} - \langle f, K_y\rangle_{\mathcal{H}}| \\ &= |\langle f, K_x - K_y\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \|K_x - K_y\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}} \sqrt{k(x,x) + k(y,y) - 2k(x,y)} \\ &\leq \|f\|_{\mathcal{H}} \sqrt{2\rho \|x - y\|} \\ &\leq \frac{2B\sqrt{\rho\kappa D}}{\lambda} \end{aligned} \tag{67}$$

where the first inequality follows from Cauchy–Schwarz inequality, the second from the Lipschitz continuity of the kernel, and the last one from Proposition 1 together with $\|x\| \leq D$ for $x \in \mathcal{X}$. □

Leveraging the above characterization of the function $f_t(\cdot)$ returned by L-ARC, we now show that the threshold function $g_t(\cdot)$ has maximum and minimum values that are uniformly bounded.

Proposition 3. *For every $t \geq 1$ and $x \in \mathcal{X}$ we have $g_t(x) \in [G_{min}, G_{max}]$ with*

$$G_{max} = S_{max} + \frac{2B\sqrt{\rho\kappa D}}{\lambda} + \eta_1 B(2\kappa + 1) \tag{68}$$

and

$$G_{min} = -\frac{2B\sqrt{\rho\kappa D}}{\lambda} - \eta_1 B(2\kappa + 1). \tag{69}$$

Proof. We now prove the upper bound (68). The proof is by contradiction and it start by assuming that there exists a $t > 1$ and $x \in \mathcal{X}$ such that $g_t(x) \geq G_{\max}$ while $g_{t'}(\cdot) < G_{\max}$ for all $t' < t$. From the update rule (18) we have that

$$\begin{aligned} g_{t-1}(x) &= g_t(x) + \eta_{t-1}(\alpha - L_t)(k(X_{t-1}, x) + 1) - \lambda\eta_{t-1} f_{t-1}(x) \\ &\geq G_{\max} - \eta_1 B(\kappa + 1) - \lambda\eta_1 |f_{t-1}(x)| \\ &\geq G_{\max} - \eta_1 B(2\kappa + 1). \end{aligned} \tag{70}$$

From Proposition 2 we also have

$$g_{t-1}(X_{t-1}) \geq g_{t-1}(x) - \frac{2B\sqrt{\rho\kappa D}}{\lambda} \geq G_{\max} - \eta_1 B(2\kappa + 1) - \frac{2B\sqrt{\rho\kappa D}}{\lambda} \geq S_{\max}, \tag{71}$$

where the last inequality follows from $G_{\max}$ being defined as (68). From Assumption 2, for all $x \in \mathcal{X}$,

$$g_{t-1}(X_{t-1}) \geq S_{\max} \implies \alpha \geq L_{t-1} \implies g_t(x) \leq (1 - \lambda\eta_{t-1}) g_{t-1}(x) \leq G_{\max}, \tag{72}$$

which contradicts with the original assumption that there exists x such that $g_t(x) \geq G_{\max}$.

The proof of the lower bound (69) follows similarly. Assume there exists $t > 1$ and $x \in \mathcal{X}$ such that $g_t(x) \leq G_{\min}$ while $g_{t'}(\cdot) > G_{\min}$ for $t' < t$. From the update rule (18) we have that

$$\begin{aligned} g_{t-1}(x) &= g_t(x) + \eta_{t-1}(\alpha - L_t)(k(X_{t-1}, x) + 1) - \lambda\eta_{t-1} f_{t-1}(x) \\ &\leq G_{\min} + \eta_1 B(\kappa + 1) + \lambda\eta_1 |f_{t-1}(x)| \\ &\leq G_{\min} + \eta_1 B(2\kappa + 1) \end{aligned} \tag{73}$$

From Proposition 2 we also have

$$g_{t-1}(X_{t-1}) \leq g_{t-1}(x) + \frac{2B\sqrt{\rho\kappa D}}{\lambda} \leq G_{\min} + \eta_1 B(2\kappa + 1) + \frac{2B\sqrt{\rho\kappa D}}{\lambda} \leq 0 \tag{74}$$

where the last inequality follows from $G_{\min}$ being defined as (69). From Assumption 2, for all $x \in \mathcal{X}$,

$$g_{t-1}(X_{t-1}) \leq 0 \implies L_{t-1} \geq \alpha \implies g_t(x) \geq (1 - \lambda\eta_{t-1}) g_{t-1}(x) \geq G_{\min} \tag{75}$$

which contradicts the assumption that there exists $\min_{x \in \mathcal{X}} g_t(x) \leq G_{\min}$. □

Having established an upper and lower bound on the maximum value of the function $g_t(\cdot)$ generated by (18) we can now bound (60). Define $\Delta_t = \eta_t^{-1} - \eta_{t-1}^{-1}$ and $\Delta_1 = \eta_1^{-1}$ and note that

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{T}(L_t-\alpha)\right| &\leq \max_{x\in\mathcal{X}}\left|\frac{1}{T}\sum_{t=1}^{T}(L_t-\alpha)(k(X_t,x)+1)\right| \\
&= \left\|\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{r=1}^{t}\Delta_r\right)\eta_t(L_t-\alpha)(k(X_t,\cdot)+1)\right\|_\infty \\
&= \left\|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(\sum_{t=r}^{T}\eta_t(L_t-\alpha)(k(X_t,\cdot)+1)\right)\right\|_\infty \\
&= \left\|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(\sum_{t=r}^{T}f_{t+1}+c_{t+1}-(1-\lambda\eta_t)f_t-c_t\right)\right\|_\infty \\
&= \left\|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(g_{T+1}-g_r+\lambda\sum_{t=r}^{T}\eta_t f_t\right)\right\|_\infty \\
&\leq \left\|\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left(g_{T+1}-g_r\right)\right\|_\infty + \left\|\frac{\lambda}{T}\sum_{r=1}^{T}\Delta_r\sum_{t=r}^{T}\eta_t f_t\right\|_\infty \\
&\leq \underbrace{\frac{1}{T}\sum_{r=1}^{T}\Delta_r\left\|g_{T+1}-g_r\right\|_\infty}_{:=E_1} + \underbrace{\frac{\lambda}{T}\sum_{t=1}^{T}\left\|f_t\right\|_\infty}_{:=E_2}.
\end{aligned} \tag{76}$$

The first term can be bounded based on Proposition (3) as

$$E_1 \leq \frac{1}{T}\max_r\left\|g_{T+1}-g_r\right\|_\infty\sum_{r=1}^{T}\Delta_r = \frac{1}{\eta_T T}\left(S_{\text{max}}+\frac{4B\sqrt{\rho\kappa D}}{\lambda}+2\eta_1 B(2\kappa+1)\right), \tag{77}$$

and similarly, for the second term, we have

$$E_2 \leq \frac{\lambda}{T}\sum_{t=1}^{T}\frac{\kappa B}{\lambda} = \kappa B. \tag{78}$$

Fix a decreasing learning rate $\eta_t = \eta_1 t^{-\omega}$ and a regularization parameter $\lambda = \lambda_0 T^{-\xi}$, then the E_1 becomes

$$E_1 = \frac{S_{\text{max}}}{\eta_1 T^{1-\omega}} + \frac{4B\sqrt{\rho\kappa D}}{\eta_1\lambda T^{1-\omega}} + \frac{2B(2\kappa+1)}{T^{1-\omega}} \tag{79}$$

For any $\omega < 1$, it follows

$$\lim_{T\to\infty}\left|\frac{1}{T}\sum_{t=1}^{T}(L_t-\alpha)\right| = \kappa B. \tag{80}$$

C Additional Experiments

C.1 Beam Selection

C.1.1 Simulation Details

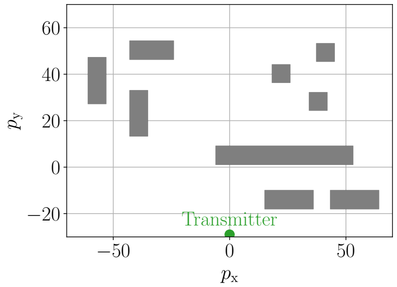

Figure 6: Network deployment assumed in the simulations. A single transmitter (green circle) communicates with receivers that are uniformly distributed in a scene containing multiple buildings (grey rectangles).

For the beam selection experiment, we consider the network deployment depicted in Figure 6, in which a transmitter (green circle) communicates with users in an urban environment with multiple buildings (grey rectangles). We assume that communication occurs at a frequency $f_c = 2.14$ GHz and that the transmitter is equipped with $N_t = 8$ transmitting antennas while receiving users have single-antenna equipment. The transmitter adopts a discrete Fourier transform beamforming codebook of size $B_{\max} = 11$, with each beam b_i given by

$$b_i = \frac{1}{\sqrt{N_t}}[1, e^{j2\pi \frac{2\pi i}{B_{\max}}}, \dots, e^{j(N_t-1)\frac{2\pi i}{B_{\max}}}] \in \mathbb{C}^{N_t}, \quad \text{for } i \in \{0, \dots, B_{\max}-1\}, \tag{81}$$

where $j = \sqrt{-1}$. The wireless channel response $h^R \in \mathbb{C}^{N_t}$ between the transmitter and a receiver located at $X_t = [p_{\mathrm{x}}, p_{\mathrm{y}}] \in \mathbb{R}^2$, is modeled using Sionna ray-tracer [Hoydis et al., 2023], and we account for small scale fading using a Rayleigh noise model [Goldsmith, 2005]. The resulting channel vector is distributed as

$$h_t \sim h^R(X_t) + Rayleigh(\sigma). \tag{82}$$

where $h^R(X_t)$ is the ray tracer output and $Rayleigh(\sigma)$ is a Rayleigh distributed random variable with parameter $\sigma = 10^{-4}$. Assuming unit power transmit symbols and receiver noise, for a channel vector h_t the communication signal-to-noise ratio (SNR) obtained using the beamformer b_i is given by

$$Y_{t,i} = h_t^{\mathrm{T}} b_i. \tag{83}$$

Beam sets are obtained calibrating an SNR predictor $\hat{Y}_t = f_{\mathrm{SNR}}(X_t)$ realized using a 3-layer fully connected neural network that is trained on 2500 samples with the user location generated uniformly at random within the deployment area.

C.1.2 Effect of the Length Scale

In Figure 7, we study the effect of the length scale l of the kernel function on the time-averaged threshold function $\bar{g}_T(X)$ returned by L-ARC. We report the value of the L-ARC time-averaged threshold, $\bar{g}_T(X)$, in (11), for the same experimental set-up as in Section 3.3, and for increasing localization of the kernel function. As the length scale parameter l decreases, corresponding to a more localized kernel, the value of the threshold is allowed to vary more across the deployment area. In particular, the threshold function reduces its value around areas where the beam selection problem becomes more challenging, such as building clusters, in order to create larger beam selection sets.

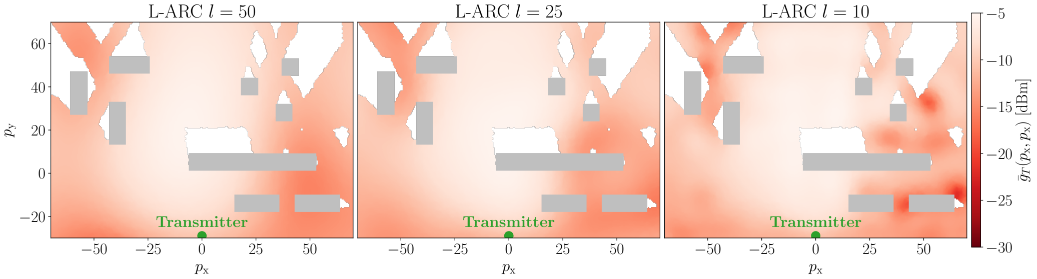

Figure 7: Time-averaged threshold function $\bar{g}_T$ for different values of localization parameter l.

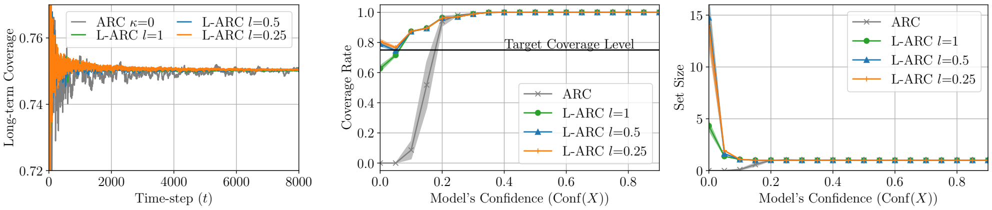

Figure 8: Long-term coverage (left), coverage rate (center), and prediction set size (right) versus model's confidence for ARC and L-ARC for different values of the localization parameter l.

C.2 Image Classification with Calibration Requirements

In this section, we consider an image classification task under calibration requirements based on the fruit-360 dataset [Muresan and Oltean, 2018]. For this problem, the feature vector X_t is an image of size 100×100, and the corresponding label $Y_t \in \mathcal{Y} = \{1, \dots, 130\}$ is one of 130 types of fruit, vegetable, or nut in image X_t. We study the online calibration of a pre-trained ResNet18 model [He et al., 2016]. For an input image X_t, the prediction set is obtained from the model's predictive distribution $\hat{p}(y|X_t)$ as

$$C(X_t, g_t) = \{y \in \mathcal{Y} : \hat{p}(y|X_t) > g_t(X_t)\}, \tag{84}$$

and we target the miscoverage loss (3) with a target miscoverage rate $\alpha = 0.25$. In order to capture calibration requirements, we impose coverage constraints that are localized in the model's confidence. The model's confidence indicator is given by the maximum value of the model's predictive distribution $\hat{p}(y|X_t)$, i.e.,

$$\text{Conf}(X_t) = \max_{y \in \mathcal{Y}} \hat{p}(y|X_t). \tag{85}$$

Accordingly, we run ARC and L-ARC calibration with a sequence of $T = 8000$ samples and we instantiate L-ARC using the exponential kernel $k(x, x') = \kappa \exp(-\|\phi(x) - \phi(x')\|^2 / l)$, where the feature vector is given by the model's uncertainty, i.e., $\phi(x) = \text{Conf}(x)$.

In the left-most panel of Figure 8 we report the long-term coverage of ARC and L-ARC for an increasing level of localization obtained by decreasing the length scale l. All methods guarantee long-term coverage. In the middle panel, we use hold-out data to evaluate the coverage of the calibrated model conditioned on the model's confidence level. For small length scale l, L-ARC yields prediction sets that satisfy the coverage requirement across different levels of the model's confidence. In contrast, ARC, due to its inability to adapt the threshold function, has a large miscoverage rate for small model confidence levels. As illustrated in the right panel, this is achieved by producing a larger set size when the model's confidence is low.

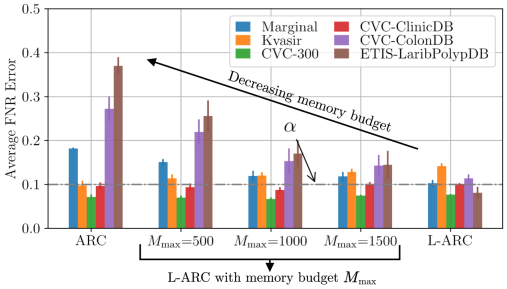

Figure 9: FNR obtained by ARC, L-ARC, and L-ARC with limited memory budget $M_{\text{max}} \in \{500, 1000, 1500\}$. As the memory budget increases, the localized risk control performance of L-ARC interpolates between ARC and L-ARC.

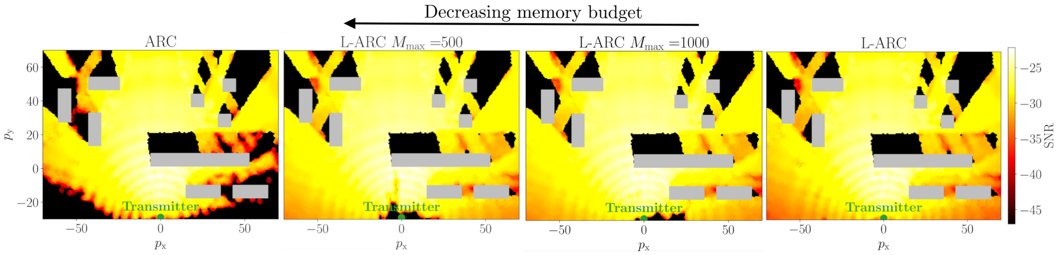

Figure 10: SNR across the deployment attained by L-ARC with limited memory budget M_{max}.

C.3 On the memory efficiency of L-ARC

In a manner similar to [Kivinen et al., 2004], it is possible to obtain a memory-efficient version of L-ARC that adopts a truncated version of L-ARC threshold (19) given by

$$g_{t+1}(\cdot) = \sum_{i=\max\{1,t-M_{\text{max}}\}}^{t} a^i_{t+1} k(X_i, \cdot) + c_{t+1}. \tag{86}$$

Unlike the threshold (19), which has a *linear* memory requirement, the truncated version (86) requires a *constant* memory and computational load that are proportional to the number of coefficients M_{max}. It is known that in online non-parametric learning, there exists a trade-off between memory efficiency and performance. In the following, we empirically study the trade-off between the localized risk control of L-ARC and its memory requirements by varying the parameteR M_{max}.

C.3.1 Tumor Segmentation

Using the setup described in Section 3.2, we now consider calibrating the image segmentation model using L-ARC with a truncated threshold (86). In Figure 9, we report the average FNR conditioned on different data sources for $M_{\text{max}} \in \{500, 1000, 1500\}$. As a benchmark, we also compare against ARC and L-ARC without truncation. By adjusting the value of M_{max}, it is possible to trade off localized risk control for memory efficiency. In fact, the effect of truncation on L-ARC's performance is minimal when the number of coefficients in the truncation is large ($M_{\text{max}} = 1500$). However, for greater memory savings ($M_{\text{max}} = 500$), L-ARC's performance becomes similar to that of ARC. In all cases, L-ARC provides better localized risk control than ARC.

C.3.2 Beam Selection

We consider the beam selection problem discussed in Section 3.3. In Figure 10, we report the SNR levels across the deployment attained by ARC, L-ARC, and L-ARC with a truncated threshold with a maximum number of coefficients $M_{\text{max}} \in \{500, 1000\}$. As the number of coefficients M_{max} and the memory requirement reduce, the localized risk control performance of L-ARC also decreases. Nonetheless, even for small M_{max}, L-ARC delivers a more consistent SNR level across the deployment compared to ARC.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The theoretical claims about the proposed calibration algorithm are supported by Theorem 1 and Theorem 2, while experimental results in Section 3 demonstrate its capability to control long-term risk and to improve fairness across data subpopulations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5, we highlight two primary limitations of L-ARC: its memory requirements and the necessity of specifying a suitable kernel. Additionally, we suggest potential directions for future research to address these issues. A memory-efficient version of L-ARC is given in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Both theorems are preceded by a necessary set of assumptions. In the proofs, provided in the appendix, we reference these assumptions when using them.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in the appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the main text, we provide details of all the experiments, including factors influencing the proposed solution, such as datasets and algorithm parameters like the choice of kernel, localization parameters, and learning rate.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In the supplementary material, we include the source code of the experiments along with a concise guide containing all the necessary information to replicate the results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the information, such as the type of datasets, the learning rate, the type of kernel function, and localization parameters, is reported in the main text. In the appendix, we provide additional details about the data generation for the beam selection experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where possible we report 95% confidence intervals in the form of error bars or shaded areas.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate that the experiments are run on a Mac Mini.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully read and comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper provides a new calibration scheme that offers long-term risk control and statistical localized risk guarantees. We do not see any negative societal impacts associated with the proposed algorithm.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used in the experiments are properly credited by citing the corresponding papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not use crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.