# CROSS-LEARNING FOR PERFORMATIVE PREDICTION WITH CHANGING OBJECTIVES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study contextual performative prediction, where a learner repeatedly chooses a model to deploy, the model influences the environment, and the learner incurs a context-dependent loss. Crucially, we assume the learning objective might change with the context, while the distribution map remains unaffected by the learners strategy of selecting which model to deploy. In this setting, deploying a model in one context also provides information about its behavior in other contexts. We formalize this structure through continuous-armed bandits with cross-learning, and introduce the cross-context zooming dimension, a complexity measure that captures the difficulty of minimizing performative regret across varying contexts. We establish a lower bound on the regret in terms of this dimension and design matching algorithms, first for the general cross-learning setting and then tailored to contextual performative prediction. More broadly, our results expose the mechanism of performative shifts as a shared component across varying ways of assessing model quality, and suggest a way to reuse information across learning problems.

## 1 INTRODUCTION

Performative prediction (Perdomo et al., 2020) refers to the problem of learning predictive models in settings where the predictions influence the distribution of observations, a form of feedback which arises naturally in social context. In credit scoring, for instance, the interest rate offered to a borrower influences their ability to repay, thereby altering the very outcome the model was trained to predict. Similarly, in dynamic pricing, the price set by a platform affects the customers' willingness to pay, and thus the demand, and prices in the future. These and similar observations have been discussed in economic forecasting for decades (Morgenstern, 1928; Grunberg and Modigliani, 1954) and recently sparked many related questions in machine learning (Hardt and Mendler-Dünner, 2025).

In machine learning, performativity surfaces in that the deployed model induces a shift in the data distribution it has been trained on. We write $\mathcal{D}(\theta)$ for the data induced by model $\theta$. This notion of a model-dependent distribution gives raise to the performative risk as a central object of study:

$$\mathcal{PR}(\theta) := \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z\,;\theta)].$$

The performative risk measures a model's observed loss after deployment. Finding minima of the performative risk is challenging as the impact of a model on the distribution, and hence $\mathcal{D}(\theta)$, is typically unknown ahead of time. Furthermore, the performative risk function can become arbitrary complex, even for simple loss functions (Miller et al., 2021). Embracing these unknowns, recent work has framed performative risk minimization as a bandit problem where the goal is to sequentially select models to minimize cumulative performative risk (Jagadeesan et al., 2022; Chen et al., 2023). They show that when leveraging the feedback structure inherent to the performative prediction problem, sublinear regret can be achieved with only weak regularity assumptions on the problem.

**The problem.** Assume you can find a performatively optimal point for your learning problem and now you wish to tweak your loss function $\ell$, e.g., by changing the weighting of data points or the regularization strength, based on some new side-information you have obtained. As a result an action that was optimal under the previous objective might become suboptimal under the new objective. A naive application of existing performative risk minimization methods, e.g. (Jagadeesan et al., 2022; Cyffers et al., 2024; Miller et al., 2021; Izzo et al., 2021), would treat each problem as an



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

independent learning problem with a different reward function that needs to be optimized. This raises a natural question: How can the gained knowledge from one problem inform the solution of the new problem? In particular, we want to leverage the structure of the performative prediction problem, and the fact that the distribution map is unaffected by the learner's objective used to assess the quality of a model. Thus, samples collected in previous instances of the problem remain informative about the model-induced shift. We formalize and address this scenario as follows.

**Our contributions.** We cast the problem of finding performative optima under changing loss functions as a contextual bandit problem with performative feedback. The distribution map is unknown but fixed across time and the context determines the loss associated with the deployed predictions in each time step.

**Problem CBPF** (Contextual bandit with performative feedback). *At round $t \leq T$, the learner observes a context $c_t \in \mathcal{C}$chosen by an oblivious adversary, chooses a model $\theta_t \in \Theta$, observes $Z_t \sim \mathcal{D}(\theta_t)$, , and incurs a context-dependent loss $\ell_{c_t}(Z_t, \theta_t)$. The learner's goal is to minimize the cumulative performative regret*

$$\mathcal{R}_T^{CBPF} = \sum\nolimits_{t \leq T} \mathcal{PR}_{c_t}(\theta_t) - \inf_{\theta \in \Theta} \mathcal{PR}_{c_t}(\theta),$$

*for the performative risk $\mathcal{PR}_c(\theta) = \mathbb{E}_{Z \sim \mathcal{D}(\theta)} [\ell_c(Z, \theta)]$.*

Building on this equivalence our contributions can be summarized as follows:

- We show that, under standard assumptions on the learning problem, the contextual performative prediction problem CBPF reduces to the more general problem of **Lipschitz-continuous bandits with cross-learning** (CBCL) (Balseiro et al., 2019). This constitutes an extension of multi-armed bandit with cross learning to continuous sets of action.
- We study Problem CBCL. We design a minimax-optimal algorithm called CL-ZOOMING (Section 3.1), with matching upper (Section 3.3) and lower (Section 3.4) bounds on its regret. This allows us to precisely characterize the dependence of the regret on the set of contexts, through the novel notion of **cross-context zooming dimension**.
- We leverage the new concepts and methods to address Problem CBPF. We first derive the regret bound of Algorithm CL-ZOOMING for Problem CBPF, and then show how to leverage the specifics of learning under performative feedback to further improve the algorithm, in particular when the distribution shift is small, or when the loss $\ell_c(Z, \theta)$ varies substantially with $\theta$. Finally, we illustrate our insights on a problem of dynamic pricing with costs.

## 2 BACKGROUND, HYPOTHESES AND REDUCTION

In performative prediction, the risk of a deployed model $\theta$ is measured by the expected loss $\ell(Z_t, \theta_t)$, across observations $Z_t \sim \mathcal{D}(\theta_t)$ that are itself impacted by the deployed model. To ensure learnability in a setting where the distribution can shift based on the deployed model some basic regularity assumptions are necessary. We adopt the following sensitivity assumption of (Perdomo et al., 2020).

**Assumption 1** ($\epsilon$-sensitivity of $\mathcal{D}$)**.** *The distribution map $\mathcal{D}$ is $\epsilon$-sensitive: for all pairs $(\theta, \theta') \in \Theta$, $\mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \epsilon \|\theta - \theta'\|$, where $\mathcal{W}$ is the Wasserstein-1 distance, and $\|\cdot\|$ the euclidean norm.*

Under Assumption 1, two similar models $\theta, \theta'$ induce close distributions. Among others it implies that observations generated under one model can be used to estimate the expected risk of nearby models. This regularity assumption also underlies fundamental results about the stability of retraining (e.g., Mendler-Dünner et al., 2020; Drusvyatskiy and Xiao, 2023). We refer to Hardt and Mendler-Dünner (2025) for a comprehensive survey of the related results. We refrain from making any other structural assumption on the distribution map to tackle the problem. Model-based approaches to performative risk minimization have been explored in (e.g., Lin and Zrnic, 2024; Mendler-Dünner et al., 2022).

To cope with the dependence of the performative risk on the model $\theta$ through the distribution argument and the predictions to be evaluated, it is useful to expose the **decoupled performative risk** $\mathcal{DPR}_c$ (Perdomo et al., 2020): for $c \in \mathcal{C}$ and $\theta, \theta' \in \Theta$,

$$\mathcal{DPR}_c(\theta', \theta) = \mathbb{E}_{Z \sim \mathcal{D}(\theta')}[\ell_c(Z, \theta)].$$

It corresponds to the risk of model $\theta$ when the observations are generated under model $\theta'$. To ensure that $\mathcal{DPR}_c(\theta', \theta)$ is close to the performative risk $\mathcal{PR}_c(\theta)$ for models $\theta'$ close to $\theta$, we assume

Lipschitz-continuity of the losses with regards to $z$. Moreover, we also assume Lipschitz-continuity of the losses with regards to $\theta$.

**Assumption 2** (Lipschitz losses)**.** *$\forall c \in \mathcal{C}$, the loss $\ell_c(z, \theta)$ is $L_z$-Lipschitz in $z$ and $L_\theta$-Lipschitz in $\theta$.*

Jagadeesan et al. (2022) show that, under these assumptions, performative prediction can be cast as *Lipschitz bandits* (Agrawal, 1995; Kleinberg, 2004; Kleinberg et al., 2008; Bubeck et al., 2011), *i.e.* a sequential decision-making problem over continuous action spaces where the expected reward is Lipschitz-continuous in the action. More specifically, they show (Lemma 2.1) that $|\mathcal{PR}_c(\theta) - \mathcal{DPR}_c(\theta', \theta)| \leq \epsilon L_z \|\theta - \theta'\|$. Then, the performative risk is $(L_\theta + \epsilon L_z)$-Lipschitz continuous: for all $c \in \mathcal{C}$, for all $\theta \in \Theta$, $|\mathcal{PR}_c(\theta) - \mathcal{PR}_c(\theta')| \leq (L_\theta + \epsilon L_z)\|\theta - \theta'\|$.

**Context-dependent loss function.** We address Problem CBPF under the assumption that the loss can change during the learning process. Throughout, we use $c \in \mathcal{C}$ to index the loss functions $\ell_c$. This setting relates to contextual Lipschitz bandits (Slivkins, 2011), which assume Lipschitzness jointly over action-context pairs, and to non-stationary Lipschitz bandits (Nguyen et al., 2026), where rewards may drift over time independently of any context. However, we emphasize that we impose no regularity on the drift of the loss function $\ell_c$ with respect to the context $c$: $c \approx c'$ does not imply that $\ell_c(z, \theta) \approx \ell_{c'}(z, \theta)$ but only that both functions are similarly smooth in both their arguments. Following standard assumptions both in bandit and performative prediction literature, and w.l.o.g., we consider bounded losses and parameter spaces $\Theta = [0, 1]^D$ and $\ell_c(z, \theta) \in [0, 1]$. While related work has considered shifts in the distribution map (Lee and Zrnic, 2026; Hardt, 2026), methods to tackle changes to the learning objective have not been explored before, to the best of our knowledge.

**Reduction to bandits with cross-learning.** The key observation is that under the performative feedback assumptions above, deploying a model $\theta$ in context $c$ induces observations $z \sim \mathcal{D}(\theta)$ which can be reused in context $c'$ to estimate another loss $\ell_{c'}(z, \theta)$ for the same model. This allows us to connect our problem more generally to **bandits with cross-learning** (Balseiro et al., 2019; Hao et al., 2022; Schneider and Zimmert, 2023; Tullii et al., 2024), in which observations collected in one context can inform decisions in other contexts. To capture our performative prediction problem, we define **continuous bandit with cross-learning** with Lipschitz reward:

**Problem CBCL** (Continuous bandit with cross learning.)**.** *At round $t \leq T$, the learner observes a context $c_t \in \mathcal{C}$ and chooses an action $\theta_t \in \Theta$. Then, they observe a random process $\{M_t(c)\}_{c \in \mathcal{C}}$ independent of the past conditionally on $\theta_t$ with mean $\mathbb{E}[M_t(c)|\theta_t] = m(\theta_t, c)$, and receive reward $M_t(c_t)$. Their goal is to minimize the regret*

$$\mathcal{R}_T^{CBCL} = \sum\nolimits_{t \leq T} \sup_{\theta \in \Theta} m(\theta, c_t) - m(\theta_t, c_t).$$

Similarly to Problem CBPF, we consider Lipschitz-continuous reward functions bounded in $[0, 1]$.

**Assumption 3** (Lipschitz reward)**.** *For all $c \in \mathcal{C}$ and all $\theta, \theta' \in \Theta$, $|m(\theta, c) - m(\theta', c)| \leq L\|\theta - \theta'\|$.*

Problem CBCL extends cross-learning from discrete to continuous action spaces, a new problem of independent interest that we study in Section 3, before returning in Section 4 to Problem CBPF. In particular, we provide a novel characterization of the complexity of the problem that gives a more nuanced picture from what was previously known. In the $K$-armed contextual bandit setting, Balseiro et al. (2019) showed that the regret is independent of the number of contexts, but extensions to structured reward in continuous action spaces remained poorly understood.

**Notations.** We denote by $\mathcal{C}_T = \{c_t\}_{t \leq T}$ the set of observed contexts. For each context $c \in \mathcal{C} - t$, when considering Problem CBCL, we define the optimal mean reward in this context as $m^\star(c) := \sup_{\theta \in \Theta} m(\theta, c)$; when considering Problem CBPF we similarly define $\mathcal{PR}_c^* := \inf_{\theta \in \Theta} \mathcal{PR}_c(\theta)$. Abusing notations, we define $\Delta(\theta, c) = m^\star(c) - m(\theta, c)$ when considering Problem CBCL, and $\Delta(\theta, c) = \mathcal{PR}_c(\theta) - \mathcal{PR}_c^*$ when considering Problem CBPF. For a set $A \subseteq \Theta$ and $\rho > 0$, let $N_\rho(A)$ denote the $\rho$-packing number of $A$, *i.e.* the largest cardinality of a subset of $A$ whose distinct elements are pairwise at distance at least $\rho$. We denote by $\mathcal{B}(x, r)$ the Euclidian ball centred in $x$ with radius $r$. For $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of $a$ and $b$, and $a \vee b$ their maximum.

## 3 Continuous bandit with cross learning

Since no regularity assumption is made on $m$ with respect to $c$, a first naive approach to Problem CBCL would be to treat each context as an independent Lipschitz bandit, at a prohibitive cost when the

**Algorithm 1** CL-ZOOMING

**Input:** horizon $T$, confidence radius $r_t$, Lipschitz constant $L$, grid $\mathcal{G}$
**Init:** Initialize active set $\mathcal{A}_1 \leftarrow \emptyset$.
**for** $t = 1, \dots, T$ **do**
  Observe context $c_t$
  **if** there exists $\theta \in \Theta$ not covered by $\mathcal{A}_t$ **then**
    $\theta_t \in \arg\min_{\theta' \in \mathcal{G}} \|\theta - \theta'\|$
    $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{\theta_t\}$ ▷ Activate $\theta_t$
  **else**
    $\theta_t \in \arg\max_{\theta \in \mathcal{A}_t} \{\widehat{m}_t(\theta, c_t) + 2r_t(\theta)\}$ ▷ Upper Confidence Bound
  **end if**
  Observe $\{M_t(c)\}_{c \in \mathcal{C}}$
**end for**

number of contexts is large. Fortunately, the structure of $M_t$ allows to transfer information collected in one context to other contexts. This raises the following question: can we exploit this structure to achieve lower regret, and if so, how should the problem's complexity depend on the context set? In this section, we introduce the *cross-contextual zooming dimension*, and show that the regret of CL-ZOOMING (Alg. 1) scales optimally with this complexity measure.

### 3.1 CL-ZOOMING ALGORITHM

To address Problem CBPF, we design Algorithm CL-ZOOMING, inspired by the ZOOMING algorithm of Kleinberg et al. (2008) and adapted to the cross-learning contextual feedback. At each step $t \leq T$, the algorithm maintains a set $\mathcal{A}_t$ of active actions that belong to a pre-specified grid of actions $\mathcal{G}$. Each active action $\theta \in \mathcal{A}_t$ has a confidence radius

$$r_t(\theta) := \min\left\{1, \sqrt{\frac{2(3+D)\log(T)}{N_t(\theta) \vee 1}}\right\} \qquad \text{where} \qquad N_t(\theta) = \sum_{s<t} \mathbb{1}\{\theta_s = \theta\}.$$

We say that an active action $\theta' \in \mathcal{A}_t$ covers an action $\theta \in \Theta$ at time $t$ if $L\|\theta - \theta'\| < r_t(\theta')$. At each round $t$, if some action is not covered, the algorithm activates the nearest action on the grid (ties broken arbitrarily) and plays it. Otherwise, the algorithm chooses an active action maximizing an upper confidence bound. By contrast to classical contextual bandits where only observations associated with action $\theta$ in the current context are used to estimate its reward, here we leverage all previous observations corresponding to $\theta$ and set

$$\widehat{m}_t(\theta, c) = \frac{1}{N_t(\theta)} \sum_{s=1}^{t-1} \mathbb{1}\{\theta_s = \theta\} M_s(c).$$

This allows us to estimate the reward at a substantially faster rate: rather than decreasing as $\left(\sum_{s=1}^{t-1} \mathbb{1}\{\theta_s = \theta, c_s = c_t\}\right)^{-1/2}$, the estimation error now decays as $N_t(\theta)^{-1/2}$. In the next section, we show that this improvement allows us to avoid paying for the number of contexts, and that we instead pay only for the cross-context zooming dimension.

### 3.2 ZOOMING DIMENSIONS FOR PROBLEM CBCL

Regret in continuous bandits is driven by near-optimal actions, which are hard to rule out yet costly if missed. The *zooming dimension* (Kleinberg et al., 2008) captures this by measuring the metric complexity of the near-optimal action set. A natural question for Problem CBCL is whether cross-learning combined with a low zooming dimension can yield low regret — but as we will see, the answer is far from straightforward. A first attempt is to consider as complexity measure the worst zooming dimension over contexts, leading to the following notion of uniform zooming dimension.

**Definition 1** ($L$-Uniform zooming dimension)**.** *Let $\alpha > 0$ and $d \geq 0$. The sequence of contextual reward functions has $L$-uniform zooming dimension at most $d$ with multiplier $\alpha$ if, for all $r \in (0, 1]$ and all $c \in \mathcal{C}_T$, $N_{r/12}(\Theta_{c,Lr}) \leq \alpha r^{-d}$, where $\forall r > 0$, $\Theta_{c,r} = \left\{\, \theta \in \Theta : \Delta(\theta, c) \leq r \,\right\}$ is the set of near-optimal actions.*

On the one hand, applying the ZOOMING algorithm to each context independently, we obtain

$$\mathbb{E}[\mathcal{R}_T^{\text{ZOOMING}}] \leq C \left( |\mathcal{C}_T|^{\frac{1}{d'+2}} \alpha^{\frac{1}{d'+2}} T^{\frac{d'+1}{d'+2}} \right) \tag{1}$$

where $d'$ is the $L$-Uniform zooming dimension with multiplier $\alpha$. On the other hand, the worst-case regret of Problem CBCL cannot be smaller than that of the hardest sub-problem, and so we expect the regret to be at least $\alpha^{\frac{1}{d'+2}} T^{\frac{d'+1}{d'+2}}$, where $d'$ is the $L$-Uniform zooming dimension with multiplier $\alpha'$. Obviously, when $\mathcal{C}_T$ is large, these two bounds can be far apart.

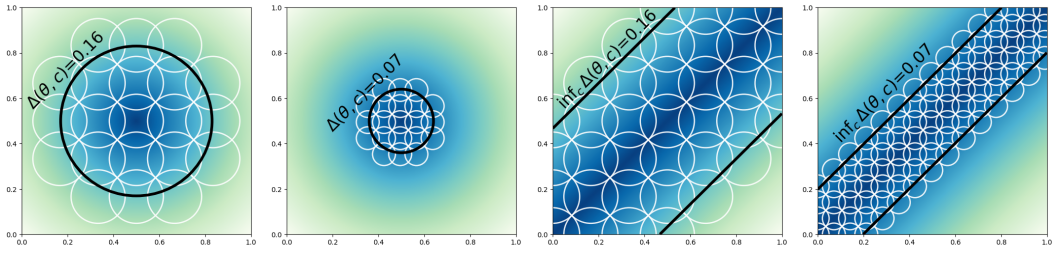
We show below that neither of these bounds is tight for cross-learning, thus disqualifying the uniform zooming dimension as a valid complexity measure for our problem. First, theorem 1 show that algorithm 1 can achieve lower regret than the independent application of ZOOMING. Second, theorem 2 shows that there exist instances in which each context has small uniform zooming dimension, yet the overall regret remains large. Indeed, to estimate $m(\theta, \cdot)$ faster, an algorithm must select $\theta$ often in more than one context, which only happens when $\theta$ has high reward. So the same action needs to be near-optimal across contexts for cross-learning to help lower the regret. The relevant complexity measure must therefore jointly capture the volume of near-optimal actions across all contexts — leading to the notion of cross-context zooming dimension.

**Definition 2** ($L$-cross-context zooming dimension)**.** *Let $\alpha > 0$ and $d \geq 0$. The sequence of contextual reward functions has $L$-cross-context zooming dimension at most $d$ with multiplier $\alpha$ if, $\forall r \in (0, 1]$, $N_{r/12}\left(\Theta^{\cup}_{Lr}\right) \leq \alpha r^{-d}$ where $\Theta^{\cup}_r = \left\{\theta \in \Theta : \min_{c \in \mathcal{C}_T} \Delta(\theta, c) \leq r\right\}$ is the set of near-optimal actions in at least one context.*

The two notions of complexity are related, as $\Theta^{\cup}_r = \cup_{c \in \mathcal{T}} \Theta_{c,r}$: if the problem has $L$-uniform zooming dimension $d'$ with multiplier $\alpha'$, $d'$ is an admissible choice of $L$-cross-context zooming dimension with multiplier $|\mathcal{C}_T|\alpha'$. When $|\mathcal{C}_T|\alpha'$ is small, both notions coincide; however the bound can be loose when different contexts share similar near-optimal actions, or when $|\mathcal{C}_T|\alpha'$ is large. Indeed, the packing number of $\Theta = [0, 1]^D$ at scale $r$ is at most order $r^{-D}$, so $d$ is also always at most the ambient dimension $D$, with $\alpha$ a constant depending only on $D$ (not $\mathcal{C}_T$). We conclude this discussion with examples of small cross-context zooming dimension.

**Clustered contexts.** Assume the contexts can be partitioned into clusters $\mathcal{K}$: there exists a mapping $\kappa : \mathcal{C} \to \mathcal{K}$ such that $m(\theta, c) = m'(\theta, \kappa(c))$, and for every $k \in \mathcal{K}$, $m'(\cdot, k)$ has zooming dimension $d$ with multiplier $\alpha$. Then the cross-contextual zooming dimension is at most $d$, with multiplier $|\mathcal{K}|\alpha$.

**Low-dimensional manifold.** Assume $\mathcal{C} = [0, 1]$ and $m(\theta, c) = 1 - \|\theta - f(c)\|$ for some unknown $f : [0, 1] \to [0, 1]^D$. The optimal action for context $c$ is $f(c)$, so the set of $r$-suboptimal actions lies in a $r$-neighborhood of the one-dimensional manifold $\{f(c) : c \in [0, 1]\}$. Under suitable regularity on $f$, covering this set requires $r^{-1}$ balls of radius $r$, independent of the ambient dimension $D$. Hence the cross-contextual zooming dimension is at most 1 (see example in Figure 1).




(a) (b) (c) (d)

Figure 1: **Zooming dimensions.** Let $m(\theta, c) = \frac{1}{2}(1 - \|\theta - c\mathbf{u}\|)$ for $\mathbf{u} = \binom{1}{1}$ and $c \in [0, 1]$. The *ambient dimension* is 2. **Figures 1a and 1b**: Heat-map of $\Delta(\cdot, c)$ for $c = 0.5$, with near-optimal set $\Theta_{r,c}$ for $r = 0.16$ (1a) and $r = 0.07$ (1b), and its covering by balls of radius $\propto r$. The *uniform zooming dimension* is 0: for all $r > 0$ and $c \in [0, 1]$, $N_{r/12}(\Theta_{c,r})$ is constant. **Figures 1c and 1d**: Heat-map of $\inf_{c \in [0,1]} \Delta(\cdot, c)$, with near-optimal set $\Theta^{\cup}_r$ for $r = 0.16$ (1c) and $r = 0.07$ (1d), and its covering by balls of radius $\propto r$. The *cross-context zooming dimension* is 1: for all $r > 0$, $N_{r/12}(\Theta^{\cup}_r) \propto r^{-1}$.

### 3.3 UPPER BOUND ON THE REGRET OF CL-ZOOMING

We now state our bound on the regret of Algorithm CL-ZOOMING, proved in Appendix A.

**Theorem 1** (Regret upper bound for CBCL)**.** *Consider an instance of Problem CBCL verifying Assumption 3, and with L-cross-context zooming dimension $(d, \alpha)$. Run Algorithm 1 with grid $\mathcal{G} = \{0, \frac{1}{G}, ..., \frac{G}{G}\}^D$ for $G = \lceil \sqrt{DT} \rceil$. Then, for some absolute constant $C$,*

$$\mathbb{E}[\mathcal{R}_T^{CBCL}] \leq C \left( (\alpha \log T)^{1/(d+2)} L^{\frac{d}{d+2}} T^{(d+1)/(d+2)} \right) .$$

Theorem 1 shows that CL-ZOOMING always outperforms the naive baseline equation 1 because of the relation between the zooming dimensions. On the one hand, when different contexts share similar optimal arms, and $\alpha$ is much smaller than $\alpha'|\mathcal{C}_T|$, CL-ZOOMING achieves even lower regret. On the other hand, since the cross-context zooming dimension is at most $D$ with multiplier independent of $|\mathcal{C}_T|$, the regret of CL-ZOOMING is always $\tilde{O}(T^{\frac{D+1}{D+2}})$. Thus, when $|\mathcal{C}_T|$ is large, the baseline may suffer large — even linear — regret, while CL-ZOOMING efficiently exploits the cross-learning structure and is never worse than a non-contextual Lipschitz bandit algorithm in dimension $D$.

*Sketch of proof.* Classical concentration arguments, combined with Lipschitz-continuity, show that the estimation error for any action covered by $\theta$ at round $t$ is at most $r_t(\theta)$ w.h.p. Unlike in classical Lipschitz bandits, this implies that $\theta$ can be selected at round $t$ only if $\Delta(\theta, c_t) \lesssim r_t(\theta) \approx \sqrt{\log(T)/N_t(\theta)}$, where $\Delta(\theta, c_t)$ is the sub-optimality gap of $\theta$ *in context* $c_t$. In other words, as an action is selected more often, it can only keep being chosen in contexts where its gap is small enough — so the contexts with the smallest gap contribute most to the regret. We show in Appendix A that, for any $\theta$, the cumulative regret from rounds where its gap exceeds $2^{-i}$ is at most of order $\sqrt{\log(T)}\, 2^i$.

To bound the number of actions for which the sub-optimality gap is of order $2^{-i}$ in some context, note that if $2^{-i} \approx \Delta(\theta, c_t) \lesssim r_t(\theta)$ and another action $\theta'$ is activated at round $t$, then necessarily $\|\theta - \theta'\| \geq 2^{-i}$: any two such actions must be separated by at least $2^{-i}$. By definition of the cross-contextual zooming dimension, there are at most $2^{id}$ such actions. Summing their contributions up to scale $2^{-i} \approx T^{-\frac{1}{d+2}}$ gives cumulative regret of order $T^{\frac{d+1}{d+2}}$. The remaining regret, from rounds where the gap is below $2^{-i} \approx T^{-\frac{1}{d+2}}$, is classically of order $2^{-i}T \approx T^{\frac{d+1}{d+2}}$. □

### 3.4 LOWER BOUND FOR CROSS-CONTEXT LIPSCHITZ BANDITS

The following theorem (proved in Appendix B) provides a lower bound on the worst-case regret of Problem CBCL in terms of its cross-context zooming dimension.

**Theorem 2** (Lower bound)**.** *For all $d \geq 1$, there exist $\beta_d, T_d > 0$ (depending on d) such that for every $T \geq T_d$, there are a deterministic context sequence $(c_t)_{t=1}^T$ and a 1-Lipschitz reward function, with 1-uniform zooming dimension 0 and 1-cross-context zooming dimension d (multipliers depending only on d), for which*

$$\mathbb{E}[\mathcal{R}_T^{CBCL}] \geq \beta_d\, T^{(d+1)/(d+2)}.$$

Theorem 2 shows that CL-ZOOMING is worst-case-optimal up to logarithmic factors, and that the cross-context zooming dimension captures the problem's intrinsic difficulty: even when each individual context has zooming dimension 0, the resulting cross-context problem can be much harder.

*Sketch of proof.* To construct hard instances, assume each context $c \in \mathcal{C}_T := \{1, \ldots, C\}$ appears for $T/C$ consecutive rounds, with $C = T^{\frac{d}{d+2}}$. Set $\epsilon = \sqrt{C/T} = T^{-\frac{1}{d+2}}$ and partition $[0,1]^d$ into $C$ hypercubes of length $\epsilon$. For each $z \in \{-1, 1\}^C$, define a reward function $m^z$ as follows: in context $c$, the optimal action lies within a bump of size $\epsilon$ on either the left or right half of the $c$-th hypercube (depending on $z_c$), with the other half sub-optimal by $\epsilon$. Outside this hypercube, the reward drops sharply — ensuring uniform zooming dimension 0 — and does not depend on $z_c$. Locating the optimal action for context $c$ thus requires sampling within the $c$-th hypercube, which is costly in any other context. Observations from other contexts are therefore uninformative about $z_c$, making cross-context learning ineffective. The bump size $\epsilon = \sqrt{C/T}$ is chosen so that, without prior information about $z_c$, the learner incurs the standard two-arm bandit regret of order $\sqrt{T/C} = T^{\frac{1}{d+2}}$ over the $T/C$ rounds in context $c$. Summing over the $C$ contexts yields the desired results. □

**Conclusion** Previous work on cross-learning highlighted a striking phenomenon(Balseiro et al., 2019; Schneider and Zimmert, 2023): worst-case regret is independent of the number of contexts, so contextual bandits with cross-learning are, in the worst case, no harder than classical multi-armed bandits. A natural question is whether this holds for Lipschitz bandits too — do worst-case rates match those of the classical Lipschitz bandit? The answer, as this section shows, is *yes and no*. Yes: without assumptions on the zooming dimension, both worst-case rates are of order $T^{\frac{D+1}{D+2}}$, independent of the number of contexts — so cross-learning costs nothing in the worst case. No: the zooming dimension of each individual context may be small while the cross-context zooming dimension is large, in which case the optimal regret exceeds that of the non-contextual problem. This reveals a phenomenon new to cross-learning: with continuous action spaces, the difficulty does depend on the set of contexts, through the metric complexity of the near-optimal actions they share.

# 4 Contextual bandit with performative feedback

We now return to Problem CBPF and propose two solutions: in Section 4.1, we establish regret upper bounds for CL-Zooming; then in Section 4.2, we design an improved algorithm specifically tailored to leverage the structure of performative prediction. Finally, Section 4.3 illustrates our method on a dynamic pricing problem with costs.

## 4.1 CL-Zooming with performative feedback

Under Assumptions 1 and 2, Problem CBPF is an instance of Problem CBCL with payoff $M_t(c_t) = 1 - \ell_{c_t}(z_t, \theta_t)$ and observation process $\{M_t(c)\}_{c\in\mathcal{C}} = \{1 - \ell_c(z_t, \theta_t)\}_{c\in\mathcal{C}}$. Moreover, the reward function $m(\theta, c) = 1 - \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell_c(z, \theta)]$ verifies Assumption 3 with Lipschitz constant $L_\theta + \epsilon L_z$.

**Corollary 1.** *Under Assumptions 1 and 2, Algorithm 1 applied to Problem CBPF with grid* $\mathcal{G} = \{0, \frac{1}{G}, ..., \frac{G}{G}\}^D$ *for* $G = (DT)^{-1/2}$ *has regret*

$$\mathbb{E}[\mathcal{R}_T^{CBPF}] \leq C(\alpha \log(T))^{\frac{1}{d+2}} (L_\theta + \epsilon L_z)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}$$

*where* $(\alpha, d)$ *is the* $(L_\theta + \epsilon L_z)$*-cross contextual zooming dimension and* $C$ *is an absolute constant.*

The regret upper bound stated in Corollary 1 treats the performative risk $\mathrm{PR}_c$ as a black-box Lipschitz function. In particular, it pays for the full Lipschitz constant $L_\theta + \epsilon L_z$. However, when the distribution shift $\epsilon$ is small enough, or when $L_\theta$ is large, improved regret bounds can be achieved by exploiting the structure of the performative risk minimization problem, as is shown in the following section.

## 4.2 Improved algorithm for performative prediction

To exploit the performative feedback structure, we build on Section 2: under Assumption 1, sufficiently close actions $\theta, \theta'$ induce a distribution shift between $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$ small enough to be negligible relative to statistical error. This motivates recursively partitioning the action space into bins within which the decoupled performative risk $\mathcal{DPR}_c(\theta, \theta')$ is a good estimator of the performative risk $\mathcal{PR}_c(\theta')$ — in effect, treating the distribution as constant within each bin. Once a bin has been sampled enough that the statistical error falls below this approximation error, the partition must be refined further, and children bin be activated in turn.

**Covering of $\Theta$ by recursive partitions** We cover $\Theta = [0, 1]^D$ using a recursive partition. At depth $i$, $\Theta$ is covered by regular bins $B_{i,z} = \bigtimes_{k\leq D:} \left[\frac{z_k - 1}{2^i \lceil\sqrt{D}\rceil}, \frac{z_k}{2^i \lceil\sqrt{D}\rceil}\right)$ with $z \in \left\{1, \ldots, 2^i \left\lceil\sqrt{D}\right\rceil\right\}^D$. For $i \geq 1$, we say that $B_{i,z}$ is a child of $B_{i-1,z'}$ if $B_{i,z} \subset B_{i-1,z'}$. We denote by $\mathrm{Pa}(B_{i,z})$ its parent (i.e. the bin $B_{i-1,z'}$ such that $B_{i-1,z} \subset B_{i',z'}$). We denote by $\mathcal{B}_t$ the set of bins active at time $t$, by $\tau(B_{i,z})$ (resp. $\tau'(B_{i,z})$) the round at which bin $B_{i,z}$ is activated (resp. deactivated), and by $N_t(B_{i,z})$ the number of actions sampled in this bin between time $\tau(B_{i,z})$ and $(t \wedge \tau'(B_{i,z})) - 1$ (included). The estimation and approximation confidence radiuses associated with $B_{i,z}$ are respectively given by

$$r_t(B_{i,z}) = \sqrt{\frac{4(2 + D)\log(T)}{N_t(B_{i,z})}} \wedge 1 \quad \text{and} \quad r(B_{i,z}) = r_i := 2^{-i}.$$

Once $r_t(B_{i,\theta}) \leq \epsilon L_z r(B_{i,z})$, $B_{i,z}$ is deactivated, and its children are activated.

**Algorithm 2** ZOOMING FOR CONTEXTUAL PERFORMATIVE PREDICTION (ZOOC)

**Input:** Horizon $T$, confidence radius $r_t$, Lipschitz constant $\epsilon L_z$, grid $\mathcal{G}$

**Init:** $\mathcal{B}_1 \leftarrow \left\{ \bigtimes_{k \le D} \left[ \frac{z_k - 1}{\lceil \sqrt{D} \rceil}, \frac{z_k}{\lceil \sqrt{D} \rceil} \right) \ : \ z \in \left\{ 1, ..., \left\lceil \sqrt{D} \right\rceil \right\}^D \right\}$.

**for** $t = 1, \dots, T$ **do**
  Observe context $c_t$
  Select $\theta_t \in \arg\min_{\theta \in \mathcal{G}} \widehat{\mathcal{PR}}_t^{\text{lcb}}(\theta)$ ▷ Estimate $\mathcal{PR}(\theta)$ using Equation 2
  Observe $Z_t \sim \mathcal{D}(\theta_t)$
  **if** $r_{t+1}(B_t(\theta_t)) \le \epsilon L_z r(B_t(\theta_t))$ **then**
    $\tau'(B_t(\theta_t)) = t + 1, \ \mathcal{B}_{t+1} \leftarrow \mathcal{B}_t \setminus \{B_t(\theta_t)\}$ ▷ Deactivate $B_t(\theta_t)$.
    Let $\{B_{i,z^{(k)}}\}_{k \le 2^D}$ be the children of $B_t(\theta_t)$.
    Set $\mathcal{B}_{t+1} \leftarrow \mathcal{B}_t \cup \{B_{i,z^{(k)}}\}_{k \le 2^D}$, and $\tau(B_{i,z^{(k)}}) = t + 1$ ▷ Activate new bins.
  **else**
    Set $\mathcal{B}_{t+1} \leftarrow \mathcal{B}_t$
  **end if**
**end for**

**Performative risk estimation** We estimate the decoupled performative risk of action $\theta \in \mathcal{G}$ as:

$$\widehat{\mathcal{DPR}}_{c,t}(B_{i,z}, \theta) = \frac{1}{N_t(B_{i,z})} \sum_{\tau(B_{i,z}) \le s < t < \tau'(B_{i,z}):\ \theta_s \in B_{i,z}} \ell_c(Z_s, \theta).$$

Then, we obtain a lower bound on the performative risk as[1] :

$$\widehat{\mathcal{PR}}_t^{\text{lcb}}(\theta) := \begin{cases} \widehat{\mathcal{DPR}}_{c_t,t}(B_t(\theta), \theta) - 2r_t(B_t(\theta)) & \text{if} \quad r_t(B_t(\theta)) \le 2\epsilon L_z r(\text{Pa}(B_t(\theta)) \\ \widehat{\mathcal{DPR}}_{c_t,t}(\text{Pa}(B_t(\theta)), \theta) - 2\epsilon L_z r(\text{Pa}(B_t(\theta)) & \text{else.} \end{cases} \tag{2}$$

**Upper bound on the regret of ZOOC** Before stating our bound, we introduce the cross-context performative zooming dimension, which captures the difficulty of Problem CBPF without relying on $(L_\theta + \epsilon L_z)$-Lipschitz-continuity of the loss — since $L_\theta$ may be large, this constant does not properly reflect the problem's difficulty.

**Definition 3** (Cross-context performative zooming dimension)**.** *A performative prediction problem instance has a $L$-cross-context performative zooming dimension equal to $d$ with multiplier $\alpha$ if for all $i \in \mathbb{N}$, $\cup_{t \le T} \Theta_{c_t, 10\sqrt{2} L r_i}$ is covered by at most $\alpha / r_i^d$ bins $B_{i,z}$ at depth $i$.*

Like the cross-contextual zooming dimension, cross-context performative zooming dimension is bounded by $D$, and captures the same source of difficulty: it is small when few actions are good at least one context, large when near-optimal actions vary widely across contexts. The key difference is that performative risk can vary much faster than Lipschitz-continuous reward functions, driven by the potentially large constant $L_\theta$, so the volume near a good action can be large even when the near-optimal set is intrinsically small. To capture the true difficulty, we count the number of bins in the $2^D$-adic partition containing at least one near-optimal action.

**Theorem 3.** *Assume that the problem has $\epsilon L_z$-cross-context performative zooming dimension equal to $d$ with multiplier $\alpha$, and that Assumption 1 and Assumption 2 hold. There exists a universal constant $C$ such that for $T$ large enough, Algorithm 2 has a regret bounded as*

$$\mathbb{E}[\mathcal{R}_T^{CBPF}] \le C \left( (4 + 2D) 2^D \right)^{\frac{1}{d+2}} \left( \alpha \log(T) \right)^{\frac{1}{d+2}} \left( \epsilon L_z \right)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}.$$

Theorem 3 shows that ZOOC exploits the structure of cross-context performative prediction to achieve lower regret. For $T$ large enough, the bound depends only on $\epsilon L_z$, which can be much smaller than the full constant $\epsilon L_z + L_\theta$ when $L_\theta$ is large or $\epsilon$ is small.

*Sketch of proof.* The algorithm treats the distribution $\mathcal{D}(\theta)$ as constant within a bin, and thus incurs an approximation error $\epsilon L_z r_i$. But we show that as long as this term is dominated by the estimation

[1] If $B_t(\theta) = B_{0,z}$ for some $z$, we set $\widehat{\mathcal{PR}}_t^{\text{lcb}}(\theta) := \widehat{\mathcal{DPR}}_{c_t,t}(B_t(\theta), \theta) - 2r_t(B_t(\theta))$

error on the reward $r_{t+1}(B_{z,i})$, the regret incurred by any selected action $\theta \in B_{z,i}$ is bounded w.h.p. by $\Delta(\theta, c_t) \lesssim r_{t+1}(B_{z,i}) \wedge \epsilon L_z r(\text{Pa}(B_{z,i}))$.

First, we notice that selecting an action $\theta \in B_{z,i}$ in a active bin $B_{z,i}$ implies that that $\Delta(\theta, c_t) \leq r_{t+1}(B_{z,i}) \approx \sqrt{\log(T)}/\sqrt{N_t(B_{i,z})}$. So a bin $B_{z,i}$ is only selected at most $O(\log(T)(\epsilon L_z r_i)^{-2})$ times, contributing $O(\log(T)(\epsilon L_z r_i)^{-1})$ to the regret. Now, for bin $B_{z,i}$ to be activated, its parent $\text{Pa}(B_{z,i})$ must have been selected until deactivation. This implies that $\text{Pa}(B_{z,i})$ contained an action that was, for some context, sub-optimal by a gap $O(\epsilon L_z r_{i-1})$. By the definition of the cross-context performative zooming dimension, there are approximately $r_i^{-d}$ such bins, leading to a total contribution of order $\log(T)(\epsilon L_z)^{-1} r_i^{-(d+1)}$ to the regret. Summing over all levels $i < i_0$, the total regret contribution of bins at these levels is of order $\log(T)(\epsilon L_z)^{-1} r_{i_0}^{-(d+1)}$. Second, selecting action $\theta \in B_{z,i_0}$ after the bin has been deactivated implies that $\Delta(\theta, c_t) \leq \epsilon L_z r_{i_0}$, since $r_{i_0} \geq r(\text{Pa}(B_t(\theta)))$. Thus, the actions selected in bins at depth larger than $i_0$ contribute at most $T \times \epsilon L_z r_{i_0}$ to the regret. Choosing $i_0$ such that $r_{i_0} \approx (\epsilon L_z)^{\frac{-2}{d+2}} \log(T)^{\frac{1}{d+2}} T^{\frac{-1}{d+2}}$, we obtain a regret bound of order $(\epsilon L_z)^{\frac{d}{d+2}} \log(T)^{\frac{1}{d+2}} T^{\frac{d+1}{d+2}}$. □

### 4.3 Application to dynamic pricing with varying costs

Finally, we illustrate our algorithm and results on dynamic pricing under varying production costs.

**Example 1** (Dynamic pricing with varying costs)**.** *At round $t \leq T$, a buyer arrives with a hidden valuation $v_t$ drawn from a distribution $\nu$. The seller posts a price $p_t$ at which they offer to manufacture the good. The trade takes place if $p_t \leq v_t$, in which case the seller incurs a production cost $c_t$, which we assume is known to the seller beforehand. The seller's reward is given by $(p_t - c_t)\mathbb{1}\{p_t \leq v_t\}$.*

The cost $c_t$ affects the seller's profitability, but not the buyer's willingness to pay, and information gathered in previous rounds remains useful for estimating $\nu$, even if the optimal price changes over time. This problem fits the framework of Problem CBPF: the model deployed is $\theta_t = p_t \in [0, 1]$, the binary observation (whether the sale happens or not) is $Z_t = \mathbb{1}\{p_t \leq v_t\} \sim \text{Bernoulli}(\nu[\theta_t, \infty))$, which depends on the price posted, and the context-dependent loss is $\ell_c(z, \theta) = 1 - (\theta - c)z \in [0, 1]$.

Assume that costs and valuations are bounded (standard in dynamic pricing), and w.l.o.g. that $c_t, v_t \in [0, 1]$. Then, 2 hold: since $(\theta - c) \in [0, 1]$, with we may take $L_z = 1$, and since $z \in \{0, 1\}$, we have $L_\theta = 1$. If the p.d.f. of $\nu$ is bounded by $\nu_{\max}$ – an assumption standard in dynamic pricing Fan et al. (2024); Luo et al. (2024); Tullii et al. (2024); Luo et al. (2022) – Assumption 1 holds with $\epsilon = \nu_{\max}$, as $\mathcal{W}(\text{Bernoulli}(\nu([\theta, \infty))), \text{Bernoulli}(\nu([\theta', \infty)))) \leq \nu_{\max} \|\theta - \theta'\|$.

Then, theorem 3 implies that using Algorithm ZOOC on this , when applied to the dynamic pricing problem with varying costs, incurs regret at most $C(\nu_{\max})^{1/3} \log(T)^{1/3} T^{2/3}$. Note that, up to a numerical constant, this is the same rate as that obtained by a Lipschitz-continuous bandit algorithm on the dynamic pricing problem with the same valuation distribution and price upper bound, for which the Lipschitz constant is $\nu_{\max}$. This shows that, perhaps surprisingly, the additional difficulty introduced by the varying costs translates, at worst, into a larger numerical constant in the regret bound, without affecting its dependence on $T$ or on the Lipschitz constant.

## 5 Conclusion

This paper draws connections between performative prediction and Lipschitz bandits with cross-learning. Along the way we make contributions to both fields. We define the cross-context zooming dimension that quantifies the hardness of a problem in terms of the structure of the losses across contexts. We propose a first formal treatment of a common scenario where some parameters of the loss function may change over time due to external factors. Our findings suggest that dedicated algorithms, leveraging information across deployment cycles, should be developed for practical applications in this context.

## AI USE STATEMENT

We used generative AI tools to improve the grammar and presentation of portions of the manuscript, and to proofread spelling and typos. We have not used AI tools for any other task, including those in the required-disclosure list (e.g., developing theoretical models, formulating or proving mathematical claims, designing methodology, or interpreting results). We have reviewed all AI-assisted text for accuracy. We take full responsibility for the final content of this work, including any text edited with the assistance of generative AI.

## REFERENCES


Rajeev Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, and Jon Schneider. Contextual bandits with cross-learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Yatong Chen, Wei Tang, Chien-Ju Ho, and Yang Liu. Performative prediction with bandit feedback: Learning through reparameterization. *arXiv preprint arXiv:2305.01094*, 2023.

Edwige Cyffers, Muni S Pydi, Jamal Atif, and Oliver Cappé. Optimal classification under performative distribution shift. *Advances in Neural Information Processing Systems*, 37:68144–68160, 2024.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

Jianqing Fan, Yongyi Guo, and Mengxin Yu. Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association*, 119(545):552–564, 2024. doi: 10.1080/01621459.2022.2128359.

Emile Grunberg and Franco Modigliani. The predictability of social events. *Journal of Political Economy*, 62(6):465–478, 1954. ISSN 00223808, 1537534X.

Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8446–8464. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hao22b.html.

Moritz Hardt. Retraining seeks stable signals. *arXiv preprint arXiv:2607.15623*, 2026.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4641–4650. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/izzo21a.html.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pages 9760–9785. PMLR, 2022.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

Jaewook Lee and Tijana Zrnic. Partially performative prediction. *arXiv preprint arXiv:2606.07890*, 2026.

Licong Lin and Tijana Zrnic. Plug-in performative optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30546–30565. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lin24ab.html.

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Contextual dynamic pricing with unknown noise: Explore-then-ucb strategy and improved regrets. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37445–37457. Curran Associates, Inc., 2022. doi: 10.52202/068431-2714. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f38d1fd25c15a0ad9ba758de4e7b1819-Paper-Conference.pdf.

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Distribution-free contextual dynamic pricing. *Mathematics of Operations Research*, 49(1):599–618, 2024. doi: 10.1287/moor.2023.1369. URL https://doi.org/10.1287/moor.2023.1369.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. In *Advances in Neural Information Processing Systems*, 2022.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

Oskar Morgenstern. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Springer, 1928.

Nicolas Nguyen, Solenne Gaucher, and Claire Vernade. Non-stationary lipschitz bandits. *Advances in Neural Information Processing Systems*, 38:15129–15180, 2026.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Jon Schneider and Julian Zimmert. Optimal cross-learning for contextual bandits with unknown context distributions. *Advances in Neural Information Processing Systems*, 36:51862–51880, 2023.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

A. Tsybakov. Introduction to nonparametric estimation. In *Springer Series in Statistics*, 2008. URL https://api.semanticscholar.org/CorpusID:42933599.

Matilde Tullii, Solenne Gaucher, Nadav Merlis, and Vianney Perchet. Improved algorithms for contextual dynamic pricing. *Advances in Neural Information Processing Systems*, 37:126088–126117, 2024.

Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997. ISBN 978-1-4612-1880-7. doi: 10.1007/978-1-4612-1880-7_29. URL https://doi.org/10.1007/978-1-4612-1880-7_29.

# A PROOF OF REGRET UPPER BOUND OF CL−ZOOMING (THEOREM 1)

For ease of notation, we henceforth denote $R_T = \mathcal{R}_T^{\text{CBCL}}$. For each active action $\theta \in \mathcal{A}_t$, define the empirical mean in context $c$ by

$$\widehat{m}_t(\theta, c) = \frac{1}{N_t(\theta)} \sum_{s=1}^{t-1} \mathbb{1}\{\theta_s = \theta\} M_s(c).$$

when $N_t(\theta) > 0$, and $\widehat{m}_t(\theta, c) = 0$ otherwise. Recall that for $\beta = 2(3 + D)\log(T)$,

$$r_t(\theta) = \min\left\{1, \sqrt{\frac{\beta}{N_t(\theta) \vee 1}}\right\}.$$

Our first step is to lower bound the probability of the following good event

$$\mathcal{E} := \{\forall t \in [T],\ \forall \theta \in \mathcal{G},\ \forall c \in \mathcal{C}_T : |\widehat{m}_t(\theta, c) - m(\theta, c)| \le r_t(\theta)\}.$$

**Lemma 1.** *For $\beta = 2(3 + D)\log(T)$,*

$$\mathbb{P}(\mathcal{E}) \ge 1 - 2T^{-1}.$$

In the next Lemma, we leverage the Lipschitz-continuity of the reward function to show that, on the good event $\mathcal{E}$, the UCB index of some active action dominates the optimal value in the current context.

**Lemma 2.** *On the event $\mathcal{E}$, for every round $t \in [T]$,*

$$m(\theta_t^*, c_t) \le \max_{\theta \in \mathcal{A}_t} \{\widehat{m}_t(\theta, c_t) + 2r_t(\theta)\}\ .$$

Next, we leverage this result to show that if no new action is activated at round $t$, the instantaneous regret is bounded by a constant time the confidence radius.

**Lemma 3.** *On the event $\mathcal{E}$, for every round $t$,*

$$\Delta(\theta_t, c_t) \le 3r_t(\theta_t).$$

A direct consequence is that if $\Delta(\theta_t, c_t) > 0$, then $r_t(\theta_t) \ge \frac{1}{3}\Delta(\theta_t, c_t)$. Since $r_t(\theta_t) = \sqrt{\frac{\beta}{N_t(\theta_t)\vee 1}}$, this implies that on the good event $\mathcal{E}$,

$$N_t(\theta_t) \le \frac{9\beta}{\Delta(\theta_t, c_t)^2}.$$

Note that by contrast to usual bounds on the number of pulls of a sub-optimal action, here $N_t(\theta_t)$ depends implicitly on the context $c_t$ through the gap $\Delta(\theta_t, c_t)$. In other words, although an action may be discarded due to it being sub-optimal for some contexts, in other contexts it can continue to be selected. To bound the regret of the CL-ZOOMING algorithm, we must therefore take into account not only the number of times an action is selected, but also the contexts in which it is selected. Moreover, we also must bound the number of active actions corresponding to a regret at a given scale *for a given context*.

We begin by extending the packing argument of Kleinberg et al. (2008) to our cross-context learning problem. Recall that

$$\Theta_{c,r} = \left\{\, \theta \in \Theta : \Delta(\theta, c) \le r \,\right\}.$$

Let $j_0 = \left\lfloor \log_2\left((T/64\alpha\beta L^d)^{\frac{1}{d+2}}\right)\right\rfloor$, and assume that $T$ is large enough so that $j_0 \le \log_2\left(\frac{0.8\sqrt{T}}{6L}\right) - 1$. For $j \le j_0$, define $r_j = 2^{-j}$ and $\Theta^j$ as the set of actions active at the end of some round $t$ that are sub-optimal by a factor less than $r_j$ in the context $c_t$

$$\Theta^j = \left\{\theta \in \Theta :\ \exists t \in [T],\ \theta \in \mathcal{A}_{t+1} \cap \Theta_{c_t, r_j}\right\}.$$

Finally, we define $\tilde{\Theta}^{j_0} = \Theta^{j_0}$, and for $j < j_0$, $\tilde{\Theta}^j = \Theta^j \setminus \{\Theta^j \cap \bigcup_{j'>j} \Theta^{j'}\}$. In words, if $\theta \in \tilde{\Theta}^j$, then $j$ is the largest index such that $\theta \in \Theta_j$, so for any $t$ such that $\theta_t \in \mathcal{A}_{t+1}$, we have $\Delta(\theta, c_t) > r_{j+1}$. The following result bound the number of actions in $\tilde{\Theta}_j$ for $j \ge 0$.

**Lemma 4.** *If the cross-context zooming condition holds with parameters $(d, \alpha)$, then for $j \geq 0$,*

$$|\tilde{\Theta}_j| \leq \alpha L^d r_j^{-d}.$$

We are now ready to prove Theorem 1. Denoting $\mathbb{E}_{\mathcal{E}}$ the expectation conditionally on the event $\mathcal{E}$, and $\bar{\mathcal{E}}$ the complementary event, we have

$$\begin{aligned} R_T &\leq \mathbb{E}_{\mathcal{E}}\left[\sum_{t\leq T}\Delta(\theta_t, c_t)\right] + \mathbb{P}(\bar{\mathcal{E}})\mathbb{E}_{\bar{\mathcal{E}}}\left[\sum_{t\leq T}\Delta(\theta_t, c_t)\right] \\ &\leq \mathbb{E}_{\mathcal{E}}\left[\sum_{t\leq T}\Delta(\theta_t, c_t)\right] + 2 \end{aligned}$$

where we used Lemma 1 to bound $\mathbb{P}(\bar{\mathcal{E}})$. Then, we have

$$\begin{aligned} \sum_{t\leq T}\Delta(\theta_t, c_t) &\leq \sum_{t:\Delta(\theta_t,c_t)\leq r_{j_0}}\Delta(\theta_t, c_t) + \sum_{t:\Delta(\theta_t,c_t)>r_{j_0}}\Delta(\theta_t, c_t) \\ &\leq T r_{j_0} + \sum_{t:\Delta(\theta_t,c_t)>r_{j_0}}\Delta(\theta_t, c_t). \end{aligned}$$

Now, note that for action $\theta_t$ to be chosen at time $t$, it must be that $\theta \in \mathcal{A}_{t+1}$. Moreover, if $\Delta(\theta_t, c_t) > r_{j_0}$, there must exist $j \leq j_0$ such that $\theta \in \Theta_{c_t,j}$. Thus, there exists $j \geq 0$ such that $\theta \in \tilde{\Theta}_j$. Then, we have

$$\sum_{t:\Delta(\theta_t,c_t)>r_{j_0}}\Delta(\theta_t, c_t) \leq \sum_{j=0}^{j_0-1}\sum_{\theta\in\tilde{\Theta}_j}\sum_{t:\theta_t=\theta}\Delta(\theta, c_t).$$

We now bound $\sum_{t:\theta_t=\theta}\Delta(\theta, c_t)$ on the event $\mathcal{E}$ for $\theta \in \tilde{\Theta}_j$ and $j < j_0$. For $i \leq j$, let us denote

$$\mathcal{T}_i(\theta) = \left\{t \leq T : \theta_t = \theta \text{ and } \Delta(\theta, c_t) \in \left[\frac{r_i}{2}, r_i\right]\right\}$$

the set of rounds where the action $\theta$ is selected and is sub-optimal by a factor $r_i \geq r_j$. On the event $\mathcal{E}$, Lemma 2 implies that

$$\begin{aligned} \sum_{t:\theta_t=\theta}\Delta(\theta, c_t) &= \sum_{i\leq j}\sum_{t\in\mathcal{T}_i(\theta)}\Delta(\theta, c_t) \\ &\leq \sum_{i\leq j}|\mathcal{T}_i(\theta)|r_i \end{aligned}$$

By Lemma 3, on the good event $\mathcal{E}$, if $\theta$ is chosen at round $t$, it must be that $N_t(\theta_t) \leq \frac{9\beta}{\Delta(\theta_t,c_t)^2}$. Since for all $t \in \mathcal{T}_i(\theta)$, $\Delta(\theta, c_t) \geq r_i/2$, we have $|\mathcal{T}_i(\theta)| \leq \frac{36\beta}{r_i^2}$. Therefore,

$$\begin{aligned} \sum_{t:\theta_t=\theta}\Delta(\theta, c_t) &\leq \sum_{i\leq j}\frac{36\beta}{r_i} \\ &\leq 64\beta r_j^{-1} \end{aligned}$$

Then,

$$\sum_{t:\Delta(\theta_t,c_t)>r_{j_0}}\Delta(\theta_t, c_t) \leq 64\beta\sum_{j=0}^{j_0-1}|\tilde{\Theta}_j|r_j^{-1}.$$

Using Lemma 4, we obtain

$$\begin{aligned} \sum_{t:\Delta(\theta_t,c_t)>r_{j_0}}\Delta(\theta_t, c_t) &\leq 64\beta\sum_{j=0}^{j_0-1}\alpha L^d r_j^{-(d+1)} \\ &\leq 64\alpha\beta L^d r_{j_0}^{-(d+1)}. \end{aligned}$$

Thus,

$$R_T \leq 2 + Tr_{j_0} + 64\alpha\beta L^d r_{j_0}^{-(d+1)}.$$

Choosing $j_0 = \left\lfloor \log_2 \left( (T/64\alpha\beta L^d)^{\frac{1}{d+2}} \right) \right\rfloor$, we have $r_{j_0} \in \left[ (T/64\alpha\beta)^{\frac{-1}{d+2}}, 2\,(T/64\alpha\beta)^{\frac{-1}{d+2}} \right]$, and we obtain

$$R_T \leq C \left( \alpha \log(T) L^d \right)^{\frac{1}{d+2}} T^{\frac{d+1}{d+2}}$$

for some absolute constant $C$.

## B Proof for Regret Lower Bound of CBCL

For ease of notation, we henceforth denote $R_T = \mathcal{R}_T^{\text{CBCL}}$. We consider a reward function on $[0,1]^D$ for $D = d$. Then, as noted previously, the cross-context zooming dimension is at most the ambient dimension, with a dimension-dependent multiplier.

Let $C = T^{\frac{D}{D+2}}$, $\epsilon = T^{\frac{-1}{D+2}}$, and assume (w.l.o.g. up to a rounding argument) that $C \in \mathbb{N}$ and that $T/C \in \mathbb{N}$, and that $T$ and $C$ are large enough. Partition $\Theta = [0,1]^D$ into $C = \epsilon^{-D}$ hypercubes $(Q_c)_{c \leq C}$ of side length equal to $\epsilon$. For $i \leq C$, denote $x_c$ the center of hypercube $Q_c$. Let us also define $x_c^{-1} = x_c - \epsilon/4 e_1$ and $x_c^{+1} = x_c + \epsilon/4 e_1$, where $e_1$ is the first vector of the canonical basis in $\mathbb{R}^D$, and $Q_c^{+1} = Q_c \cap \{\theta : d_\infty(\theta, x_c^{+1}) < d_\infty(\theta, x_c^{-1})\}$ and $Q_c^{-1} = Q_c \cap \{\theta : d_\infty(\theta, x_c^{+1}) \geq d_\infty(\theta, x_c^{-1})\}$ respectively the right and left half of hypercube $Q_c$. We denote by $d_\infty(x, x') = \max_{i \leq d} |x_i - x_i'|$ the infinity norm, and by $d'_\infty(x, x') = \max\{2|x_1 - x_1'|, \max_{1<i\leq d} |x_i - x_i'|\}$ its scaled variant.

We assume the rewards to be Bernoulli random variables, and the contexts to be in $\mathcal{C}_T = \{1, ..., C\}$. More precisely, we assume that each context $c \in \{1, ..., C\}$ appears exactly $T/C$ times. For $z \in \{-1, 1\}^C$, we define the reward function $m^z$ in context $c \in \{1, ..., C\}$ as follows :

$$m^z(\theta, c) = \begin{cases} \frac{1}{2} - \frac{d_\infty(\theta, Q_c)}{4\sqrt{D}} & \textbf{if} \quad \theta \notin Q_c \\ \frac{1}{2} + \frac{\epsilon}{8\sqrt{D}} \left( 1 - \frac{2d'_\infty(\theta, x_c^{z_c})}{\epsilon} \right) & \textbf{if} \quad d_\infty(\theta, x_c^{z_c}) < d_\infty(\theta, x_c^{-z_c}) \\ \frac{1}{2} + \frac{\epsilon}{16\sqrt{D}} \left( 1 - \frac{2d'_\infty(\theta, x_c^{-z_c})}{\epsilon} \right) & \textbf{else.} \end{cases}$$

The following result shows that for all $z \in \mathcal{Z}$, the reward function $m^z$ has a 1-uniform zooming dimension 1 with multiplier $\alpha_D$ depending on the dimension.

**Lemma 5.** *For all $z \in \{-1, 1\}^C$, the reward function $m^z$ verifies Assumption 3 with $L = 1$. Moreover, there exists a constant $\alpha_D$ depending on $d$ such that this function has 1-uniform zooming dimension 0 with multiplier $\alpha_D$.*

In context $c$ reward function $m^z$ has two bumps in the hypercube $Q_c$, with the highest bump located in the left half of the hypercube if $z_c = -1$, and in the right half of the hypercube otherwise. Choosing an action in the wrong half of the hypercube leads to an instantaneous regret at least $\frac{\epsilon}{8\sqrt{D}}$. Moreover, choosing any action outside of this interval leads to an instantaneous regret at least $\frac{\epsilon}{4\sqrt{D}}$.

Standard techniques for deriving regret lower bounds in bandit problems (such as those based on Bretagnolle–Huber's inequality) do not directly apply to our setting, as the cross-learning structure induces additional dependencies among rewards observed in different contexts. To establish a lower bound on the regret of any strategy against the worst-case reward function $m^z$, we rely on the celebrated Fano's lemma.

**Proposition 1** (Fano's Lemma (as stated in Yu (1997) )**.** *Let $(\mathcal{Z}, \rho)$ be a metric space and $\{\mathbb{P}_z : z \in \mathcal{Z}\}$ be a collection of probability measures. For any totally bounded $Z \subset \mathcal{Z}$, define the Kullback–Leibler diameter by $d_{\text{KL}}(Z) = \sup_{z,z' \in Z} D(\mathbb{P}_z \| \mathbb{P}_{z'})$. Then, for all $\tau > 0$,*

$$\inf_{\hat{z}} \sup_{z \in \mathcal{Z}} \mathbb{P}_z \left\{ \rho(\hat{z}(X), z)^2 \geq \frac{\tau^2}{4} \right\} \geq 1 - \frac{d_{\text{KL}}(Z) + \log 2}{\log \mathcal{M}(\tau, Z, \rho)}$$

*where $\mathcal{M}(\tau, Z, \rho)$ is the $\tau$-packing number of the set $Z$ for the distance $\rho$.*

We apply Proposition 1 to the metric space $\mathcal{Z} = [-1,1]^C$ equipped with the scaled euclidean distance $\rho(z,z')^2 = \frac{1}{4}\sum_{c\le C}\|z_c - z'_c\|^2$. To build a totally bounded subset $Z$, we rely on the Varshamov-Gilbert Lemma.

**Proposition 2** (Varshamov-Gilbert bound (see,e.g., Lemma 2.9 in Tsybakov (2008))). *There exists a subset $Z$ of $\mathcal{Z} = \{-1,1\}^C$ of size $N \ge \exp(C/8)$ and such that $\forall z, z' \in Z$, $\sum_{c\le C}\mathbb{1}\{z_c \ne z'_c\} \ge \frac{C}{4}$.*

Note that for $z, z' \in \mathcal{Z}$, $\rho(z,z')^2 = \sum_{c\le C}\mathbb{1}\{z_c \ne z'_c\}$, so for $\tau = \sqrt{\frac{C}{4}}$, the $N$ elements of $Z$ are separated by a distance $\rho$ at least $\tau$. Thus, $\log \mathcal{M}(\tau, Z, \rho) \ge \log(N) \ge \frac{C}{8}$. We now turn to bounding the diameter of $Z$ in Kullback-Leibler divergence. For all $z, z' \in Z$, we denote by $\mathbb{P}_z$ the distribution when the mean reward function is $m^z$, and by $\mathrm{kl}(p,q)$ the Kullback-Leibler divergence between Bernoulli random variables with parameters $p$ and $q$. We have

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_z, \mathbb{P}_{z'}) &= \sum_{t\le T}\mathbb{E}_{\mathbb{P}_z}\left[\mathrm{kl}(m^z(\theta_t, c_t), m^{z'}(\theta_t, c_t))\right] \\ &\le \frac{T\epsilon^2}{32D} \end{aligned}$$

since $m^z(\theta, c) \in [1/4, 3/4]$ and $\max_{c\le C, \theta\in[0,1]^D, z, z'\in\mathcal{Z}} |m^z(\theta,c) - m^{z'}(\theta,c)| \le \frac{\epsilon}{16\sqrt{D}}$. Thus, Proposition 1 implies that

$$\inf_{\hat{z}}\sup_{z\in\mathcal{Z}}\mathbb{P}_z\left\{\rho^2(\hat{z}(X), z) \ge \frac{C}{16}\right\} \ge 1 - \frac{8(\frac{T\epsilon^2}{32D} + \log 2)}{C}.$$

For $T$ large enough, $\frac{T\epsilon^2}{64} \ge \log 2$. Since $T\epsilon^2 = C$, this implies that

$$\inf_{\hat{z}}\sup_{z\in\mathcal{Z}}\mathbb{P}_z\left\{\rho^2(\hat{z}(X), z) \ge \frac{C}{16}\right\} \ge \frac{1}{2}.$$

Now, we define the estimator $\hat{z}(X)$ has follows : for $c \le C$, we set $z_c = 1$ if $\sum_{t\le T: c_t = c}\mathbb{1}\{\theta_t \in Q_c^{+1}\} \ge \sum_{t\le T: c_t = c}\mathbb{1}\{\theta_t \in Q_c^{-1}\}$, and $z_c = -1$ otherwise. The lower bound $\rho^2(\hat{z}(X), z) \ge \frac{C}{16}$ implies that in at least $\frac{C}{16}$ contexts, the optimal half-hypercube $Q_c^{z_c}$ was selected less than half of the times in context $c$. Thus, denoting $N_Q^c$ the number of times an action was selected in the set $Q$ in context $c$,

$$\begin{aligned} R_T &\ge \mathbb{E}\left[\sum_{c\le C}(T/C - N^c_{Q_c^{z_c}})\frac{\epsilon}{16\sqrt{D}}\right] \\ &\ge \frac{1}{2}\times\frac{C}{16}\times\frac{T}{2C}\times\frac{\epsilon}{16\sqrt{D}} \\ &\ge \frac{T^{\frac{d+1}{d+2}}}{1024\sqrt{D}}. \end{aligned}$$

This concludes the proof of Theorem 2.

## C Proof of Regret Upper Bound of ZOOC

For ease of notation, we henceforth denote $R_T = \mathcal{R}_T^{\mathrm{CBPF}}$. In the following lemma, we define a control the probability of a good event where our estimators have small errors. In our statement, we adopt the convention $\frac{\cdot}{0} = +\infty$. Recall that $r_t(B_{i,z}) = r_t(B_{i,z}) = \sqrt{\frac{2(4+2D)\log(T)}{N_t(B_{i,z})}} \wedge 1$.

**Lemma 6.** *Let $\mathcal{E}$ be the following good event :*

$$\begin{aligned} \mathcal{E} := \Big\{ &\forall\, t\in[T],\ \forall\,\theta\in\mathcal{G},\ \forall\, c\in\mathcal{C}_T, \left|\widehat{\mathcal{DPR}}_{c,t}(B_t(\theta),\theta) - \mathcal{PR}_c(\theta)\right| \le 2r_t(B_t(\theta)) \\ &\textit{and } \left|\widehat{\mathcal{DPR}}_{c,t}(Pa(B_t(\theta)),\theta) - \mathcal{PR}_c(\theta)\right| \le 2\epsilon L_z r(Pa(B_t(\theta))) \Big\}. \end{aligned}$$

*The event $\mathcal{E}$ holds with probability at least $1 - 2/T$.*

Let us define, for $\theta \in \mathcal{G}$, $r_t(\theta) = \min_{B_{i,z} \in \mathrm{Pa}(B_t(\theta))} \{2\epsilon L_z r_i\} \wedge r_t(B_t(\theta))$.

**Lemma 7.** *On the event $\mathcal{E}$, for every round $t \in [T]$,*

$$\mathcal{PR}_{c_t}(\theta_t) - 5\min\{\epsilon L_z r(Pa(B_t(\theta_t))), r_t(B_t(\theta_t))\} \leq \mathcal{PR}_{c_t}(\theta_t^*).$$

Let

$$\Delta(\theta, c) = \mathcal{PR}_c(\theta) - \min_{\theta'} \mathcal{PR}_c(\theta').$$

On the event $\mathcal{E}$, for every round $t$, Lemma 7 implies that

$$\Delta(\theta, c) \leq 5\min\{\epsilon L_z r(\mathrm{Pa}(B_t(\theta_t))), r_t(B_t(\theta_t))\}.$$

Let us define the set $\mathcal{T}(B_{i,z}) = \{t : \tau(B_{i,z}) \leq t \leq \tau(B_{i,z}),\ \theta_t \in B_{i,z}\}$ of rounds where we played in bin $B_{i,z}$ when it is active, and define $R(B_{i,z})$ the regret for playing in bin $B_{i,z}$ when it is active:

$$R(B_{i,z}) = \sum_{t \in \mathcal{T}(B_{i,z})} \Delta(\theta_t, c).$$

Similarly, let us define the set $\mathcal{T}^{\mathrm{Ch}}(B_{i,z}) = \{t : \tau(B_{i,z}) \leq t,\ \theta_t \in B_{i,z}\}$ of rounds where we played in bin $B_{i,z}$ after it was actived, and define $R^{\mathrm{Ch}}(B_{i,z})$ the regret for playing in bin $B_{i,z}$ or in one of its children (or descendent):

$$R^{\mathrm{Ch}}(B_{i,z}) = \sum_{t \in \mathcal{T}^{\mathrm{Ch}}(B_{i,z})} \Delta(\theta_t, c).$$

**Lemma 8.** *On the event $\mathcal{E}$,*

$$R(B_{i,z}) \leq \frac{11(4+2D)\log(T)}{\epsilon L_z r_i}.$$

*Moreover,*

$$R^{Ch}(B_{i,z}) \leq \frac{11(4+2D)\log(T)}{\epsilon L_z r_i} + 5\epsilon L_z r_i |\mathcal{T}^{Ch}(B_{i,z})|.$$

To bound the regret of Algorithm ZOOC, it remains to bound the number of bins that are activated at some point during the game.

**Lemma 9.** *On the good event $\mathcal{E}$, for $i > 0$,*

$$\sum_z \mathbb{1}\{|\mathcal{T}(B_{i,z})| > 0\} \leq 2^D \alpha 2^{(i-1)d}.$$

Note that for $i = 0$, by definition we must have $\alpha \geq |\mathcal{B}_1|$, so this inequality also holds.

We are now ready to bound the regret of Algorithm ZOOC. We begin by the classical decomposition of the regret

$$\begin{aligned} R_T &\leq \mathbb{E}\left[\sum_{t \leq T} \Delta(\theta_t, c_t) | \mathcal{E}\right] + T\mathbb{P}(\bar{\mathcal{E}}) \\ &\leq \mathbb{E}\left[\sum_{t \leq T} \Delta(\theta_t, c_t) | \mathcal{E}\right] + 2. \end{aligned}$$

We now bound the regret on the good event $\mathcal{E}$. For some $i_0$ to be defined later, we have

$$R_T = \sum_{i < i_0} \sum_z \sum_{t \in \mathcal{T}(B_{i,z})} \Delta(\theta_t, c_t) + \sum_z \sum_{t \in \mathcal{T}^{\mathrm{Ch}}(B_{i_0,z})} \Delta(\theta_t, c_t).$$

Combining Lemma 9 with Lemma 8, we find that on the event $\mathcal{E}$,

$$R_T \leq \sum_{i \leq i_0} \frac{11(4+2D)\log(T)}{\epsilon L_z r_i} 2^D \alpha 2^{(i-1)d} + 2^D \alpha 2^{(i_0-1)d} \frac{11(4+2D)\log(T)}{\epsilon L_z r_i} + 5\epsilon L_z r_{i_0} T$$
$$\leq \frac{33(4+2D)2^D \alpha \log(T)}{\epsilon L_z} 2^{i_0(d+1)} + 5\epsilon L_z 2^{-i_0} T.$$

Choosing $i_0 = \left\lceil \frac{1}{d+2} \log_2 \left( \frac{5(\epsilon L_z)^2 T}{33(4+2D)2^D \alpha \log(T)} \right) \right\rceil$, we find that there exists a universal constant $C$ such that

$$R_T \leq C \left( \frac{((4+2D)2^D \alpha \log(T))^{\frac{1}{d+2}} T^{\frac{d+1}{d+2}}}{(\epsilon L_z)^{\frac{d+1}{d+2}}} (\epsilon L_z)^{\frac{1}{d+2}} \right)$$
$$\leq C \left((4+2D)2^D\right)^{\frac{1}{d+2}} (\alpha \log(T))^{\frac{1}{d+2}} (\epsilon L_z)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}.$$

# D Proof of auxilviary Lemmas

## D.1 Proof of Lemma 1

The proof of Lemma 1 follows from the classical concentration result for martingale sequence (see, e.g., Lemma 12 in Tullii et al. (2024) for a proof of this result).

**Lemma 10.** *Let $(y_t)_{t \geq 1}$ be a sequence of random variables adapted for a filtration $\mathcal{F}_t$, such that $y_t - \mathbb{E}[y_t | \mathcal{F}_{t-1}] \in [m, M]$. Assume that for $t \in \mathbb{N}_*$, $\iota_t \in \{0,1\}$ is $\mathcal{F}_{t-1}$-measurable, and define $N_t = \sum_{s \leq t} \iota_s$, and $\widehat{\mu}_t = \frac{\sum_{s \leq t} \iota_s (y_s - \mathbb{E}[y_s | \mathcal{F}_{s-1}])}{N_t}$ if $N_t \geq 1$. Then, for any $t \in \mathbb{N}_*$ and $\alpha \in (0,1)$,*

$$\mathbb{P}\left( N_t = 0 \text{ or } |\widehat{\mu}_t| \leq (M-m)\sqrt{\frac{\log(1/\alpha)}{2N_t}} \right) \geq 1 - 2t\alpha.$$

Let $t \in [T]$, $\theta \in \mathcal{G}$ and $c \in \mathcal{C}_T$. Applying Lemma 10 with the natural filtration $\mathcal{F}_t = \sigma\left(\{\theta_s, M_s\}_{s<t}\right)$, the random variables $y_t = M_t(c)$, and the indicator variable $\iota_t = \mathbb{1}\{\theta_t = \theta\}$, we find that for all $t \in \mathcal{N}_*$ and all $\alpha \in (0,1)$,

$$\mathbb{P}\left( N_t(\theta) = 0 \text{ or } |\widehat{m}_t(\theta, c) - m(\theta, c)| \leq \sqrt{\frac{2\log(1/\alpha)}{N_t(\theta)}} \right) \geq 1 - 2t\alpha.$$

Note that the grid $\mathcal{G}$ has size at most $\sqrt{DT}^D \leq T^D$. Moreover, there are at most $T$ different contexts in $\mathcal{C}_T$, and at most $T$ timesteps. Thus, choosing $\alpha = T^{-(3+D)}$, an union bound over at most $T^{2+D}$ different values for the triplet $(\theta, t, c)$ yields that

$$\mathbb{P}\left( \exists \theta \in \mathcal{G}, c \in \mathcal{C}_T, t \leq T : N_t(\theta) \neq 0 \text{ and } |\widehat{m}_t(\theta, c) - m(\theta, c)| \leq \sqrt{\frac{2(3+D)\log(T)}{N_t(\theta)}} \right) \leq 1 - 2T^{-1}$$

where we used $T \geq D$. Then, the result follows by noticing that when $N_t(\theta) = 0$, the bound $|\widehat{m}_t(\theta, c) - m(\theta, c)| \leq r_t(\theta)$ holds trivially.

## D.2 Proof of Lemma 2

Fix a round $t$. Let

$$\theta_t^\star \in \arg\max_{\theta \in \Theta} m(\theta, c_t).$$

If the algorithm activates an action at time $t$, then that action is played and $\mathrm{UCB}_t(\theta_t) = 1$. Since rewards lie in $[0,1]$, the inequality $m(\theta_t^*, c_t) \leq \max_{\theta \in \mathcal{A}_t} \mathrm{UCB}_t(\theta)$ is trivially verified. We therefore consider the case where the algorithm does not activate a new action, and instead selects an action

already activated with maximal UCB. In this case, the active set covers $\Theta$. Hence there exists $\theta' \in \mathcal{A}_t$ such that

$$L\|\theta' - \theta_t^\star\| < r_t(\theta').$$

By Assumption 3,

$$m(\theta_t^\star, c_t) - m(\theta', c_t) \le L\|\theta_t^\star - \theta'\| < r_t(\theta').$$

Now, on the event $\mathcal{E}$,

$$\widehat{m}_t(\theta', c_t) \ge m(\theta', c_t) - r_t(\theta').$$

Therefore,

$$\begin{aligned}\mathrm{UCB}_t(\theta') &= \widehat{m}_t(\theta', c_t) + 2r_t(\theta') \\ &\ge m(\theta', c_t) + r_t(\theta') \\ &\ge m(\theta_t^\star, c_t).\end{aligned}$$

Taking the maximum over $\mathcal{A}_t$ gives the result.

### D.3 Proof of Lemma 3

If at time $t$ the algorithm activates a new action, then $r_t(\theta_t) = 1$ and the bound holds trivially. Thus, we can consider a round in which no action is activated. By Lemma 2,

$$m(\theta_t^*, c_t) \le \mathrm{UCB}_t(\theta_t),$$

because $\theta_t$ maximizes the UCB index over active actions. Hence

$$\begin{aligned}m(\theta_t^*, c_t) &\le \widehat{m}_t(\theta_t, c_t) + 2r_t(\theta_t) \\ &\le m(\theta_t, c_t) + 3r_t(\theta_t),\end{aligned}$$

where the second inequality uses the good event $\mathcal{E}$. Subtracting $m(\theta_t, c_t)$ from both sides yields

$$\Delta(\theta_t, c_t) \le 3r_t(\theta_t).$$

### D.4 Proof of Lemma 4

Let $\theta, \theta' \in \tilde{\Theta}_j$ be two distinct actions. Without loss of generality, suppose $\theta$ was activated before $\theta'$. Let $t_1$ be the last round where $\theta$ was selected while $\theta'$ was not active. At that time, by Lemma 3 we had

$$r_{t_1}(\theta) \ge \frac{\Delta(\theta, c_{t_1})}{3} \ge \frac{r_{j+1}}{3}.$$

Moreover, let $t_2$ denote the activation time of $\theta'$. At that time, there existed an action $\theta''$ such that $\theta''$ was not covered by any previously active action, and $\theta'$ is the action in $\mathcal{G}$ that is the closest to $\theta''$. In particular, $\theta''$ was not covered by $\theta$. Therefore,

$$\|\theta - \theta''\| \ge \frac{r_{t_2}(\theta)}{L}.$$

Now, $r_t(\theta) := \min\left\{1, \sqrt{\frac{2(3+D)\log T}{N_t(\theta) \vee 1}}\right\}$ and $N_{t_2}(\theta) = N_{t_1}(\theta) + 1$, so $r_{t_1}(\theta) \le \sqrt{\frac{9}{8}} r_{t_2}(\theta)$. Hence,

$$\|\theta - \theta''\| \ge \sqrt{\frac{8}{9}} \frac{r_{t_1}(\theta)}{L} \ge \sqrt{\frac{8}{9}} \frac{r_{j+1}}{3L}.$$

Moreover, our choice of grid $\mathcal{G}$ ensures that

$$\|\theta' - \theta''\| \le T^{-1/2},$$

therefore

$$\|\theta - \theta'\| \ge \sqrt{\frac{8}{9}} \frac{r_{j+1}}{3L} - \sqrt{1/T} \ge \frac{r_{j+1}}{6L} \left(2\sqrt{\frac{8}{9}} - 6L\sqrt{1/T}2^{j+1}\right).$$

For $j \le j_0 \le \log_2\left(\frac{0.8\sqrt{T}}{6L}\right) - 1$, $2\sqrt{\frac{8}{9}} - 6L\sqrt{1/T}2^{j+1} \ge 1$. Thus

$$\|\theta - \theta'\| \ge \frac{r_{j+1}}{6L} = \frac{r_{j+2}}{3L}.$$

This proves that any two distinct actions $\theta, \theta' \in \tilde{\Theta}_j$ must be separated by at least $r_{j+2}/3L$. Since $\Theta_j \subset \cup_{t\le T}\Theta_{c_t,r_j}$, its cardinality is at most the packing number of $\cup_{t\le T}\Theta_{c_t,r_j}$ at scale $r_{j+2}/3L$. This cardinality is controlled by the cross-context zooming condition at scale $r = r_j/L$:

$$N_{r_{j+2}/3L}\left(\cup_{t\le T}\Theta_{c_t,r_j}\right) = N_{r/12}\left(\cup_{t\le T}\Theta_{c_t,rL}\right).$$

where we used $r_{j+2} = r_j/4$. Then,

$$|\tilde{\Theta}_j| \le \alpha r_j^{-d} L^d.$$

### D.5 Proof of Lemma 5

We first show that for all $z \in \{-1,1\}^C$, the reward function $m^z$ verifies Assumption 3 with $L = 1$. To do so, note that for all $c \le C$, on the boundary of $Q_c^{z_c}$, $2d'_\infty(\theta, x_c^{z_c}) = \epsilon$, so $m^z(\theta, c) = \frac{1}{2}$. Similarly, on the boundary of $Q_c^{-z_c}$, $2d'_\infty(\theta, x_c^{-z_c}) = \epsilon$, so $m^z(\theta, c) = \frac{1}{2}$. Thus, the function $\theta \mapsto m^z(\theta, c)$ is continuous on the boundary of the hypercubes $Q_c^{+1}$ and $Q_c^{-1}$. The triangle inequality implies that it is $\frac{1}{4\sqrt{D}}$-Lipschitz continuous w.r.t. the $d_\infty$-norm, so it is $\frac{1}{4}$-Lipschitz continuous w.r.t. the euclidean norm.

Now, for $r \le \frac{\epsilon}{8\sqrt{D}}$, $\Theta_{c,r} \subset \{\theta : d'_\infty(\theta, x_c^{z_c}) \in [2\sqrt{D}r, 4\sqrt{D}r]\} \cup \{\theta : d'_\infty(\theta, x_c^{-z_c}) \in, [(4\sqrt{D}r - \epsilon)_+, (8\sqrt{D}r - \epsilon)_+]\}$. The volume of this set is at most $(4\sqrt{D}r)^D + (8\sqrt{D}r)^D$, so it contains at most $(12\sqrt{D}r)^D/(r/24)^D = (288\sqrt{D})^D$ bins of radius $r/24$. Thus, the $r/12$-packing number of this set is at most $(288\sqrt{D})^D$. Similarly, for $r > \frac{\epsilon}{8\sqrt{D}}\Theta_{c,r} \subset \{\theta : d_\infty(\theta, x_c) \le 4\sqrt{D}r]\}$, which has a $r/12$-packing number at most $(96\sqrt{D})^D$. Thus, for $\alpha_D = (288\sqrt{D})^D$, $m^z$ has 1-uniform zooming dimension 0 with multiplier $\alpha_D$.

### D.6 Proof of Lemma 6

The proof of Lemma 6 requires going beyond standard estimation arguments in bandit settings. On the one hand, we must control the supremum of the estimation error over a class of loss functions. On the other hand, unlike classical bandit settings, estimating the risk requires aggregating observations generated under different actions. As a result, neither standard martingale arguments (which control the error for a fixed function) nor the classical symetrization trick (which relies on independence of observations) can help us bound the supremum of the errors.

Instead, we apply Lemma 10 for a given $t \le T$, $c \in \mathcal{C}_t$, $\theta \in \mathcal{G}$, and $\{i, z\}$ such that $\theta \in B_{i,z}$. We define let $\iota_s = \mathbb{1}\{\tau(B_{i,z}) \le s < \tau'(B_{i,z})\}\mathbb{1}\{\theta_s \in B_{i,z}\}$, and $y_s = \ell_c(Z_s, \theta)$. Then, Lemma 10 implies that

$$\mathbb{P}\left(N_t(B_{i,z}) = 0 \text{ or } \left|\frac{\sum_{\tau(B_{i,z})\le s<t:\ \theta_s\in B_{i,z}} \mathcal{DPR}(\theta_s,\theta)}{N_t(B_{i,z})} - \widehat{\mathcal{DPR}}(B_{i,z},\theta)\right| \le \sqrt{\frac{2\log(1/\alpha)}{N_t(B_{i,z})}}\right) \ge 1-2t\alpha.$$

Now, for all $\theta'' \in B_{i,z}$, $\|\theta'' - \theta\| \le r_i$, so $|\mathcal{DPR}(\theta'', \theta) - \mathcal{DPR}(\theta, \theta)| \le \epsilon L_z r_i$. Thus,

$$\mathbb{P}\left(N_t(B_{i,z}) = 0 \text{ or } \left|\mathcal{PR}(\theta) - \widehat{\mathcal{DPR}}(B_{i,z},\theta)\right| \le \sqrt{\frac{2\log(1/\alpha)}{N_t(B_{i,z})}} + \epsilon L_z r_i\right) \ge 1 - 2t\alpha.$$

Let us bound the number of bins $B_{i,z}$ over which we need to control the probability. The deactivation condition ensures that a bin at level $i$ must be chosen at least $(\epsilon L_z r_i)^2$ times before de-activation. Thus, the maximum depth that the algorithm can reach is $I_{\max} = \log_4(T/(\epsilon L_z)^2) - 1$. Now, the number of bins at level $i \le I_{\max}$ is $\left\lceil\sqrt{D}\right\rceil^D \sum_{i\le I_{\max}} 2^{iD} = \left\lceil\sqrt{D}\right\rceil^D 2^{(I_{\max}+1)D} = \left\lceil\sqrt{D}\right\rceil^D \sqrt{T^D/(\epsilon L_z)^{2D}}$.

Assuming that $\left\lceil \sqrt{D} \right\rceil / (\epsilon L_z) \leq T$, we find that there are at most $T^D$ such bins. Then, applying a union bound over the $|\mathcal{G}|$ possible values for $\theta$, the $T^D$ possible values for $\{i, z\}$, the $T$ rounds, and the $T$ contexts, we find that

$$\mathbb{P}\Bigg(\forall \theta \in \mathcal{G},\ t \leq T, B_{i,z} \text{ such that } \tau(B_{i,z}) < t \text{ and } \theta \in B_{i,z} : N_t(B_{i,z}) = 0$$
$$\text{or } \left|\mathcal{PR}(\theta) - \widehat{\mathcal{DPR}}(B_{i,z}, \theta)\right| \leq \sqrt{\frac{2\log(T^{4+D}|\mathcal{G}|/\alpha)}{N_t(B_{i,z})}} + \epsilon L_z r(B_{i,z})\Bigg) \geq 1 - 2\alpha.$$

Let us take $\alpha = T^{-1}$ and notice that $|\mathcal{G}| \leq ((L_\theta + \epsilon L_z)\sqrt{TD})^D \leq T^D$ when $T \geq D(L_\theta + \epsilon L_z)^2$. Then, our choice $r_t(B_{i,z}) = \sqrt{\frac{2(4+2D)\log(T)}{N_t(B_{i,z})}}$ implies that with probability at least $1 - 2T^{-1}$,$\forall\, t \in [T],\ \forall\, \theta \in \mathcal{G},\ \forall\, c \in \mathcal{C}_T$, $B_{i,z}$ such that $\tau(B_{i,z}) < t$ and $\theta \in B_{i,z}$

$$\left|\widehat{\mathcal{DPR}}_{c,t}(B_{i,z}, \theta) - \mathcal{PR}_c(\theta)\right| \leq r_t(B_{i,z}) + \epsilon L_z r(B_{i,z})$$

or $N_t(B_{i,z}) = 0$. We conclude by noticing that when $N_t(B_{i,z}) = 0$, we have $r_t(B_t(\theta)) = 1$, so the inequality is immediately verified. Moreover, by the design of our algorithm, if $t < \tau'(B_{i,z})$, $\epsilon L_z r(B_{i,z}) \leq r_t(B_{i,z})$. Finally, if $t \geq \tau'(B_{i,z})$, $\epsilon L_z r_i \geq r_t(B_{i,z})$.

### D.6.1 Proof of Lemma 7

On the event $\mathcal{E}$, for all $\theta \in \Theta$,

$$\left|\widehat{\mathcal{DPR}}_{c,t}(B_t(\theta), \theta) - \mathcal{PR}_{c_t}(\theta)\right| \leq 2r_t(B_t(\theta))$$

and

$$\left|\widehat{\mathcal{DPR}}_{c,t}(\text{Pa}(B_t(\theta)), \theta) - \mathcal{PR}_{c_t}(\theta)\right| \leq 2\epsilon L_z r(\text{Pa}(B_t(\theta)))$$

On the one hand, in implies that

$$\mathcal{PR}_{c_t}(\mathcal{G}(\theta_t^*)) \geq \widehat{\mathcal{PR}}_t^{\text{lcb}}(\mathcal{G}(\theta_t^*))$$

where we recall that $\theta_t^* \in \arg\min_\theta \mathcal{PR}_{c_t t}(\theta)$ and we define $\mathcal{G}(\theta_t^*) = \arg\min_{\theta \in cG} \|\theta_t^* - \theta\|$. Since $\mathcal{PR}_{c_t}$ is $(L_\theta + \epsilon L_z)$-Lipschitz continuous, and since our choice of the grid $\mathcal{G}$ implies $\|\mathcal{G}(\theta_t^*) - \theta_t^*\| \leq 1/\sqrt{T}$, we obtain

$$\widehat{\mathcal{PR}}_t^{\text{lcb}}(\mathcal{G}(\theta_t^*)) \leq \mathcal{PR}_{c_t}(\theta_t^*) + 1/\sqrt{T}.$$

Now, by carefully distinguishing the cases $\epsilon L_z r(\text{Pa}(B_t(\theta_t))) \leq r_t(B_t(\theta_t))$ and $\epsilon L_z r(\text{Pa}(B_t(\theta_t))) > r_t(B_t(\theta_t))$, we can show similarly that

$$\mathcal{PR}_{c_t}(\theta_t) - 4\min\{\epsilon L_z r(\text{Pa}(B_t(\theta_t))), r_t(B_t(\theta_t))\} \leq \widehat{\mathcal{PR}}_t^{\text{lcb}}(\theta_t).$$

By design of our algorithm, $\widehat{\mathcal{PR}}_t^{\text{lcb}}(\theta_t) \leq \mathcal{PR}_{c_t}(\theta_t^*)$, which implies that

$$\mathcal{PR}_{c_t}(\theta_t) - 4\min\{\epsilon L_z r(\text{Pa}(B_t(\theta_t))), r_t(B_t(\theta_t))\} \leq \mathcal{PR}_{c_t}(\theta_t^*) + 1/\sqrt{T}.$$

Noticing that $\min\{\epsilon L_z r(\text{Pa}(B_t(\theta_t))), r_t(B_t(\theta_t))\} \geq \sqrt{1/T}$, we conclude that

$$\mathcal{PR}_{c_t}(\theta_t) - 5\min\{\epsilon L_z r(\text{Pa}(B_t(\theta_t))), r_t(B_t(\theta_t))\} \leq \mathcal{PR}_{c_t}(\theta_t^*).$$

### D.6.2 Proof of Lemma 8

Let us first bound $R(B_{i,z})$ as follows. On the event $\mathcal{E}$, we have by Lemma 7,

$$\Delta(\theta_t, c_t) \leq 5\min\{r_t(B_t(\theta_t)), \epsilon L_z r(\text{Pa}(B_t(\theta_t)))\}.$$

Moreover, for $t \in \mathcal{T}(B_{i,z})$, $B_t(\theta_t) = B_{i,z}$. Thus,

$$\begin{aligned} R(B_{i,z}) &= \sum_{t \in \mathcal{T}(B_{i,z})} \Delta(\theta_t, c) \\ &\leq 5 \sum_{t \in \mathcal{T}(B_{i,z})} r_t(B_{i,z}) \\ &\leq 5\sqrt{2(4+2D)\log(T)} \sum_{t \in \mathcal{T}(B_{i,z})} \sqrt{\frac{1}{N_t(B_{i,z})}} \\ &\leq 5\sqrt{2(4+2D)\log(T)|\mathcal{T}(B_{i,z})|}. \end{aligned}$$

By design of our algorithm, the deactivation condition ensures that $\sqrt{|\mathcal{T}(B_{i,z})|} \leq \frac{\sqrt{2(4+2D)\log(T)}}{\epsilon L_z r_i} + 1$. Thus,

$$R(B_{i,z}) \leq \frac{11(4+2D)\log(T)}{\epsilon L_z r_i}.$$

To bound $R^{\text{Ch}}(B_{i,z})$ on the event $\mathcal{E}$, we decompose it as follows:

$$R^{\text{Ch}}(B_{i,z}) = \sum_{t \in \mathcal{T}(B_{i,z})} \Delta(\theta_t, c) + \sum_{t \in \mathcal{T}^{\text{Ch}}(B_{i,z}), \tau'(B_{i,z}) \leq t} \Delta(\theta_t, c).$$

Using the previous result as well as Lemma 7, together with the remark that for $t \in \mathcal{T}^{\text{Ch}}(B_{i,z})$ such that $\tau'(B_{i,z}) \leq t$, $r(\text{Pa}(B_t(\theta_t)) \leq r_i$, we obtain

$$R^{\text{Ch}}(B_{i,z}) \leq \frac{11(4+2D)\log(T)}{\epsilon L_z r_i} + 5\epsilon L_z r_i |\mathcal{T}^{\text{Ch}}(B_{i,z})|.$$

### D.6.3 Proof of Lemma 9

We assume henceforth that $i \geq 1$, and bound the number of bins $B_{i,z}$ that can be activated during the game. For the bin $B_{i,z}$ to be activated, it implies that there exists $z'$ such that the bin $B_{i,z}$ is a child of $B_{i-1,z'}$, and such that $B_{i-1,z'}$ has been selected enough time to be deactivated. In particular, at $t = \tau'(B_{i-1,z'})$, $B_{i-1,z'}$ is selected, and $r_{t+1}(B_{i-1,z'}) \leq 2\epsilon L_z r_{i-1}$. Now,

$$r_{t+1}(B_{i-1,z'}) = \sqrt{\frac{2(4+2D)\log(T)}{N_t(B_{i,z})}} \geq \frac{1}{\sqrt{2}} \sqrt{\frac{2(4+2D)\log\left(|\mathcal{G}|T\right)}{N_t(B_{i-1,z'})}} = \frac{r_t(B_{i-1,z'})}{\sqrt{2}}.$$

Thus, $\frac{r_t(B_{i-1,z'})}{\sqrt{2}} \leq 2\epsilon L_z r_{i-1}$ Moreover, Lemma 7 implies that on event $\mathcal{E}$,

$$\Delta(\theta_t, c_t) \leq 5 r_t(B_{i-1,z'}) \leq 10\sqrt{2}\epsilon L_z r_{i-1}.$$

Thus, for a bin $B_{z,i}$ to be activated, it must be the child of a bin $B_{z',i-1}$ covering an action $\theta'' \in \subset \bigcup_{t \leq T} \Theta_{c_t, 12\sqrt{2}\epsilon L_z r_{i-1}}$. Under our assumption, there are at most $\alpha / r_{i-1}^d = \alpha 2^{-(i-1)d}$ such bins. Since each of these bins, when deactivated, are replaced by $2^D$ children, we find that

$$\sum_z \mathbb{1}\{|\mathcal{T}(B_{i,z})| > 0\} \leq 2^D \alpha 2^{-(i-1)d}.$$