A Unified Model and Dimension for Interactive Estimation

Nataly Brukhim
Princeton University
nbrukhim@princeton.edu

Miro Dudik
Microsoft Research
mdudik@microsoft.com

Aldo Pacchiano
Broad Institute of MIT and Harvard & Boston University
apacchia@broadinstitute.org

Robert Schapire
Microsoft Research
schapire@microsoft.com

Abstract

We study an abstract framework for interactive learning called *interactive estimation* in which the goal is to estimate a target from its "similarity" to points queried by the learner. We introduce a combinatorial measure called *dissimilarity dimension* which is used to derive learnability bounds in our model. We present a simple, general, and broadly-applicable algorithm, for which we obtain both regret and PAC generalization bounds that are polynomial in the new dimension. We show that our framework subsumes and thereby unifies two classic learning models: statistical-query learning and structured bandits. We also delineate how the dissimilarity dimension is related to well-known parameters for both frameworks, in some cases yielding significantly improved analyses.

1 Introduction

We study a general interactive learning protocol called *interactive estimation*. In this model, the learner repeatedly queries the environment with an element from a set of *alternatives*, and observes a stochastic reward whose expectation is given by an arbitrary measure of the "similarity" between the queried alternative and the unknown ground truth. Thus, in rough terms, the goal is to estimate a target from its similarity to queried alternatives. By studying such a general abstraction of interactive learning, we are able to reason about the properties of a very broad family of learning settings, and to make connections across a variety of contexts.

Our results are based on a combinatorial complexity measure we introduce called the *dissimilarity dimension*, which is used to derive learnability bounds in our model. Intuitively, this measure corresponds to the length of the longest sequence of alternatives in which each one has a similar suboptimal value of similarity to all its predecessors. We then use the measure to analyze the performance of a simple, broadly-applicable class of algorithms which repeatedly make new queries that best fit the preceding observations. We prove both regret bounds and PAC generalization bounds that are all polynomial in the dissimilarity dimension.

We show that our learning framework subsumes two classic learning models that were seemingly unrelated prior to this work:

First, our model subsumes the statistical query (SQ) model, introduced by Kearns [19] for designing noise-tolerant learning algorithms. In the SQ model, the learner can sequentially ask certain queries of an oracle, who responds with answers that are only approximately correct, with the goal of correctly estimating a target. Despite its simplicity, it has been proven to be a powerful model. Indeed, a wide range of algorithmic techniques in machine learning are implementable using SQ learning. Thus, it has been proven useful, not only for designing noise-tolerant algorithms, but also for its connections to other noise models, and as an explanatory tool to prove hardness of many problems (see the survey

of Reyzin [23]). We show that our framework subsumes the SQ model, and furthermore that the dissimilarity dimension generalizes well-known parameters that characterize SQ learnability.

Second, our model captures structured bandits, in which the learner repeatedly chooses actions which yield stochastic rewards, with the goal of minimizing regret relative to the best action in hindsight. Over more than a decade, the eluder dimension [24] has been a central technique for analyzing regret for contextual bandits and reinforcement learning (RL) with function approximation [30, 22, 29, 10]. We will see that the dissimilarity dimension is upper-bounded by the eluder dimension, and that there can in fact be a large gap between the two. This sometimes leads to an improved analysis when relying on the proposed dissimilarity measure rather than the eluder dimension.

Because SQ and bandits are both subsumed by our framework, all the results mentioned above directly apply to those settings as well, including the applicability of our general-purpose algorithms.

To summarize, our main contributions are as follows:

- **Unified framework.** We derive a general framework which captures various interactive learning settings, including specifically SQ and bandits.
- **Novel dimension, performance bounds.** We introduce the dissimilarity dimension that is used to derive learnability bounds in our model. We study a general, simple algorithm, and give a novel analysis that results in both regret and PAC generalization bounds that are polynomial in the new dimension. We also give lower bounds in the SQ and bandit settings.
- **Improved analysis.** We show instances in which the standard analysis of a certain class of algorithms using the eluder dimension yields bounds that are arbitrarily large, but in which an analysis using our dimension yields low regret bounds.

Related work. The interactive estimation model we consider in this work is defined with respect to an evaluation function that can be thought of as an arbitrary measure of the "similarity" between the queried alternative and the target. Previously, Balcan and Blum [3] developed a theory of similarity-based learning that generalizes kernel methods, providing sufficient conditions for a similarity function to be useful for learning. Chen et al. [8] review several approaches to classification based on similarity between examples, including, for instance, kernels and nearest neighbors. Ben-David et al. [5] studied a learning-by-distances model that resembles ours using a metric as a measure of similarity. In comparison to these works, our model admits an arbitrary similarity measure for which we derive a general dimension, algorithm, and bounds.

In the context of bandit and reinforcement learning, a parameter called the decision-estimation coefficient (DEC) has recently been proposed by Foster et al. [15] to characterize learnability in interactive decision making. Unlike DEC, our dimension is combinatorial in nature, and applies to settings like SQ, which are not captured by DEC.

As discussed above, our model subsumes SQ and bandits, both of which have been extensively studied (see the references above as well as various surveys [20, 23]).

2 Setting

In this paper, we study an interactive learning protocol called *interactive estimation*. In this protocol, the learner is provided with a set $\mathcal{Z}$ of *alternatives*, and an *evaluation function* $\rho : \mathcal{Z} \times \mathcal{Z} \to [-1, 1]$. Intuitively, ρ can be viewed as a measure of "similarity," though it need not be symmetric. There is also a distinguished alternative $z^* \in \mathcal{Z}$ called the *target*, fixed throughout the interaction, and unknown to the learner. In each of a sequence of steps $t = 1, \ldots, T$, the learner selects one alternative $z_t \in \mathcal{Z}$ and receives a stochastic *reward* $r_t \in [-1, 1]$ drawn independently, conditioned on z_t, with expectation satisfying $\mathbb{E}[r_t \mid z_t] = \rho(z_t \mid z^*)$. Informally, by choosing alternatives and observing their similarity to z^*, the learner aims to get close to the target. The special case when $r_t = \rho(z_t \mid z^*)$, that is, when rewards are deterministic functions of the queried alternatives, is referred to as the *deterministic setting*.

We generally assume $\rho(z^* \mid z^*) \geq \rho(z \mid z^*)$ for all $z \in \mathcal{Z}$ and denote this optimal value as $\alpha^* := \rho(z^* \mid z^*)$. We will assume that the value of α^* is known to the learner or that we are provided with an alternate *optimality level* $\alpha \leq \alpha^*$ such that the task is to identify z with $\rho(z \mid z^*) \geq \alpha$. At the end of Section 3 we discuss how this assumption can be relaxed.

We consider two alternative goals for a learner in this model: *sublinear regret* and *PAC generalization*. A learner achieves *sublinear regret* relative to an optimality level $\alpha \leq \alpha^*$ if $\text{Regret}(T, \alpha) = o(T)$, where

$$\text{Regret}(T, \alpha) = \sum_{t=1}^{T} \Big(\alpha - \rho(z_t \mid z^*) \Big).$$

We say that a learner achieves *PAC generalization* if for any $\epsilon, \delta > 0$ and $\alpha \leq \alpha^*$, with probability at least $1 - \delta$ (over the randomness of the query responses and the learner's own randomization), after $m(\epsilon, \delta, \alpha)$ interactions in the protocol above, the learner outputs $\hat{z}$ such that $\rho(\hat{z} \mid z^*) \geq \alpha - \epsilon$. The function $m(\epsilon, \delta, \alpha)$ is referred to as sample complexity. We recover standard notions of regret and PAC generalization by setting $\alpha = \alpha^*$.

Example 1 (Point on a sphere)**.** *Let $\|\cdot\|$ denote the standard Euclidean norm in $\mathbb{R}^n$ and let $\mathcal{Z} = \mathcal{S}_{n-1} = \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\| = 1\}$ be the unit sphere in $\mathbb{R}^n$. The goal is to estimate an unknown point $\mathbf{z}^* \in \mathcal{S}_{n-1}$ based on rewards equal to the inner product between the queries and the target, that is, $r_t = \rho_{\text{sphere}}(\mathbf{z}_t \mid \mathbf{z}^*) \coloneqq \langle \mathbf{z}_t, \mathbf{z}^* \rangle$.*

We now introduce the two main examples corresponding to classic learning models that are subsumed by the interactive estimation model.

Example 2 (Structured bandits)**.** *Let $\mathcal{A}$ be an action set, $\mathcal{F}$ a space of reward functions $f : \mathcal{A} \to [-1, 1]$, and $f^* \in \mathcal{F}$ the target reward function. In step t, the learner chooses an action $a_t \in \mathcal{A}$ and receives reward $r_t \in [-1, 1]$ with $\mathbb{E}[r_t \mid a_t] = f^*(a_t)$. The goal is to maximize the sum of rewards. Let $a^* = \text{argmax}_{a \in \mathcal{A}} f^*(a)$ be an optimal action. To represent bandits in our formalism, we let $\mathcal{Z} = \mathcal{F} \times \mathcal{A}$, $z^* = (f^*, a^*)$, and $\rho_{\text{bandits}}\big((f, a) \;\big|\; (f^*, a^*)\big) = f^*(a)$.*

The structured bandit problem has been extensively studied, and Example 2 captures its expressiveness within our framework. For example, it recovers the possibly simplest case of K-armed bandits, by considering $\mathcal{A} = \{1, ..., K\}$ and $\mathcal{F} = [0, 1]^K$. At each round, the learner chooses arm $a_t \in \mathcal{A}$ and observes a reward r_t which is drawn from a distribution with mean $f^*(a_t)$. See Appendix D for a more concrete example of K-armed bandits instantiated within our framework. In Section 5 we also give concrete bounds for other example classes including linear bandits and GLM bandits.

We remark that although the protocol in which the learner submits pairs (f, a) in each round rather than an action a, may seem more complicated than the standard bandits protocol, it is in fact equivalent as the function f is ignored in the evaluation. Moreover, this formulation naturally captures any algorithms for realizable bandits. Such algorithms often keep track of a version space (set of functions f consistent with the data), and so at each point of interaction, there is an implicit f_t that is associated with a_t produced at that time. The above protocol simply makes the choice of f_t explicit.

Example 3 (SQ learning)**.** *Given a domain $\mathcal{X}$, the goal is to learn a binary classifier $h^* : \mathcal{X} \to \{\pm 1\}$ from some hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, based on training examples (x, y) drawn from some distribution D such that $y = h^*(x)$. In step t, the learner produces a hypothesis h_t and observes the accuracy of h_t on a fresh finite sample. In this case, $\mathcal{Z} = \mathcal{H}$, the evaluation function is equal to the expected accuracy $\rho_{\text{SQ}}(h \mid h^*) = \mathbb{E}_{x \sim D}[h(x)h^*(x)]$, and the reward is the empirical accuracy on a fresh sample.*

The SQ learning model considered in this work (Example 3) differs from the original model of Kearns [19] because it is restricted, as in previous works [7, 13, 31], to so-called *correlational* queries (called CSQs) and assumes stochastic responses, as opposed to allowing arbitrary queries and adversarial responses. We discuss relationships between various SQ variants in Appendix B.

We finish this section by introducing a central concept of this paper, a new combinatorial complexity measure called the *dissimilarity dimension* which, as we will see, allows us to derive learnability bounds for the interactive estimation protocol.

Definition 1 (Dissimilarity dimension)**.** *For a set $\mathcal{Z}$, scalars $\alpha \in \mathbb{R}$, $\epsilon > 0$, and evaluation function $\rho : \mathcal{Z} \times \mathcal{Z} \to [-1, 1]$, the* dissimilarity dimension $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon)$ *is the largest integer d for which there exist $z_1, \ldots, z_d \in \mathcal{Z}$ with $\rho(z_i \mid z_i) \geq \alpha$, and a scalar $c \leq \alpha - \epsilon$, such that for all $i < j$,*

$$\Big| \rho(z_i \mid z_j) - c \Big| \leq \frac{\epsilon}{\sqrt{d}}.$$

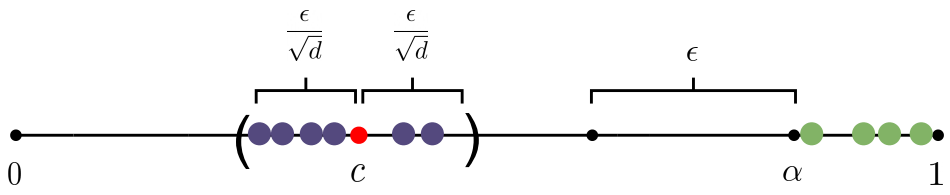

Figure 1: An illustration of the dissimilarity dimension $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) = 4$. The figure shows ρ values on the $[0, 1]$ interval for a sequence of elements z_1, z_2, z_3, z_4. The four green points represent self-evaluation values $\rho(z_i \mid z_i)$ for $i = 1, \ldots, 4$; all are greater than or equal to α. The six blue points represent values $\rho(z_i \mid z_j)$ for $i < j$; all are within the distance $\epsilon/\sqrt{d}$ from the value $c \leq \alpha - \epsilon$.

Furthermore, denote the monotonic dissimilarity dimension *as* $\overline{\mathtt{d}}_\rho(\mathcal{Z}, \alpha, \epsilon) := \max_{\epsilon' \geq \epsilon} \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon')$. [1]

Note that $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) = 0$ if there is no z such that $\rho(z \mid z) \geq \alpha$, and otherwise $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) \geq 1$. In particular, if $\alpha \leq \alpha^*$ then $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) \geq 1$.

In rough terms, this dimension corresponds to the longest sequence of points with α-large self-evaluation, such that the evaluation $\rho(z_i \mid z_j)$ of each point z_i relative to every successive point z_j is "small" (significantly less than α), and also tightly clustered around some value c. Thus, each point is similar to itself, but dissimilar from all successive points to about the same degree. The idea is illustrated in Figure 1. The monotonic dissimilarity dimension is the tightest upper bound on the dissimilarity dimension that is non-increasing in ϵ.

Various concrete examples where the dissimilarity dimension can be bounded are provided in Section 5. For instance, using a general bound for linear bandits from Section 5, we can show that for the task of finding a point on a sphere based on inner products (Example 1), the dimension $\overline{\mathtt{d}}_{\rho_{\text{sphere}}}(\mathcal{Z}, \alpha, \epsilon) \leq 4n + 3$, a bound that is independent of both α and ϵ.

3 Algorithms and upper bounds

In this section we analyze algorithms for the interactive estimation protocol, which we call *interactive estimation algorithms*. We show that when an interactive estimation algorithm satisfies two properties, *large self-evaluations* and *decaying estimation error*, then its regret can be bounded using the dissimilarity dimension. We introduce a simple algorithm (Algorithm 1), which satisfies these properties for many standard classes of alternatives. The first property requires that the algorithm only select alternatives that would achieve the expected reward of at least α if they were the target:

Definition 2 (α-large self-evaluations)**.** *An interactive estimation algorithm has α-large self-evaluations if at every time step $t = 1, \ldots, T$, it selects a query z_t such that $z_t \in \mathcal{Z}_\alpha$, where*

$$\mathcal{Z}_\alpha = \{z \in \mathcal{Z} : \rho(z \mid z) \geq \alpha\}. \tag{1}$$

Algorithm 1 satisfies this property, with the optimality level α provided as input. At the end of the section, we discuss the case when α is not provided and α^* is unknown. We derive an optimistic version of Algorithm 1 that achieves α^*-large self-evaluations with high probability.

The second property states that the queries produced by the algorithm provide increasingly good estimates of the expected rewards in the previous rounds (that is they are good estimators *with the benefit of hindsight*), as quantified by the square loss.

Definition 3 (Decaying estimation error)**.** *An interactive estimation algorithm has decaying estimation error if there exists $C_{T,\delta} \geq 0$ growing sublinearly in T, that is, $C_{T,\delta} = o(T)$, such that with*

[1] We remark that the term $\sqrt{d}$ in Definition 1 can be generalized to any function $g(d)$ and state our main results in terms of g. However, for ease of presentation we pick $g(d) = \sqrt{d}$ as it simplifies the comparison between the dissimilarity dimension and eluder dimension (e.g. Thm. 11 and Prop. 12).

Algorithm 1 Interactive Estimation via Least Squares

1: **Input:** set of alternatives $\mathcal{Z}$, evaluation function ρ, optimality level α, number of steps T.
2: **for** $t = 1, \ldots, T$ **do**
3: Submit the query $z_t = \operatorname{argmin}_{z \in \mathcal{Z}_\alpha} \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - r_i\Big)^2$.
4: Observe reward r_t.
5: **end for**

probability at least $1 - \delta$ *the sequence of queries* $z_1, \ldots, z_T$ *produced by the algorithm satisfies*

$$\sum_{i=1}^{t-1} \Big(\rho(z_i \mid z_t) - \rho(z_i \mid z^*)\Big)^2 \le C_{T,\delta} \tag{2}$$

for all $t \in \{1, \ldots, T\}$ *simultaneously.*

Algorithm 1 optimizes an empirical version of Eq. (2), with the observed rewards r_i in place of the expectations $\rho(z_i \mid z^*)$. Thus, in the deterministic setting, with $r_i = \rho(z_i \mid z^*)$, Algorithm 1 satisfies this property with $C_{T,\delta} = 0$. It can also be shown that it satisfies this property when the set of alternatives is finite:

Theorem 4. *Assume that* $|\mathcal{Z}| < \infty$. *Then Algorithm 1 satisfies the decaying estimation error property with* $C_{T,\delta} = O\big(\ln(T|\mathcal{Z}|/\delta)\big)$.

In the general case, when $\mathcal{Z}$ is infinite, we show that Algorithm 1 satisfies the decaying estimation error property with $C_{T,\delta} = O\big(\log(TN/\delta)\big)$ where N is a suitable covering number of $\mathcal{Z}$ (see Corollary 18 in Appendix A.1). For example, for *linear bandits*, which is an instance of Example 2 in which the action set and function class correspond to a subset of the unit ball in $\mathbb{R}^n$, we obtain $C_{T,\delta} = O\big(n \log(1/\epsilon) + \log(T/\delta)\big)$.

In Appendix A.3, we discuss an approach in which we have access to an online regression oracle for the least squares problem in step 3. We show that a suitably modified version of Algorithm 1 has a decaying estimation error as long as the online regression oracle achieves a sublinear regret (but without further dependence on a covering number).

To develop some intuition how Algorithm 1 works, we can again consider the K-armed bandit problem (a special case of Example 2), and suppose that $\alpha = 0.75$. In each step, the algorithm picks a pair (f_t, a_t), where $f_t \in [0,1]^K$ is the vector of mean reward estimates and a_t is the arm with the largest mean estimate. The estimates $f_t(a)$, $a = 1, \ldots, K$, are formed by optimizing the least squares error of the observed rewards, under the constraint that at least one of the mean estimates must be above 0.75. As a result, the algorithm pulls the arm with the largest average reward as long as that average is above 0.75 (arms that have not been pulled are assumed to have averages above 0.75). If all the averages are below 0.75 then the algorithm selects the arm a with the smallest value $n_a(0.75 - \hat{\mu}_a)^2$, where n_a is how many times the arm has been pulled so far and $\hat{\mu}_a$ is its average reward; it can be verified that this solves the least squares problem subject to the constraint that at least one of the mean estimates is above 0.75.

We next state our main results: a regret bound and a PAC generalization guarantee. They are both based on bounding how many "bad" queries any algorithm with large self-evaluations and decaying estimation error can make. Concretely, we say that a query $z \in \mathcal{Z}$ is ϵ-*bad* if its suboptimalty gap is greater than ϵ, that is, if

$$\rho(z \mid z^*) < \alpha - \epsilon.$$

The next lemma shows that the number of ϵ-bad queries is upper bounded polynomially in the dissimilarity dimension.

Lemma 5 (Few bad queries)**.** *Let* $\epsilon, \delta > 0$, *and let* $d = \overline{\mathsf{d}}_\rho(\mathcal{Z}, \alpha, \epsilon) < \infty$ *for some set* $\mathcal{Z}$, *evaluation function* ρ *and* $\alpha \le \alpha^*$. *Let* Alg *be an interactive estimation algorithm with* α*-large self-evaluations and a decaying estimation error with some* $C_{T,\delta}$. *Then, with probability at least* $1 - \delta$, *the number of* ϵ*-bad queries that* Alg *makes in* T *steps is at most* $2d^{1.5}\ln(4/\epsilon) + 12d^{2.5}C_{T,\delta}/\epsilon^2$. *Consequently, if* $C_{T,\delta} \ge \ln(2T)$, *then with probability at least* $1 - \delta$, *the number of* ϵ*-bad queries is at most* $36d^{2.5}C_{T,\delta}/\epsilon^2$, *and if* $C_{T,\delta} = 0$ *then it is at most* $2d^{1.5}\ln(4/\epsilon)$.

Algorithm 2 PAC Interactive Estimation

Input: set of alternatives $\mathcal{Z}$, evaluation function ρ, optimality level α, base interactive estimation algorithm Alg, parameters T, n_1, n_2.
Run algorithm Alg with the provided $\mathcal{Z}, \rho, \alpha, T$.
Sample n_1 indices $t_1, \ldots, t_{n_1}$ uniformly at random from $\{1, \ldots, T\}$.
For each $\ell = 1, \ldots, n_1$, submit query z_{t_ℓ} for n_2 times; denote the average response $\bar{r}_{t_\ell}$.
Let $\hat{\ell} = \operatorname{argmax}_{\ell \in \{1,\ldots,n_1\}} \bar{r}_{t_\ell}$.
Output $\hat{z} = z_{t_{\hat{\ell}}}$.

The above result is the core component of our main theorems. The proof is given in Appendix A.4; here we sketch the main ideas. The goal is to show that the "bad" interval $[-1, \alpha - \epsilon]$ cannot contain too many queries made by Alg. The proof starts by partitioning this interval into disjoint subintervals and then bounds the number of queries in each subinterval. It does so by constructing a graph with nodes corresponding to queries, which are connected by an edge if they satisfy the dimension conditions. The decaying errors that imply a certain minimum number of edges (as a function of number of queries). On the other hand, the dissimilarity dimension bounds the size of the largest clique, which implies an upper bound on the number of edges (using Turán's Theorem [26], a standard result from extremal graph theory). Combining the bounds yields an upper bound on the number of queries in the subinterval. Summing across subintervals proves the lemma.

The following theorems use Lemma 5 to bound both the regret and PAC sample complexity. The proofs are deferred to Appendices A.6 and A.7.

Theorem 6 (Regret)**.** *Let $\delta, T > 0$, and let $d = \overline{\mathsf{d}}_\rho(\mathcal{Z}, \alpha, 1/T)$ for some set $\mathcal{Z}$, evaluation function ρ and $\alpha \leq \alpha^*$. Let* Alg *be an interactive estimation algorithm with α-large self-evaluations and a decaying estimation error with some $C_{T,\delta}$. If $C_{T,\delta} \geq \ln(2T)$ then with probability at least $1 - \delta$, the regret of* Alg *satisfies*

$$\operatorname{Regret}(T, \alpha) \leq 1 + 12 d^{1.25} \sqrt{C_{T,\delta} T}.$$

In the deterministic setting, $\operatorname{Regret}(T, \alpha) \leq 1 + 12 d^{1.5}$.

For an algorithm with a decaying estimation error, the term $C_{T,\delta}$ is sublinear in T, implying a sublinear regret in Theorem 6. For example, Algorithm 1 has a decaying estimation error with $C_{T,\delta}$ that scales logarithmically with T/δ for many standard function classes, and so the overall regret scales as $O(\sqrt{T \log T})$ (see Corollary 18 in Appendix A.1).

To derive PAC generalization guarantees, we apply a variant of online-to-batch reduction to any algorithm with large self-evaluations and a decaying estimation error. The resulting approach, shown in Algorithm 2, satisfies the following guarantee (proved in Appendix A.7):

Theorem 7 (PAC generalization)**.** *Let $\epsilon, \delta > 0$, and let $d = \overline{\mathsf{d}}_\rho(\mathcal{Z}, \alpha, \epsilon)$ for some set $\mathcal{Z}$, evaluation function ρ and $\alpha \leq \alpha^*$. Let* Alg *be an interactive estimation algorithm with α-large self-evaluations and a decaying estimation error with $C_{T,\delta} \geq \ln(2T)$, and suppose that we run Algorithm 2 with* Alg *as the base algorithm, $T \geq 64 d^{2.5} (C_{T,\delta/2})/\epsilon^2$, $n_1 = \lceil \log_2(4/\delta) \rceil$, and $n_2 = \lceil 128 \ln(8 n_1/\delta)/\epsilon^2 \rceil$. Then, with probability at least $1 - \delta$, the output $\hat{z} \in \mathcal{Z}$ satisfies*

$$\rho(\hat{z} \mid z^*) \geq \alpha - \epsilon,$$

and the overall number of issued queries is $O\big(\frac{d^{2.5}(C_{T,\delta/2}) + \ln^2(1/\delta)}{\epsilon^2}\big)$.

In the deterministic setting, it suffices to run Alg *with $T > 2 d^{1.5} \ln(4/\epsilon)$ and return $\hat{z} = z_{\hat{t}}$ where $\hat{t} = \operatorname{argmax}_{t \in \{1,\ldots,T\}} r_t$ is the index of the largest observed reward. Then, with probability 1, we obtain $\rho(\hat{z} \mid z^*) \geq \alpha - \epsilon$ and issue at most $O(d^{1.5} \ln(4/\epsilon))$ queries.*

Unknown α^* and optimism. Algorithms 1 and 2 achieve performance guarantees with respect to a provided optimality level $\alpha \leq \alpha^*$. When it is not easy to provide a non-trivial α (for example, when α^* is unknown and cannot be non-trivially bounded), Algorithm 3 uses the optimistic least squares algorithmic template (see, e.g., [24]) to ensure α^*-large self-evaluations with high probability and to achieve a sublinear $\operatorname{Regret}(T, \alpha^*)$. Algorithm 3 takes as input a confidence radius parameter R of the same order as the decaying estimation error parameter $C_{T,\delta}$ for Algorithm 1. We can then show that $z^* \in \mathcal{Z}_t$ with high probability for all $t \in \{1, \ldots, T\}$. Therefore, z_t must satisfy

Algorithm 3 Optimistic Interactive Estimation via Least Squares

1: **Input:** set of alternatives $\mathcal{Z}$, evaluation function ρ, number of steps T, confidence-set radius R.
2: **for** $t = 1, \ldots, T$ **do**
3: Compute confidence set

$$\hat{z}_t = \operatorname*{argmin}_{z \in \mathcal{Z}} \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - r_i\Big)^2,$$

$$\mathcal{Z}_t = \left\{z \in \mathcal{Z} : \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - \rho(z_i \mid \hat{z}_t)\Big)^2 \leq R\right\}.$$

4: Submit the query $z_t = \operatorname{argmax}_{z \in \mathcal{Z}_t} \rho(z \mid z)$.
5: Observe reward r_t.
6: **end for**

$\rho(z_t \,|\, z_t) \geq \rho(z^* \,|\, z^*) = \alpha^*$. In Appendix A.1 we show this modified version of the algorithm satisfies the decaying estimation error property. This technique allows us to achieve a sublinear Regret(T, α^*) without knowing α^* beforehand. Similar to the case of fixed α, it is possible to derive a version of Algorithm 3 that leverages an online regression oracle. (See Appendix A.3 for details.)

4 Statistical queries

In this section we consider the statistical query (SQ) model, as defined in Example 3. In particular, we study the connection between our generalized framework and SQ learning, showing specifically that the dissimilarity dimension can be used to recover generalization bounds based on a known combinatorial parameter that characterizes SQ learning, called the *strong SQ dimension*. There are several notions of such a dimension [12, 25]. Here we focus on the one due to Szörényi [25]:

Definition 4 (Strong SQ dimension, [25])**.** *For a fixed distribution D over $\mathcal{X}$, the* strong SQ dimension *of a hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ with respect to some $\epsilon > 0$, denoted $\mathtt{dim}_{\mathrm{SQ}}(\mathcal{H}, \epsilon)$, is the largest number d for which there exist $h_1, \ldots, h_d \in \mathcal{H}$ such that:*

(a) $|\langle h_i, h_j \rangle| \leq 1 - \epsilon$ *for all* $1 \leq i < j \leq d$*, and*
(b) $|\langle h_i, h_j \rangle - \langle h_{i'}, h_{j'} \rangle| \leq \frac{1}{d}$ *for all* $1 \leq i < j \leq d$, $1 \leq i' < j' \leq d$,

where $\langle h, h' \rangle := \mathbb{E}_{x \sim D}[h(x)h'(x)]$.

The dissimilarity and strong SQ dimensions are closely related to one another in the sense of each providing a kind of polynomial bound on the other, as stated in the next proposition (see Appendix B.1 for the proof).

Proposition 8. *Let D be a fixed distribution over $\mathcal{X}$, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypotheses class. For $\epsilon > 0$, let $d_{\mathrm{SQ}}(\epsilon) = \mathtt{dim}_{\mathrm{SQ}}(\mathcal{H}, \epsilon)$, and let $d_\rho(\epsilon) = \mathtt{d}_{\rho_{\mathrm{SQ}}}(\mathcal{H}, 1, \epsilon)$.*

If $d_\rho(\epsilon) \geq 2$ then

$$\min\Big\{d_{\mathrm{SQ}}(\epsilon),\ \big\lfloor 4\epsilon^2 \left(d_{\mathrm{SQ}}(\epsilon)\right)^2 \big\rfloor\Big\} \leq d_\rho(\epsilon) \leq \max\Big\{d_{\mathrm{SQ}}(\epsilon/4),\ 4\epsilon^2 \left(d_{\mathrm{SQ}}(\epsilon/4) + 1\right)^2\Big\}. \tag{3}$$

Similarly, if $d_\rho(4\epsilon) \geq 2$ then

$$\min\left\{d_\rho(4\epsilon),\ \left\lfloor \frac{\sqrt{d_\rho(4\epsilon)}}{8\epsilon} \right\rfloor\right\} \leq d_{\mathrm{SQ}}(\epsilon) \leq \max\left\{d_\rho(\epsilon),\ \frac{\sqrt{d_\rho(\epsilon) + 1}}{2\epsilon}\right\}. \tag{4}$$

We next give a lower bound based on the strong SQ dimension, which together with Proposition 8 will allow us to lower bound sample complexity of any interactive estimation algorithm in the SQ setting in terms of the dissimilarity dimension.

Theorem 9 (SQ lower bound)**.** *Let $\epsilon > 0$, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypothesis class with strong SQ dimension $d_{\mathrm{SQ}} = \mathtt{dim}_{\mathrm{SQ}}(\mathcal{H}, 2\epsilon) \geq 11$. Let* Alg *be any interactive estimation algorithm with the property that for any target $h^* \in \mathcal{H}$,* Alg *outputs an ϵ-approximation to h^* with probability at least $2/3$ using at most m queries. Then $m > \sqrt[3]{d_{\mathrm{SQ}}}/12$.*

The proof relies on a reduction to a lower bound of Szörényi [25]. However, the lower bound of Szörényi [25] holds within an SQ model that differs from ours, in that it allows adversarial query responses. Therefore, we first need to show how to obtain a learning algorithm Alg' that can be used with an adversarial oracle from an interactive estimation algorithm Alg that uses an unbiased stochastic query oracle (as we assume in this work). To do this, we apply the reduction technique developed by Feldman et al. [13]. (See Appendix B.2 for the full proof and additional details.)

Combining Theorem 9 and Proposition 8 yields a lower bound on the sample complexity of interactive estimation in the SQ setting, for a sufficiently small ϵ, in terms of the dissimilarity dimension:

Corollary 10. *Let $\epsilon > 0$, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypothesis class with strong SQ dimension $\mathtt{dim}_{\mathrm{SQ}}(\mathcal{H}, 2\epsilon) \geq 11$. Let $d_\rho(\epsilon) = \mathtt{d}_{\rho_{\mathrm{SQ}}}(\mathcal{H}, 1, \epsilon)$. Assume $\epsilon \leq 1/\big(2\sqrt{d_\rho(\epsilon)}\big)$. Let Alg be any interactive estimation algorithm with the property that for any target $h^* \in \mathcal{H}$, Alg outputs an ϵ-approximation to h^* with probability at least $2/3$ using at most m queries. Then $m > \sqrt[3]{d_\rho(\epsilon)}/12$.*

5 Bandits

In this section we focus on the bandits setting described in Example 2. We study the relationship between the dissimilarity dimension and the *eluder dimension* [24], a common combinatorial dimension for bounding regret of bandit algorithms. We show that eluder dimension can be used to upper bound the dissimilarity dimension, and we also highlight the cases when dissimilarity dimension leads to a tighter analysis.

Throughout this section we follow the setup introduced in Example 2. We consider an action set $\mathcal{A}$, a class $\mathcal{F}$ of reward functions $f : \mathcal{A} \to [-1, 1]$, and a target reward function $f^* \in \mathcal{F}$. We map this to our setting by considering the set of alternatives $\mathcal{Z} = \mathcal{F} \times \mathcal{A}$, evaluation function $\rho_{\text{bandits}}\big((f, a) \,\big|\, (f', a')\big) = f'(a)$ and the target (f^*, a^*), where $a^* = \operatorname{argmax}_{a \in \mathcal{A}} f^*(a)$.

5.1 Comparison with eluder dimension

We start by describing the relationship between our dimension and the eluder dimension. Following Russo and Van Roy [24], we define ϵ-dependence and ϵ-eluder dimension as follows:

Definition 5 (ϵ-dependence)**.** *An action $a \in \mathcal{A}$ is* ϵ-dependent *on actions $\{a_1, \ldots, a_n\} \subseteq \mathcal{A}$ with respect to $\mathcal{F}$ if any pair of functions $f, f' \in \mathcal{F}$ satisfying $\sqrt{\sum_{i=1}^n (f(a_i) - f'(a_i))^2} \leq \epsilon$ also satisfies $|f(a) - f'(a)| \leq \epsilon$. Furthermore, an action a is* ϵ-independent *of $\{a_1, \ldots, a_n\}$ with respect to $\mathcal{F}$ if it is not ϵ-dependent on $\{a_1, \ldots, a_n\}$.*

Definition 6 (ϵ-eluder dimension)**.** *The* ϵ-eluder *dimension $\mathtt{dim}_{\mathrm{E}}(\mathcal{F}, \epsilon)$ is the length d of the longest sequence of elements in $\mathcal{A}$ such that every element is ϵ-independent of its predecessors. Moreover, the* monotone eluder dimension *is defined as $\overline{\mathtt{dim}}_{\mathrm{E}}(\mathcal{F}, \epsilon) \coloneqq \max_{\epsilon' \geq \epsilon} \mathtt{dim}_{\mathrm{E}}(\mathcal{F}, \epsilon')$.*

The next theorem shows that the dissimilarity dimension is upper bounded by the eluder dimension (see Appendix C.1 for a proof):

Theorem 11. *Let $\mathcal{Z} = \mathcal{F} \times \mathcal{A}$, $\rho = \rho_{\text{bandits}}$, $\epsilon > 0$, $\alpha \leq \alpha^*$. Then $\overline{\mathtt{d}}_\rho(\mathcal{Z}, \alpha, 3\epsilon/2) \leq 9\, \overline{\mathtt{dim}}_{\mathrm{E}}(\mathcal{F}, \epsilon)$.*

Nevertheless, as the next example shows, the eluder dimension can be arbitrarily large, while the dissimilarity dimension remains constant. In this example, the action set is a circle in $\mathbb{R}^2$, that is, $\mathcal{A} = \mathcal{C} \coloneqq \{\mathbf{v} \in \mathbb{R}^2 : \|\mathbf{v}\| = 1\}$. We fix two open semicircles $U_0, U_1 \subseteq \mathcal{C}$ with positive x and y coordinates, respectively, and for any $N \in \mathbb{N}$ and $\epsilon > 0$, construct a function class $\mathcal{F}_{N,\epsilon}$ with all the functions $f : \mathcal{A} \to [-1, 1]$ obtained by the following process. First, pick one of the semicircles U_j and any N points from U_j. On each of these points, f can equal either $+\epsilon$ or $-\epsilon$. Everywhere else in U_j, f equals zero, and everywhere outside U_j, it equals the linear function $\langle \mathbf{v}, \mathbf{a} \rangle$ parameterized by some $\mathbf{v} \in \mathcal{C} \setminus U_j$. Thus, the functions are constructed to be "simple" (namely, linear) near the optimal action $\mathbf{v}$, but complex far from it. The eluder dimension is large to capture overall complexity, whereas the dissimilarity dimension is small to capture the simplicity near the optimum. (See Appendix C.5 for the formal construction of $\mathcal{F}_{N,\epsilon}$ and the proof of Proposition 12.)

Proposition 12. *Let $\epsilon \in (0, 1/2)$, $N \in \mathbb{N}$ and consider the action set $\mathcal{A} = \mathcal{C}$. Then, there is a function class $\mathcal{F}_{N,\epsilon} \subseteq [-1, 1]^{\mathcal{A}}$, such that for $\mathcal{Z}_{N,\epsilon} \coloneqq \mathcal{F}_{N,\epsilon} \times \mathcal{A}$, $\rho = \rho_{\text{bandits}}$, it holds that $\mathtt{d}_\rho(\mathcal{Z}_{N,\epsilon}, 1, \epsilon) \leq 16$, but the eluder dimension is lower bounded as $\mathtt{dim}_{\mathrm{E}}(\mathcal{F}_{N,\epsilon}, \epsilon) \geq N$.*

Thus, our regret bound based on the dissimilarity dimension implies that (optimistic) least squares algorithms have a regret independent of N. The same analysis with the eluder dimension [24] yields

a regret bound scaling polynomially with N. This shows that in the cases when the function classes are simple near the optimum, but complex far from it, the dissimilarity dimension can better capture the statistical complexity of bandit optimization than the eluder dimension.

5.2 Dissimilarity dimension bounds

We next derive dissimilarity dimension bounds for several standard bandit classes. Existing bounds on eluder dimension can be used to immediately bound the dissimilarity dimension, but in several cases we are able to obtain tighter bounds.

We first consider *linear bandits*. Let $\mathcal{B}_n = \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| \leq 1\}$ be the unit ball in $\mathbb{R}^n$. Actions are chosen from a set $\mathcal{A} \subseteq \mathcal{B}_n$; the reward function class is $\mathcal{F}^{lb} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subseteq \mathcal{B}_n$ and $f_{\boldsymbol{\theta}}(\mathbf{a}) = \langle \boldsymbol{\theta}, \mathbf{a} \rangle$. The corresponding set of alternatives is denoted $\mathcal{Z}^{lb} = \mathcal{F}^{lb} \times \mathcal{A}$. In this case we obtain the following bound (see Appendix C.2 for a proof):

Theorem 13 (Linear bandits)**.** *Let $\mathcal{Z}^{lb}$ be as defined above, let $\rho = \rho_{\text{bandits}}$, and let $\epsilon > 0$, $\alpha \leq \alpha^*$. Then $\mathtt{d}_\rho(\mathcal{Z}^{lb}, \alpha, \epsilon) \leq 4n + 3$. Moreover, when $\alpha = 1$, then $\mathtt{d}_\rho(\mathcal{Z}^{lb}, \alpha, \epsilon) \leq 2n + 1$.*

The proof proceeds by deriving an upper bound as well as a lower bound on the rank of the matrix $\mathbf{M}$ with entries $M_{ij} = \rho(z_i \mid z_j) - c$ obtained from elements $z_1, \ldots, z_d$ that satisfy the dimension condition for $d = \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon)$ with a scalar c. The upper bound on the rank is $n + 1$, and the lower bound is $d/4$ (which can be tightened to $d/2$ when $\mathbf{M}$ is symmetric). The upper bound is obtained by basic linear algebra and the lower bound from a standard result on ranks of perturbed identity matrices [2, Lemma 2.2]. Combining these bounds then yields the claim of Theorem 13. Similar to existing bounds on eluder dimension [24, Proposition 6], our bound in Theorem 13 is linear in n. However, the eluder dimension bound has an additional dependence on $1/\epsilon$, while our bound does not.

Next, we consider *generalized linear model* (GLM) bandits. Similar to linear bandits, the action set is $\mathcal{A} \subseteq \mathcal{B}_n$, but the function class includes a nonlinearity. Specifically, we are provided with a function $g : \mathbb{R} \to \mathbb{R}$ that is differentiable and strictly increasing, and consider the function class $\mathcal{F}^{glm} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ where $\Theta \subseteq \mathcal{B}_n$ and $f_{\boldsymbol{\theta}}(\mathbf{a}) = g(\langle \boldsymbol{\theta}, \mathbf{a} \rangle)$. Furthermore, we assume that there are $\underline{h}, \overline{h} > 0$ such that for all $\mathbf{a} \in \mathcal{A}$, $\boldsymbol{\theta} \in \Theta$, we have $\underline{h} \leq g'(\langle \boldsymbol{\theta}, \mathbf{a} \rangle) \leq \overline{h}$. Define $r = \overline{h}/\underline{h}$. We again denote $\mathcal{Z}^{glm} = \mathcal{F}^{glm} \times \mathcal{A}$. Using an existing bound on the eluder dimension for GLM bandits ([24], Proposition 7) and the fact that our dimension is bounded by the eluder dimension (Theorem 11), we obtain the following bound (see Appendix C.3 for a proof):

Theorem 14 (GLM bandits)**.** *Let $\mathcal{Z}^{glm}$ be as defined above, let $\rho = \rho_{\text{bandits}}$, and let $\epsilon > 0$, $\alpha \leq \alpha^*$. Then $\mathtt{d}_\rho(\mathcal{Z}^{glm}, \alpha, \epsilon) \leq O(nr^2 \log(\overline{h}/\epsilon))$.*

By considering a different proof technique, along the lines of Theorem 13, it might be possible to tighten this bound. We leave this extension for future work.

Next, we consider a bandit setting that is similar to GLMs, but in this case the non-linearity is provided by the non-differentiable rectified linear unit (ReLU) activation function $\mathit{relu}(x) = \max\{x, 0\}$. We consider the action set $\mathcal{A} = \mathcal{B}_n$, and the set of reward functions $\mathcal{F}^{\mathit{relu}}$ consisting of all functions of the form $f_{\boldsymbol{\theta},b}(\mathbf{a}) = \mathit{relu}(\langle \boldsymbol{\theta}, \mathbf{a} \rangle - b)$ for some $\boldsymbol{\theta} \in \mathcal{B}_n$ and $b \in [0, 1)$. The subset of $\mathcal{F}^{\mathit{relu}}$ with a fixed value of b is denoted $\mathcal{F}^{\mathit{relu}}_b$, and we consider the set of alternatives $\mathcal{Z}^{\mathit{relu}}_b = \mathcal{F}^{\mathit{relu}}_b \times \mathcal{B}_n$.

Unlike the classes considered above, this setting can be shown to be challenging to learn in the general case. Indeed, it turns out that eluder dimension (as well as a related measure called star dimension) is growing at least exponentially with n [21, 10]. The same lower bound can be shown for the dissimilarity dimension by a similar proof technique. The following theorem also provides an exponential upper bound, showing that in certain regimes the exponential dependence is tight (see Appendix C.4 for a proof):

Theorem 15 (ReLU bandits)**.** *Let $\mathcal{Z}^{\mathit{relu}}$ be as defined above, let $\rho = \rho_{\text{bandits}}$, and let $\epsilon, b > 0$ such that $b \leq 1 - \epsilon$. Then $\mathtt{d}_\rho(\mathcal{Z}^{\mathit{relu}}_b, 1 - b, \epsilon) = O\big(\epsilon^{-n/2}\big)$, and $\mathtt{d}_\rho(\mathcal{Z}^{\mathit{relu}}_{1-\epsilon}, \epsilon, \epsilon) = \Omega\big(\epsilon^{-n/2}\big)$.*

We note that previous work ([10], Theorem 5.1) has shown that for a function class of one-layer neural networks with ReLU activations, obtaining sublinear regret requires $T = \Omega(\epsilon^{-(n-2)})$.

6 Conclusion

In this paper, we have introduced a new model for interactive estimation and proposed a new combinatorial dimension, called dissimilarity dimension, to study the hardness of learning in this

model. In (stochastic, correlational) statistical query learning, our dimension is polynomially related to the strong SQ dimension. In bandits, our dimension is upper bounded by the eluder dimension, and there are examples where the dissimilarity dimension leads to much tighter regret bounds.

While this work provides an initial investigation of the dissimilarity dimension, many open questions remain. For example, our regret bound for the general setting scales as $d^{1.25}$. Is it possible to tighten this to linear dependence, as is the case, for example, for eluder dimension? On the algorithmic side, we currently require solving a least squares problem of size t in iteration t. Although we also introduce an algorithm that leverages an online regression oracle (see Appendix A.3), the oracle-based approach still requires solving a least squares problem (on the data smoothed by the oracle). Is it possible to derive dissimilarity-dimension-based regret bounds directly for the predictions produced by the oracle? Ultimately, we hope investigations of relationships between dissimilarity dimension and related notions may help us understand the hardness of learning in interactive settings.

Acknowledgments and Disclosure of Funding

Aldo Pacchiano would like to thank the support of the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. This work was supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard.

References

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[2] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability and Computing*, 18(1-2):3–15, 2009.

[3] Maria-Florina Balcan and Avrim Blum. On a theory of learning with similarity functions. In *Proceedings of the 23rd international conference on Machine learning*, pages 73–80, 2006.

[4] Shai Ben-David and Eli Dichterman. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.

[5] Shai Ben-David, Alon Itai, and Eyal Kushilevitz. Learning by distances. *Information and Computation*, 117(2):240–250, March 1995.

[6] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[7] Nader H Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2(Feb):359–395, 2002.

[8] Yihua Chen, Eric K Garcia, Maya R Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(3), 2009.

[9] Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

[10] Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34:26168–26182, 2021.

[11] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[12] Vitaly Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer and System Sciences*, 78(5):1444–1459, 2012.

[13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2): 1–37, 2017.

[14] Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[15] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[16] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[17] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[18] Adam Tauman Kalai and Santosh Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, pages 423–440, 2002.

[19] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[21] Gene Li, Pritish Kamath, Dylan J Foster, and Nathan Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

[22] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

[23] Lev Reyzin. Statistical queries and statistical algorithms: Foundations and applications. *arXiv preprint arXiv:2004.00557*, 2020.

[24] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[25] Balázs Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pages 186–200. Springer, 2009.

[26] P Turán. On an extremal problem in graph theory (in Hungarian). *Mat. Fiz. Lapok*, 48:436–452, 1941.

[27] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[28] Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[29] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

[30] Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. *Advances in Neural Information Processing Systems*, 26, 2013.

[31] Ke Yang. New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, 70(4):485–509, 2005.

A Missing proofs of Section 3

A.1 Analysis of Least Squares Algorithms (Algorithms 1 and 3)

Our analysis relies on the following variant of Freedman's inequality [16] (see Agarwal et al. [1, Lemma 9] and Beygelzimer et al. [6, Theorem 1]).

Lemma 16 (Simplified Freedman's inequality)**.** *Let $R > 0$ and let $X_1, \ldots, X_n$ be a sequence of real-valued random variables, such that for all $i \in [n]$ it holds that $X_i \le R$ and $\mathbb{E}[X_i | X_1, \ldots, X_{i-1}] = 0$. For any $\delta \in (0,1)$, and $\eta \in (0, 1/R)$, with probability at least $1-\delta$,*

$$\sum_{i=1}^{n} X_i \le \eta \sum_{i=1}^{n} \mathbb{E}[X_i^2 \mid X_1, \ldots, X_{i-1}] + \frac{\ln(1/\delta)}{\eta}. \tag{5}$$

Next we define an ϵ-cover of a set $\mathcal{Z}$, that will be used in the bound of Theorem 17.

Definition 7 (ϵ-cover)**.** *Let ψ be the pseudometric over the set $\mathcal{Z}$ defined, for any $z_1, z_2 \in \mathcal{Z}$, as*

$$\psi(z_1, z_2) = \sup_{z \in \mathcal{Z}} \Big| \rho(z \mid z_1) - \rho(z \mid z_2) \Big|. \tag{6}$$

We say a set $N \subseteq \mathcal{Z}$ is an ϵ-cover of $\mathcal{Z}$ with respect to ψ if for every $z \in \mathcal{Z}$ there exists some $z' \in N$ such that $\psi(z, z') \le \epsilon$. We denote by $\mathcal{N}(\mathcal{Z}, \epsilon)$ the minimum cardinality of any ϵ-cover of $\mathcal{Z}$.

For example, in the case of linear bandits (see Section 5.2) when $\mathcal{Z} = \mathcal{Z}^{lb}$ and $\rho = \rho_{\text{bandits}}$, it can be shown that $\mathcal{N}(\mathcal{Z}^{lb}, \epsilon)$ is upper bounded by the ℓ_2-covering number of the n-dimensional unit ball. This is because for any $z, z_1, z_2 \in \Theta \times \mathcal{A}$,

$$\big|\rho(z \mid z_1) - \rho(z \mid z_2)\big| = \big|\langle \boldsymbol{\theta}_1, \mathbf{a}\rangle - \langle \boldsymbol{\theta}_2, \mathbf{a}\rangle\big| \le \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \|\mathbf{a}\| \le \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

The bound on $\mathcal{N}(\mathcal{Z}^{lb}, \epsilon)$ now follows because the ℓ_2-covering number of the unit ball with radius ϵ is $O\big((3/\epsilon)^n\big)$ (see, for example, Lemma D.1 of Du et al. [11]).

We next show that Algorithm 1 satisfies the decaying estimation error property with $C_{T,\delta}$ that scales logarithmically with the covering number with respect to ψ.

Theorem 17 (LS guarantee)**.** *Consider the setting from Section 2, where the learner sequentially issues the queries $z_1, \ldots, z_T$ and receives responses $r_1, \ldots, r_T$. Assume there is $\beta \ge 0$ such that $\big|r_t - \mathbb{E}[r_t \mid z_t]\big| \le \beta$ for all t, and $\beta' \ge 2\beta$ such that for all z, z', $\big|\rho(z \mid z') - \rho(z \mid z^*)\big| \le \beta'$. Let $\tilde{\mathcal{Z}}$ be a set of alternatives such that $z^* \in \tilde{\mathcal{Z}}$ and let $\hat{z}_t$ be defined as the least squares optimizer,*

$$\hat{z}_t = \operatorname*{argmin}_{z \in \tilde{\mathcal{Z}}} \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - r_i\Big)^2.$$

Then, for any sequence of queries $z_1, \ldots, z_T \in \tilde{\mathcal{Z}}$ (possibly equal to $\hat{z}_1, \ldots, \hat{z}_T$), we have with probability $1-\delta$, for all $t \in [T]$ simultaneously,

$$\sum_{i=1}^{t-1} \Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \le C_{T,\delta}, \text{ and}$$

$$z^* \in \left\{ z \in \tilde{\mathcal{Z}} : \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - \rho(z_i \mid \hat{z}_t)\Big)^2 \le C_{T,\delta} \right\},$$

where $C_{T,\delta} = 16\beta\beta' \ln\big(2T\mathcal{N}(\tilde{\mathcal{Z}}, \beta'/T) \,/\, \delta\big)$.

Proof. For $i = 1, \ldots, T$, let $h_i = (z_1, r_1, \ldots, z_{i-1}, r_{i-1}, z_i)$ denote the history of interaction up to the query z_i, but excluding the response r_i, and let $\xi_i = r_i - \rho(z_i \mid z^*)$. In the interactive estimation setting, we then have $\mathbb{E}[\xi_i \mid h_i] = 0$, and by the lemma assumption, $\mathbb{E}[\xi_i^2 \mid h_i] \le \beta^2$.

Since $\hat{z}_t$ is the minimizer of the least squares loss up to time t, we have

$$\sum_{i=1}^{t-1} \Big(\rho(z_i \mid \hat{z}_t) - r_i\Big)^2 \le \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z^*) - r_i\Big)^2,$$

which can be rewritten, substituting $r_i = \rho(z_i \mid z^*) + \xi_i$, as

$$\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*) - \xi_i\Big)^2 \le \sum_{i=1}^{t-1}\xi_i^2.$$

Therefore, by re-arranging terms, we get

$$\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \le 2\sum_{i=1}^{t-1}\xi_i\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big). \tag{7}$$

Set $\epsilon_1 = \beta'/T$, and let N be a minimal ϵ_1-cover of $\tilde{\mathcal{Z}}$ with respect to the pseudometric ψ (see Eq. 6). Furthermore, let $\hat{z}_t^\epsilon \in N$ be an element of this cover that is ϵ_1-close to $\hat{z}_t$ (with respect to ψ). Then,

$$\begin{aligned}
\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 &\le 2\sum_{i=1}^{t-1}\xi_i\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big) \\
&= 2\sum_{i=1}^{t-1}\xi_i\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid \hat{z}_t^\epsilon) \\
&\qquad\qquad + \rho(z_i \mid \hat{z}_t^\epsilon) - \rho(z_i \mid z^*)\Big) \\
&\le 2t\beta\epsilon_1 + 2\sum_{i=1}^{t-1}\xi_i\Big(\rho(z_i \mid \hat{z}_t^\epsilon) - \rho(z_i \mid z^*)\Big),
\end{aligned} \tag{8}$$

where the first inequality follows from Eq. (7), and the last inequality follows because $|\xi_i| \le \beta$ and $\hat{z}_t^\epsilon$ is ϵ_1-close to $\hat{z}_t$.

Now, for any $z \in N$ and $i \in [T]$, define

$$K_i^z = \xi_i\Big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\Big).$$

Since $\mathbb{E}[\xi_i \mid h_i] = 0$, we have, for any *fixed* $z \in N$, $\mathbb{E}[K_i^z \mid h_i] = 0$. This means that for any fixed $z \in N$, $K_1^z, \ldots, K_T^z$ is a martingale difference sequence. By the lemma assumptions, $|K_i^z| \le \beta\beta'$. Also,

$$\mathbb{E}\big[(K_i^z)^2 \bigm| h_i\big] \le \beta^2\mathbb{E}\Big[\big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\big)^2 \Big| h_i\Big] = \beta^2\big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\big)^2. \tag{9}$$

Thus, by Freedman's inequality (Lemma 16) with $\eta = 1/(4\beta\beta')$ and $\delta' = \frac{\delta}{T|N|}$, we obtain that for any fixed $z \in N$ and $t \in [T]$, with probability at least $1 - \delta'$,

$$\begin{aligned}
\sum_{i=1}^{t-1}\xi_i\Big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\Big) &\le \frac{1}{4\beta\beta'}\sum_{i=1}^{t-1}\beta^2\Big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\Big) + 4\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big) \\
&= \frac{\beta}{4\beta'}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid z) - \rho(z_i \mid z^*)\Big)^2 + 4\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big).
\end{aligned} \tag{10}$$

Taking a union bound over all $z \in N$ and $t \in [T]$, we obtain that Eq. (10) holds with probability at least $1 - \delta$ simultaneously for all $z \in N$ and $t \in [T]$. Henceforth, we assume that we are in the event when Eq. (10) holds for all $z \in N$ and $t \in [T]$.

Applying the bound of Eq. (10) with $z = \hat{z}_t^\epsilon$ to the sum on the right-hand side of Eq. (8) then yields

$$\begin{aligned}
\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \le 2t\beta\epsilon_1 + \frac{\beta}{2\beta'}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t^\epsilon) - \rho(z_i \mid z^*)\Big)^2 \\
+ 8\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big).
\end{aligned} \tag{11}$$

Using the inequality $(a+b)^2 \le 2a^2 + 2b^2$, which holds for any $a, b \in \mathbb{R}$, and the fact that $\hat{z}_t^\epsilon$ and $\hat{z}_t$ are ϵ_1-close, we obtain, for every $i = 1, \ldots, t-1$,

$$\Big(\rho(z_i \mid \hat{z}_t^\epsilon) - \rho(z_i \mid z^*)\Big)^2 = \Big(\big[\rho(z_i \mid \hat{z}_t^\epsilon) - \rho(z_i \mid \hat{z}_t)\big] + \big[\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\big]\Big)^2$$

$$\leq 2\epsilon_1^2 + 2\big(\rho(z_i \mid z^*) - \rho(z_i \mid \hat{z}_t)\big)^2.$$

Plugging this into the right-hand side of Eq. (11) yields

$$\begin{aligned}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 &\leq 2t\beta\epsilon_1 + \frac{\beta}{\beta'}t\epsilon_1^2 + \frac{\beta}{\beta'}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \\ &\qquad + 8\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big) \\ &\leq 2t\beta\epsilon_1 + \frac{t\beta\epsilon_1^2}{\beta'} + \frac{1}{2}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \\ &\qquad + 8\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big),\end{aligned}$$

where the last inequality follows by the assumption that $\beta' \geq 2\beta$. Then, by re-arranging terms and multiplying by 2, we get

$$\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 \leq 4t\beta\epsilon_1 + \frac{2t\beta\epsilon_1^2}{\beta'} + 16\beta\beta'\ln\Big(\frac{T|N|}{\delta}\Big).$$

Recall that we set $\epsilon_1 = \beta'/T$, and N is a minimal ϵ_1-cover of $\tilde{\mathcal{Z}}$, so $|N| = \mathcal{N}(\tilde{\mathcal{Z}}, \beta'/T)$. Plugging these values in the previous equation, we thus obtain that with probability at least $1-\delta$, for all $t \in [T]$,

$$\begin{aligned}\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z^*)\Big)^2 &\leq 4\beta\beta' + \frac{2\beta\beta'}{T} + 16\beta\beta'\ln\Big(\frac{T\mathcal{N}(\tilde{\mathcal{Z}}, \beta'/T)}{\delta}\Big) \\ &\leq 16\beta\beta'\ln\Big(\frac{2T\mathcal{N}(\tilde{\mathcal{Z}}, \beta'/T)}{\delta}\Big), \qquad (12)\end{aligned}$$

where the last inequality follows because $4 + 2/T \leq 16\ln 2$ for $T \geq 1$. Finally, when Eq. (12) holds, we also have

$$z^* \in \Big\{z \in \tilde{\mathcal{Z}} : \sum_{i=1}^{t-1}\Big(\rho(z_i \mid z) - \rho(z_i \mid \hat{z}_t)\Big)^2 \leq 16\beta\beta'\ln\Big(\frac{2T\mathcal{N}(\tilde{\mathcal{Z}}, \beta'/T)}{\delta}\Big)\Big\}. \qquad \square$$

Considering Algorithm 1 and using Theorem 17 with $\tilde{\mathcal{Z}} = \mathcal{Z}_\alpha$, $z_t = \hat{z}_t$, $\beta = 2$ and $\beta' = 4$ then immediately yields the following corollary (α-large self-evaluations follow because $\hat{z}_t \in \mathcal{Z}_\alpha$):

Corollary 18. *Consider the setting from Section 2 with a set of alternatives $\mathcal{Z}$ and an evaluation function ρ. Let α be an optimality level such that $\mathcal{N}(\mathcal{Z}_\alpha, 4/T) = e^{o(T)}$. Then Algorithm 1 has α-large self-evaluations and satisfies the decaying error property with $C_{T,\delta} = 128\ln\big(2T\mathcal{N}(\mathcal{Z}_\alpha, 4/T)/\delta\big)$.*

Similarly, Theorem 17 also implies that Algorithm 3 satisfies the decaying error property as well as α^*-large self-evaluations, although α^* is not known:

Corollary 19. *Consider the setting from Section 2 with a set of alternatives $\mathcal{Z}$ and an evaluation function ρ, and assume that $\mathcal{N}(\mathcal{Z}, 4/T) = e^{o(T)}$. Then Algorithm 3 with $R = 128\ln\big(2T\mathcal{N}(\mathcal{Z}, 4/T)/\delta\big)$ has α^*-large self-evaluations and satisfies the decaying error property with $C_{T,\delta} = 4R = 512\ln\big(2T\mathcal{N}(\mathcal{Z}, 4/T)/\delta\big)$.*

Proof. We apply Theorem 17 with $\tilde{\mathcal{Z}} = \mathcal{Z}$, $\beta = 2$ and $\beta' = 4$. Our choice of R in Algorithm 3 coincides with the value of $C_{T,\delta}$ appearing in Theorem 17, and therefore the theorem implies that $z^* \in \mathcal{Z}_t$ with probability at least $1-\delta$ for all $t \in [T]$. In that case the queries z_t issued by the algorithm satisfy $\rho(z_t \mid z_t) \geq \rho(z^* \mid z^*) = \alpha^*$ and thus the algorithm has α^*-large self-evaluations.

For the second part, the triangle inequality implies that with probability at least $1-\delta$ for all $t \in [T]$,

$$\begin{aligned}\sqrt{\sum_{i=1}^{t-1}\Big(\rho(z_i \mid z) - \rho(z_i \mid z_t)\Big)^2} &\leq \sqrt{\sum_{i=1}^{t-1}\Big(\rho(z_i \mid z) - \rho(z_i \mid \hat{z}_t)\Big)^2} + \sqrt{\sum_{i=1}^{t-1}\Big(\rho(z_i \mid \hat{z}_t) - \rho(z_i \mid z_t)\Big)^2} \\ &\leq \sqrt{R} + \sqrt{R},\end{aligned}$$

where the bound on the first term on the right-hand side follows by Theorem 17 and the bound on the second term by the fact that $z_t \in \mathcal{Z}_t$. $\square$

A.2 Proof of Theorem 4

The theorem follows immediately from Corollary 18, because $\mathcal{N}(\mathcal{Z}_\alpha, \epsilon) \le |\mathcal{Z}_\alpha| \le |\mathcal{Z}| < \infty$ for any α and ϵ.

A.3 Online Regression Oracles

We assume access to an *online regression oracle* $\mathcal{R}eg$, which solves a regression problem over a function class $\Phi = \{\phi_z : z \in \mathcal{Z}\}$ indexed by $z \in \mathcal{Z}$, where $\phi_z : \mathcal{Z} \to \mathbb{R}$ is defined as $\phi_z(z') = \rho(z' \mid z)$ for all $z' \in \mathcal{Z}$; that is, functions ϕ_z evaluate ρ in its first argument.

The oracle operates in the following protocol: In each time step, the oracle receives an observation z_t, produces a prediction $\widehat{\rho}_t \in \mathbb{R}$, and finally receives a response r_t and incurs square loss $(\widehat{\rho}_t - r_t)^2$. We assume that for any T and sequence of observations and responses (even if generated adaptively), the oracle satisfies the following regret bound:

$$\sum_{t=1}^{T}(\widehat{\rho}_t - r_t)^2 - \inf_{z\in\mathcal{Z}}\sum_{t=1}^{T}(\rho(z_t \mid z) - r_t)^2 \le \mathrm{Regret}_{\mathcal{R}eg}(T), \tag{13}$$

where $\mathrm{Regret}_{\mathcal{R}eg}(\cdot)$ is a non-decreasing sublinear function (that typically also depends on various properties of ρ, $\mathcal{Z}$, and the range of responses r_t). For many function classes, there are well-known constructions of online regression oracles that satisfy Eq. (13) [9, 28, 18]. For example, if Φ is finite, there are oracles with $\mathrm{Regret}_{\mathcal{R}eg}(T) = O(\ln|\Phi|)$ and for parametric classes, such as linear functions, there are oracles with $\mathrm{Regret}_{\mathcal{R}eg}(T) = O(d\log(T/d))$. More examples can be found in Section 2.2 of Foster and Rakhlin [14].

Algorithm 4 Interactive Estimation via Least Squares

1: **Input:** online regression oracle $\mathcal{R}eg$, optimality level α.
2: Initialize z_1 to an arbitrary element of $\mathcal{Z}_\alpha$.
3: **for** $t = 1 \dots T$ **do**
4: Use $\mathcal{R}eg$ to predict $\widehat{\rho}_t$ given the observation z_t.
5: Observe reward r_{t+1} and pass it to $\mathcal{R}eg$.
6: Set $z_{t+1} = \operatorname*{argmin}_{z\in\mathcal{Z}_\alpha}\sum_{i=1}^{t}\Big(\rho(z_i \mid z) - \widehat{\rho}_i\Big)^2$.
7: **end for**

We now analyze Algorithm 4 under the assumption of access to an online regression oracle. This algorithm takes as input an online regression oracle $\mathcal{R}eg$. Algorithm 4 can also be modified using the optimistic least squares template of Russo and Van Roy [24] to handle the case when α^* is unknown. This is done by replacing step 6 of Algorithm 4 with the following two steps:

$$\mathcal{Z}_{t+1} = \Big\{z \in \mathcal{Z} : \sum_{i=1}^{t}(\rho(z_i \mid z) - \widehat{\rho}_i)^2 \le R\Big\},$$
$$z_{t+1} = \operatorname*{argmax}_{z\in\mathcal{Z}_{t+1}}\rho(z \mid z),$$

where $R = 8\mathrm{Regret}_{\mathcal{R}eg}(T) + 64\beta\max\{\beta,\beta'\}\ln\big(\frac{T}{\delta}\big)$ and β, β' are defined as in Lemma 20. The results of Lemma 20 justify the validity of these choices (using a similar reasoning as in Corollary 19) and imply that Algorithm 4 satisfies the decaying estimation error property (Definition 3), provided that the regression oracle $\mathcal{R}eg$ satisfies the regret bound of Eq. (13).

Lemma 20. *Consider the setting defined in Section 2 with $\alpha \le \alpha^*$, and assume there are $\beta, \beta' \ge 0$ such that $|r_t - \mathbb{E}[r_t \mid z_t]| \le \beta$ for all t and there is $\beta' \ge 2\beta$ s.t. for all $z \in \mathcal{Z}$ and $\widehat{\rho} \in \Gamma_z$, $\big|\widehat{\rho} - \rho(z \mid z^*)\big| \le \beta'$ where $\Gamma_z \subset \mathbb{R}$ is the space of plausible responses of $\mathcal{R}eg$ for input z. The sequence of queries $z_1, \dots, z_T$ as defined in Algorithm 4 satisfies with probability at least $1-\delta$ for all $t \in [T]$ simultaneously,*

$$\sum_{i=1}^{t}\left(\rho(z_i \mid z_{t+1}) - \rho(z_i \mid z^*)\right)^2 \le C_{T,\delta} \quad \text{and} \quad z^* \in \Big\{z \in \mathcal{Z}_\alpha : \sum_{i=1}^{t}(\rho(z_i \mid z) - \widehat{\rho}_i)^2 \le C_{T,\delta}\Big\}$$

where $C_{T,\delta} = 8\mathrm{Regret}_{\mathcal{R}eg}(T) + 64\beta\max\{\beta,\beta'\}\ln\big(\frac{T}{\delta}\big)$.

Proof. Let $\xi_i = r_i - \rho(z_i \mid z^*)$. Recall that $\mathbb{E}[r_i \mid z_i] = \rho(z_i \mid z^*)$ and therefore, by assumption, $|\xi_i| \leq \beta$ for all i. By definition, the online regression oracle satisfies

$$\begin{aligned}\sum_{i=1}^{t}(\widehat{\rho}_i - r_i)^2 &\leq \inf_{z\in\mathcal{Z}}\sum_{i=1}^{t}(\rho(z_i \mid z) - r_i)^2 + \mathrm{Regret}_{\mathcal{R}eg}(t)\\ &\overset{(i)}{\leq} \sum_{i=1}^{t}(\rho(z_i \mid z^*) - r_i)^2 + \mathrm{Regret}_{\mathcal{R}eg}(t)\\ &= \sum_{i=1}^{t}\xi_i^2 + \mathrm{Regret}_{\mathcal{R}eg}(t). \end{aligned} \tag{14}$$

Inequality (i) holds because $z^* \in \mathcal{Z}$. Expanding the LHS,

$$\sum_{i=1}^{t}(\widehat{\rho}_i - r_i)^2 = \sum_{i=1}^{t}\Big[(\widehat{\rho}_i - \rho(z_i \mid z^*))^2 - 2\xi_i(\widehat{\rho}_i - \rho(z_i \mid z^*)) + \xi_i^2\Big].$$

Plugging this back into Eq. (14) and rearranging, we obtain

$$\sum_{i=1}^{t}(\widehat{\rho}_i - \rho(z_i \mid z^*))^2 \leq \sum_{i=1}^{t} 2\xi_i(\widehat{\rho}_i - \rho(z_i \mid z^*)) + \mathrm{Regret}_{\mathcal{R}eg}(t). \tag{15}$$

For any $i \in [T]$ we define:

$$K_i = \xi_i(\widehat{\rho}_i - \rho(z_i \mid z^*))$$

Observe that

$$\mathbb{E}\left[K_i \mid \{z_\ell, \widehat{\rho}_\ell\}_{\ell=1}^{i}\right] = 0.$$

Thus $K_1, \ldots, K_T$ is a martingale difference sequence. Notice that $|K_i| \leq \beta\beta'$ and that

$$\mathbb{E}\big[K_i^2 \;\big|\; \{z_\ell, \widehat{\rho}_\ell\}_{\ell=1}^{i}\big] \leq \beta^2\mathbb{E}\Big[\big(\widehat{\rho}_i - \rho(z_i \mid z^*)\big)^2 \;\Big|\; \{z_\ell, \widehat{\rho}_\ell\}_{\ell=1}^{i}\Big] = \beta^2\big(\widehat{\rho}_i - \rho(z_i \mid z^*)\big)^2.$$

Then, by plugging this into Freedman's inequality (Lemma 16) with $\eta := \frac{1}{4}\min(1/\beta^2, 1/\beta\beta')$ and $\delta' := \delta/T$, we get that for any fixed $t \in [T]$, with probability at least $1-\delta$,

$$\sum_{i=1}^{t}\xi_i\big(\widehat{\rho}_i - \rho(z_i \mid z^*)\big) \leq \frac{1}{4}\sum_{i=1}^{t}\big(\widehat{\rho}_i - \rho(z_i \mid z^*)\big)^2 + 4\beta\max\{\beta, \beta'\}\ln\left(\frac{T}{\delta}\right). \tag{16}$$

Plugging Eq. (16) back into Eq. (15) and rearranging terms yields

$$\sum_{i=1}^{t}(\widehat{\rho}_i - \rho(z_i \mid z^*))^2 \leq 2\mathrm{Regret}_{\mathcal{R}eg}(t) + 16\beta\max\{\beta, \beta'\}\ln\left(\frac{T}{\delta}\right) \tag{17}$$

with probability at least $1-\delta$ for all $t \in [T]$. Thus we conclude that with probability at least $1-\delta$, for all $t \in [T]$,

$$z^* \in \Big\{z \in \mathcal{Z}_\alpha : \sum_{i=1}^{t}(\rho(z_i \mid z) - \widehat{\rho}_i)^2 \leq C_{T,\delta}\Big\}.$$

By the triangle inequality,

$$\sqrt{\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \rho(z_i \mid z_{t+1}))^2} \leq \sqrt{\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \widehat{\rho}_i)^2} + \sqrt{\sum_{i=1}^{t}(\widehat{\rho}_i - \rho(z_i \mid z_{t+1}))^2}.$$

Since by definition $z_{t+1} = \mathrm{argmin}_{z\in\mathcal{Z}_\alpha}\sum_{i=1}^{t}\Big(\rho(z_i \mid z) - \widehat{\rho}_i\Big)^2$, we have

$$\sum_{i=1}^{t}\Big(\rho(z_i \mid z_{t+1}) - \widehat{\rho}_i\Big)^2 \leq \sum_{i=1}^{t}\Big(\rho(z_i \mid z^*) - \widehat{\rho}_i\Big)^2.$$

Substituting back into the triangle inequality above,

$$\sqrt{\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \rho(z_i \mid z_{t+1}))^2} \leq 2\sqrt{\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \widehat{\rho}_i)^2},$$

implying

$$\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \rho(z_i \mid z_{t+1}))^2 \leq 4\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \widehat{\rho}_i)^2.$$

Plugging Eq. (17) on the right-hand side yields

$$\sum_{i=1}^{t}(\rho(z_i \mid z^*) - \rho(z_i \mid z_{t+1}))^2 \leq 8\mathrm{Regret}_{\mathcal{R}eg}(t) + 64\beta \max\{\beta, \beta'\} \ln\left(\frac{T}{\delta}\right).$$

The result follows by using the monotonicity of $\mathrm{Regret}_{\mathcal{R}eg}(t)$. □

A.4 Proof of Lemma 5

The proof uses Turán's Theorem [26], a standard result from extremal graph theory that bounds the number of edges of a graph that does not contain a clique of a given size:

Theorem 21 (Turán's Theorem)**.** *Let $G = (V, E)$ be an undirected graph without self-loops and whose largest clique is of size at most d. Then*

$$|E| \leq \left(1 - \frac{1}{d}\right)\frac{|V|^2}{2}.$$

We now turn to the proof of Lemma 5. First note that if $\epsilon \geq 1 + \alpha$ then no query is ϵ-bad, because $\alpha - \epsilon \leq -1 \leq \rho(z \mid z^*)$ for every $z \in \mathcal{Z}$, and therefore the lemma holds. In the remainder of the proof, we assume that $0 < \epsilon < 1 + \alpha$.

Consider the queries $z_1, \ldots, z_T$ and their corresponding values relative to z^*, denoted as $v_t = \rho(z_t|z^*)$ for $t \in [T]$. A query z_t is ϵ-bad if its corresponding value v_t is in the interval $I = [-1, \alpha-\epsilon)$. The proof proceeds by partitioning the interval I into subintervals and separately bounding the number of values v_t in each subinterval.

To define these subintervals, let $q = 1 + \frac{1}{\sqrt{d}}$, and consider the sequence of suboptimality gaps $\epsilon_i = q^{i-1}\epsilon$ for $i = 1, \ldots, n+1$, where

$$n = \left\lceil \log_q\left(\frac{1+\alpha}{\epsilon}\right)\right\rceil.$$

The gaps ϵ_i form an increasing sequence $\epsilon, q\epsilon, q^2\epsilon, \ldots$ such that the last element satisfies

$$\epsilon_{n+1} = q^n\epsilon \geq \left(\frac{1+\alpha}{\epsilon}\right)\epsilon = 1 + \alpha.$$

Using these gaps we define intervals $I_i = [\alpha - \epsilon_{i+1}, \alpha - \epsilon_i)$ for $i = 1, \ldots, n$. Since $\epsilon_{n+1} \geq 1 + \alpha$, the union $I_1 \cup \cdots \cup I_n = [\alpha - \epsilon_{n+1}, \alpha - \epsilon)$ covers the interval I. We bound the number of values v_t in each interval I_i.

Let S_i be the set of query indices with values in I_i, that is $S_i = \left\{t \in [T] : \rho(z_t \mid z^*) \in I_i\right\}$, let $m_i = |S_i|$, and assume that $m_i \geq 2$ (the case $m_i \leq 1$ will be dealt with later). Furthermore, let $c_i = \alpha - (\epsilon_{i+1} + \epsilon_i)/2$ be the midpoint of the interval I_i. Since the width of the interval I_i is $\epsilon_{i+1} - \epsilon_i = (q-1)\epsilon_i = \epsilon_i/\sqrt{d}$, we obtain

$$\left|\rho(z_t \mid z^*) - c_i\right| \leq \frac{\epsilon_i}{2\sqrt{d}} \tag{18}$$

for all $t \in S_i$.

Let $d_i = \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon_i)$ be the (non-monotonic) dissimilarity dimension with respect to the suboptimality gap ϵ_i. Since $\alpha \leq \alpha^*$ and $\epsilon_i \geq \epsilon$, we have $1 \leq d_i \leq d$. We construct an upper bound on m_i, exploiting the fact that d_i is the dissimilarity dimension with respect to ϵ_i.

In the rest of the proof we refer to a pair of queries with indices $s, t \in S_i$ such that $s < t$ as *dissimilar* if

$$\left|\rho(z_s \mid z_t) - c_i\right| \leq \frac{\epsilon_i}{\sqrt{d_i}}.$$

This is exactly the property appearing in the definition of the dissimilarity dimension with respect to ϵ_i, and so there cannot be more than d_i queries such that every pair is dissimilar (note that all the queries z_t satisfy $\rho(z_t \mid z_t) \geq \alpha$ thanks to α-large self-evaluations).

The derivation of the bound on m_i proceeds in several steps. First, we identify pairs of dissimilar queries and construct a graph where each edge corresponds to a dissimilar pair. Second, we use Turán's Theorem (Theorem 21) to upper bound the number of such pairs, using the fact that the graph cannot contain a clique of size greater than d_i. Finally, using the bound on the number of dissimilar pairs, we bound m_i.

To start, let $t_1 < t_2 < \ldots < t_{m_i}$ be the query indices included in S_i, and let $k \in \{2, \ldots, m_i\}$. Consider a uniform distribution over $\ell \in [k-1]$. Then by Markov's inequality and the decaying estimation error property, we obtain

$$\begin{aligned}
\frac{1}{k-1}\sum_{\ell=1}^{k-1} \mathbf{1}\left[\left(\rho(z_{t_\ell} \mid z_{t_k}) - \rho(z_{t_\ell} \mid z^*)\right)^2 \geq \frac{\epsilon_i^2}{4d}\right] &\leq \left[\frac{1}{k-1}\sum_{\ell=1}^{k-1}\left(\rho(z_{t_\ell} \mid z_{t_k}) - \rho(z_{t_\ell} \mid z^*)\right)^2\right] \cdot \frac{4d}{\epsilon_i^2} \\
&\leq \left[\frac{1}{k-1}\sum_{s=1}^{t_k-1}\left(\rho(z_s \mid z_{t_k}) - \rho(z_s \mid z^*)\right)^2\right] \cdot \frac{4d}{\epsilon_i^2} \\
&\leq \frac{C_{T,\delta}}{k-1} \cdot \frac{4d}{\epsilon_i^2}.
\end{aligned}$$

Multiplying by $k-1$, we therefore obtain

$$\left|\left\{s \in S_i : \; s < t_k \text{ and } \left|\rho(z_s \mid z_{t_k}) - \rho(z_s \mid z^*)\right| \geq \frac{\epsilon_i}{2\sqrt{d}}\right\}\right| \leq \frac{4dC_{T,\delta}}{\epsilon_i^2},$$

and summing across all $k \in \{2, \ldots, m_i\}$ then yields

$$\left|\left\{s, t \in S_i : \; s < t \text{ and } \left|\rho(z_s \mid z_t) - \rho(z_s \mid z^*)\right| \geq \frac{\epsilon_i}{2\sqrt{d}}\right\}\right| \leq m_i \frac{4dC_{T,\delta}}{\epsilon_i^2}. \tag{19}$$

We next construct an undirected graph without self-loops, $G_i = (V_i, E_i)$. The vertex set of the graph is $V_i = S_i$. The edge set is defined to be

$$E_i = \left\{\{t, s\} \subseteq S_i : \; s < t \text{ and } \left|\rho(z_s \mid z_t) - \rho(z_s \mid z^*)\right| < \frac{\epsilon_i}{2\sqrt{d}}\right\}.$$

By comparing with Eq. (19), we obtain

$$|E_i| \geq \frac{m_i(m_i-1)}{2} - m_i \frac{4dC_{T,\delta}}{\epsilon_i^2}. \tag{20}$$

Note that any pair of vertices $s < t$ connected by an edge corresponds to a dissimilar pair of queries:

$$\begin{aligned}
\left|\rho(z_s \mid z_t) - c_i\right| &\leq \left|\rho(z_s \mid z_t) - \rho(z_s \mid z^*)\right| + \left|\rho(z_s \mid z^*) - c_i\right| \\
&< \frac{\epsilon_i}{2\sqrt{d}} + \frac{\epsilon_i}{2\sqrt{d}} \\
&\leq \frac{\epsilon_i}{\sqrt{d_i}},
\end{aligned}$$

where the first inequality is the triangular inequality, the second inequality follows by combining the definition of E_i and Eq. (18), and the final one is from the fact that $d_i \leq d$. From the definition of the dissimilarity coefficient, the largest clique in G_i is of size at most d_i. Using Turán's Theorem, we thus must have

$$|E_i| \leq \left(1 - \frac{1}{d_i}\right) \cdot \frac{m_i^2}{2} \leq \left(1 - \frac{1}{d}\right) \cdot \frac{m_i^2}{2}.$$

Combining with the lower bound on $|E_i|$ from Eq. (20), we obtain

$$\frac{m_i(m_i-1)}{2} - m_i\frac{4dC_{T,\delta}}{\epsilon_i^2} \le \left(1-\frac{1}{d}\right)\cdot\frac{m_i^2}{2}.$$

Dividing by m_i, multiplying by $2d$, and rearranging then yields

$$m_i \le 2d\left(\frac{1}{2}+\frac{4dC_{T,\delta}}{\epsilon_i^2}\right) = d + \frac{8d^2C_{T,\delta}}{\epsilon_i^2}.$$

We have originally assumed that $m_i \ge 2$, but the bound that we have just derived also holds when $m_i \le 1$ (because $d \ge 1$).

To complete the proof it suffices to sum up the upper bounds on m_i across $i = 1, \ldots, n$:

$$\begin{aligned}\sum_{i=1}^n m_i &= \sum_{i=1}^n \left[d + \frac{8d^2C_{T,\delta}}{\epsilon^2}\cdot\left(1/q^2\right)^{i-1}\right] \\ &\le nd + \frac{8d^2C_{T,\delta}}{\epsilon^2}\cdot\frac{1}{1-(1/q^2)}. \end{aligned} \tag{21}$$

To bound n, we use the fact that $\alpha \le 1$, the inequality $\ln(1+x) \ge \frac{x}{1+x}$ (which holds for $x \ge 0$), and the fact that $d \ge 1$:

$$\begin{aligned} n &\le 1 + \log_q(2/\epsilon) \\ &= 1 + \frac{\ln(2/\epsilon)}{\ln\left(1+\frac{1}{\sqrt{d}}\right)} \le 1 + [\ln(2/\epsilon)]\cdot\frac{1+\frac{1}{\sqrt{d}}}{\frac{1}{\sqrt{d}}} = 1 + (\sqrt{d}+1)\ln(2/\epsilon) \\ &\le 2\ln 2 + 2\sqrt{d}\ln(2/\epsilon) \le 2\sqrt{d}\ln(4/\epsilon). \end{aligned}$$

Also,

$$1 - \frac{1}{q^2} = 1 - \frac{1}{1+\frac{2}{\sqrt{d}}+\frac{1}{d}} \ge 1 - \frac{1}{1+\frac{2}{\sqrt{d}}} = \frac{\frac{2}{\sqrt{d}}}{1+\frac{2}{\sqrt{d}}} = \frac{2}{\sqrt{d}+2} \ge \frac{2}{3\sqrt{d}}.$$

Plugging these back in Eq. (21) yields

$$\sum_{i=1}^n m_i \le 2d^{1.5}\ln(4/\epsilon) + \frac{12d^{2.5}C_{T,\delta}}{\epsilon^2},$$

completing the proof of the main claim of the lemma.

The second claim holds vacuously when $T = 0$, so assume that $T \ge 1$. If $C_{T,\delta} \ge \ln(2T)$, and using the fact that $(\ln x) \le x$ and $2 \le 3\ln 2$, we can write

$$2d^{1.5}\ln(4/\epsilon) = d^{1.5}\ln(16/\epsilon^2) \le \frac{16d^{1.5}}{\epsilon^2} \le \frac{24(\ln 2)d^{1.5}}{\epsilon^2} \le \frac{24d^{2.5}}{\epsilon^2}\cdot\ln(2T) \le \frac{24d^{2.5}C_{T,\delta}}{\epsilon^2},$$

which yields the first part of the second claim. The second part is immediate by plugging in $C_{T,\delta} = 0$ in the main claim.

A.5 Useful lemmas

In this subsection we prove two lemmas that will be needed for the proofs of the main results (Theorems 6 and 7) in Appendices A.6 and A.7. They both rely on a standard technique of bounding a sum by a definite integral:

Proposition 22. *Let $f : \mathbb{R} \to \mathbb{R}$ be a non-increasing function and $T \ge 1$. Then*

$$\sum_{t=1}^T f(t) \le f(1) + \int_1^T f(t)\mathrm{d}t.$$

Proof. The proof is immediate by noting that $f(t) \le \int_{t-1}^t f(t)\mathrm{d}t$. □

In the lemmas below we write $\mathbb{R}_+$ to denote $[0, +\infty)$.

Lemma 23. *Let $q_1, \dots, q_T$ be a sequence in $\mathbb{R}_+$, and let $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-increasing function such that for all $\epsilon > 0$,*

$$\sum_{t=1}^{T} \mathbf{1}(q_t \geq \epsilon) \leq \frac{\kappa(\epsilon)}{\epsilon^2}.$$

Then, for any $\tau \geq 0$,

$$\sum_{t=1}^{T} q_t \leq T\tau + 2\sqrt{\kappa(\tau)T}.$$

Proof. First, since we are only concerned with bounding the sum $\sum_t q_t$, we assume without loss of generality that the sequence is in descending order, i.e., $q_1 \geq \cdots \geq q_T$. Then, for any $\tau \geq 0$,

$$\sum_{t=1}^{T} q_t = \sum_{t=1}^{T} q_t \mathbf{1}(q_t \leq \tau) + \sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau) \leq T\tau + \sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau). \tag{22}$$

Consider any k such that $q_k > \tau$. Since the sequence $q_1, \dots, q_T$ is non-increasing, we have

$$k \leq \sum_{t=1}^{T} \mathbf{1}(q_t \geq q_k) \leq \frac{\kappa(q_k)}{q_k^2} \leq \frac{\kappa(\tau)}{q_k^2},$$

where the last inequality follows by the monotonicity of κ. This in turn implies that $q_k \leq \sqrt{\frac{\kappa(\tau)}{k}}$. Therefore,

$$\sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau) \leq \sum_{t=1}^{T} \sqrt{\frac{\kappa(\tau)}{t}}. \tag{23}$$

By Proposition 22,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + 2\sqrt{T} - 2\sqrt{1} < 2\sqrt{T}. \tag{24}$$

Combining Eqs. (22), (23) and (24), we get

$$\sum_{t=1}^{T} q_t \leq T\tau + 2\sqrt{\kappa(\tau)T}.$$

which concludes the proof. □

Lemma 24. *Let $a > 0$, let $q_1, \dots, q_T$ be a sequence of reals in $[0, a]$, and let $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-increasing function such that for all $\epsilon \in (0, a]$,*

$$\sum_{t=1}^{T} \mathbf{1}(q_t \geq \epsilon) \leq \kappa(\epsilon) \ln\left(\frac{a}{\epsilon}\right).$$

Then, for any $\tau \geq 0$,

$$\sum_{t=1}^{T} q_t \leq T\tau + a[1 + \kappa(\tau)] \exp\left(-\frac{1}{\kappa(\tau)}\right).$$

Proof. We follow a similar proof strategy as in Lemma 23 and start by bounding the sum $\sum_t q_t$. We assume without loss of generality that the sequence is in descending order, i.e., $q_1 \geq \cdots \geq q_T$. Then, for any $\tau \geq 0$,

$$\sum_{t=1}^{T} q_t = \sum_{t=1}^{T} q_t \mathbf{1}(q_t \leq \tau) + \sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau) \leq T\tau + \sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau). \tag{25}$$

Consider any k such that $q_k > \tau$. Then

$$k \leq \sum_{t=1}^{T} \mathbf{1}(q_t \geq q_k) \leq \kappa(q_k) \ln\left(\frac{a}{q_k}\right) \leq \kappa(\tau) \ln\left(\frac{a}{q_k}\right),$$

where the last inequality follows by the monotonicity of κ and the fact that $\ln(a/q_k) \geq 0$. This in turn implies that $q_k \leq a \exp\left(-\frac{k}{\kappa(\tau)}\right)$. Therefore,

$$\sum_{t=1}^{T} q_t \mathbf{1}(q_t > \tau) \leq \sum_{t=1}^{T} a \exp\left(-\frac{t}{\kappa(\tau)}\right). \tag{26}$$

By Proposition 22,

$$\begin{aligned}\sum_{t=1}^{T} \exp\left(-\frac{t}{\kappa(\tau)}\right) &\leq \exp\left(-\frac{1}{\kappa(\tau)}\right) - \kappa(\tau)\left(\exp\left(-\frac{T}{\kappa(\tau)}\right) - \exp\left(-\frac{1}{\kappa(\tau)}\right)\right) \\ &\leq [1+\kappa(\tau)] \exp\left(-\frac{1}{\kappa(\tau)}\right). \end{aligned} \tag{27}$$

Combining Eqs. (25), (26) and (27), we get

$$\sum_{t=1}^{T} q_t \leq T\tau + a[1+\kappa(\tau)] \exp\left(-\frac{1}{\kappa(\tau)}\right),$$

which concludes the proof. □

A.6 Proof of Theorem 6

Throughout the proof we use the shorthand $d_\epsilon = \overline{\mathrm{d}}_\rho(\mathcal{Z}, \alpha, \epsilon)$, so $d = d_{1/T}$. The proof proceeds by applying Lemmas 23 and 24 to the bounds on the number of bad queries from Lemma 5. Specifically, let $q_t = [\alpha - \rho(z_t \mid z^*)]_+$ denote the suboptimality of each query z_t made by the algorithm. Then, for any $\epsilon > 0$, the number of ϵ-bad queries can be written as $\sum_{t=1}^{T} \mathbf{1}(q_t \geq \epsilon)$.

First consider the case $C_{T,\delta} \geq \ln(2T)$. By Lemma 5, with probability at least $1-\delta$, the number of ϵ-bad queries is at most $36 d_\epsilon^{2.5} C_{T,\delta}/\epsilon^2$. Setting $\kappa(\epsilon) = 36 d_\epsilon^{2.5} C_{T,\delta}$, we apply Lemma 23, with $\tau = 1/T$, to obtain that with probability at least $1-\delta$,

$$\mathrm{Regret}(T, \alpha) \leq \sum_{t=1}^{T} q_t \leq 1 + 12 d^{1.25} \sqrt{C_{T,\delta} T}.$$

If $C_{T,\delta} = 0$, then by Lemma 5, the number of ϵ-bad queries is at most $2 d_\epsilon^{1.5} \ln(4/\epsilon)$. Setting $a = 4$ and $\kappa(\epsilon) = 2 d_\epsilon^{1.5}$, we apply Lemma 24, with $\tau = 1/T$, to obtain

$$\mathrm{Regret}(T, \alpha) \leq \sum_{t=1}^{T} q_t \leq 1 + 4(1 + 2d^{1.5}) \exp\left(-\frac{1}{2d^{1.5}}\right) \leq 1 + 12 d^{1.5},$$

completing the proof.

A.7 Proof of Theorem 7

First, we consider the deterministic setting. By Lemma 5, at most $2d^{1.5} \ln(4/\epsilon)$ of queries issued by Alg are ϵ-bad. Setting $T > 2d^{1.5} \ln(4/\epsilon)$ implies that at least one query is not ϵ-bad. Thus, returning $\hat{z}$ for which the observed reward is the largest guarantees that $\rho(\hat{z} \mid z^*) \geq \alpha - \epsilon$, as needed.

Next, we prove the result for the case $C_{T,\delta} \geq \ln(2T)$. By Lemma 5, with probability at least $1 - \delta/2$, there are at most $\frac{16}{9\epsilon^2} \cdot 36 d^{2.5}(C_{T,\delta/2})$ queries that are $3\epsilon/4$-bad. Setting $T \geq 64 d^{2.5}(C_{T,\delta/2})/\epsilon^2$, implies that at least half of the queries are not $3\epsilon/4$-bad. In the remainder of the proof, we only consider the high-probability event in which this is the case.

For $n_1 = \lceil \log_2(4/\delta) \rceil$ the probability that all n_1 samples are $3\epsilon/4$-bad is at most $(1/2)^{n_1} \leq \delta/4$.

For $n_2 = \lceil 128 \ln(8n_1/\delta)/\epsilon^2 \rceil$, by applying Hoeffding's inequality and union bound over each of the n_1 rounds we get that with probability at most $\delta/4$ there is some index $\ell \leq n_1$ for which $|\bar{r}_{t_\ell} - \rho(z_{t_\ell} \mid z^*)| > \epsilon/8$.

Overall, with probability at least $1 - \delta$ we get that there is at least one index j of the n_1 sampled indices that is not $3\epsilon/4$-bad, and that $|\bar{r}_{t_\ell} - \rho(z_{t_\ell} \mid z^*)| \leq \epsilon/8$ for all $\ell = 1, \ldots, n_1$. Therefore,

$$\bar{r}_{t_j} \geq \rho(z_{t_j} \mid z^*) - \epsilon/8 \geq \alpha - 3\epsilon/4 - \epsilon/8 = \alpha - 7\epsilon/8.$$

For all indices k that are ϵ-bad we have

$$\bar{r}_{t_k} \leq \rho(z_{t_k} \mid z^*) + \epsilon/8 < \alpha - \epsilon + \epsilon/8 = \alpha - 7\epsilon/8.$$

Thus, for all of the ϵ-bad queries we have $\bar{r}_{t_k} < \bar{r}_{t_j}$, and so Algorithm 2 will not return any of the ϵ-bad queries, because it is choosing the index with maximum value of $\bar{r}_{t_\ell}$. In other words, the returned query $z_{t_{\hat{\ell}}}$ satisfies

$$\rho(z_{t_{\hat{\ell}}} \mid z^*) \geq \alpha - \epsilon.$$

B Missing proofs of Section 4

First we discuss the connection between our SQ setting and the SQ model of Kearns [19]. We focus on two aspects in which they appear to differ and explain why these models are equivalent.

Correlational *vs* general statistical queries. The restriction of the SQ model in which the oracle may only output the approximate correlation between a query and the target function, termed *correlational statistical query* (CSQ), was studied by Bshouty and Feldman [7]. The CSQ oracle can be viewed as providing something akin to a negative distance between the query and the target. This is equivalent to the *learning by distances* framework of Ben-David et al. [5], who defined their model independently of Kearns [19]. Bshouty and Feldman [7] showed that an arbitrary statistical query can be answered by asking two SQs that are independent of the target and two CSQs. That is, in the distribution-dependent learning model (i.e., when the learner has access to the distribution over $\mathcal{X}$), correlational queries can simulate general queries.

Adversarial *vs* statistical noise. The setting we consider in this work assumes stochastic query responses, similar to several previous works [13, 31, 4]. On the other hand, the original SQ model [19] assumed that the query oracle can respond with an adversarial (rather than statistical) noise, up to a pre-specified tolerance parameter $\tau > 0$. The previous works have shown that the two noise models are equivalent [13, 31, 4]

B.1 Proof of Proposition 8

We first prove the first inequality of Eq. (3). Let $\epsilon > 0$ and let $d = d_{\mathrm{SQ}}(\epsilon)$. Then there exists a sequence $h_1, \ldots, h_d \in \mathcal{H}$ satisfying both conditions of Definition 4. Let d' be equal to the leftmost expression of Eq. (3). We aim to show $d_\rho(\epsilon) \geq d'$. Note that $d' \leq d$.

Let c be the midpoint between $c_{min} = \min_{i<j} \langle h_i, h_j \rangle$ and $c_{max} = \max_{i<j} \langle h_i, h_j \rangle$. Then $c \leq 1 - \epsilon$. Moreover, for all $i \neq j$,

$$|\langle h_i, h_j \rangle - c| \leq \frac{1}{2}|c_{max} - c_{min}| \leq \frac{1}{2d} \leq \frac{\epsilon}{\sqrt{d'}}$$

where the last inequality follows from our choice of d' (which ensures $d' \leq 4(d\epsilon)^2$). Thus, $h_1, \ldots, h_{d'}$, the first d' elements of the original sequence of hypotheses, satisfy Definition 1, proving the claim.

We prove the first inequality of Eq. (4) in a similar way. Let us re-define $d = d_\rho(4\epsilon)$ and let d' be equal to the leftmost expression of Eq. (4). As before, $d' \leq d$. Then there exists a sequence $h_1, \ldots, h_d \in \mathcal{H}$ satisfying the conditions of Definition 1 for some $c \leq 1 - 4\epsilon$. Then for all $i \neq j$, $\langle h_i, h_j \rangle \leq c + \frac{4\epsilon}{\sqrt{d}} \leq 1 - \epsilon$, since $d \geq 2$. Moreover, for all $i \neq j$ and $i' \neq j'$,

$$|\langle h_i, h_j \rangle - \langle h_{i'}, h_{j'} \rangle| = |\langle h_i, h_j \rangle - c + c - \langle h_{i'}, h_{j'} \rangle| \leq \frac{8\epsilon}{\sqrt{d}} \leq \frac{1}{d'},$$

with the last inequality following from our choice of d'. Thus, $h_1, \ldots, h_{d'}$, the first d' hypotheses in the original sequence, satisfy Definition 4.

The second inequality of Eq. (4), now follows from the first inequality of Eq. (3), since if the second inequality of Eq. (4) does not hold then the leftmost expression of Eq. (3) must be at least $d_\rho(\epsilon)$, a contradiction. Likewise, the second inequality of Eq. (3), now follows from the first inequality of Eq. (4).

B.2 Lower bound setting

Definition 8 (SQ oracle (adversarial))**.** *Let D be the input distribution over the domain X. For a* tolerance *parameter $\tau > 0$, $\mathcal{O}^{adv}(\tau) := \mathcal{O}^{adv}_{D,h^*}(\tau)$ oracle is the oracle that for any query function $h \in \mathcal{H}$, returns a value $v \in [\mu - \tau, \mu + \tau]$, where $\mu = \mathbb{E}_{x\sim D}[h(x)h^*(x)]$.*

Definition 9 (Sample oracle (statistical))**.** *Let D be the input distribution over the domain X. The Sample oracle $\mathcal{O} := \mathcal{O}_{D,h^*}$ oracle is the oracle that given any function $h \in \mathcal{H}$, takes an independent random sample x from D and returns the value $v = h(x)h^*(x)$.*

We will need the following results for our proof. The first is a reduction from an adversarial noise oracle to a statistical one. Specifically, consider the learning setting defined in Section 2, for a sample oracle $\mathcal{O}$. Let Alg be a (possibly randomized) algorithm for that setting. The following theorem shows a simulation of $\mathcal{O}$ via $\mathcal{O}^{adv}$ the SQ oracle.

Theorem 25 ([13], Theorem 3.13)**.** *Assume that* Alg *outputs a ϵ-approximation to h^* with probability at least δ, using m samples from $\mathcal{O}$. Then, for any $\delta' \in (0, 1/4]$, there exists a SQ algorithm* Alg$'$ *that uses at most m queries to $\mathcal{O}^{adv}(\delta'^2/m)$ and outputs an ϵ-approximation to h^* with probability at least $\delta - \delta'$.*

Their result is obtained by simulating Alg using $\mathcal{O}^{adv}$ as follows: for any query of Alg to $\mathcal{O}$, the response of $\mathcal{O}^{adv}$ to that query is used as bias for a coin flip, which is then given to the learner as the simulated outcome of $\mathcal{O}$. They then prove that the true m samples of $\mathcal{O}$ and the simulated coin flips are statistically close by bounding their distributional distance. This implies that the success probability of Alg$'$, the simulated algorithm, is not much worse than that of Alg, the original algorithm.

We note that the result originally stated in [13] differs from Theorem 25 above in two ways. First, it reduces to a *variant* of $\mathcal{O}^{adv}(\tau)$ with a tolerance $\tau' \in [\tau, \sqrt{\tau}]$. Thus, it holds for $\mathcal{O}^{adv}(\tau)$ as well. Second, it is phrased in a more general setting of search problems over distributions, which captures the SQ model, as detailed in [13], Section 6.

The second result that is needed for our proof is the following lower bound due to [25].

Theorem 26 ([25], Theorem 8)**.** *Let $\epsilon > 0$, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypothesis space with strong SQ dimension $d_{\text{SQ}} := \texttt{dim}_{\text{SQ}}(\mathcal{H}, 2\epsilon) \geq 3$ (see Definition 4). Then for any SQ algorithm* Alg *using m queries to $\mathcal{O}^{adv}(\tau)$ with tolerance $\tau \geq 2/\sqrt{d_{\text{SQ}}}$, there exist $h^* \in \mathcal{H}$ such that if* Alg *outputs an ϵ-approximation to h^*, then $m > d_{\text{SQ}}\tau^2/3$.*

B.2.1 Proof of Theorem 9

Set $\delta = 2/3$. Let D be a distribution over $\mathcal{X}$. Assume towards contradiction that there exists a learning algorithm Alg such that for any $h^* \in \mathcal{H}$, given oracle access to $\mathcal{O} := \mathcal{O}_{D,h^*}$ and using $m \leq \sqrt[3]{d_{\text{SQ}}}/12$ samples from $\mathcal{O}$, the algorithm Alg outputs an ϵ-approximation to h^* with probability at least δ.

We then apply Theorem 25 for $\delta' = \delta/2$ to simulate the algorithm using $\mathcal{O}^{adv} := \mathcal{O}^{adv}_{D,h^*}$. The resulting algorithm uses $m \leq \sqrt[3]{d_{\text{SQ}}}/12$ queries to $\mathcal{O}^{adv}(\tau)$ for $\tau = \delta'^2/m > 4/(3\sqrt[3]{d_{\text{SQ}}}) \geq 2/\sqrt{d_{\text{SQ}}}$ and has success probability of at least $\delta - \delta' = \delta/2 > 1/3$. By Theorem 26 we obtain a contradiction, as $m > d_{\text{SQ}}\tau^2/3 > \sqrt[3]{d_{\text{SQ}}}/2$.

C Missing proofs of Section 5

C.1 Proof of Theorem 11

Let $d = \overline{\texttt{d}}_\rho(\mathcal{Z}, \alpha, 3\epsilon/2)$. Note that the eluder dimension is always at least 1, so the theorem trivially holds if $d \leq 9$. In the remainder of the proof assume that $d \geq 10$.

From the definition of the monotonic dissimilarity dimension, there exists $\tau \geq 3\epsilon/2$ such that $d = \texttt{d}(\mathcal{Z}, \alpha, \tau)$. Let $(f_1, a_1), \ldots, (f_d, a_d)$ be a sequence satisfying the dimension conditions for τ.

We will show that the first $\lceil d/9 \rceil$ elements of this sequence also satisfy the conditions of the eluder dimension for some $\epsilon' \geq \epsilon$. Specifically, we will show that there is some $\epsilon' \geq \epsilon$ such that every element a_j with $j \leq \lceil d/9 \rceil$ in the sequence above is ϵ'-independent of its predecessors. That is, we will show that for every such element a_j, there exists a pair of functions $f, f' \in \mathcal{F}$ that satisfy

$$\sqrt{\sum_{i=1}^{j-1} \big(f(a_i) - f'(a_i)\big)^2} \leq \epsilon',$$

yet it also holds that $f(a_j) - f'(a_j) > \epsilon'$.

By definition of the dissimilarity dimension, there exists $c \leq \alpha - \tau$ such that for all $i < j$,

$$\Big|f_j(a_i) - c\Big| = \Big|\rho\big((f_i, a_i) \mid (f_j, a_j)\big) - c\Big| \leq \frac{\tau}{\sqrt{d}}. \tag{28}$$

Then, by the triangle inequality,

$$\big|f_j(a_i) - f_{j+1}(a_i)\big| = \big|f_j(a_i) - c + c - f_{j+1}(a_i)\big| \leq \big|f_j(a_i) - c\big| + \big|f_{j+1}(a_i) - c\big| \leq \frac{2\tau}{\sqrt{d}}. \tag{29}$$

Therefore,

$$\big(f_j(a_i) - f_{j+1}(a_i)\big)^2 \leq \frac{4\tau^2}{d}, \tag{30}$$

and so for all $j \leq \lceil d/9 \rceil$ it holds that,

$$\sum_{i=1}^{j-1} (f_j(a_i) - f_{j+1}(a_i))^2 < \frac{4\tau^2}{9}. \tag{31}$$

Next, recall that for all $j \leq d$ we have $f_j(a_j) \geq \alpha \geq c + \tau$ and $f_{j+1}(a_j) \leq c + \frac{\tau}{\sqrt{d}}$. Thus,

$$f_j(a_j) - f_{j+1}(a_j) \geq c + \tau - c - \frac{\tau}{\sqrt{d}} > \frac{2\tau}{3}, \tag{32}$$

where the last inequality holds for $d \geq 10$. Overall, Eqs. (31) and (32) then demonstrate that for $\epsilon' = 2\tau/3 \geq \epsilon$, the element a_j is ϵ'-independent of its predecessors, finishing the proof.

C.2 Proof of Theorem 13

Our proof uses the following result on ranks of perturbed identity matrices (see [2, Lemma 2.2]):

Lemma 27. *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a symmetric matrix such that $A_{ii} = 1$ for all i and $|A_{ij}| \leq 1/\sqrt{d}$ for all $i \neq j$. Then* $\operatorname{rank}(\mathbf{A}) > d/2$.

The proof begins by constructing a matrix $\mathbf{M}$ whose entries are derived from the evaluation values of elements that satisfy the dimension condition. Then we bound the rank of $\mathbf{M}$ from above as well as from below. The lower bound is expressed in terms of the dimension $d = \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon)$ while the upper bound is expressed in terms of n. Combining the bounds then yields the result of the theorem.

Construction of $\mathbf{M}$. Let $(f_{\boldsymbol{\theta}_1}, \mathbf{a}_1), \ldots, (f_{\boldsymbol{\theta}_d}, \mathbf{a}_d)$ denote the alternatives that satisfy the dimension conditions, with respect to some value c such that $c \leq \alpha - \epsilon$ (see Definition 1). Define $\mathbf{M}$ to be the $d \times d$ matrix with entries $M_{ij} = \langle \boldsymbol{\theta}_i, \mathbf{a}_j \rangle - c$ for $i, j \leq d$. Note that all diagonal entries of $\mathbf{M}$ are at least $\alpha - c \geq \epsilon$, and all other entries are in $\big[-\frac{\epsilon}{\sqrt{d}}, \frac{\epsilon}{\sqrt{d}}\big]$.

Upper bound on $\operatorname{rank}(\mathbf{M})$. Let $\mathbf{K} \in \mathbb{R}^{d \times d}$ be the matrix of inner products, $K_{ij} = \langle \boldsymbol{\theta}_i, \mathbf{a}_j \rangle$, and let $\mathbf{U}$ be the Gram matrix for the set of vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \mathbf{a}_1, \ldots, \mathbf{a}_d$. Then $\mathbf{U}$ is a $2d \times 2d$ matrix of the rank at most n, because the vectors are of the dimension n (see, e.g., [17, Theorem 7.2.10]), and $\mathbf{K}$ is a submatrix of $\mathbf{U}$, so $\operatorname{rank}(\mathbf{K}) \leq \operatorname{rank}(\mathbf{U}) \leq n$. Moreover, $\mathbf{M} = \mathbf{K} - c\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^d$. Therefore, by subadditivity of rank,

$$\operatorname{rank}(\mathbf{M}) = \operatorname{rank}(\mathbf{K} - c\mathbf{1}\mathbf{1}^\top) \leq \operatorname{rank}(\mathbf{K}) + \operatorname{rank}(-c\mathbf{1}\mathbf{1}^\top) \leq n + 1. \tag{33}$$

Lower bound on $\mathrm{rank}(\mathbf{M})$**.** Let $\mathbf{D} \in \mathbb{R}^{d\times d}$ be the diagonal matrix with entries $D_{ii} = 1/\sqrt{M_{ii}}$. Since $M_{ii} \geq \epsilon > 0$, we have $0 < D_{ii} \leq 1/\sqrt{\epsilon}$. Consider the matrix $\mathbf{M}' = \mathbf{DMD}$. Then

$$\mathrm{rank}(\mathbf{M}') = \mathrm{rank}(\mathbf{DMD}) = \mathrm{rank}(\mathbf{M}), \tag{34}$$

because the matrix $\mathbf{D}$ is non-singular (see [17, Section 0.4.6(b)]). Furthermore, matrix $\mathbf{M}'$ satisfies $M'_{ii} = 1$ for all i and

$$|M'_{ij}| = |D_{ii}M_{ij}D_{jj}| \leq \frac{1}{\sqrt{\epsilon}} \cdot \frac{\epsilon}{\sqrt{d}} \cdot \frac{1}{\sqrt{\epsilon}} = \frac{1}{\sqrt{d}}$$

for all $i \neq j$. Consider the symmetric matrix $\mathbf{S} = (\mathbf{M}' + (\mathbf{M}')^\top)/2$. Then, we also have $S_{ii} = 1$ for all i, and $|S_{ij}| \leq 1/\sqrt{d}$ for all $i \neq j$. Thus, by Lemma 27, $\mathrm{rank}(\mathbf{S}) > d/2$. Moreover, by the subadditivity of the rank

$$d/2 < \mathrm{rank}(\mathbf{S}) \leq \mathrm{rank}(\mathbf{M}'/2) + \mathrm{rank}\big((\mathbf{M}')^\top/2\big) = 2\,\mathrm{rank}(\mathbf{M}'). \tag{35}$$

Combining Eqs. (35), (34) and (33), we therefore obtain

$$d/2 < 2\,\mathrm{rank}(\mathbf{M}') = 2\,\mathrm{rank}(\mathbf{M}) \leq 2n + 2,$$

and so $d < 4n + 4$. Since d is an integer, we must have $d \leq 4n + 3$.

In the special case that $\alpha = 1$, we have $\langle \boldsymbol{\theta}_i, \mathbf{a}_i \rangle \geq 1$ for all $i \leq d$, which is only possible when $\boldsymbol{\theta}_i = \mathbf{a}_i$ for all $i \leq d$. As a result, the matrices $\mathbf{M}$ and $\mathbf{M}'$ are both symmetric, and thus $\mathbf{S} = \mathbf{M}'$ and

$$d/2 < \mathrm{rank}(\mathbf{S}) = \mathrm{rank}(\mathbf{M}') = \mathrm{rank}(\mathbf{M}) \leq n + 1,$$

implying that $d \leq 2n + 1$.

C.3 Proof of Theorem 14

Using an existing bound on the eluder dimension for GLM bandits ([24], Proposition 7), and the fact that our dimension is bounded by the eluder dimension (Theorem 11) the result follows.

C.4 Proof of Theorem 15

Denote $d = \mathtt{d}_\rho(\mathcal{Z}_b^{relu}, 1-b, \epsilon)$. Notice that since $b < 1$, for any $\boldsymbol{\theta}, \mathbf{a} \in \mathcal{B}_n$ such that $\boldsymbol{\theta} \neq \mathbf{a}$ it holds that $f_{\boldsymbol{\theta},b}(\mathbf{a}) < f_{\boldsymbol{\theta},b}(\boldsymbol{\theta}) = 1 - b$. Let $(f_{\boldsymbol{\theta}_1,b}, \boldsymbol{\theta}_1), \ldots, (f_{\boldsymbol{\theta}_d,b}, \boldsymbol{\theta}_d)$ be a sequence of elements satisfying the dimension definition, with respect to a corresponding scalar $c \leq 1 - b - \epsilon$. Since the evaluation is symmetric for ReLU functions, we can view this sequence as a set, and denote $U = \{\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_d\}$. In addition, note that by the dimension definition, for all $\boldsymbol{\theta} \in U$, $\|\boldsymbol{\theta}\| = 1$.

We start by proving an upper bound on d. Assume $d \geq 9$. First, consider the case $c \leq \epsilon/3$. Let U_0 be any subset of the unit sphere such that for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ in U_0 it holds that $\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle \leq b + 2\epsilon/3$. Observe that for all such $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ we have $f_{\boldsymbol{\theta},b}(\mathbf{u}') = f_{\boldsymbol{\theta}',b}(\boldsymbol{\theta}) \in [0, 2\epsilon/3]$. Thus, we get that $d \leq |U_0|$.

A standard sphere covering argument shows that the size of such a set is upper bounded as follows. The δ-covering number of the unit sphere is at most $(3/\delta)^n$ ([27], Cor. 4.2.13). Thus, there are at most $(3/\delta)^n$ points such that each pair $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ satisfies $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \geq \delta$, or equivalently $\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle \leq 1 - \delta^2/2$. By setting $\delta = \sqrt{2(1 - b - 2\epsilon/3)}$ we get that $|U_0| \leq (\frac{3}{\delta})^n \leq (\frac{3}{2\sqrt{2(\epsilon - 2\epsilon/3)}})^n \leq (4/\sqrt{\epsilon})^n$, which yields the desired bound.

Now, consider the case $c > \epsilon/3$. In this case, for all $i \neq j$, we have that $f_{\boldsymbol{\theta}_j,b}(\boldsymbol{\theta}_i) \geq c - \frac{\epsilon}{\sqrt{d}} \geq c - \epsilon/3 > 0$, and so $\langle \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \rangle > b$. Let $c' = c + b$. Thus, it must also hold that, $|\langle \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \rangle - c'| \leq \frac{\epsilon}{\sqrt{d}}$ for all $i \neq j$. Note that $c' < 1 - \epsilon$. Then, by applying Lemma 13 we get that d is upper bounded by $2n + 4$. Overall, the bound in the claim holds.

Next, we show a lower bound on $d = \mathtt{d}_\rho(\mathcal{Z}_{1-\epsilon}^{relu}, \epsilon, \epsilon)$, by lower bounding the size of the set U defined above. We now apply a sphere *packing* argument, which shows that there exists such a set U with size $|U| \geq (1/2\epsilon)^{n/2}$. We follow a similar argument as above. Specifically, the δ-packing number of the unit sphere is at most $(1/\delta)^n$ ([27], Cor. 4.2.13). By plugging in $\delta = \sqrt{2\epsilon}$, yields the desired bound.

C.5 Proof of Proposition 12

We start with an auxiliary lemma:

Lemma 28 (dissimilarity subadditivity)**.** *Let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be two sets, and let $\alpha \in \mathbb{R}$ and $\epsilon > 0$. Denote $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2$ and let $\rho : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be an evaluation function. Then,*

$$\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) \le \mathtt{d}_\rho(\mathcal{Z}_1, \alpha, \epsilon) + \mathtt{d}_\rho(\mathcal{Z}_2, \alpha, \epsilon)$$

and

$$\overline{\mathtt{d}}_\rho(\mathcal{Z}, \alpha, \epsilon) \le \overline{\mathtt{d}}_\rho(\mathcal{Z}_1, \alpha, \epsilon) + \overline{\mathtt{d}}_\rho(\mathcal{Z}_2, \alpha, \epsilon).$$

Proof. Let $z_1, \ldots, z_d \subseteq \mathcal{Z}$ such that there exists $c \le \alpha - \epsilon$ with $|\rho(z_i|z_j) - c| \le \frac{\epsilon}{\sqrt{d}}$ for all $i < j$, and $\rho(z_i|z_i) \ge \alpha$. Let $I_1, I_2 \subseteq [d]$ be disjoint sets of indices with $\{z_i\}_{i \in I_1} \subseteq \mathcal{Z}_1$ and $I_2 = [d] \setminus I_1$. Consider the sub-sequence $z_{\ell_1}, \ldots, z_{\ell_{|I_1|}}$ ordered by appearance in $z_1, \ldots, z_d$ of elements in I_1. By definition for all $i < j$,

$$|\rho(z_{\ell_i}|z_{\ell_j}) - c| \le \frac{\epsilon}{\sqrt{d}}$$

Therefore the sequence $z_{\ell_1}, \ldots, z_{\ell_{|I_1|}}$ satisfies $|\rho(z_{\ell_i}|z_{\ell_j}) - c| \le \frac{\epsilon}{\sqrt{|I_1|}}$ for all $i < j$ and therefore $|I_1| \le \mathtt{d}_\rho(\mathcal{Z}_1, \alpha, \epsilon)$. The same logic implies $|I_2| \le \mathtt{d}_\rho(\mathcal{Z}_2, \alpha, \epsilon)$.

To get the monotonic version, note that if $\epsilon^* = \arg\max_{\epsilon' \ge \epsilon} \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon')$,

$$\begin{aligned}
\overline{\mathtt{d}}_\rho(\mathcal{Z}, \alpha, \epsilon) &= \mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon^*) \\
&\le \mathtt{d}_\rho(\mathcal{Z}_1, \alpha, \epsilon^*) + \mathtt{d}_\rho(\mathcal{Z}_2, \alpha, \epsilon^*) \\
&\le \overline{\mathtt{d}}_\rho(\mathcal{Z}_1, \alpha, \epsilon) + \overline{\mathtt{d}}_\rho(\mathcal{Z}_2, \alpha, \epsilon)
\end{aligned}$$

where the first inequality is by the first statement of the lemma which was proved above, the next inequality is by definition of the monotonic dimension, and the last inequality follows by $\epsilon \le \epsilon^*$. □

We can now construct the classes that demonstrate the separation of the eluder and dissimilarity dimension.

We consider two overlapping semicircles, indexed by $j \in \{0, 1\}$, and defined as

$$U_0 = \Big\{(\cos x, \sin x) : \ x \in \Big(-\frac{\pi}{2}, \frac{\pi}{2}\Big)\Big\}, \text{ and } U_1 = \Big\{(\cos x, \sin x) : \ x \in \big(0, \pi\big)\Big\}.$$

For each $j \in \{0, 1\}$, and any $N \in \mathbb{N}$ and $\epsilon > 0$, we define the function class

$$\mathcal{F}_{j,N,\epsilon} \coloneqq \Big\{f_{\mathbf{v},S,\sigma} : \ \mathbf{v} \in \mathcal{C} \setminus U_j, \ S \subseteq U_j, \ |S| = N, \sigma \in \{\pm\epsilon\}^S\Big\},$$

containing functions

$$f_{\mathbf{v},S,\sigma}(\mathbf{a}) = \begin{cases} 0 & \text{if } \mathbf{a} \in U_j \setminus S, \\ \sigma(\mathbf{a}) & \text{if } \mathbf{a} \in S, \\ \langle \mathbf{v}, \mathbf{a} \rangle & \text{if } \mathbf{a} \in \mathcal{C} \setminus U_j. \end{cases} \tag{36}$$

In words, the functions in the class $\mathcal{F}_{j,N,\epsilon}$ are linear outside of the semicircle U_j, and zero in the semicircle U_j, except for a set of size N, where they can take any combination of values $+\epsilon$ and $-\epsilon$. For any $N \in \mathbb{N}$ and $\epsilon > 0$, we define the class $\mathcal{F}_{N,\epsilon} \coloneqq \bigcup_{j \in \{0,1\}} \mathcal{F}_{j,N,\epsilon}$ and show that this class has a constant dissimilarity dimension but its eluder dimension is at least N.

Finally, consider the action set $\mathcal{A} = \mathcal{C}$ and the function class $\mathcal{F}_{N,\epsilon}$ as defined above. Let $\mathcal{Z}_{N,\epsilon} = \mathcal{F}_{N,\epsilon} \times \mathcal{A}$, $\rho = \rho_{\text{bandits}}$ and $\epsilon \in (0, 1/2)$. We how show that $\mathtt{dim}_{\mathrm{E}}(\mathcal{F}_{N,\epsilon}, \epsilon) \ge N$, but $\mathtt{d}_\rho(\mathcal{Z}_{N,\epsilon}, 1, \epsilon) \le 16$. First we prove the following lower bound on the eluder dimension of $\mathcal{F}_{j,N,\epsilon}$.

Lemma 29. *The eluder dimension of $\mathcal{F}_{j,N,\epsilon}$ satisfies $\mathtt{dim}_{\mathrm{E}}(\mathcal{F}_{j,N,\epsilon}, \epsilon) \ge N$ for all $j \in \{0, 1\}$.*

Proof. Let $\mathbf{a}_1, \ldots, \mathbf{a}_N$ be an arbitrary set of points in U_j and consider the functions $\{f_i\}_{i=1}^{N+1} \subseteq \mathcal{F}_{j,N,\epsilon,S,\mathbf{v}}$ that for all $i, i' \le N$ satisfy:

$$f_i(\mathbf{a}_{i'}) = \begin{cases} \epsilon & \text{if } i' \ne i \\ -\epsilon & \text{if } i' = i, \end{cases}$$

and $f_{N+1}(\mathbf{a}_i) = \epsilon$ for all $i \leq N$. We now show that for all $i \leq N$, the action $\mathbf{a}_i$ is ϵ-independent of $\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}$ with respect to $\mathcal{F}_{j,N,\epsilon,S,\mathbf{v}}$. This holds since $\sqrt{\sum_{j=1}^{i-1}(f_i(\mathbf{a}_j) - f_{N+1}(\mathbf{a}_j))^2} = 0$ while $|f_i(\mathbf{a}_i) - f_{N+1}(\mathbf{a}_i)| = 2\epsilon > \epsilon$. This finalizes the proof. □

Lemma 30. *Denote $\mathcal{Z}_{N,\epsilon} = \mathcal{F}_{N,\epsilon} \times \mathcal{A}$ and let $\epsilon \in (0, 1/2)$. Then, $\mathtt{d}_\rho(\mathcal{Z}_{N,\epsilon}, 1, \epsilon) \leq 16$.*

Proof. Denote by $\mathcal{Z}_{j,N,\epsilon} = \mathcal{F}_{j,N,\epsilon} \times \mathcal{A}$ for all $j \in \{0, 1\}$. We start by showing that for all $j \in \{0, 1\}$, $\mathtt{d}_\rho(\mathcal{Z}_{j,N,\epsilon}, 1, \epsilon) \leq 8$. Let $z_1, \ldots, z_d$ be a maximal sequence certifying the dissimilarity dimension $\mathtt{d}_\rho(\mathcal{Z}_{N,\epsilon}, 1, \epsilon) \geq d$, i.e., it holds that,

$$|\rho(z_i|z_{i'}) - c| \leq \frac{\epsilon}{\sqrt{d}} \text{ for all } i < i', \qquad \text{while } \rho(z_i|z_i) \geq 1.$$

Since $\rho(z_i|z_i) \geq 1$, it must be the case that $z_i = (f_{\mathbf{v}_i,S_i,\sigma_i}, \mathbf{v}_i)$ with $\mathbf{v}_i \notin U_j$ (since otherwise the self evaluations would be strictly less than 1). This implies that $\rho(z_i|z_j) = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ for all $i < j$. Consequently, the score evaluations of all $z_1, \ldots, z_d$ are equivalent to the score evaluations of the linear problem defined by $\mathbf{v}_1, \ldots, \mathbf{v}_d$. Thus Theorem 13 implies the maximum length of such a sequence can be of size at most $2 \times 2 + 4 = 8$. Finally the sub-additivity of the dissimilarity dimension (see Lemma 28) implies,

$$\mathtt{d}_\rho(\mathcal{Z}_{N,\epsilon}, 1, \epsilon) \leq \sum_{j=0}^{1} \mathtt{d}_\rho(\mathcal{Z}_{j,N,\epsilon}, 1, \epsilon) \leq 16. \qquad \square$$

Combining the results of Lemmas 29 and 30 finalizes the proof of Proposition 12.

D Multi-Armed Bandits

In this section we explore the dissimilarity dimension of the K-armed bandit problem. In this setting the learner interacts with a set of K arms and at every step of a sequential interaction pulls an arm $a_t \in [K]$ and receives a reward r_t such that $\mathbb{E}[r_t] = \mu_{i_t}$ where μ_{a_t} is the mean reward of arm a_t. For simplicity we will assume $\mu_a \in [0, 1]$ for all $a \in [K]$ and that $|r_t| \leq 1$.

The K-armed bandit problem is an instance of structured bandits where $\mathcal{A} = [K]$ and $\mathcal{F} = [0, 1]^K$. The dissimilarity dimension of the K-armed bandit problem satisfies,

Proposition 31. *Consider the action set $\mathcal{A} = [K]$ and the function class $\mathcal{F} = [0, 1]^K$ as defined above. Let $\alpha \in [0, 1]$ and $\mathcal{Z} = \mathcal{F} \times \mathcal{A}$, $\rho = \rho_{\text{bandits}}$ and $\epsilon \in (0, 1/2)$. Then $\mathtt{d}_\rho(\mathcal{Z}, \alpha, \epsilon) \leq K$.*

Proof. Let $c \leq \alpha - \epsilon$ and $z_1, \ldots, z_d \in \mathcal{Z}$ with $z_i = (f_i, a_i)$ be a maximal sequence such that, $\rho(z_i|z_i) \geq \alpha$ while

$$|\rho(z_i|z_j) - c| \leq \frac{\epsilon}{\sqrt{d}}.$$

For $i < j$. Substituting the definition of ρ, this implies $f_i(a_i) \geq \alpha$ for all $i \in [K]$ while $|f_j(a_i) - c| \leq \frac{\epsilon}{\sqrt{d}}$ for all $i < j$. By definition of c, if $d \geq 2$

$$f_j(a_i) \leq \alpha - \epsilon + \frac{\epsilon}{\sqrt{d}} \leq \alpha - \left(1 - \frac{1}{\sqrt{2}}\right)\epsilon < \alpha - \frac{\epsilon}{4}, \quad \text{for all } i < j. \tag{37}$$

Let $I_i = \{a_\ell\}_{\ell=1}^{i}$ be the set of actions up to index i in the tuple sequence $z_1, \ldots, z_i$. Equation 37 implies that $f_j(a) \leq \alpha - \frac{\epsilon}{4}$ for all $j > i$. Since a_j satisfies $f_j(a_j) \geq \alpha > \alpha - \frac{\epsilon}{4}$ this implies $a_j \notin I_i$. We conclude that $a_i \neq a_j$ for all $i < j$. Since there are at most K different arm values, this implies $d \leq K$. □

D.1 Structured Bandits

We will now explain in detail how what Algorithm 1 reduces to in the structured bandits setting from Example 2. We write $z_i = (f_i, a_i)$ for all $i \in [T]$. The large evaluation set $\mathcal{Z}_\alpha$ can be reduced to the following set of functions,

$$\mathcal{F}_\alpha = \{f \in \mathcal{F} \text{s t.} \max_{a \in \mathcal{A}} f(a) \geq \alpha\}.$$

Algorithm 5 Interactive Estimation for Structured Bandits

1: **Input:** action set $\mathcal{A}$, function class $\mathcal{F}$, optimality level α, number of steps T. Compute large-evaluation function-action set $\mathcal{F}_\alpha = \{f \in \mathcal{F} \text{ s.t. } \max_{a\in\mathcal{A}} f(a) \geq \alpha\}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Compute regression function

$$f_t = \operatorname*{argmin}_{f\in\mathcal{F}_\alpha} \sum_{i=1}^{t-1} \left(f(a_i) - r_i\right)^2 .$$

4: Submit the query $a_t = \operatorname{argmax}_{a\in\mathcal{A}} f_t(a)$.
5: Observe reward r_t.
6: **end for**

Since the ρ_{bandits} function $\rho_{\text{bandits}}(z_i|z)$ is independent on a for $z = (f, a)$, the least squares equation $\operatorname{argmin}_{z\in\mathcal{Z}_\alpha} \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - r_i\Big)^2$ reduces to,

$$f_t = \operatorname*{argmin}_{f\in\mathcal{F}_\alpha} \sum_{i=1}^{t-1} \left(f(a_i) - r_i\right)^2 .$$

finally, to ensure the action query has a self evaluation of at least α, we output $a_t = \operatorname{argmax}_{a\in\mathcal{A}} f_t(a)$.

We will now explain in detail what the Optimistic Interactive Estimation Algorithm 3 reduces to in the structured bandits setting from Example 2. We write $z_i = (f_i, a_i)$ for all $i \in [T]$.

The least squares objective ($z_t = \operatorname{argmin}_{z\in\mathcal{Z}_\alpha} \sum_{i=1}^{t-1} \Big(\rho(z_i \mid z) - r_i\Big)^2$) can be written as,

$$\hat{f}_t = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{t-1} \left(f(a_i) - r_i\right)^2 .$$

The action component of the z element in this objective can be ignored since the ρ_{bandits} evaluation function does not depend on it. The confidence ball $\mathcal{Z}_t = \left\{ z \in \mathcal{Z} : \sum_{i=1}^{t-1} \Big(\rho(z_i \,|\, z) - \rho(z_i \,|\, \hat{z}_t)\Big)^2 \leq R \right\}$ reduces to,

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} \Big(f(a_i) - \hat{f}_t(a_i)\Big)^2 \leq R \right\}.$$

The query can be reduced to the action component of z_t,

$$a_t = \operatorname*{argmax}_{f,a\in\mathcal{F}_t\times\mathcal{A}} f(a).$$

Algorithm 6 summarizes this reduction and corresponds to the standard optimistic least squares for sturctured bandit problems from [24].

Algorithm 6 Optimistic Interactive Estimation for Structured Bandits

Input: action set $\mathcal{A}$, function class $\mathcal{F}$, confidence-set radius R, number of steps T.
for $t = 1, \ldots, T$ **do**
    Compute confidence set

$$\hat{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \left(f(a_i) - r_i\right)^2 .$$

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} \Big(f(a_i) - \hat{f}_t(a_i) \Big)^2 \leq R \right\}.$$

    Submit the query $a_t = \operatorname{argmax}_{f, a \in \mathcal{F}_t \times \mathcal{A}} f(a)$.
    Observe reward r_t.
end for