Understanding the Eluder Dimension

Gene Li
Toyota Technological
Institute at Chicago
gene@ttic.edu

Pritish Kamath
Google Research
pritish@alum.mit.edu

Dylan J. Foster
Microsoft Research
dylanfoster@microsoft.com

Nathan Srebro
Toyota Technological
Institute at Chicago
nati@ttic.edu

Abstract

We provide new insights on eluder dimension, a complexity measure that has been extensively used to bound the regret of algorithms for online bandits and reinforcement learning with function approximation. First, we study the relationship between the eluder dimension for a function class and a generalized notion of *rank*, defined for any monotone "activation" $\sigma : \mathbb{R} \to \mathbb{R}$, which corresponds to the minimal dimension required to represent the class as a generalized linear model. It is known that when σ has derivatives bounded away from 0, σ-rank gives rise to an upper bound on eluder dimension for any function class; we show however that eluder dimension can be exponentially smaller than σ-rank. We also show that the condition on the derivative is necessary; namely, when σ is the relu activation, the eluder dimension can be exponentially larger than σ-rank. For binary-valued function classes, we obtain a characterization of the eluder dimension in terms of star number and threshold dimension, quantities which are relevant in active learning and online learning respectively.

1 Introduction

Russo and Van Roy [37] introduced the notion of **eluder dimension** for a function class and used it to analyze algorithms (based on the *Upper Confidence Bound (UCB)* and *Thompson Sampling* paradigms) for the multi-armed bandit problem with function approximation. Since then, eluder dimension has been extensively used to construct and analyze the regret of algorithms for contextual bandits and reinforcement learning (RL) with function approximation [see, e.g., 46, 36, 44, 6, 14, 19, 12, 16, 26, 24]. Even though the eluder dimension has become a central technique for reinforcement learning theory, little is known about when exactly it is bounded. This paper makes progress toward filling this gap in our knowledge.

Russo and Van Roy [37] established upper bounds on eluder dimension for (i) function classes for which inputs have finite cardinality (the "tabular" setting), (ii) linear functions over $\mathbb{R}^d$ of bounded norm, and (iii) generalized linear functions over $\mathbb{R}^d$ of bounded norm, with any activation that has derivatives bounded away from 0. Apart from these function classes (and those that can be embedded into these), understanding of eluder dimension has been limited. Indeed, one might wonder whether a function class has "small" eluder dimension only if it can be realized as a class of (generalized) linear functions! This leads us to our first motivating question:

Question 1. *Are all function classes with small eluder dimension essentially generalized linear models?*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Answering this question has substantial ramifications on the scope of prior work. An answer of "yes" would imply that the results in the aforementioned work which gives regret guarantees in terms of eluder dimension do not go beyond already-established regret guarantees for generalized linear bandit or RL settings [see, e.g. 17, 32, 45, 31]. An answer of "no" can be construed as a positive result for RL theory, as it would indicate that existing (and future) results which use the eluder dimension apply to a richer set of function classes than generalized linear models.

To answer Question 1, we first formally define what it means for a function class to be written as a generalized linear model (GLM). Informally, for an activation $\sigma : \mathbb{R} \to \mathbb{R}$ and a function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, we define the **$\sigma$-rank** to be the smallest dimension d needed to express every function in $\mathcal{F}$ as a generalized linear function in $\mathbb{R}^d$ with activation σ (see Definition 3). Intuitively, the σ-rank captures the best possible upper bound on eluder dimension that the results from Russo and Van Roy [37] can give for a given $\mathcal{F}$ by treating it as a GLM with activation σ. We ask how the eluder dimension of any function class relates to its σ-rank for various activations σ. We show that the answer to Question 1 is indeed "no", i.e., there exists a function class with eluder dimension d but the σ-rank *for any* monotone σ is at least $\exp(\Omega(d))$. Thus, while Russo and Van Roy [37] show that the set of function classes with small eluder dimension is a *superset* of the set of GLM function classes, we (roughly speaking) show that the set of function classes with small eluder dimension is *strictly larger* than the set of GLM function classes.

We also prove that the requirement from Russo and Van Roy [37] that the derivative of the activation is bounded away from 0 is necessary in order to bound the eluder dimension of GLMs. The upper bound in the paper [37] becomes vacuous when the activation function has zero derivative; we show a lower bound which indicates this requirement cannot be dropped. Namely, when σ is the relu activation, we show that eluder dimension can be *exponentially larger* than σ-rank.

In a second line of inquiry, we study a *combinatorial* version of eluder dimension. The original definition of Russo and Van Roy [37] is defined for real-valued function classes, but one can specialize the definition to binary-valued function classes, leading to a so-called **combinatorial eluder dimension**. Thus, our second motivating question is:

Question 2. *Can we bound the combinatorial eluder dimension, perhaps in terms of more familiar learning-theoretic quantities?*

One might wonder: if the combinatorial eluder dimension is just a special case of the scale-sensitive version, why study it at all? Our reasons are threefold.

- The first and most immediate reason is that we are able to show new characterizations of eluder dimension once we move to the combinatorial definition. We elucidate a fundamental connection between eluder dimension and two other well-studied learning-theoretic quantities: (1) star number, a quantity that characterizes the label complexity of *pool-based active learning* [21], (2) threshold dimension, a quantity that characterizes the regret of *online learning* [3].[1] We believe that this new result may help us better understand how different learning tasks relate to each other.

- The second reason is that the combinatorial eluder dimension (or a multi-class variant of it) has already been studied for policy-based learning for contextual bandits and RL [see, e.g., 19, 34]. Thus, understanding the combinatorial eluder dimension may shed light on the challenges of policy-based RL.

- Our last reason has more philosophical bent. Historically, the discovery of VC dimension placed statistical learning theory on solid footing. Specifically, the fundamental theorem of statistical learning allows us to precisely characterize the statistical complexity of PAC learnability in terms of the combinatorial VC dimension. The insights from understanding the role of VC dimension in classification have led researchers to develop *scale-sensitive* complexity measures such as fat shattering dimension to provide sharper guarantees on learning. While we do not claim that the eluder dimension is fundamental to online RL and bandit settings, we believe that a better combinatorial understanding can lead to a better understanding of online decision making. In some sense, we are "working backwards" from the original scale-sensitive definition of eluder dimension to understand its combinatorial properties.

[1] Finiteness of the threshold dimension is equivalent to finiteness of Littlestone dimension [40, 22, 3].

1.1 Main contributions

In this work, we provide several results which show when eluder dimension can be bounded (or is unbounded). Our results can be separated into two categories: (1) those pertaining to the (scale-sensitive) eluder dimension (as originally defined by Russo and Van Roy [37]); (2) those pertaining to the combinatorial eluder dimension, defined for binary-valued function classes.

First, we investigate the relationship between eluder dimension and our notion of *generalized rank* that captures the realizability of any function class as a GLM. In Section 2, we formally introduce the eluder dimension (as well as a related quantity of the scale-sensitive star number). In Section 3, we formalize the notion of "generalized rank". In Section 4, we provide several results.

1. We show that eluder dimension can be *exponentially smaller* than σ-rank for any monotone activation σ, not just those with derivatives bounded away from 0 (Theorem 6).
2. We show that the condition on the derivative being bounded away from 0 is necessary for σ-rank to be an upper bound on eluder dimension. Namely, when σ is the relu activation, we show that eluder dimension can be *exponentially larger* than σ-rank (Theorem 7).[2]

In Section 5, we specialize the eluder dimension to the binary-valued setting and present our results on the combinatorial eluder dimension.

1. We show that eluder dimension is finite if and only if both star number and threshold dimension are finite. Specifically, in Theorem 8 we show the following:

$$\max\{\mathsf{Sdim}(\mathcal{F}), \mathsf{Tdim}(\mathcal{F})\} \leq \mathsf{Edim}(\mathcal{F}) \leq \exp(O(\max\{\mathsf{Sdim}(\mathcal{F}), \mathsf{Tdim}(\mathcal{F})\})).$$

   Furthermore, we demonstrate that both inequalities can be tight (see Theorem 9 and discussion above it).
2. We investigate the comparison between eluder dimension and sign-rk and prove stronger separations than in the scale-sensitive case; namely we show examples where one quantity is finite while the other is infinite (Theorem 10).

1.2 Related work

In this section, we highlight several related works.

Bounds on eluder dimension. Several papers provide bounds on eluder dimension for various function classes. The original bounds on tabular, linear, and generalized linear functions were proved by Russo and Van Roy [37] (and later generalized by Osband and Van Roy [36]). Mou et al. [34] provide several bounds for the combinatorial eluder dimension, mostly focusing on linear function classes. When the function class lies in an RKHS, Huang et al. [25] show that the eluder dimension is equivalent to the notion of *information gain* [41, 38], which can be seen as an infinite dimensional generalization of the fact that the eluder dimension for linear functions over $\mathbb{R}^d$ is $\tilde{\Theta}(d)$. In concurrent work, Dong et al. [12] also prove an exponential lower bound on the eluder dimension for ReLU networks.

Applications of eluder dimension. The main application of the eluder dimension is to design algorithms and prove regret guarantees for contextual bandits and reinforcement learning. A few examples include the papers [46, 36, 44, 6, 14, 19, 28, 16, 26, 24, 34]. While the majority of papers prove upper bounds via eluder dimension, Foster et al. [19] provided lower bounds for contextual bandits in terms of eluder dimension, if one is hoping for instance-dependent regret bounds. In addition, several works observe that eluder dimension sometimes does not characterize the sample complexity, as the guarantee via eluder dimension can be too loose [23, 20]. Beyond the online RL setting, eluder dimension has been applied to risk sensitive RL [15], Markov games [24, 29], representation learning [47], and active online learning [10].

Other complexity measures for RL. We also touch upon other complexity measures which have been suggested for RL. One category is Bellman/Witness rank approach [see e.g. 27, 11, 42], which is generalized to bilinear classes [13]. These complexity measures capture an interplay between

[2] This result was independently established by Dong et al. [12].

the MDP dynamics and the function approximator class; in contrast, eluder dimension is purely a property of the function approximator class and can be stated (and studied) without referring to an online RL problem. Jin et al. [28] define a *Bellman-Eluder dimension* which captures function classes which have small Bellman rank or eluder dimension. Lastly, Foster et al. [20] propose a *Decision-Estimation Coefficient* and prove that it is necessary and sufficient for sample-efficient interactive learning.

Notions of rank. The notion of rank we propose is a generalization of the classical notion of *sign rank*, also called *dimension complexity*. Sign rank has been studied extensively in combinatorics, learning theory, and communication complexity [see e.g. 1, 18, 5, 2, and references therein]. The norm requirements in our definition of σ-rank are related to the notion of *margin complexity* [5, 7, 30].

2 Eluder dimension and star number

Eluder dimension is a "sequential" notion of complexity for function classes, originally defined by Russo and Van Roy [37]. Informally speaking, it characterizes the longest sequence of *adversarially chosen* points one must observe in order to accurately estimate the function value at any other point. We consider a variant of the original definition, proposed by Foster et al. [19], that is never larger and is sufficient to analyze all the applications of eluder dimension in literature.

Definition 1. *For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, $f^\star \in \mathcal{F}$, and scale $\varepsilon \geq 0$, the* ***exact eluder dimension*** $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon)$ *is the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$|f_i(x_i) - f^\star(x_i)| > \varepsilon, \quad \textit{and} \quad \sum_{j<i} (f_i(x_j) - f^\star(x_j))^2 \leq \varepsilon^2. \tag{1}$$

Then for all $\varepsilon > 0$:

- *the* ***eluder dimension*** *is* $\mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon')$.
- $\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon)$ *and* $\mathsf{Edim}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon)$.

This definition is never larger than the original definition of Russo and Van Roy [37], which asks for a witnessing *pair* of functions $f_i, f_i' \in \mathcal{F}$ (the above restricts $f_i' = f^\star$). Hence, all lower bounds on our variant of eluder dimension immediately apply to the original definition. Moreover, all upper bounds on eluder dimension in this paper can also be shown to hold for the other definition (unless stated otherwise).

We also consider the closely related notion of *star number* defined by Foster et al. [19], which generalizes a combinatorial parameter introduced in the active learning literature by Hanneke and Yang [21] (we will denote it as Sdim for consistency). We study the combinatorial star number in more detail in Section 5. The *only* difference between the definitions of eluder dimension and star number is that $\sum_{j<i}$ is replaced by $\sum_{j \neq i}$, which makes the star number a "non-sequential" notion of complexity.

Definition 2. *For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, $f^\star \in \mathcal{F}$, and scale $\varepsilon \geq 0$, the* ***exact star number*** $\underline{\mathsf{Sdim}}_{f^\star}(\mathcal{F}, \varepsilon)$ *is the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$|f_i(x_i) - f^\star(x_i)| > \varepsilon, \quad \textit{and} \quad \sum_{j \neq i} (f_i(x_j) - f^\star(x_j))^2 \leq \varepsilon^2.$$

Then for all $\varepsilon > 0$:

- *the* ***star number*** *is* $\mathsf{Sdim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\mathsf{Sdim}}_{f^\star}(\mathcal{F}, \varepsilon')$.
- $\underline{\mathsf{Sdim}}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \underline{\mathsf{Sdim}}_{f^\star}(\mathcal{F}, \varepsilon)$ *and* $\mathsf{Sdim}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \mathsf{Sdim}_{f^\star}(\mathcal{F}, \varepsilon)$.

It immediately follows from these definitions that the star number is never larger than eluder dimension. On the other hand, the star number can be arbitrarily smaller than eluder dimension.

Proposition 1. *For all $\mathcal{F}$, $f^* \in \mathcal{F}$ and scale $\varepsilon \geq 0$, it holds that*[3]

$$\underline{\mathsf{Sdim}}_{f^*}(\mathcal{F}, \varepsilon) \;\leq\; \underline{\mathsf{Edim}}_{f^*}(\mathcal{F}, \varepsilon) \;\leq\; \min\{|\mathcal{X}|, |\mathcal{F}| - 1\}\ .$$

Proposition 2 (simplified from Foster et al. [19, Prop 2.3])**.** *For the class of threshold functions given as $\mathcal{F}_n^{\mathrm{th}} := \{f_i : [n] \to \{0,1\} \mid i \in [n+1]\}$, where $f_i(x) := \mathbb{1}\{x \geq i\}$, and for $f^\star = f_{n+1}$, it holds for all $\varepsilon < 1$ that $\underline{\mathsf{Sdim}}_{f^\star}(\mathcal{F}, \varepsilon) = 2$ and $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon) = n$.*

3 Generalized rank

Dimension complexity has been studied extensively in combinatorics, learning theory, and communication complexity [see e.g. 1, 18, 5, 2]. The classical notion of dimension complexity, also known as *sign rank*, corresponds to the smallest dimension required to embed the input space such that all hypotheses in the function class under consideration are realizable as halfspaces. We consider a generalized notion of rank that is specified for any particular activation $\sigma : \mathbb{R} \to \mathbb{R}$, and captures to the smallest dimension required to represent the function class as a GLM when σ is the activation. In what follows, we let $\mathcal{B}_d(R) := \left\{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\right\}$.

Definition 3. *For any $\sigma : \mathbb{R} \to \mathbb{R}$, the* **$\sigma$-rank** *of a function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$ at scale $R > 0$, denoted as σ-$\mathsf{rk}(\mathcal{F}, R)$, is the smallest dimension d for which there exists mappings $\phi : \mathcal{X} \to \mathcal{B}_d(1)$ and $w : \mathcal{F} \to \mathcal{B}_d(R)$ such that*[4]

$$\textit{for all } (x, f) \in \mathcal{X} \times \mathcal{F} \;\;:\;\; f(x) \;=\; \sigma(\langle w(f), \phi(x)\rangle), \tag{2}$$

or ∞ if no such d exists. For a collection of activation functions $\Sigma \subseteq (\mathbb{R} \to \mathbb{R})$, the **$\Sigma$-rank** *is*

$$\Sigma\text{-}\mathsf{rk}(\mathcal{F}, R) \;:=\; \min_{\sigma \in \Sigma} \sigma\text{-}\mathsf{rk}(\mathcal{F}, R).$$

Examples. We present some examples of Σ-rk that motivate our definition above.

1. **Threshold activation.** $\mathsf{sign}(z)$ yields the classic notion of *sign-rank* (equivalent to *dimension complexity*, as already mentioned). In this case, the scale R is irrelevant, so we denote $\mathsf{sign\text{-}rk}(\mathcal{F}) := \mathsf{sign\text{-}rk}(\mathcal{F}, R)$ for any R. Note that this quantity is meaningful only for $\mathcal{F} \subseteq (\mathcal{X} \to \{-1, 1\})$.
2. **Identity activation.** For $\mathsf{id}(z) := z$, $\mathsf{id\text{-}rk}(\mathcal{F}, R)$ is the smallest dimension needed to represent each $f \in \mathcal{F}$ as a (norm-bounded) linear function. We abbreviate $\mathsf{rk} := \mathsf{id\text{-}rk}$, as this corresponds to the standard notion of rank of the matrix $(f(x))_{x,f}$ (albeit with the additional norm constraint).
3. **Monotone activations.** For $L \geq \mu \geq 0$, $\mathcal{M}_\mu^L$ consists of all activations σ such that for all $z < z'$, it holds that $\mu \leq \frac{\sigma(z') - \sigma(z)}{z' - z} \leq L$ (for differentiable σ, this is equivalent to $\mu \leq \sigma'(z) \leq L$ for all $z \in \mathbb{R}$).[5] An important special case is when $\mu = 0$, and a particular $\sigma \in \mathcal{M}_0^1$ of interest is the rectified linear unit (ReLU) defined as $\mathsf{relu}(z) := \max\{z, 0\}$. For ease of notation, we denote $\mathcal{M}_\mu := \mathcal{M}_\mu^1$.
   While we will always be explicit about the Lipschitz constant, note that the scale of the Lipschitz constant L (and μ) is interchangable with the scale of R. In particular,
   $$\mathcal{M}_\mu^L\text{-}\mathsf{rk}(\mathcal{F}, R) \;=\; \mathcal{M}_{\mu/L}\text{-}\mathsf{rk}(\mathcal{F}, RL). \tag{3}$$
4. **All activations.** Σ^{all} consists of all activations σ. We mention this notion of rank only in passing, and we will not focus on it for the rest of the paper.

We present a result which relates the aforementioned quantities (proof in Appendix A).

[3] For the definition of eluder dimension considered by [37], an upper bound of $\min\{|\mathcal{X}|, \binom{|F|}{2}\}$ holds, which can be tight. This upper bound holds because the witnessing pair of functions (f_i, f_i') has to be distinct for each i.

[4] Note that only the product of the scales of ϕ and w is relevant. The definition remains equivalent if we let $\phi : \mathcal{X} \to \mathcal{B}_d(R_\phi)$ and $w : \mathcal{F} \to \mathcal{B}_d(R_w)$ for any R_ϕ and R_w such that $R = R_\phi \cdot R_w$.

[5] To prove upper bounds on eluder dimension, it suffices for this condition to hold only when $|z| \leq R$, [see e.g. 37]. Since we fix σ in our definition first and then consider σ-rank at different scales R, this weaker condition complicates our definitions. Note that at any specific scale R, we can always modify σ to satisfy the required constraint everywhere by extending it linearly whenever $|z| > R$.

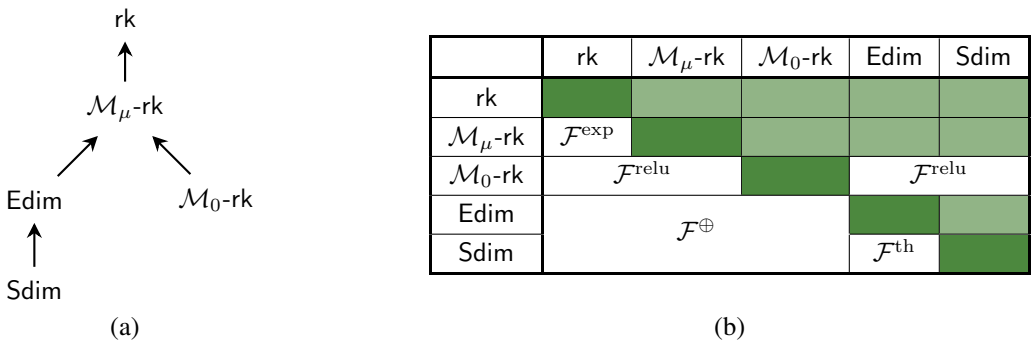


Figure 1: (a) Each arrow $M_1 \to M_2$ indicates that $M_1(\mathcal{F}) \lesssim M_2(\mathcal{F})$ for all $\mathcal{F}$, where the dependence on R and ε is suppressed for clarity (see Propositions 1, 3, 4 for precise bounds). Whenever $M_2 \to M_1$ arrow is missing, there is an example of a class $\mathcal{F}$ where $M_1(\mathcal{F}) \ll M_2(\mathcal{F})$. (b) An entry $\mathcal{F}$ in (M_1, M_2) means that $M_1(\mathcal{F}) \ll M_2(\mathcal{F})$. Green cells indicate that $M_1(\mathcal{F}) \gtrsim M_2(\mathcal{F})$ for all $\mathcal{F}$.

Proposition 3. *For all $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, $R > 0$ and $\mu \in (0, 1]$, we have:*

$$\mathsf{rk}(\mathcal{F}, R) \;\geq\; \mathcal{M}_\mu\text{-}\mathsf{rk}(\mathcal{F}, R) \;\geq\; \mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}, R) \;\geq\; \mathsf{sign}\text{-}\mathsf{rk}(\mathcal{F}) - 1,$$

where the last inequality is meaningful only for $\mathcal{F} \subseteq (\mathcal{X} \to \{-1, 1\})$. Moreover, for each inequality above, there exists a function class $\mathcal{F}$ which exhibits an infinite gap between the two quantities.

4 Eluder dimension versus generalized rank

In this section, we compare eluder dimension and star number with each notion of generalized rank: rk, $\mathcal{M}_\mu$-rk (for $\mu > 0$) and $\mathcal{M}_0$-rk. Our results are summarized in Figure 1.

Eluder vs. rk and $\mathcal{M}_\mu$-rk. Russo and Van Roy [37] and Osband and Van Roy [36] provided upper bounds on eluder dimension for linear and generalized linear function classes. For completeness, we restate this result, with a slight improvement and include the proof with precise dependence on problem parameters in Appendix B. Intuitively, the generalized linear rank allows us to capture the tightest possible upper bound that the guarantees in the papers [37, 36] can provide on eluder dimension.

Proposition 4 (cf. [37], Prop. 6, 7; [36], Prop. 2, 4)**.** *For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$ and $\varepsilon > 0$:*

(i) For all $R > 0$, $\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) \;\leq\; \mathsf{rk}(\mathcal{F}, R) \cdot O\left(\log \frac{R}{\varepsilon}\right)$.

(ii) For all $L, \mu, R > 0$, $\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) \;\leq\; \mathcal{M}_\mu^L\text{-}\mathsf{rk}(\mathcal{F}, R) \cdot O\left(\frac{L^2}{\mu^2} \log\left(\frac{RL}{\varepsilon}\right)\right)$.

This result has been used to prove upper bounds on eluder dimension of various function classes beyond GLMs; for example, the class of bounded degree polynomials, by taking the feature map $\phi(x)$ to be the vector of low degree monomials. The upper bound in Part (i) is in fact tight (up to constants) for the class of linear functions, as shown in Proposition 5 below. This trivially implies the optimality of the bound in Part (ii) up to the factor of $(L/\mu)^2$ which to the best of our knowledge is open.

Proposition 5 ([33])**.** *For any $R > 0$, $\mathcal{X} := \mathcal{B}_d(1)$ and $\mathcal{F} := \{f_\theta : x \mapsto \langle \theta, x \rangle \mid \theta \in \mathcal{B}_d(R)\}$, it holds that:*

$$\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) \;\geq\; \Omega\left(d \log\left(\frac{R}{\varepsilon}\right)\right) .$$

For completeness, we include the proof in Appendix C.

Eluder vs. $\mathcal{M}_0$-rk. It turns out that eluder dimension and $\mathcal{M}_0$-rk are incomparable. That is, there exists a function class for which eluder dimension is exponentially smaller than $\mathcal{M}_0$-rk (and hence $\mathcal{M}_\mu$-rk and rk by Proposition 3). Moreover, there exists a different function class for which eluder dimension (even the star number) is exponentially larger than relu-rk (and hence $\mathcal{M}_0$-rk).

First, we show that the eluder dimension can be exponentially smaller than $\mathcal{M}_0$-rk for the class of parities over d bits. Thus, parities over d bits exhibits an example where the eluder dimension is exponentially smaller than the best possible bound one can derive using the existing results of Proposition 4.

Theorem 6. *For $\mathcal{X} = \{-1,1\}^d$ and $\mathcal{F}^{\oplus} := \left\{ f_S : x \mapsto \prod_{i \in S} x_i \mid S \subseteq [d] \right\}$, it holds that*

(i) $\mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}^{\oplus}, R) \;\geq\; 2^{d/2} - 1$ *for all* $R > 0$.

(ii) $\underline{\mathsf{Sdim}}(\mathcal{F}^{\oplus}, \varepsilon) \;\leq\; \underline{\mathsf{Edim}}(\mathcal{F}^{\oplus}, \varepsilon) \;\leq\; d$ *for all* $\varepsilon \geq 0$.

Proof. Part (i). From Proposition 3, we have that $\mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}^{\oplus}, R) \geq \mathsf{sign\text{-}rk}(\mathcal{F}^{\oplus}) - 1$ for any $\sigma \in \mathcal{M}_0$. The proof is now complete by noting a well known result that $\mathsf{sign\text{-}rk}(\mathcal{F}^{\oplus}) \geq 2^{d/2}$ [18].

Part (ii). For any $x \in \{-1,1\}^d$ consider its 0-1 representation $\widetilde{x} \in \mathbb{F}_2^d$ (representing $+1$ by 0 and -1 by 1). All functions in $\mathcal{F}^{\oplus}$ can be simply viewed as linear functions over $\mathbb{F}_2$. Namely, any parity function is indexed by a vector $a \in \mathbb{F}_2^d$, with $f_a(x) := (-1)^{\langle a, \widetilde{x} \rangle}$.

Note that $\underline{\mathsf{Edim}}(\mathcal{F}^{\oplus}, \varepsilon) = 0$ for all $\varepsilon \geq 2$. For any $\varepsilon < 2$, suppose $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}^{\oplus}, \varepsilon) = m$, witnessed by $(x_1, f_{a_1}), \ldots, (x_m, f_{a_m}) \in \{-1,1\}^d$ and $f^\star = f_{a^\star}$. We have

- $f_{a_i}(x_i) \neq f_{a^\star}(x_i)$, and
- $f_{a_i}(x_j) = f_{a^\star}(x_j)$ for all $j < i$: since $\sum_{j<i} (f_{a_i}(x_j) - f_{a^\star}(x_j))^2 < \varepsilon^2 < 4$ iff all terms are 0.

Thus, we have $\langle a_i - a^\star, \widetilde{x} \rangle = 0$ for all $\widetilde{x} \in \mathbb{F}_2\text{-span}(\{\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}\})$. But $\langle a_i - a^\star, \widetilde{x}_i \rangle = 1$ and hence $\widetilde{x}_i$ is linearly independent of $\{\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}\}$ over $\mathbb{F}_2^d$. Thus, $\{\widetilde{x}_1, \ldots, \widetilde{x}_m\}$ are all linearly independent over $\mathbb{F}_2^d$, and hence $m \leq d$. □

The bound in part (ii) of Theorem 6 was also calculated by Mou et al. [34, Prop. 3].

Next, we show a separation in the other direction for eluder dimension vs. $\mathcal{M}_0$-rk using the ReLU function class. Thus, we cannot hope to remove the requirement for the activation function σ to be strictly monotonically increasing in Proposition 4 (ii) for bounding the eluder dimension.

Theorem 7. *Let $R > 0$ and $\mathcal{X} = \mathcal{B}_d(1)$. Define*

$$\mathcal{F}^{\mathrm{relu}} := \left\{ f_{\theta,b} : x \mapsto \mathsf{relu}(\langle \theta, x \rangle + b) \mid \|\theta\|^2 + b^2 \leq R^2 \right\}.$$

It holds that

(i) $\mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}^{\mathrm{relu}}, R) \leq \mathsf{relu\text{-}rk}(\mathcal{F}^{\mathrm{relu}}, R) \leq d + 1$,

(ii) $\underline{\mathsf{Edim}}(\mathcal{F}^{\mathrm{relu}}, \varepsilon) \geq \underline{\mathsf{Sdim}}(\mathcal{F}^{\mathrm{relu}}, \varepsilon) \geq \left(\frac{R}{4\varepsilon}\right)^{d/2}$ *for all* $\varepsilon \in (0, \frac{R}{4})$.

Proof. Part (i) is immediate from the definition. We show Part (ii) in the special case of $R = 2$; the general case follows by relatively scaling ε, since relu is *homogeneous*, namely, $\mathsf{relu}(ax) = a \cdot \mathsf{relu}(x)$. Consider any $U \subseteq \mathcal{X}$ such that $\|u\| = 1$ and $\langle u, v \rangle \leq 1 - \varepsilon$ for all $u, v \in U$. It holds that $\underline{\mathsf{Sdim}}_{f^\star}(\mathcal{F}^{\mathrm{relu}}, \varepsilon) \geq |U|$ when $f^\star$ is the identically zero function, since the function $f_u(x) = \mathsf{relu}(\langle u, x \rangle - (1 - \varepsilon))$ is such that $f_u(v) = 0$ for all $v \in U \setminus \{u\}$, whereas $f_u(u) = \varepsilon$. A standard sphere packing argument shows that such a set U exists with $|U| \geq (1/2\varepsilon)^{d/2}$ for all $\varepsilon < 1/2$. In particular, the δ-packing number of the unit sphere is at least $(1/\delta)^d$ [43, Cor. 4.2.13]. Thus, we can find $(1/\delta)^d$ points such that each pair u, v satisfies $\|u - v\| \geq \delta$, or equivalently $\langle u, v \rangle \leq 1 - \delta^2/2$. Setting $\delta = \sqrt{2\varepsilon}$ proves the claimed lower bound. □

Theorem 7 was independently shown by Dong et al. [12, Thm. 5.1].

We remark that while we considered the variant of eluder dimension as defined by Foster et al. [19], the lower bound on eluder dimension in Theorem 7 immediately holds for the notion defined by Russo and Van Roy [37]. On the other hand, the upper bound on eluder dimension in Theorem 6 can be shown to hold even with the definition of Russo and Van Roy [37] (by replacing every instance of $a^\star$ by a_i').

	f_1	f_2	f_3	f_4	f_5	f_6
x_1	1	0	0	0	0	0
x_2	*	1	0	0	0	0
x_3	*	*	1	0	0	0
x_4	*	*	*	1	0	0
x_5	*	*	*	*	1	0
x_6	*	*	*	*	*	1

Eluder sequence

	f_1	f_2	f_3	f_4	f_5	f_6
x_1	1	0	0	0	0	0
x_2	0	1	0	0	0	0
x_3	0	0	1	0	0	0
x_4	0	0	0	1	0	0
x_5	0	0	0	0	1	0
x_6	0	0	0	0	0	1

Star sequence

	f_1	f_2	f_3	f_4	f_5	f_6
x_1	1	0	0	0	0	0
x_2	1	1	0	0	0	0
x_3	1	1	1	0	0	0
x_4	1	1	1	1	0	0
x_5	1	1	1	1	1	0
x_6	1	1	1	1	1	1

Threshold sequence

Figure 2: Illustration of witnessing sequences of length 6 for eluder dimension, star number and threshold dimension with respect to $f^\star = 0$ (we use the 0/1 representation of functions for clarity). '*' in the eluder witness sequence refers to a free value, either 0 or 1.

5 Relationships between combinatorial measures

In this section, we specialize the eluder dimension to binary-valued function classes and prove several additional characterizations that relate this combinatorial eluder dimension to other learning-theoretic quantities. First, we (re)define these learning-theoretic quantities. The first two definitions (of combinatorial eluder dimension and star number respectively) are the specialization of the scale-sensitive versions (cf. Definition 1 and 2) to the binary-valued function setting. We abuse notation by dropping the argument ε from previous definitions in order to be consistent.

Definition 4. *Fix any function class $\mathcal{F} \subseteq (\mathcal{X} \to \{1, -1\})$ and any $f^\star \in \mathcal{F}$.*

- *The **combinatorial eluder dimension** w.r.t. $f^\star$, denoted $\mathsf{Edim}_{f^\star}(\mathcal{F})$, is defined as the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$f_i(x_i) \neq f^\star(x_i), \quad \text{and} \quad \text{for all } j < i: \quad f_i(x_j) = f^\star(x_j).$$

- *The **star number** w.r.t. $f^\star$, denoted $\mathsf{Sdim}_{f^\star}(\mathcal{F})$, is defined as the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$f_i(x_i) \neq f^\star(x_i), \quad \text{and} \quad \text{for all } j \neq i: \quad f_i(x_j) = f^\star(x_j).$$

- *The **threshold dimension** w.r.t. $f^\star$, denoted $\mathsf{Tdim}_{f^\star}(\mathcal{F})$, is defined as the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$\text{for all } k \geq i: \quad f_i(x_k) \neq f^\star(x_k), \quad \text{and} \quad \text{for all } j < i: \quad f_i(x_j) = f^\star(x_j).$$

As before, we define the combinatorial eluder dimension (resp. star number and threshold dimension) to be $\mathsf{Edim}(\mathcal{F}) := \sup_{f \in \mathcal{F}} \mathsf{Edim}_f(\mathcal{F})$ ($\mathsf{Sdim}(\mathcal{F})$ and $\mathsf{Tdim}(\mathcal{F})$ are defined similarly).

Let us pause to unpack these definitions and give some background. In fact, the star number definition stated above is the original definition of Hanneke and Yang [21], who give tight upper and lower bounds on the label complexity of *pool-based active learning* via the star number $\mathsf{Sdim}(\mathcal{F})$ and show that almost every previously proposed complexity measure for active learning takes a worst case value equal to the star number. Roughly speaking, the star number corresponds to the number of "singletons" one can embed in a function class; that is, the maximum number of functions that differ from a base function $f^\star$ at exactly one point among a subset of the domain $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$.

The threshold dimension has recently gained attention due to its role in proving an equivalence relationship between private PAC learning and online learning [see, e.g., 3, 9]. We slightly generalize the definition of Alon et al. [3] to allow for any base function $f^\star$, in the spirit of the other two definitions. A classical result in model theory provides a link between the threshold dimension and Littlestone dimension (which we denote Ldim), a quantity which is both necessary and sufficient for online learnability [8, 4]. In particular, results by Shelah [40] and Hodges et al. [22] show that for any binary-valued $\mathcal{F}$ and any $f^\star \in \mathcal{F}$:

$$\lfloor \log \mathsf{Tdim}_{f^\star}(\mathcal{F}) \rfloor \leq \mathsf{Ldim} \leq 2^{\mathsf{Tdim}_{f^\star}(\mathcal{F})}.$$

A combinatorial proof of this fact can be found in Thm. 3 of Alon et al. [3].[6] Thus, finiteness of threshold dimension is necessary and sufficient for online learnability (albeit in a much weaker, "qualitative" sense).

5.1 A qualitative equivalence

Now, we prove a rather surprising characterization that closely ties all three quantities in Definition 4 together. Our result implies that for any binary-valued function class, finiteness of the combinatorial eluder dimension is *equivalent* to finiteness of both star number and threshold dimension.

Theorem 8. *For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \{1, -1\})$ and any $f^\star \in \mathcal{F}$, the following holds:*

$$\max\{\mathsf{Sdim}_{f^\star}(\mathcal{F}), \mathsf{Tdim}_{f^\star}(\mathcal{F})\} \leq \mathsf{Edim}_{f^\star}(\mathcal{F}) \leq 4^{\max\{\mathsf{Sdim}_{f^\star}(\mathcal{F}), \mathsf{Tdim}_{f^\star}(\mathcal{F})\}}.$$

The proof of Theorem 8 can be found in Appendix D. The lower bound is trivial by examining Definition 4. To prove the upper bound, we rely on a novel connection to Ramsey theory. In particular, we show that sequences $(x_1, f_1), \ldots (x_m, f_m)$ which witness $\mathsf{Edim}_{f^\star} = m$ form a bijection with red-blue colorings of the complete graph K_m, while subsequences of the witnessing eluder sequence that are valid star number witnesses or threshold dimension witnesses can be interpreted as monochromatic colorings of subgraphs of K_m (see Figure 2). Thus, applying classical bounds from Ramsey theory implies the result.

Theorem 8 has an exponential gap between the upper and lower bounds. Can we improve either of the inequalities? The lower bound cannot be improved, by considering the simple examples over $\mathcal{X} = [n]$ of the singleton class $\mathcal{F}_n^{\text{sing}} := \{x \mapsto \mathbb{1}\{x = i\} \mid i \in [n+1]\}$ and the threshold class $\mathcal{F}_n^{\text{th}} := \{x \mapsto \mathbb{1}\{x \geq i\} \mid i \in [n+1]\}$. We also show that the upper bound cannot be improved, for example, to $\mathsf{Edim}(\mathcal{F}) \leq \text{poly}(\mathsf{Sdim}(\mathcal{F}), \mathsf{Tdim}(\mathcal{F}))$ in general.

Theorem 9. *For every $N > 0$, there exists a function class $\mathcal{F}_N$ such that $\mathsf{Edim}(\mathcal{F}_N) = N$ and $\max\{\mathsf{Sdim}(\mathcal{F}), \mathsf{Tdim}(\mathcal{F})\} < c \cdot \log_2 N$, where $c > 1/2$ is some absolute numerical constant.*

The proof of Theorem 9 can be found in Appendix E. It relies on the probabilistic method to show the existence of a randomly constructed $\mathcal{F}$ which satisfies the desired properties.

5.2 Comparisons with sign-rk

In this section, we investigate the comparison of these combinatorial quantities (eluder, star, threshold) with sign-rk, and examine whether we can prove stronger separations. One direction is clear; for the class of linear classifiers in $\mathbb{R}^d$, the combinatorial eluder dimension, star number, and threshold dimension are all infinite. However, we ask if the other separation is also possible: can we construct $\mathcal{F}$ where eluder/star/threshold are finite but sign-rk $= \infty$? We have already provided an explicit exponential separation: Theorem 6 shows that for the function class $\mathcal{F}^\oplus$, we had $\mathsf{Edim}(\mathcal{F}^\oplus) = \mathsf{Sdim}(\mathcal{F}^\oplus) = d$ but $\mathsf{sign\text{-}rk}(\mathcal{F}^\oplus) \geq 2^{d/2}$. (One can also show that $\mathsf{Tdim}(\mathcal{F}^\oplus) = d$.) We are able to show stronger (but nonconstructive) separations by extending the probabilistic techniques of Alon et al. [2], who recently provided similar separations for VC dimension versus sign-rk. In view of Proposition 3, this result also provides a separation for the scale-sensitive Definition 1.

Theorem 10. *For every $N > 0$, there exists a function class $\mathcal{F}_N \subseteq ([N] \to \{1, -1\})$ such that $\mathsf{Edim}_1(\mathcal{F}_N) = 4$ and $\mathsf{sign\text{-}rk}(\mathcal{F}_N) \geq \Omega(N^{1/9}/\log N)$, where 1 is shorthand for the all 1s function.*

The proof of Theorem 10 can be found in Appendix F. It is straightforward to replace the reference function $f^\star(x) = 1$ with any fixed reference function $f^\star : [N] \to \{1, -1\}$. First, we use Lemma 22 of Alon et al. [2] which bounds the number of distinct matrices with sign-rk $= r$; then using a probabilistic argument we show that there must be many (more) matrices with $\mathsf{Edim}_1 = 4$, so at least one of them must have large sign-rk.

The careful reader might notice that we do not prove the existence of a function class where $\mathsf{Edim}(\mathcal{F})$ is constant and the sign-rk is infinite; instead we prove the weaker statement that a function class exists with Edim w.r.t. any *fixed* function $f^\star$ is bounded. We conjecture that the stronger statement holds; see Appendix F.1 for more details.

[6] Alon et al. [3] prove the result for $f^\star(x) = -1$, but it is easy to extend their proof to hold for any $f^\star$.

5.3 Back to scale-sensitive?

It is natural to ask if our results can be extended back to the scale-sensitive definitions. First, we require a scale-sensitive version of threshold dimension. One proposal is the following:

Definition 5. *For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, $f^\star \in \mathcal{F}$, and scale $\varepsilon \geq 0$, the **exact threshold dimension** $\underline{\mathsf{Tdim}}_{f^\star}(\mathcal{F}, \varepsilon)$ is the largest m such that there exists $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ satisfying for all $i \in [m]$:*

$$\forall j \geq i \ : |f_i(x_j) - f^\star(x_j)| > \varepsilon, \quad \textit{and} \quad \sum_{j<i} (f_i(x_j) - f^\star(x_j))^2 \leq \varepsilon^2.$$

Then for all $\varepsilon > 0$:

- *the **threshold dimension** is $\mathsf{Tdim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \underline{\mathsf{Tdim}}_{f^\star}(\mathcal{F}, \varepsilon')$.*
- *$\underline{\mathsf{Tdim}}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \underline{\mathsf{Tdim}}_{f^\star}(\mathcal{F}, \varepsilon)$ and $\mathsf{Tdim}(\mathcal{F}, \varepsilon) := \sup_{f^\star \in \mathcal{F}} \mathsf{Tdim}_{f^\star}(\mathcal{F}, \varepsilon)$.*

Definition 5 mirrors the scale-sensitive definitions for eluder and star; it also recovers the combinatorial definition when $\mathcal{F}$ is binary-valued. By definition, the relationship that $\mathsf{Edim}_{f^\star}(\mathcal{F}, \varepsilon) \geq \max\{\mathsf{Sdim}_{f^\star}(\mathcal{F}, \varepsilon), \mathsf{Tdim}_{f^\star}(\mathcal{F}, \varepsilon)\}$ for every $\mathcal{F}, f^\star \in \mathcal{F}, \varepsilon > 0$ is trivial. However, one cannot hope to prove the corresponding upper bound under this definition. For example, for any $\varepsilon > 0$, take the function class which is represented by the $N \times N$ matrix:

$$f_j(x_i) = \begin{cases} 0 & i < j \\ \varepsilon & i = j \\ 0.99\varepsilon & i > j. \end{cases}$$

It is easy to see that $\mathsf{Edim}_0(\mathcal{F}, \varepsilon) = N$, while $\mathsf{Sdim}_0(\mathcal{F}, \varepsilon) = 2$ and $\mathsf{Tdim}_0(\mathcal{F}, \varepsilon) = 1$. Notice that this class is still "threshold-like", but Definition 5 does not capture this for said value of ε. Generally speaking, it is unclear how to carry over the intuition from Ramsey theory that applies in the combinatorial case to the scale-sensitive case; we leave this to future work.

Acknowledgments and Disclosure of Funding

We thank Gaurav Mahajan for allowing us to include the proof of Proposition 5 [33]. We thank Akshay Krishnamurthy, Tengyu Ma, and Ruosong Wang for helpful discussions. GL was partially supported by NSF award IIS-1764032. PK was partially supported by NSF BIGDATA award 1546500. Part of this work was done while GL, PK, and DF were participating in the Simons program on the Theoretical Foundations of Reinforcement Learning.

References

[1] N. Alon, P. Frankl, and V. Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *26th Annual Symposium on Foundations of Computer Science*, pages 277–280. IEEE Computer Society, 1985. URL https://doi.org/10.1109/SFCS.1985.30.

[2] N. Alon, S. Moran, and A. Yehudayoff. Sign rank versus VC dimension. In *29th Annual Conference on Learning Theory*, volume 49 of *PMLR*, pages 47–80, 2016. URL http://proceedings.mlr.press/v49/alon16.html.

[3] N. Alon, R. Livni, M. Malliaris, and S. Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 852–860, New York, NY, USA, 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3313276.3316312.

[4] N. Alon, O. Ben-Eliezer, Y. Dagan, S. Moran, M. Naor, and E. Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 447–455, 2021.

[5] R. I. Arriaga and S. S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Mach. Learn.*, 63(2):161–182, 2006. URL https://doi.org/10.1007/s10994-006-6265-7.

[6] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pages 463–474, 2020. URL http://proceedings.mlr.press/v119/ayoub20a.html.

[7] S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in euclidean half spaces. *Journal of Machine Learning Research*, 3(Nov):441–461, 2002.

[8] S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

[9] M. Bun, R. Livni, and S. Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

[10] Y. Chen, H. Luo, T. Ma, and C. Zhang. Active online learning with hidden shifting domains. In *International Conference on Artificial Intelligence and Statistics*, pages 2053–2061. PMLR, 2021.

[11] K. Dong, J. Peng, Y. Wang, and Y. Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

[12] K. Dong, J. Yang, and T. Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *arXiv*, 2102.04168, 2021. URL https://arxiv.org/abs/2102.04168.

[13] S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[14] S. S. Du, J. D. Lee, G. Mahajan, and R. Wang. Agnostic Q-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. In *Advances in Neural Information Processing Systems*, volume 33, pages 22327–22337. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/fd5c905bcd8c3348ad1b35d7231ee2b1-Paper.pdf.

[15] Y. Fei, Z. Yang, and Z. Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021.

[16] F. Feng, W. Yin, A. Agarwal, and L. Yang. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pages 3263–3273. PMLR, 2021.

[17] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

[18] J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002. URL https://doi.org/10.1016/S0022-0000(02)00019-3.

[19] D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv*, 2010.03104, 2020. URL https://arxiv.org/abs/2010.03104.

[20] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[21] S. Hanneke and L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(109):3487–3602, 2015. URL http://jmlr.org/papers/v16/hanneke15a.html.

[22] W. Hodges et al. *A shorter model theory*. Cambridge university press, 1997.

[23] B. Huang, K. Huang, S. Kakade, J. D. Lee, Q. Lei, R. Wang, and J. Yang. Going beyond linear rl: Sample efficient neural function approximation. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] B. Huang, J. D. Lee, Z. Wang, and Z. Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

[25] K. Huang, S. M. Kakade, J. D. Lee, and Q. Lei. A short note on the relationship of information gain and eluder dimension. *arXiv preprint arXiv:2107.02377*, 2021.

[26] H. Ishfaq, Q. Cui, V. Nguyen, A. Ayoub, Z. Yang, Z. Wang, D. Precup, and L. Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

[27] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/jiang17c.html.

[28] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *arXiv*, 2102.00815, 2021. URL https://arxiv.org/abs/2102.00815.

[29] C. Jin, Q. Liu, and T. Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

[30] P. Kamath, O. Montasser, and N. Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory*, pages 2236–2262. PMLR, 2020.

[31] B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

[32] L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/li17c.html.

[33] G. Mahajan and S. Lovett. Personal communication. 2021.

[34] W. Mou, Z. Wen, and X. Chen. On the sample complexity of reinforcement learning with policy space generalization. *arXiv preprint arXiv:2008.07353*, 2020.

[35] D. Mubayi and A. Suk. A survey of quantitative bounds for hypergraph ramsey problems. arxiv e-prints (july 2017). *arXiv preprint math.CO/1707.04229*, 2017.

[36] I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf.

[37] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf.

[38] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[39] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. URL https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/.

[40] S. Shelah. *Classification theory: and the number of non-isomorphic models*. Elsevier, 1990.

[41] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[42] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[43] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html.

[44] R. Wang, R. R. Salakhutdinov, and L. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/440924c5948e05070663f88e69e8242b-Paper.pdf.

[45] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

[46] Z. Wen and B. Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/bad5f33780c42f2588878a9d07405083-Paper.pdf.

[47] Z. Xu and A. Tewari. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] *Since this is a theory paper regarding several notions of complexity for learning problems, we do not foresee any immediate potential negative impacts.*
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Proposition 3

The proof uses the following result, which is straightforward by definition.

Proposition 11. *Σ-rank satisfies the following for all $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$.*

(i) For all $R < R'$: $\Sigma\text{-}\mathsf{rk}(\mathcal{F}, R) \geq \Sigma\text{-}\mathsf{rk}(\mathcal{F}, R')$.

(ii) For all $\Sigma_1 \subseteq \Sigma_2$ and $R > 0$: $\Sigma_1\text{-}\mathsf{rk}(\mathcal{F}, R) \geq \Sigma_2\text{-}\mathsf{rk}(\mathcal{F}, R)$.

Proof of Proposition 3. The first two inequalities follow from immediately from Proposition 11. For the last inequality, let $\phi : \mathcal{X} \to \mathbb{R}^d$ and $w : \mathcal{F} \to \mathbb{R}^d$ be the mappings that witness $\sigma\text{-}\mathsf{rk}(\mathcal{F}, R) = d$ for some $\sigma \in \mathcal{M}$. Thus, we have that for all $(x, f) \in \mathcal{X} \times \mathcal{F}$, it holds that, $f(x) = \sigma(\langle w(f), \phi(x)\rangle)$. Let $t \in \mathbb{R}$ be any value such that $\sigma(t) = 0$. From monotonicity of σ and the fact that $\mathcal{F}$ is $\{-1, 1\}$-valued, we have that for all $(x, f) \in \mathcal{X} \times \mathcal{F}$, $f(x) = \mathsf{sign}(\langle w(f), \phi(x)\rangle - t)$. Thus, $\mathsf{sign}\text{-}\mathsf{rk}(\mathcal{F}) \leq \sigma\text{-}\mathsf{rk}(\mathcal{F}, R) + 1$. We next move to the examples witnessing the separations.

- **$\mathsf{rk} \gg \mathcal{M}_\mu\text{-}\mathsf{rk}$**: Consider $\mathcal{X} = [0, 1]$ and $\mathcal{F}^{\exp} := \{f_\theta : x \mapsto \sigma(\theta \cdot x) \mid |\theta| \leq 1\}$, where $\sigma(\cdot)$ is defined piecewise as follows: $\sigma(z) = e^z$ for $z \in [0, 1]$; outside of $[0, 1]$, we extend the function linearly with slope 1 when $z < 0$ and slope e when $z > 1$. Since $\sigma \in \mathcal{M}_1^e$, we have $\mathcal{M}_1^e\text{-}\mathsf{rk}(\mathcal{F}^{\exp}, 1) = 1$ and hence $\mathcal{M}_{1/e}\text{-}\mathsf{rk}(\mathcal{F}, e) = 1$ (from Equation 3).

  Consider the points $\{x_j := j/d \mid j \in \{1, \ldots, d\}\}$ and the functions $\{f_{\theta_i} := f_{i/d} \mid j \in \{1, \ldots, d\}\}$. The matrix A given by $A_{ij} := f_{\theta_i}(x_j) = (e^{j/d^2})^i$ is a Vandermonde matrix, and hence $\mathrm{rank}(A) = d$. Since d can be chosen to be arbitrarily large, it follows that $\mathsf{rk}(\mathcal{F}^{\exp}, R) = \infty$ for all $R > 0$.

- **$\mathcal{M}_\mu\text{-}\mathsf{rk} \gg \mathcal{M}_0\text{-}\mathsf{rk}$**: Consider $\mathcal{X} = [0, 1]$ and $\mathcal{F}^{\mathrm{relu}} := \left\{f_{a,b} : x \mapsto \mathsf{relu}(ax + b) \mid a^2 + b^2 \leq 1\right\}$, the class of ReLUs with biases in 1 dimension. Clearly, $\mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}^{\mathrm{relu}}, \sqrt{2}) = 2$.

  Suppose for contradiction that for some $\sigma \in \mathcal{M}_\mu$, it holds that $\sigma\text{-}\mathsf{rk}(\mathcal{F}^{\mathrm{relu}}, R) = d$ with mappings $\phi : \mathcal{X} \to \mathcal{B}_d(1)$ and $w : \mathcal{F}^{\mathrm{relu}} \to \mathcal{B}_d(R)$ for some $R > 0$. Consider $n = d + 2$ points $0 < x_1 < x_2 < \cdots < x_n < 1$. For each i, $f_i(x) := \mathsf{relu}(x - x_{i-1}) \in \mathcal{F}^{\mathrm{relu}}$ (let $x_0 := 0$) satisfies $f_i(x_j) = \sigma(\langle w(f_i), \phi(x_j)\rangle) = 0$ for all $j < i$, and $f_i(x_j) = \sigma(\langle w(f_i), \phi(x_j)\rangle) > 0$ for all $j \geq i$. That is, $\langle w(f_i), \phi(x_j)\rangle = \sigma^{-1}(0)$ for all $j < i$ (since $\sigma : \mathbb{R} \to \mathbb{R}$ is strictly monotone, $\sigma^{-1}(0)$ is uniquely defined) and $\langle w(f_i), \phi(x_j)\rangle > \sigma^{-1}(0)$ for all $j \geq i$. Consider the matrix $A \in \mathbb{R}^{n \times n}$ given by $A_{ij} := \langle w(f_i), \phi(x_j)\rangle$. By definition, $\mathrm{rank}(A) \leq d$. On the other hand, we have that $A - \sigma^{-1}(0) \cdot J$ is an upper-triangular matrix with non-zero diagonals, where $J \in \mathbb{R}^{n \times n}$ is the all-1s matrix. It follows that $\mathrm{rank}(A) \geq n - 1$ and hence $n \leq d + 1$. This is a contradiction and hence $\mathcal{M}_\mu(\mathcal{F}^{\mathrm{relu}}, R) = \infty$ for all $\mu, R > 0$.

- **$\mathcal{M}_0\text{-}\mathsf{rk} \gg \mathsf{sign}\text{-}\mathsf{rk}$**: Consider $\mathcal{X} = [0, 1]$ and $\mathcal{F}^{\mathrm{th}} := \{f_t : x \mapsto \mathsf{sign}(x - t) \mid t \in [0, 1]\}$, the class of Thresholds. Clearly, $\mathsf{sign}\text{-}\mathsf{rk}(\mathcal{F}^{\mathrm{th}}) = 2$.

  We briefly sketch the argument showing $\mathcal{M}_0\text{-}\mathsf{rk}(\mathcal{F}, R) = \infty$ for any $R > 0$. Suppose for some $\sigma \in \mathcal{M}_0$ it holds that $\sigma\text{-}\mathsf{rk}(\mathcal{F}^{\mathrm{th}}, R) = d$. Then it is possible to realize $\mathcal{F}^{\mathrm{th}}$ as halfspaces *with margin*, since σ is 1-Lipschitz. This implies that $\mathcal{F}^{\mathrm{th}}$ is online learnable with a finite mistake bound (via the Perceptron algorithm). This is a contradiction since $\mathcal{F}^{\mathrm{th}}$ is not online learnable with a finite mistake bound [see e.g. 39, Lemma 21.6 and Ex. 21.4].

B Proof of Proposition 4

Proposition 4 follows from Claim 13 and Claim 14 below, where for clarity, we keep track of the norms of ϕ and w separately. Our improvement comes about from the following lemma, which is inspired by a similar step in [37].

Lemma 12 (Inspired by [37])**.** *For all $k \geq 1$ and $\alpha, \beta > 0$, if $(1 + \alpha)^k \leq 1 + \beta k$, then $k \leq \frac{e}{e-1} \cdot \frac{1+\alpha}{\alpha} \cdot \ln\left(\frac{2\beta(1+\alpha)}{\alpha}\right)$.*

Proof. We consider two cases. In both cases, we use that $\ln(1+\alpha) \geq \frac{\alpha}{1+\alpha}$ holds for all $\alpha > 0$. Also note that $(1+\alpha)^k > 1 + k\alpha$ and hence $\beta > \alpha$.

Case 1: If $\beta k < 1$, we have $(1+\alpha)^k \leq 2$ and hence $k \leq \frac{\ln 2}{\ln(1+\alpha)} \leq \ln 2 \cdot \frac{1+\alpha}{\alpha}$. Since $\beta > \alpha$, we have that $\ln 2 \leq \frac{e}{e-1} \cdot \ln\left(\frac{2\beta(1+\alpha)}{\alpha}\right)$, thereby completing the proof for this case.

Case 2: If $\beta k \geq 1$, we have $(1+\alpha)^k \leq 2\beta k$. Taking logarithms, we have $k\ln(1+\alpha) \leq \ln k + \ln 2\beta$. Hence, we have $\frac{k\alpha}{1+\alpha} \leq \ln\left(\frac{k\alpha}{1+\alpha}\right) + \ln\left(\frac{2\beta(1+\alpha)}{\alpha}\right)$. Using $\ln x \leq \frac{x}{e}$ for all $x \geq 0$, we get for $x = \frac{k\alpha}{1+\alpha}$ that $k \cdot \frac{\alpha}{1+\alpha} \cdot \left(1 - \frac{1}{e}\right) \leq \ln\left(\frac{2\beta(1+\alpha)}{\alpha}\right)$, thereby completing the proof. □

Claim 13. *Suppose* $\mathsf{rk}(\mathcal{F}, R_\phi R_w) = d$ *is witnessed by mappings* $\phi : \mathcal{X} \to \mathcal{B}_d(R_\phi)$ *and* $w : \mathcal{F} \to \mathcal{B}_d(R_w)$. *Then* $\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) \leq \frac{3e}{e-1} \cdot d \cdot \log\left(\frac{24R_\phi^2 R_w^2}{\varepsilon^2}\right)$ *for all* $\varepsilon < R_\phi R_w$.

Proof. Suppose $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon) = m$ witnessed by the sequence $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$, for some $f^\star \in \mathcal{F}$ and $\varepsilon > 0$. Denote $w_i := w(f_i) - w(f^\star)$, and $\phi_i := \phi(x_i)$. It follows that $w_i \in \mathcal{B}_d(2R_w)$ and $\phi_i \in \mathcal{B}_d(R_\phi)$. From Equation 1, we have that for all $i \in [m]$:

$$\max_{w \in \mathbb{R}^d} \left\{ |\langle w, \phi_i \rangle| : \sum_{j<i} \langle w, \phi_i \rangle^2 \leq \varepsilon^2, \|w\|_2 \leq 2R_w \right\} > \varepsilon \tag{4}$$

Let $V_i := \lambda I + \sum_{j<i} \phi_j \phi_j^\top$. The above equation implies that for all $i \in [m]$:

$$\max_{w \in \mathbb{R}^d} \left\{ |\langle w, \phi_i \rangle| : \|w\|_{V_i}^2 \leq \varepsilon^2 + \lambda \cdot 4R_w^2 \right\} > \varepsilon,$$

which via convex duality and setting $\lambda := \varepsilon^2/(4R_w^2)$, further implies that for all $i \in [m]$:

$$\|\phi_i\|_{V_i^{-1}}^2 > \frac{\varepsilon^2}{\varepsilon^2 + \lambda \cdot 4R_w^2} = \frac{1}{2},$$

We will use a potential argument to track the quantity $\det(V_i)$. First, we have the upper bound:

$$\det(V_{m+1}) \leq \left(\frac{\mathrm{tr}(V_{m+1})}{d}\right)^d \leq \left(\frac{\lambda d + mR_\phi^2}{d}\right)^d = \lambda^d \left(1 + \frac{mR_\phi^2}{\lambda d}\right)^d = \lambda^d \left(1 + \frac{m}{d} \cdot \frac{4R_w^2 R_\phi^2}{\varepsilon^2}\right)^d,$$

using, AM-GM inequality, the linearity of trace, and the definition of λ. We also have the lower bound, using the Matrix Determinant Lemma:

$$\det(V_{m+1}) = \det(V_m + \phi_m \phi_m^\top) = \det(V_m) \cdot (1 + \|\phi_m\|_{V_m^{-1}}^2) \geq \lambda^d (3/2)^m,$$

where the last step follows by induction. Combining the upper and lower bounds, we get that $(3/2)^{m/d} \leq 1 + \frac{m}{d} \cdot \frac{4R_w^2 R_\phi^2}{\varepsilon^2}$. The claim follows from an application of Lemma 12 with $k = \frac{m}{d}$, $\alpha = \frac{1}{2}$ and $\beta = \frac{4R_\phi^2 R_w^2}{\varepsilon^2}$. □

Claim 14. *For* $\sigma \in \mathcal{M}_\mu^L$, *suppose* σ-$\mathsf{rk}(\mathcal{F}, R_\phi R_w) = d$ *is witnessed by mappings* $\phi : \mathcal{X} \to \mathcal{B}_d(R_\phi)$ *and* $w : \mathcal{F} \to \mathcal{B}_d(R_w)$. *Then* $\underline{\mathsf{Edim}}(\mathcal{F}, \varepsilon) \leq \frac{3e}{e-1} \cdot d \cdot \frac{L^2}{\mu^2} \cdot \log\left(\frac{24R_\phi^2 R_w^2 L^2}{\varepsilon^2}\right)$ *for all* $\varepsilon < R_\phi R_w L$.

Proof. Suppose $\underline{\mathsf{Edim}}_{f^\star}(\mathcal{F}, \varepsilon) = m$ witnessed by the sequence $(x_1, f_1), \ldots, (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$, for some $f^\star \in \mathcal{F}$ and $\varepsilon > 0$. Denote $w_i := w(f_i) - w(f^\star)$, and $\phi_i := \phi(x_i)$. It follows that $w_i \in \mathcal{B}_d(2R_w)$ and $\phi_i \in \mathcal{B}_d(R_\phi)$. Since $\sigma \in \mathcal{M}_\mu^L$, we have for any $w_1, w_2, x \in \mathbb{R}^d$:

$$\mu\,|\langle w_1 - w_2, x \rangle| \leq |\sigma(\langle w_1, x \rangle) - \sigma(\langle w_2, x \rangle)| \leq L\,|\langle w_1 - w_2, x \rangle|.$$

Therefore, Eq. (4) can be replaced with:

$$\max_{w\in\mathbb{R}^d}\left\{|\langle w,\phi_i\rangle| : \sum_{j<i}\langle w,\phi_i\rangle^2 \le \varepsilon^2/\mu^2, w\in\mathcal{B}_d(2R_w)\right\} > \varepsilon/L. \tag{5}$$

Following the same steps with $\lambda := \varepsilon^2/(4R_w^2\mu^2)$ and V_i defined as before, we can further show that:

$$\max_{w\in\mathbb{R}^d}\left\{|\langle w,\phi_i\rangle| : \|w\|_{V_i}^2 \le \varepsilon^2/\mu^2 + \lambda\cdot 4R_w^2\right\} > \varepsilon/L,$$

which implies the bound:

$$\|\phi_i\|_{V_i^{-1}}^2 > \frac{\varepsilon^2/L^2}{\varepsilon^2/\mu^2+\lambda\cdot 4R_w^2} = \frac{1}{2}\cdot\frac{\mu^2}{L^2}.$$

Using similar upper and lower bounds on $\det(V_m)$ gives us that $(1+\mu^2/(2L^2))^{m/d} \le 1 + \frac{m}{d}\cdot\frac{4R_w^2R_\phi^2\mu^2}{\varepsilon^2}$; the proof again concludes with an application of Lemma 12 with $k=\frac{m}{d}$, $\alpha=\frac{\mu^2}{2L^2}$ and $\beta=\frac{4R_\phi^2R_w^2\mu^2}{\varepsilon^2}$. □

Discussion of prior work. We clarify the differences between Proposition 4 and the corresponding propositions in Russo and Van Roy [37], Osband and Van Roy [36].

Proposition 6 in [37] considers the setting exactly as in Claim 13. In our notation, the stated bound has the form $\mathsf{rk}(\mathcal{F}, R_\phi R_w)\cdot O(\log\left(3+\frac{12R_w^2}{\varepsilon^2}\right))$; the factor of R_ϕ is missing inside the log term. As explained in Definition 3 (footnote 4), only the product $R_\phi R_w$ is relevant as the scale of ϕ and w is interchangeable.

Proposition 7 in [37] considers the setting exactly as in Claim 14. In our notation, the stated bound there has the form $\mathcal{M}_\mu^L\text{-}\mathsf{rk}(\mathcal{F}, R_\phi R_w)\cdot O(\frac{L^2}{\mu^2}\cdot\log\left(\frac{3L^2}{\mu^2}+\frac{L^2}{\mu^2}\cdot\frac{12R_w^2L^2}{\varepsilon^2}\right))$; again the factor of R_ϕ is missing. Also, the factor of L in $R_w^2L^2/\varepsilon^2$ is improvable to μ.

Proposition 4 in [36] considers the setting analogous to Claim 14, but for vector-valued function classes, that is, $\mathcal{F}\subseteq(\mathcal{X}\to\mathbb{R}^k)$. In the special case of $k=1$, their bound in our notation, has the form $\mathcal{M}_\mu^L\text{-}\mathsf{rk}(\mathcal{F}, R_\phi R_w)\cdot O(\frac{L^2}{\mu^2}\cdot\log\left(\frac{L^2}{\mu^2}+\frac{L^2}{\mu^2}\cdot\frac{R_\phi^2R_w^2}{\varepsilon^2}\right))$. The term $R_\phi R_w$ should be $R_\phi R_w\mu$. As shown in Equation 3, it is possible to make $R_\phi R_w$ arbitrarily small while keeping L/μ fixed.

Lastly, we note that Lemma 12 is slightly different than the corresponding inequality used in Russo and Van Roy [37], Osband and Van Roy [36] (which has $(1+\beta)$ in place of 2β). This allows us to remove the additive terms of 3 and $3\frac{L^2}{\mu^2}$ inside the log factor for the rk case and the $\mathcal{M}_\mu^L\text{-}\mathsf{rk}$ case respectively. In the $\mathcal{M}_\mu^L\text{-}\mathsf{rk}$ case, this gives a nontrivial improvement in some regime of parameters; namely the bounds of [37, 36] would only give us a term of $\log\left(\frac{L}{\mu}+\frac{R_\phi R_w L}{\varepsilon}\right)$, which can be loose when μ is very small.

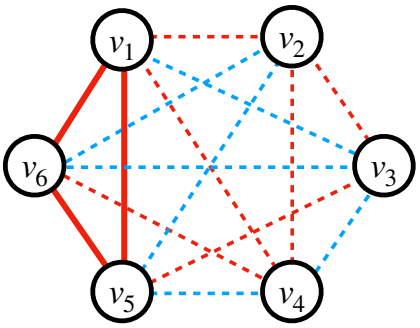
C Proof of Proposition 5

We exhibit a sequence $(x_1,\theta_1),\dots(x_m,\theta_m)$ that witnesses the claimed lower bound on eluder dimension with $\theta^\star=0$. It suffices to consider the case of $R=1$, as this is just a matter of scaling relative to ε. First, consider the case of $d=1$. For any $\alpha\in(\varepsilon,\sqrt{2}\varepsilon)$ and $k:=\lfloor\log_2(1/\alpha)\rfloor$, let $x_i := 1/2^{(k-i)}$ and $\theta_i=\alpha\cdot 2^{k-i}$ for $i\in\{0,\dots,k\}$. For each i, it holds that $\theta_i x_i=\alpha>\varepsilon$ and $\sum_{j<i}(\theta_i x_j)^2\le\alpha^2/2<\varepsilon^2$. Since $|x_i|\le 1$ and $|\theta_i|\le 1$ we get $\underline{\mathsf{Edim}}(\mathcal{F},\varepsilon)\ge k+1\ge\Omega(\log(1/\varepsilon))$.

For $d>1$, consider d copies of the above 1 dimensional sequence repeated in each dimension. Namely, consider the sequence $(x_{ij},\theta_{ij})_{i\in[d],j\in[k]}$ with $x_{ij}:=\boldsymbol{e}_i/2^{k-j}$ and $\theta_{ij}:=\alpha 2^{k-j}\cdot\boldsymbol{e}_i$ (where $\boldsymbol{e}_i$ is the i-th standard basis vector). Since $x_{ij},\theta_{ij}\in\mathcal{B}_d(1)$, we have $\underline{\mathsf{Edim}}(\mathcal{F},\varepsilon)\ge d(k+1)=\Omega\left(d\log(1/\varepsilon)\right)$.

	f_1	f_2	f_3	f_4	f_5	f_6
x_1	1	0	0	0	0	0
x_2	0	1	0	0	0	0
x_3	1	0	1	0	0	0
x_4	0	0	1	1	0	0
x_5	0	1	0	1	1	0
x_6	0	1	1	0	0	1

(a) Eluder matrix



(b) Ramsey graph

Figure 3: An example illustrating the connection between the upper bound and Ramsey numbers. Left: a sequence $\{(x_1, f_1), \ldots (x_6, f_6)\}$ witnessing $\mathsf{Edim}_0(\mathcal{F}) = 6$, represented in matrix form. (We switch to 0/1-valued function classes for clarity.) Right: In the corresponding graph K_6, we color the graph edges e_{ij} to be red if $f_j(x_i) = 0$ and blue if $f_j(x_i) = 1$. Since $R(3, 3) = 6$, we are guaranteed a subgraph K_3 which is monochromatic; in this example, the subgraph is given by the vertices $\{v_1, v_5, v_6\}$. Red subgraphs indicate sequences which witness $\mathsf{Sdim}_0(\mathcal{F})$; blue subgraphs witness $\mathsf{Tdim}_0(\mathcal{F})$. In this case, $\{(x_1, f_1), (x_5, f_5), (x_6, f_6)\}$ witnesses $\mathsf{Sdim}_0(\mathcal{F}) \geq 3$.

D Proof of Theorem 8

Proof of lower bound. The lower bound is straightforward from the definition, since any sequence of $(x_1, f_1), \ldots, (x_m, f_m)$ that witness $\mathsf{Sdim}_{f^\star}(\mathcal{F}) = m$ or $\mathsf{Tdim}_{f^\star}(\mathcal{F}) = m$ must also be valid "eluder sequences"; so the eluder dimension can only be larger.

Proof of upper bound. Fix $\mathcal{F}$ and $f^\star \in \mathcal{F}$. Let $(x_1, f_1), \ldots (x_m, f_m) \in \mathcal{X} \times \mathcal{F}$ be the sequence which witnesses $\mathsf{Edim}_{f^\star}(\mathcal{F}) = m$. To prove the bound, we will show that there exists a subset of size at least $k \geq \log_4 m$ which witnesses $\mathsf{Sdim}_{f^\star} = k$ or $\mathsf{Tdim}_{f^\star} = k$.

This follows from a connection to Ramsey theory. A visualization of the proof is depicted in Figure 3. Recall that diagonal Ramsey number $R(k, k)$ is defined as the smallest m such that every red-blue labeling of the edges of the graph K_m contains a monochromatic subgraph K_k.

In addition, define $E(k, k)$ as the smallest m such that any eluder sequence $(x_1, f_1), \ldots, (x_m, f_m)$ contains a subsequence $(x_{i_1}, f_{i_1}), \ldots (x_{i_k}, f_{i_k})$ which witnesses $\mathsf{Sdim}_{f^\star}(\mathcal{F}) \geq k$ or $\mathsf{Tdim}_{f^\star}(\mathcal{F}) \geq k$. We claim that $E(k, k) \leq R(k, k)$.

To see this, note that there exists a bijection between colorings of K_m and eluder sequences. Every eluder sequence $(x_1, f_1), \ldots (x_m, f_m)$ can be used to construct a red-blue coloring of K_m as follows. For every edge e_{ij} for $i > j \in [m]$, we color it red if $f_j(x_i) = f^\star(x_i)$ and blue otherwise. Observe that if there exists a subsequence $(x_{i_1}, f_{i_1}), \ldots (x_{i_k}, f_{i_k})$ which witnesses $\mathsf{Sdim}_{f^\star} \geq k$, then the subgraph comprised of the vertices $i_1, \ldots, i_k$ in the coloring of K_m must be monochromatic red. Likewise, if a subsequence witnesses $\mathsf{Tdim}_{f^\star} \geq k$, then the subgraph must be monochromatic blue. Thus, if m is such that if every coloring of K_m induces a monochromatic coloring K_k, then for any eluder sequence, we can always find a subsequence that witnesses $\mathsf{Sdim}_{f^\star} \geq k$ or $\mathsf{Tdim}_{f^\star} \geq k$. This shows that $E(k, k) \leq R(k, k)$.

The proof concludes by applying the classical bound $R(k, k) \leq 4^k$ [see, e.g., 35].

E Proof of Theorem 9

We will construct $\mathcal{F}_N \subseteq ([N] \to \{1, -1\})$ randomly, such that $|\mathcal{F}_N| = N + 1$ and $\mathsf{Edim}_1(\mathcal{F}_N) = N$. Note that it is equivalent to define an $N \times (N + 1)$ sign matrix B, representing the values of $\mathcal{X} \times \mathcal{F}$

with entry $B_{ij} = f_j(x_i)$. Let B be randomly drawn according to the following distribution:

$$B_{ij} \sim \begin{cases} 1 & i < j, \\ -1 & i = j, \\ \mathrm{Rad}(1/2) & i > j. \end{cases}$$

By construction, with probability 1, $(x_1, f_1) \dots (x_N, f_N)$ is a valid sequence witnessing $\mathsf{Edim}(\mathcal{F}_N) = \mathsf{Edim}_1(\mathcal{F}_N) = N$.

We now have to argue that there exists some matrix $C \in \{1, -1\}^{N \times (N+1)}$ such that the equivalent function class $\mathcal{F}_N$ has small threshold dimension and star number. We use the following two lemmas to simplify the requirement that $\mathcal{F}_N$ have small threshold dimension and star number with respect to all base functions $f^\star \in \mathcal{F}$ to just considering the base functions $f^\star(x) = -1$ and $f^\star(x) = 1$.

Lemma 15. *For any $\mathcal{F}$ and $\mathsf{dim} \in \{\mathsf{Edim}, \mathsf{Tdim}, \mathsf{Sdim}\}$, we have $\mathsf{dim}(\mathcal{F}) \le \mathsf{dim}_1(\mathcal{F}) + \mathsf{dim}_{-1}(\mathcal{F})$.*

Lemma 16. *For any $\mathcal{F}$, $\mathsf{Tdim}(\mathcal{F}) \le 2\mathsf{Tdim}_1(\mathcal{F})$.*

We set up some additional notation. Denote $\mathbf{I} = (i_1, i_2, \dots, i_k)$ and $\mathbf{J} = (j_1, j_2, \dots, j_k)$ to be k-length sequences of distinct elements from $[N]$. For any two sequences $\mathbf{I}$, $\mathbf{J}$, we use $(x_{\mathbf{I}}, f_{\mathbf{J}}) := ((x_{i_1}, f_{j_1}), \dots (x_{i_k}, f_{j_k}))$. We define valid star sequences to be any $(x_{\mathbf{I}}, f_{\mathbf{J}})$ which witness $\mathsf{Sdim}_1(\mathcal{F}_N) = k$ or $\mathsf{Sdim}_{-1}(\mathcal{F}_N) = k$, and we define valid threshold sequences to be any $(x_{\mathbf{I}}, f_{\mathbf{J}})$ which witnesses $\mathsf{Tdim}_1(\mathcal{F}_N) = k$.

Define the random variable X_k to be the number of valid star or threshold sequences, i.e., $X_k := |\{(\mathbf{I}, \mathbf{J}) \colon (x_{\mathbf{I}}, f_{\mathbf{J}}) \text{ is valid star or threshold sequence}\}|$. By linearity of expectation, we have

$$\mathbb{E}[X_k] \le N^k (N+1)^k \cdot \max_{\mathbf{I},\mathbf{J} \subset [N]} p_{\mathbf{I},\mathbf{J}}, \quad \text{where } p_{\mathbf{I},\mathbf{J}} := \mathbb{P}\left[(x_{\mathbf{I}}, f_{\mathbf{J}}) \text{ is a valid star or threshold sequence}\right].$$

Now we apply the following lemma to upper bound the expectation.

Lemma 17. *For all $\mathbf{I}$, $\mathbf{J}$, we have $p_{\mathbf{I},\mathbf{J}} \le 3 \cdot 2^{-k(k-1)/2}$.*

We apply Lemma 17 to the previous display. When $k = \Omega(\log_2 N)$, we have $\mathbb{E}[X_k] < 1$. By the probabilistic method, there must exist an $N \times (N+1)$ valued matrix C such that the corresponding $\mathcal{F}_N \subseteq ([N] \to \{1, -1\}$ has $\mathsf{Sdim}_1(\mathcal{F}_N) < O(\log_2 N)$, $\mathsf{Sdim}_{-1}(\mathcal{F}_N) < O(\log_2 N)$, and $\mathsf{Tdim}_1(\mathcal{F}_N) < O(\log_2 N)$, but $\mathsf{Edim}_1(\mathcal{F}_N) = N$. By Lemma 15 and 16, this concludes the proof. □

Proof of Lemma 15. We prove the result for star number; the result for eluder dimension and threshold dimension can be shown with a similar argument. Fix any $f^\star \in \mathcal{F}$, and let $(x_1, f_1), \dots (x_m, f_m)$ denote the sequence which witnesses $\mathsf{Sdim}_{f^\star}(\mathcal{F}) = m$. Let $I_+ \subseteq [m]$ denote the indices i for which $f^\star(x_i) = 1$ and $I_- \subseteq [m]$ denote the indices i for which $f^\star(x_i) = -1$. By definition of star number, $f_j(x_j) = -1$ for every $j \in I_+$ and $f_k(x_j) = f^\star(x_j) = 1$ for every $j \ne k \in I_+$. Thus we know that $\{(x_k, f_k) : k \in I_+\}$ is a valid sequence which witnesses $\mathsf{Sdim}_1(\mathcal{F}) \ge |I_+|$. Similarly $\{(x_k, f_k) : k \in I_-\}$ is a valid sequence which witnesses $\mathsf{Sdim}_{-1}(\mathcal{F}) \ge |I_-|$. Thus we have shown that $\mathsf{Sdim}_{f^\star}(\mathcal{F}) = m \le \mathsf{Sdim}_1(\mathcal{F}) + \mathsf{Sdim}_{-1}(\mathcal{F})$. Taking the supremum on the LHS yields the claim. □

Proof of Lemma 16. Fix any $f^\star \in \mathcal{F}$, and let $(x_1, f_1), \dots (x_m, f_m)$ denote the sequence which witnesses $\mathsf{Tdim}_{f^\star}(\mathcal{F}) = m$. Again let $I_+ \subseteq [m]$ denote the indices i for which $f^\star(x_i) = 1$ and $I_- \subseteq [m]$ denote the indices i for which $f^\star(x_i) = -1$. Either $|I_+| \ge m/2$ or $|I_-| \ge m/2$. We break into cases.

Case 1. If $|I_+| \ge m/2$, then taking the subsequence indexed by I_+ already shows that $\mathsf{Tdim}_1(\mathcal{F}) \ge m/2 = \mathsf{Tdim}_{f^\star}(\mathcal{F})/2$, and we are done.

Case 2. If $|I_-| \ge m/2$, then let us consider the subsequence indexed by I_-. We reindex it to call it $(x_1, f_1), \dots (x_k, f_k)$, where $k = |I_-|$. Observe that the sequence

$$(x_k, f^\star), (x_{k-1}, f_k), (x_{k-2}, f_{k-1}), \dots, (x_1, f_2)$$

witnesses $\mathsf{Tdim}_1(\mathcal{F}) \ge k \ge m/2 = \mathsf{Tdim}_{f^\star}(\mathcal{F})/2$.

Thus in both cases we have shown that $\mathsf{Tdim}_1(\mathcal{F}) \ge \mathsf{Tdim}_{f^\star}(\mathcal{F})/2$; taking the supremum yields the claim. □

Proof of Lemma 17. Fix any $\mathbf{I}$, $\mathbf{J}$ to be k-length subsequences of $[N]$. In order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ to be a valid star sequence w.r.t. $f^\star(x) = 1$, the following properties of the matrix B must hold:

1. For every $r \in [k]$, $B_{i_r,j_r} = -1$.
2. For every $r, s \in [k]$ such that $r \neq s$, $B_{i_r,j_s} = 1$.

In order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ to be a valid star sequence w.r.t. $f^\star(x) = -1$, we just flip the values in the above two properties.

Likewise, in order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ to be a valid threshold sequence w.r.t. $f^\star(x) = 1$, the following properties of the matrix B must hold:

1. For every $r \in [k]$, $B_{i_r,j_r} = -1$.
2. For every $r, s \in [k]$ such that $r < s$, $B_{i_r,j_s} = 1$.
3. For every $r, s \in [k]$ such that $r \geq s$, $B_{i_r,j_s} = -1$.

First, we will prove that the probability that $(x_{\mathbf{I}}, f_{\mathbf{J}})$ is a valid star sequence w.r.t. $f^\star(x) = 1$, as well as the probability that $(x_{\mathbf{I}}, f_{\mathbf{J}})$ is a valid threshold sequence w.r.t. $f^\star(x) = 1$ are both $2^{-k(k-1)/2}$. For any $r \in [k]$, if $i_r < j_r$, then by construction of B we know that $f_{j_r}(x_{i_r}) = 1$, so $(x_{\mathbf{I}}, f_{\mathbf{J}})$ cannot be a valid star sequence or threshold sequence w.r.t $f^\star(x) = 1$. Henceforth, assume $i_r \geq j_r$ for all $r \in [k]$. Now define indices $r_1, r_2, \ldots, r_k$ as the permutation of $[k]$ such that $i_{r_1} > i_{r_2} > \cdots > i_{r_k}$. For any $p < q$, we have $i_{r_p} > i_{r_q} \geq j_{r_q}$. Thus for every $p < q$, the corresponding entry (i_{r_p}, j_{r_q}) is sampled from $\mathrm{Rad}(1/2)$. In order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ to be a valid star sequence, all of these must take the value of 1; likewise in order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ to be a valid threshold sequence, all of these must take the value of -1. Since there are $k(k-1)/2$ of these, we have the desired result.

Now we bound the probability that $(x_{\mathbf{I}}, f_{\mathbf{J}})$ is a valid star sequence w.r.t. $f^\star(x) = -1$. Note that because the definition of star number is permutation-invariant, we can assume that $i_1 > i_2 > \cdots > i_k$ without loss of generality. Consider the pair (i_1, j_1). We require $B_{i_1,j_1} = 1$, so either $i_1 > j_1$ or $i_1 < j_1$. Since we require $B_{i_2,j_1} = -1$, we cannot have $i_1 < j_1$, so we must have $i_1 > j_1$. Using a similar argument, we must have $i_r > j_r$ for all $r \in [k-1]$. Thus, there must be at least $k(k-1)/2$ random entries in the submatrix given by $(\mathbf{I}, \mathbf{J})$, all of which must take value -1 in order for $(x_{\mathbf{I}}, f_{\mathbf{J}})$ is a valid star sequence w.r.t. $f^\star(x) = -1$.

By union bound we get $p_{\mathbf{I},\mathbf{J}} \leq 3 \cdot 2^{-k(k-1)/2}$, thus proving the result. □

F Proof of Theorem 10

For simplicity, let us consider only function classes of size N. It is equivalent to reason about $N \times N$ sign matrices which define the values that $\mathcal{X} \times \mathcal{F}$ take. We slightly abuse notation to define the sign-rk of an $N \times N$ matrix S to be

$$\text{sign-rk}(S) \coloneqq \left\{\mathrm{rank}(M) : M \in \mathbb{R}^{N \times N},\ \text{sign}(M_{ij}) = S_{ij} \text{ for all } i, j \in [N]\right\}.$$

We also define $\mathsf{Edim}_1(S)$ similarly: $\mathsf{Edim}_1(S)$ is the maximum k such that we can find two $\mathbf{I}$, $\mathbf{J}$ which are k-length subsequences of $[N]$ such that the matrix S restricted to $\mathbf{I}$, $\mathbf{J}$ has -1 on the diagonal and $+1$ above the diagonal.

The proof relies on the following key lemma, which bounds the number of matrices with sign-rk at most r.

Lemma 18 (e.g., Lemma 22 of [2])**.** *Let $r \leq N/2$. The number of $N \times N$ sign matrices with sign rank at most r does not exceed $2^{O(rN \log N)}$.*

In order to prove the result, we use a probabilistic argument to show that there must exist many distinct $N \times N$ matrices with $\mathsf{Edim}_1 = 4$.

Lemma 19. *The number of $N \times N$ sign matrices with $\mathsf{Edim}_1 \leq 4$ is at least $2^{\Omega(N^{10/9})}$.*

The above lemmas imply that there must exist at least one sign matrix with $\mathsf{Edim}_1(S) \leq 4$ and $\text{sign-rk}(S) \geq \Omega(N^{1/9}/\log N)$. This proves Theorem 10, assuming Lemma 19 which we now prove.

Proof of Lemma 19. Define the set of E_5-light matrices as the set of 5×5 sign matrices which are always $+1$ above the diagonal. We claim there exists an $N \times N$ sign matrix C which (1) contains no E_5-light matrices and (2) has at least $\Omega(N^{10/9})$ entries that are $+1$. Such a matrix C has $\mathsf{Edim}_1(C) \leq 4$; moreover changing any $+1$ to -1 in C will not increase Edim_1.

Let B be a random $N \times N$ sign matrix with each entry $+1$ with probability $p \coloneqq 1/(2N^{8/9})$. Define the random variable

$$X \coloneqq (\#\ +1\text{'s in } C) - (\#\ E_5\text{-light matrices in } C).$$

Then $\mathbb{E}[X] \geq N^2 p - N^{10} p^{10} \geq \Omega(N^{10/9})$. Take some matrix with value of X at least the expectation and change a $+1$ to a -1 in every E_5-light matrix to get C. Since this does not affect the value of X, we know that the resulting matrix B has $\Omega(N^{10/9})$ entries that are $+1$.

Since changing any $+1$ to -1 in C will not increase Edim_1, we see that there are at least $2^{\Omega(N^{10/9})}$ distinct sign matrices with $\mathsf{Edim} \leq 4$. □

F.1 A stronger separation?

Our result does not fully show the separation between eluder dimension and sign-rk. In the random construction used in Lemma 19, it could be the case that the matrix C we pick satisfies $\mathsf{Edim}_1(C) \leq 4$, but there could be some other $f^\star$ (column of C) such that $\mathsf{Edim}_{f^\star}(C) = \omega(1)$.

We conjecture that the stronger separation result also holds:

Conjecture 20. *There exists absolute constants $k \in \mathbb{N}$ and $c > 0$ such that the following hold. For every $N > 0$, there exists a function class $\mathcal{F}_N \subseteq ([N] \to \{1, -1\})$ such that $\mathsf{Edim}(\mathcal{F}_N) \leq k$ and $\mathsf{sign\text{-}rk}(\mathcal{F}_N) \geq \Omega(N^c / \log N)$.*

In light of Theorem 8, it suffices to show that there exists some function class $\mathcal{F}_N$ such that $\max\{\mathsf{Sdim}(\mathcal{F}_N), \mathsf{Tdim}(\mathcal{F}_N)\} \leq k$ and $\mathsf{sign\text{-}rk}(\mathcal{F}_N) \geq \Omega(N^c / \log N)$.

Consider the easier problem of showing the separation for threshold dimension. Here there is no difficulty. The key step is to apply Lemma 16 to reduce the problem to showing the result with respect to a *single* $f^\star$. It follows as a corollary of Theorem 10 that there exists a function class $\mathcal{F}_N$ such that $\mathsf{Tdim}(\mathcal{F}_N) \leq 8$ and $\mathsf{sign\text{-}rk}(\mathcal{F}_N) \geq \Omega(N^{1/9} / \log N)$. (Using a more direct analysis, it is possible to improve the constants 8 and $1/9$.)

Showing the separation for star number (and eluder dimension) is a different story. We do not have a direct analogue of Lemma 16 for star number and eluder dimension. The weaker Lemma 15 allows us to reduce two showing the separation for two functions; but it is unclear how to leverage this reduction to extend the construction in the proof of Lemma 19.