Differentially Private Nonparametric Regression Under a Growth Condition

Noah Golowich* NZG@MIT.EDU
MIT EECS

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Given a real-valued hypothesis class $\mathcal{H}$, we investigate under what conditions there is a differentially private algorithm which learns an optimal hypothesis from $\mathcal{H}$ given i.i.d. data. Inspired by recent results for the related setting of binary classification (Alon et al., 2019; Bun et al., 2020), where it was shown that *online learnability* of a binary class is necessary and sufficient for its private learnability, Jung et al. (2020) showed that in the setting of regression, online learnability of $\mathcal{H}$ is *necessary* for private learnability. Here online learnability of $\mathcal{H}$ is characterized by the finiteness of its *η-sequential fat shattering dimension,* $\mathrm{sfat}_\eta(\mathcal{H})$, for all $\eta > 0$. In terms of *sufficient* conditions for private learnability, Jung et al. (2020) showed that $\mathcal{H}$ is privately learnable if $\lim_{\eta \downarrow 0} \mathrm{sfat}_\eta(\mathcal{H})$ is finite, which is a fairly restrictive condition. We show that under the relaxed condition $\liminf_{\eta \downarrow 0} \eta \cdot \mathrm{sfat}_\eta(\mathcal{H}) = 0$, $\mathcal{H}$ is privately learnable, establishing the first nonparametric private learnability guarantee for classes $\mathcal{H}$ with $\mathrm{sfat}_\eta(\mathcal{H})$ *diverging* as $\eta \downarrow 0$. Our techniques involve a novel filtering procedure to output stable hypotheses for nonparametric function classes.

Keywords: differential privacy, nonparametric regression, sequential fat-shattering dimension

1. Introduction

In recent years there has been an increased focus on the importance of protecting the privacy of potentially sensitive users' data on which machine learning algorithms are trained (Roth and Kearns, 2019; Nissim et al., 2018). The model of *differentially private learning* (Dwork et al., 2006; Dwork and Roth, 2013; Vadhan, 2017) provides a way to formalize the accuracy-privacy tradeoffs encountered. The vast majority of work in this area focuses on the setting of private classification, namely where we must predict a $\{0, 1\}$-valued label for each data point x (Kasiviswanathan et al., 2008; Beimel et al., 2014; Bun et al., 2015; Feldman and Xiao, 2014; Beimel et al., 2013; Bun et al., 2018; Beimel et al., 2019; Alon et al., 2019; Kaplan et al., 2020; Bun et al., 2020; Neel et al., 2019; Bun, 2020). Many natural machine learning problems, however, in application domains ranging from ecology to medicine (Dua and Graff, 2017), are phrased more naturally as *regression* problems, where for each data point x we must predict a real-valued label. In this paper we study this problem of differentially private regression for nonparametric function classes.

In the setting of differentially private binary classification, a major recent development (Alon et al., 2019; Bun et al., 2020) is the result that a hypothesis class $\mathcal{F}$ consisting of binary classifiers is learnable with approximate differential privacy (Definition 3) if and only if it is *online learnable*, which is known to hold in turn if and only if the *Littlestone dimension* of $\mathcal{F}$ is finite (Littlestone, 1987; Ben-David et al., 2009). Such an equivalence, however, remains open for the setting of differentially private regression (this question was asked in Bun et al. (2020)). The combinatorial

* Supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.

© 2021 N. Golowich.

parameter characterizing online learnability for regression is the *sequential fat-shattering dimension* (Rakhlin et al., 2015b) (Definition 6), which may be viewed as a scale-sensitive analogue of the Littlestone dimension. In one direction, Jung et al. (2020) recently showed that if a class $\mathcal{F}$ consisting of bounded real-valued functions is privately learnable, then it is online learnable, i.e., the sequential fat-shattering dimension of $\mathcal{F}$ is finite at all scales. The other direction, namely whether online learnability of $\mathcal{F}$ in the regression setting implies private learnability, remains open.

1.1. Results

In this paper, we make progress towards the question of whether online learnability in the regression setting implies private learnability by exhibiting a sufficient condition for private learnability in terms of the growth of the sequential fat-shattering dimension of a class. For input space $\mathcal{X}$, a class $\mathcal{H}$ consisting of hypotheses $h : \mathcal{X} \to [-1, 1]$, and $\eta > 0$, let $\mathrm{sfat}_\eta(\mathcal{H})$ denote the η-sequential fat-shattering dimension of $\mathcal{H}$ (Definition 6). As in Jung et al. (2020); Rakhlin et al. (2015b), we work with the *absolute loss* to measure the error of a hypothesis $h : \mathcal{X} \to [-1, 1]$: for a distribution Q supported on $\mathcal{X} \times [-1, 1]$, write $\mathrm{err}_Q(h) := \mathbb{E}_{(x,y)\sim Q}\left[|h(x) - y|\right]$. Our main result is as follows:

Theorem 1 (Private nonparametric regression; informal version of Theorem 44) *Let $\mathcal{H}$ be a class of hypotheses $h : \mathcal{X} \to [-1, 1]$. For any $\varepsilon, \delta, \eta \in (0, 1)$, for some $n = \frac{2^{\tilde{O}(\mathrm{sfat}_\eta(\mathcal{H}))}}{\varepsilon\eta^4}$, there is an (ε, δ)-differentially private algorithm which, given n i.i.d. samples from any distribution Q on $\mathcal{X} \times [-1, 1]$, with high probability outputs a hypothesis $\hat{h} : \mathcal{X} \to [-1, 1]$ so that*

$$\mathrm{err}_Q(\hat{h}) \le \inf_{h\in\mathcal{H}} \mathrm{err}_Q(h) + O\left(\eta \cdot \mathrm{sfat}_\eta(\mathcal{H})\right).$$

As an immediate consequence, we obtain the following sufficient condition for *private learnability* (Definition 4) of a real-valued hypothesis class:

Corollary 2 *Suppose $\mathcal{H}$ is a class of hypotheses $h : \mathcal{X} \to [-1, 1]$ satisfying $\liminf_{\eta\downarrow 0} \eta\cdot\mathrm{sfat}_\eta(\mathcal{H}) = 0$. Then $\mathcal{H}$ is privately learnable.*

Prior to our work, essentially the strongest private learnability guarantee for a nonparametric real-valued function class was (Jung et al., 2020, Theorem 15), which established that if the *sequential pseudo-dimension* of a class $\mathcal{H}$ is finite, then $\mathcal{H}$ is privately learnable. However, the sequential pseudo-dimension of $\mathcal{H}$ is lower-bounded by $\mathrm{sfat}_\eta(\mathcal{H})$ for all $\eta > 0$ (and in fact may be defined as $\lim_{\eta\downarrow 0} \mathrm{sfat}_\eta(\mathcal{H})$), and thus its boundedness implies that $\mathrm{sfat}_\eta(\mathcal{H})$ is bounded uniformly over $\eta > 0$. Thus Corollary 2 is the first result to establish a private learnability result for a nonparametric family of classes $\mathcal{H}$ with the property that $\mathrm{sfat}_\eta(\mathcal{H})$ can *diverge* as $\eta \downarrow 0$. Even very simple function classes may have $\mathrm{sfat}_\eta(\mathcal{H})$ diverging as $\eta \downarrow 0$: for instance, the class of all single-dimensional linear functions $\mathcal{H} = \{x \mapsto ax+b : x, a, b \in \mathbb{R},\ \ |x| \le 1, |a| \le 1, |b| \le 1\}$ satisfies $\mathrm{sfat}_\eta(\mathcal{H}) = \Theta(\log(1/\eta))$.

Techniques: new filtering procedure The proof of Theorem 1 proceeds in two stages. The first, fairly straightforward, step extends the algorithm `ReduceTree` of Ghazi et al. (2020a) which was used to construct a private learner in the setting of binary classification for a class of finite Littlestone dimension; our analogue for regression is `ReduceTreeReg` (Algorithm 3). From a technical standpoint, this involves extending the notion of *irreducibility* to real-valued classes (Section 3). However, unlike for the case of classification, `ReduceTreeReg` alone is not sufficient for

our purposes. In particular, `ReduceTreeReg` leads, roughly speaking, to the following guarantee, which we informally call *weak stability*. Given any distribution Q on $\mathcal{X} \times [-1, 1]$, there is a hypothesis $\sigma^\star : \mathcal{X} \to [-1, 1]$ with low error on Q so that given some number n_0 of i.i.d. samples from Q, we can output a collection of hypotheses $\hat{g}_1, \ldots, \hat{g}_M$ so that for some $1 \leq j \leq M$ we have $\|\hat{g}_j - \sigma^\star\|_\infty \leq \eta$ with some not-to-small probability. Here $\eta > 0$ is a small value representing a lower bound on the desired error. In the setting of classification Ghazi et al. (2020a) showed the stronger guarantee (which we informally call *strong stability*) that $\hat{g}_j = \sigma^\star$ for some j. The guarantee of strong stability allowed them to perform multiple draws of n_0 samples and use a *private sparse selection procedure* (an analogue of the stable histograms procedure of Bun et al. (2016) for the selection problem; see Section 2.4) to privately output a hypothesis with low population error.

The guarantee of weak stability is, however, insufficient to apply the sparse selection procedure. Thus we introduce a new procedure, called `SOAFilter` (Algorithm 2) to upgrade the guarantee of weak stability provided by `ReduceTreeReg` to one of strong stability; this is our main technical contribution. At a high level, `SOAFilter` first "filters out" many candidate hypotheses $h : \mathcal{X} \to [-1, 1]$ which are well-approximated by some hypothesis which is not filtered out (`FilterStep`, Algorithm 1). It then assigns each hypothesis $\hat{g}_j$, $1 \leq j \leq M$, as above, to some not-too-large collection of hypotheses which are not filtered out in a careful way that can ensure strong stability. Further details are provided in Section 5.

1.2. Related work

Differentially private regression As discussed in the previous sections, the most closely related work to ours is Jung et al. (2020), which showed that finiteness of sequential pseudo-dimension (namely, $\lim_{\eta \downarrow 0} \operatorname{sfat}_\eta(\mathcal{H})$) is sufficient for private learnability. A number of other papers have studied special cases of regression: for instance, Chaudhuri and Monteleoni (2009) studied differentially private logistic regression, Chaudhuri et al. (2011); Kifer et al. (2012); Bassily et al. (2014) proved upper and lower bounds on the minimax rate of empirical misk minimization, which includes linear regression with general loss functions as a special case, Wang (2018) showed improved adaptive linear regression algorithms, Cai et al. (2019) showed improved bounds on the minimax rate of linear regression with ℓ_2 loss, Bernstein and Sheldon (2019) studied differentially private Bayesian linear regression, and Alabi et al. (2020) studied differentially private linear regression in one dimension with the goal of optimizing performance on certain empirical datasets. Our work may be viewed as orthogonal to these papers, which study linear models in finite-dimensional spaces. While the growth condition $\lim_{\eta \downarrow 0} \eta \cdot \operatorname{sfat}_\eta(\mathcal{H}) = 0$ is generally satisfied for such models,[1] Theorem 1 does not improve upon any existing sample complexity bounds in these specialized settings (where in most cases optimal minimax rates are known). On the other hand, these existing works do not address the nonparametric setting where essentially no structure is imposed on the hypothesis class.

Online learnability for nonparametric classes The sequential fat-shattering dimension was introduced by Rakhlin et al. (2015b) and shown to characterize online learnability of a real-valued hypothesis class in Rakhlin et al. (2015a). It is a sequential analogue of the fat-shattering dimension, which was introduced in Alon et al. (1997); Kearns and Schapire (1994) and was shown to characterize learnability in the i.i.d. setting. A substantial amount of work has established bounds

1. For instance, if $\mathcal{X}$ is the unit ball in $\mathbb{R}^d$ with respect to the ℓ_2 norm, and $\mathcal{H} = \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq 1\}$, then $\operatorname{sfat}_\eta(\mathcal{H}) \leq O(d \log 1/\eta)$ since $\mathcal{H}$ has a pointwise (i.e., sup-norm) η-cover of size $O(1/\eta^d)$, i.e., pointwise metric entropy $O(d \log 1/\eta)$.

on the complexity of various learning tasks in terms of the fat-shattering dimension in the i.i.d. setting (e.g., Anthony and Bartlett (2009); Mendelson (2002); Bartlett et al. (1996)), and in terms of the sequential fat-shattering dimension and related complexity measures in the online setting (e.g., Rakhlin and Sridharan (2014a, 2017); Foster and Krishnamurthy (2018)). Our work begins such a study in the setting of differentially private learning (with i.i.d. data).

Overview of the paper In Section 2 we give preliminaries. In Section 3 we introduce the notion of *irreducibility* for the setting of regression. In Section 4 we state the weak stability guarantee of the `ReduceTreeReg` algorithm, which we then upgrade to one of strong stability in Section 5 using our "filtering" algorithm. Section 6 describes how to combine the components of the previous sections to prove Theorem 1. Finally, we discuss some directions for future work in Section 7. Several lemma statements in the main body are stated informally; full and rigorous statements and proofs of all lemmas and theorems are given in the appendix.

2. Preliminaries

2.1. PAC learning & discretization of hypothesis classes

For a positive integer K, let $[K] := \{1, 2, \dots, K\}$. Let $\mathcal{X}$ denote an input space and $\mathcal{Y}$ denote an output space, which will always be a subset of the real line. We let $\mathcal{Y}^{\mathcal{X}}$ denote the space of *hypotheses* on $\mathcal{X}$, namely functions $h : \mathcal{X} \to \mathcal{Y}$. We are given a known *hypothesis class* $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. For a distribution Q on $\mathcal{X} \times \mathcal{Y}$ and $h \in \mathcal{Y}^{\mathcal{X}}$, let $\mathrm{err}_Q(h) := \mathbb{E}_{(x,y)\sim Q}[|h(x) - y|]$ denote the *population error* of h.[2] A *dataset* $S_n \in (\mathcal{X} \times \mathcal{Y})^n$ is a tuple of n elements of $\mathcal{X} \times \mathcal{Y}$; for Q as above, let Q^n be the distribution of $S_n \in (\mathcal{X} \times \mathcal{Y})^n$ consisting of n i.i.d. draws from Q. For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $\delta_{(x,y)}$ denote the point measure at (x, y), and for a dataset S_n write $\hat{Q}_{S_n} := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ to denote the *empirical measure* for S_n. The *empirical error* of a hypothesis h with respect to a dataset S_n is defined to be $\mathrm{err}_{\hat{Q}_{S_n}}(h)$. To avoid having to make technical measurability assumptions on $\mathcal{H}, \mathcal{X}$, we will assume throughout the paper that $\mathcal{H}, \mathcal{X}$ are countable (or finite).

Ultimately we aim to solve the following problem: for $\mathcal{Y} = [-1, 1]$ and some small error η_0, find some $\hat{h}$ so that $\mathrm{err}_Q(\hat{h}) \le \inf_{h\in\mathcal{H}}\{\mathrm{err}_Q(h)\} + \eta_0$ given a sample $S_n \sim Q^n$. To streamline the analysis, though, we will often work with the *discretization* of the class $\mathcal{H}$ *at scale* η, for some $\eta < \eta_0$: it is denoted $\lfloor\mathcal{H}\rfloor_\eta$ and is obtained by dividing the interval $[-1, 1]$ into $\lceil 2/\eta\rceil$ intervals each of length $2/\lceil 2/\eta\rceil \le \eta$, and rounding $h(x)$, for each $h \in \mathcal{H}, x \in \mathcal{X}$, to the interval containing $h(x)$. A formal definition of $\lfloor\mathcal{H}\rfloor_\eta$ is as follows: first, for a real number $y \in [-1, 1]$, define $\lfloor y\rfloor_\eta \in [\lceil 2/\eta\rceil]$ as follows: $\lfloor y\rfloor_\eta := 1 + \left\lfloor \frac{(y+1)}{2} \cdot \lceil 2/\eta\rceil \right\rfloor$ for $y < 1$ and $\lfloor y\rfloor_\eta := \lceil 2/\eta\rceil$ for $y = 1$.

Next, for $h \in \mathcal{H}$, define $\lfloor h\rfloor_\eta \in [\lceil 2/\eta\rceil]^{\mathcal{X}}$ by $\lfloor h\rfloor_\eta(x) = \lfloor h(x)\rfloor_\eta$, for $x \in \mathcal{X}$. Then the discretization $\lfloor\mathcal{H}\rfloor_\eta \subset \{1, 2, \dots, \lceil 2/\eta\rceil\}^{\mathcal{X}}$ is defined as $\lfloor\mathcal{H}\rfloor_\eta := \{\lfloor h\rfloor_\eta : h \in \mathcal{H}\}$. Moreover, the *discretization of a distribution* Q on $\mathcal{X} \times [-1, 1]$ *at scale* η, denoted $\lfloor Q\rfloor_\eta$, is defined to be the distribution of $(x, \lfloor y\rfloor_\eta)$, where $(x, y) \sim Q$. In Appendix A.2, we show that for $h \in [-1, 1]^{\mathcal{X}}$, $\mathrm{err}_Q(h)$ is roughly η times $\mathrm{err}_{\lfloor Q\rfloor_\eta}(\lfloor h\rfloor_\eta)$, up to an additive error of $\pm O(\eta)$ (see (3)), and that we have the bound $\mathrm{sfat}_2(\lfloor\mathcal{H}\rfloor_\eta) \le \mathrm{sfat}_\eta(\mathcal{H})$ on the sequential fat-shattering dimension of $\lfloor\mathcal{H}\rfloor_\eta$ at scale 2 (Lemma 23). We will often write $K := \lceil 2/\eta\rceil$ when considering the discretization of classes.

For any $h \in \mathbb{R}^{\mathcal{X}}$ write $\|h\|_\infty := \sup_{x\in\mathcal{X}} |h(x)|$.

2. Following Jung et al. (2020); Rakhlin et al. (2015a), we work with the absolute loss; the results may readily be generalized to any other Lipschitz loss function.

2.2. Differential privacy

In this paper we study algorithms which satisfy *approximate differential privacy*, defined as follows:

Definition 3 (Differential privacy, Dwork et al. (2006)) *Fix sets $\mathcal{Z}, \mathcal{W}$, $n \in \mathbb{N}$, $\varepsilon, \delta \in (0,1)$, and suppose $\mathcal{W}$ is countable. A randomized algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ is (ε, δ)-differentially private if the following holds: for any datasets $S_n, S_n' \in \mathcal{Z}^n$ differing in a single example[3] and for all subsets $\mathcal{E} \subset \mathcal{W}$, $\Pr[A(S_n) \in \mathcal{E}] \leq e^{\varepsilon} \cdot \Pr[A(S_n') \in \mathcal{E}] + \delta$.*

Our goal is to solve the PAC learning problem (as introduced in Section 2.1) with an algorithm that is (ε, δ)-differentially private as a function of S_n. Typically in the differential privacy literature it is assumed that $\delta = n^{-\omega(1)}$. To this end, we make the following definition:

Definition 4 (Private learnability) *A class $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$ is* privately (PAC) learnable *if for all $\varepsilon, \delta, \eta, \beta \in (0,1)$, there is a bound $n = n_{\mathcal{H}}(\varepsilon, \delta, \eta, \beta)$ so that the following holds:*

- *There is an (ε, δ)-differentially private algorithm A that takes as input a dataset $S_n \in (\mathcal{X} \times [-1,1])^n$ and outputs some $A(S_n) \in [-1,1]^{\mathcal{X}}$ so that: for any distribution Q on $\mathcal{X} \times [-1,1]$, with probability at least $1-\beta$ over $S_n \sim Q^n$, $\mathrm{err}_Q(A(S_n)) \leq \inf_{h \in \mathcal{H}} \{\mathrm{err}_Q(h)\} + \eta$.*
- *For fixed ε, η, β, the mapping $\delta \mapsto n_{\mathcal{H}}(\varepsilon, \delta, \eta, \beta)$ is $\delta^{-o(1)}$, i.e., for any constant $c > 0$ there is $\delta_0 > 0$ so that for $0 < \delta < \delta_0$ we have $n_{\mathcal{H}}(\varepsilon, \delta, \eta, \beta) \leq 1/\delta^c$.*

Our algorithms will satisfy the stronger guarantee that for fixed η and $\mathcal{H}$, the bound $n_{\mathcal{H}}(\varepsilon, \delta, \eta, \beta)$ grows polynomially in $1/\varepsilon, \log(1/\delta), \log(1/\beta)$.

2.3. Sequential fat-shattering dimension

For a positive integer K, we begin by defining K-ary $\mathcal{X}$-valued trees. For a positive integer t and a sequence $k_1, k_2, \ldots, \in [K]$, write $k_{1:t} = (k_1, \ldots, k_t)$. Let $k_{1:0}$ denote the empty sequence.

Definition 5 ($\mathcal{X}$-valued tree) *For $d, K \in \mathbb{N}$, a K-ary $\mathcal{X}$-valued tree of* depth *d is a collection of partial functions $\mathbf{x}_t : [K]^{t-1} \to \mathcal{X}$, for $1 \leq t \leq d$, each with nonempty domain, so that for all $k_{1:t}$ that lie in the domain of $\mathbf{x}_{t+1}$:*

1. *The sequence $k_{1:t-1}$ lies in the domain of $\mathbf{x}_t$ (i.e., a node's parent is a node);*
2. *For all $k_t' \in [K]$ the sequence $(k_1, \ldots, k_{t-1}, k_t')$ lies in the domain of $\mathbf{x}_{t+1}$ (i.e., each non-root node has $K-1$ siblings).*

We write $\mathbf{x} := (\mathbf{x}_1, \ldots, \mathbf{x}_d)$. We say that the tree $\mathbf{x}$ is complete *if for each t the domain of $\mathbf{x}_t$ is all of $[K]^{t-1}$. The tree $\mathbf{x}$ is* binary *if it is 2-ary (i.e., $K = 2$ in the above).*

Associated with each sequence $k_{1:t} \in [K]^t$ for which $k_{1:t-1}$ is in the domain of $\mathbf{x}_t$, for some $1 \leq t \leq d$, is a *node* of the tree. We say that this node is a *leaf* if $k_{1:t}$ is not in the domain of $\mathbf{x}_{t+1}$ (or if $t = d$). Moreover, for any non-leaf node associated with $k_{1:t} \in [K]^t$, we say that it is *labeled* by the point $\mathbf{x}_{t+1}(k_{1:t}) \in \mathcal{X}$. For any such node v, the nodes associated with $(k_1, \ldots, k_t, k_{t+1}')$, for each choice of $k_{t+1}' \in [K]$ are the *children* of v; we say that the coresponding edge between v and each child is *labeled* by k_{t+1}'. Note that a node is a leaf if and only if it has no children. Note also that any non-leaf node has exactly K children.

3. Written out, we have $S_n = (z_1, \ldots, z_n)$ and $S_n' = (z_1, \ldots, z_{n-1}, z_n')$ for some $z_1, \ldots, z_n, z_n' \in \mathcal{Z}$.

Fix $\alpha > 0$. A complete binary (i.e., 2-ary) $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is *α-shattered* by a class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there is a complete $\mathbb{R}$-valued binary tree $\mathbf{s}$ of depth d so that for all $k_{1:d} \in \{1,2\}^d$, there is some $f \in \mathcal{F}$ so that $(3 - 2k_t) \cdot (f(\mathbf{x}_t(k_{1:t-1})) - \mathbf{s}_t(k_{1:t-1})) \geq \alpha/2$ for all $1 \leq t \leq d$. The tree $\mathbf{s}$ is called the *witness to shattering*.

Definition 6 (Sequential fat-shattering dimension) *The* α-sequential fat shattering dimension *of a class $\mathcal{F}$, denoted* $\mathrm{sfat}_\alpha(\mathcal{F})$*, is the greatest positive integer d so that there is an $\mathcal{X}$-valued binary tree of depth d which is α-shattered by $\mathcal{X}$. As a convention, if $\mathcal{F}$ is empty, we write* $\mathrm{sfat}_\alpha(\mathcal{F}) = -1$.

2.4. Sparse selection procedure

A key building block in our private learning protocols is a differentially private algorithm for the following sparse selection problem from Ghazi et al. (2020b). For $m, s \in \mathbb{N}$, the (m,s)-*sparse selection* problem is defined as follows: there is some (possibly infinite) universe $\mathcal{U}$, and m users. Each user $i \in [m]$ is given some set $\mathcal{S}_i \subset \mathcal{U}$ of size $|\mathcal{S}_i| \leq s$. An algorithm is said to solve the (m,s)-sparse selection problem *with additive error* $\eta > 0$ if, given as input the sets $\mathcal{S}_1, \dots, \mathcal{S}_m$, it outputs some universe element $\hat{u} \in \mathcal{U}$ so that $|\{i : \hat{u} \in \mathcal{S}_i\}| \geq \max_{u \in \mathcal{U}} |\{i : u \in \mathcal{S}_i\}| - \eta$. We will use the following proposition, which shows that the sparse selection problem can be solved privately with error independent of the size of the universe $\mathcal{U}$:

Proposition 7 (Ghazi et al. (2020b), Lemma 36) *For $\varepsilon, \delta, \beta \in (0,1)$, there is an (ε,δ)-differentially private algorithm that, given an input dataset to the (m,s)-sparse selection problem, outputs a universe element $\hat{u}$ such that with probability at least $1-\beta$, the (additive) error of $\hat{u}$ is $O\left(\frac{1}{\varepsilon}\log\left(\frac{ms}{\varepsilon\delta\beta}\right)\right)$.*

In our application of Proposition 7, the universe $\mathcal{U}$ will be the set of hypotheses $[K]^{\mathcal{X}}$ and so the output of the sparse selection procedure will be a private hypothesis; see Section 6.

3. Irreducibility for real-valued classes

In this section we introduce the concept of *irreducibility* in the context of regression, extending the work of Ghazi et al. (2020a), which defined irreducibility for $\{0,1\}$-valued classes in the context of classification. Throughout this section, we will fix a positive integer K and an input space $\mathcal{X}$, and consider a class $\mathcal{F} \subset [K]^{\mathcal{X}}$ so that $\mathrm{sfat}_2(\mathcal{F})$ is finite. As discussed in Section 2.1, $\mathcal{F}$ will arise in the proof of Theorem 1 as the η-discretization of a real-valued class $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, where $K = \lceil 2/\eta \rceil$. We begin with the following definition which will simplify our notation.

Definition 8 (Ancestor set, depth of a node) *Let $\mathbf{x}$ be a $\mathcal{X}$-valued tree of depth d, and v be a node of $\mathbf{x}$ corresponding to the tuple $(k_1, \dots, k_t) \in [K]^t$. The* ancestor set *of v, denoted $\mathbf{A}(v)$, is the subset of $\mathcal{X} \times [K]$ given by $\mathbf{A}(v) := \{(\mathbf{x}_1, k_1), (\mathbf{x}_2(k_1), k_2), \dots, (\mathbf{x}_t(k_{1:t-1}), k_t)\}$. The integer t is referred to as the* depth *of the node v and is denoted as $t = \mathrm{depth}(v)$.*

In the context of the above definition, note that t is an upper bound on the size of $\mathbf{A}(v)$. It is possible that for some distinct s, s' we could have $(\mathbf{x}_s(k_{1:s-1}), k_s) = (\mathbf{x}_{s'}(k_{1:s'-1}), k_{s'})$ and hence the size of $\mathbf{A}(v)$ could be strictly less than t. Note that $\mathbf{A}(v)$ depends on the tree $\mathbf{x}$, though we do not explicitly notate this dependence since the tree $\mathbf{x}$ will always be clear from the node v.

For any $x \in \mathcal{X}, k \in [K]$, set $\mathcal{F}|_{(x,k)} := \{f \in \mathcal{F} : f(x) = k\}$. For a set $S = \{(x_1,k_1), \dots, (x_\ell, k_\ell)\}$, similarly set $\mathcal{F}|_S := \bigcap_{i \in [\ell]} \mathcal{F}|_{(x_i,k_i)} = \{f \in \mathcal{F} : f(x_i) = k_i \ \ \forall i \in [\ell]\}$.

Definition 9 (Irreducibility) *For an integer $\ell \geq 1$, a class $\mathcal{F} \subset [K]^{\mathcal{X}}$ is* ℓ-irreducible *if for any K-ary $\mathcal{X}$-valued tree $\mathbf{x}$ of depth at most ℓ, the tree $\mathbf{x}$ has some leaf v so that $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}(v)}) = \mathrm{sfat}_2(\mathcal{F})$.*

We say $\mathcal{F}$ is *irreducible* if it is 1-irreducible. For convenience we will say that all classes are *0-irreducible* (i.e., 0-irreducibility is vacuous); thus ℓ-irreducibility makes sense for all non-negative integers ℓ. Note that ℓ-irreducibility implies ℓ'-irreducibility for $\ell' < \ell$. The following simple, though fundamental, lemma forms the basis of a number of the stability-type results we show:

Lemma 10 *Suppose $\mathcal{G} \subset [K]^{\mathcal{X}}$ is irreducible. Then there are at most 2 values of $k \in [K]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$, and if there are 2 values, they differ by 1.*

Using Lemma 10, we next define the *SOA hypothesis* associated to an irreducible hypothesis class $\mathcal{G} \subset [K]^{\mathcal{X}}$, which assigns to each x some element $k \in [K]$ maximizing $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)})$. Such SOA hypotheses were crucial in the development of private learning algorithms for classification (Ghazi et al., 2020a; Bun et al., 2020), and they will likewise play a major role in this paper.

Definition 11 (SOA hypothesis) *Fix an irreducible class $\mathcal{G} \subset [K]^{\mathcal{X}}$. Define $\mathrm{SOA}_{\mathcal{G}} \in [K]^{\mathcal{X}}$ as follows: for each $x \in \mathcal{X}$, $\mathrm{SOA}_{\mathcal{G}}(x)$ is equal to some $k \in [K]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$. By Lemma 10, there are at most 2 such values of k. If there are 2 such values of k, i.e., there is some $k \in [K-1]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G}|_{(x,k+1)}) = \mathrm{sfat}_2(\mathcal{G})$, the tie is broken as follows:*

- *If there is some ℓ so that $\mathcal{G}|_{(x,k)}$ is ℓ-irreducible but $\mathcal{G}|_{(x,k+1)}$ is not, then set $\mathrm{SOA}_{\mathcal{G}}(x) = k$; vise versa, if $\mathcal{G}|_{(x,k+1)}$ is ℓ-irreducible but $\mathcal{G}|_{(x,k)}$ is not, then $\mathrm{SOA}_{\mathcal{G}}(x) = k+1$.*
- *If the previous item does not hold, then set $\mathrm{SOA}_{\mathcal{G}}(x) = k$.*

Lemma 12 below is similar to (Ghazi et al., 2020a, Lemma 4.3) proved in the setting of classification and is the basis for the "weak stability" results presented in Section 4. The key difference between Lemma 12 and (Ghazi et al., 2020a, Lemma 4.3) is that in the setting of classification, it can be established that $\mathrm{SOA}_{\mathcal{H}} = \mathrm{SOA}_{\mathcal{G}}$, whereas for the setting of regression we only get "approximate equality", i.e., $\| \mathrm{SOA}_{\mathcal{H}} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq 1$.

Lemma 12 *Suppose $\mathcal{H} \subset \mathcal{G}$, $\mathrm{sfat}_2(\mathcal{H}) = \mathrm{sfat}_2(\mathcal{G})$, and that $\mathcal{H}$ is irreducible. Then it holds that $\| \mathrm{SOA}_{\mathcal{H}} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq 1$.*

Following Ghazi et al. (2020a), we say that $\mathcal{G} \subset \mathcal{F}$ is a *finite restriction subclass* if it holds that $\mathcal{G} = \mathcal{F}|_{(x_1,y_1),\ldots,(x_M,y_M)}$ for some $(x_1,y_1),\ldots,(x_M,y_M) \in \mathcal{X} \times [K]$. Note that if $\mathcal{X}$ is countable, the set of finite restriction subclasses of $\mathcal{F}$ is countable. (The set of all subclasses of $\mathcal{F}$ may be uncountable; thus, by considering finite restriction subclasses we avoid having to deal with uncountable sets.)

4. The `ReduceTreeReg` algorithm: obtaining weak stability

In this section we state the weak stability guarantee afforded by the algorithm `ReduceTreeReg` (Algorithm 3). Overall the algorithm and its analysis is very similar to that of the `ReduceTree` algorithm of Ghazi et al. (2020a), so all details are given in the appendix. (Some modifications from Ghazi et al. (2020a) are necessary, though, for instance because a class with finite sequential fat-shattering dimension does not immediately give rise to one of comparable Littlestone dimension; thus we cannot use the results of Ghazi et al. (2020a) in a black-box manner.) As in Section 3 we work with the discretized problem: given $\mathcal{X}, K$ a class $\mathcal{F} \subset [K]^{\mathcal{X}}$ with $d := \mathrm{sfat}_2(\mathcal{F}) \ll K$,

$n \in \mathbb{N}$, and a distribution P on $\mathcal{X} \times [K]$, the algorithm `ReduceTreeReg` receives a dataset $S_n \in (\mathcal{X} \times [K])^n$ drawn from P^n. It also takes as input a parameter α_1, for which it is assumed that $\alpha_1 - 3d \geq \inf_{f \in \mathcal{F}} \mathrm{err}_P(f)$. The guarantee of `ReduceTreeReg` is stated (informally) as follows:

Lemma 13 (Weak stability; informal version of Lemmas 34 and 35) *Suppose $\mathcal{F}, P, \alpha_1$ are given as described above. Then there are $d+1$ hypotheses $\sigma_1^\star, \dots, \sigma_{d+1}^\star : \mathcal{X} \to [K]$, depending only on $\mathcal{F}, P$,[4] so that, for sufficiently large n, given as input a dataset $S_n \sim P^n$, `ReduceTreeReg` outputs a set $\hat{\mathcal{S}} \subset [K]^{\mathcal{X}}$ of size $|\hat{\mathcal{S}}| \leq K^{2^{\tilde{O}(d)}}$ so that:*

- *With high probability, for some $t \in [d+1]$ and $\hat{g} \in \hat{\mathcal{S}}$, it holds that $\|\hat{g} - \sigma_t^\star\|_\infty \leq 5$.*
- *With high probability, all $\hat{g} \in \hat{\mathcal{S}}$ satisfy $\mathrm{err}_P(\hat{g}) \leq \alpha_1$.*

Note that Lemma 13 only guarantees that $\|\hat{g} - \sigma_t^\star\|_\infty \leq 5$ with high probability, which we informally refer to as *weak stability*; in order to apply Proposition 7 to obtain a private learning algorithm, we would need that $\|\hat{g} - \sigma_t^\star\|_\infty = 0$ (which we refer to as *strong stability*). In the following section we discuss how to upgrade the guarantee of weak stability to one of strong stability.

5. The algorithm `SOAFilter`: from weak to strong stability

In this section we introduce the algorithm `SOAFilter` and state its main guarantee. As in Section 3, we continue on working with the discretized version of the problem, i.e., $\mathcal{X}, K$ are fixed, $\mathcal{X}$ is countable, and we are given some countable hypothesis class $\mathcal{F} \subset [K]^{\mathcal{X}}$, known to the algorithm, distribution P on $\mathcal{X} \times [K]$, unknown to the algorithm, and the goal is to find $f \in \mathcal{F}$ minimizing $\mathrm{err}_P(f)$. We will write $d := \mathrm{sfat}_2(\mathcal{F})$ throughout this section. The error bounds we establish in this section will grow as $O(d)$ (see, e.g., item 1 below); thus, if $\mathcal{F}$ arises as a discretization $\mathcal{F} = \lfloor \mathcal{H} \rfloor_\eta$, in order to ensure the error in the non-discretized version of the problem, which is $O(d)/K$, is small, we work in the regime $d \ll K$. Recalling that $K = \lceil 2/\eta \rceil$ for a discretization scale η (Section 2.1) and so $d/K = O(\eta \cdot d) \leq O(\eta \cdot \mathrm{sfat}_\eta(\mathcal{H}))$ (Lemma 23), the growth condition $O(\mathrm{sfat}_\eta(\mathcal{H}) \cdot \eta) \to 0$ arises as a sufficient condition for $d/K \to 0$.

We address the following problem: suppose there is some class $\mathcal{G} \subset \mathcal{F}$ which is ℓ-irreducible for some large $\ell \in \mathbb{N}$, and for which $\mathrm{err}_P(\mathrm{SOA}_{\mathcal{G}})$ is known to be small. Unfortunately, the algorithm does not know $\mathrm{SOA}_{\mathcal{G}}$; instead, we only know of some procedure (formalized as part of `ReduceTreeReg` described in Section 4) to produce, given i.i.d. samples from P, a collection of hypotheses $\hat{g}_1, \hat{g}_2, \dots, \hat{g}_M \in [K]^{\mathcal{X}}$, so that with some positive probability (lower bounded by $1/O(d)$) at least one such hypothesis $\hat{g}_i$ satisfies $\|\mathrm{SOA}_{\mathcal{G}} - \hat{g}_i\|_\infty \leq \chi$ for some small positive constant χ.[5] Recall that we call this guarantee *weak stability*. We can repeat this procedure many times with disjoint samples from P, thus generating many hypotheses $\hat{g}_i$ satisfying $\|\mathrm{SOA}_{\mathcal{G}} - \hat{g}_i\|_\infty \leq \chi$, with the goal of applying the sparse selection procedure of Proposition 7. However, in order to do so, we would need that for a given draw of $(\hat{g}_1, \dots, \hat{g}_M)$, some hypothesis $\hat{g}_i$ is *equal* to $\mathrm{SOA}_{\mathcal{G}}$ with positive probability, i.e., $\chi = 0$. Since we wish to avoid dependence on $|\mathcal{X}|$ in our sample complexity bounds (e.g., if $\mathcal{X}$ is infinite), given only the guarantee that $\|\mathrm{SOA}_{\mathcal{G}} - \hat{g}_i\|_\infty \leq \chi$ for some $\chi > 0$, it is nontrivial to privately output some hypothesis close to $\mathrm{SOA}_{\mathcal{G}}$.

In this section we overcome this challenge as follows: given $\mathcal{G}$ as above and $\hat{g} \in [K]^{\mathcal{X}}$ with $\|\mathrm{SOA}_{\mathcal{G}} - \hat{g}\|_\infty \leq \chi$, we introduce an algorithm, `SOAFilter` (Algorithm 2), which outputs some set $\mathscr{R}_{\hat{g}}$ consisting of many subclasses $\mathcal{L} \subset [K]^{\mathcal{X}}$, of size bounded above as a function of d and K (in

4. Each of the hypotheses $\sigma_i^\star$ is of the form $\mathrm{SOA}_{\mathcal{G}}$ for some $\mathcal{G} \subset \mathcal{F}$ which is ℓ'-irreducible for sufficiently large ℓ'.
5. We were able to establish such a guarantee for $\chi = 5$ in Section 4 (see Lemma 13).

particular, $|\mathscr{R}_{\hat{g}}| \leq K^{d^{O(d)}}$), so that the following two properties hold, which we refer to informally as *strong stability* (see Lemma 18 for a formal statement):

1. Each $\mathcal{L} \in \mathscr{R}_{\hat{g}}$ is irreducible and satisfies $\| \mathrm{SOA}_{\mathcal{L}} - \hat{g} \|_\infty \leq O(\chi \cdot d)$.
2. For some irreducible $\mathcal{L}^\star \subset \mathcal{F}$ depending only on $\mathcal{G}$, we have $\mathcal{L}^\star \in \mathscr{R}_{\hat{g}}$.

Given a collection of hypotheses $\hat{g}_1, \ldots, \hat{g}_M \in [K]^{\mathcal{X}}$ as above, if we run `SOAFilter` on each of the hypotheses $\hat{g}_i$, then the set $\hat{\mathscr{R}} := \mathscr{R}_{\hat{g}_1} \cup \cdots \cup \mathscr{R}_{\hat{g}_M}$ is of bounded size (namely, at most $M \cdot K^{d^{O(d)}}$), and as long as $\| \mathrm{SOA}_{\mathcal{G}} - \hat{g}_i \|_\infty \leq \chi$ for some $i \in [M]$ we have that $\mathcal{L}^\star \in \hat{\mathscr{R}}$ (item 2) and $\| \mathrm{SOA}_{\mathcal{L}^\star} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq O(\chi \cdot d)$ (item 1). These properties (in particular, that $\hat{\mathscr{R}}$ contains *exactly* the class $\mathcal{L}^\star$) are sufficient to apply the sparse selection procedure of Proposition 7, and thus obtain a private learning algorithm for $\mathcal{F}$. In Section 5.1, we describe a subroutine of `SOAFilter`, which we call `FilterStep`; we then describe `SOAFilter` in Section 5.2.

5.1. `FilterStep` algorithm

A challenge in achieving a strong stability guarantee as explained in the above paragraphs is that the class $\mathcal{F}$ could consist of too many functions with small oscillatory behavior: in particular, suppose that $\mathcal{F} = \{f : f(x) \in \{1,2\}\ \forall x \in \mathcal{X}\}$, so that $\mathrm{sfat}_2(\mathcal{F}) = 0$. Suppose that $\mathrm{SOA}_{\mathcal{G}}$ and $\hat{g}$ are arbitrary functions taking values in $\{1,2\}$; then $\| \hat{g} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq 1$. Moreover, each irreducible subclass $\mathcal{L} \subset \mathcal{F}$ satisfies $\| \mathrm{SOA}_{\mathcal{L}} - \hat{g} \|_\infty \leq 1$. Since we aim to have $|\mathscr{R}_{\hat{g}}| \leq K^{d^{O(d)}}$, and yet the number of irreducible subclasses $\mathcal{L} \subset \mathcal{F}$ could be much larger than this quantity, we will have to narrow down the set of subclasses $\mathcal{L}$ which can be added to $\mathscr{R}_{\hat{g}}$; this is done in the algorithm `FilterStep`, which "filters out" many $\mathcal{H} \subset \mathcal{F}$, and assigns to each $\mathcal{H}$ which is filtered out some $\mathcal{L} \subset \mathcal{F}$ which is not filtered out that is a good ℓ_∞ approximation of $\mathcal{H}$.

To describe the algorithm `FilterStep`, fix a class $\mathcal{F} \subset [K]^{\mathcal{X}}$. For $\ell \geq 0$ and $0 \leq b \leq d$, set

$$\mathscr{I}_{\ell,b}(\mathcal{F}) := \left\{ \mathcal{H} \subset \mathcal{F} : \substack{\mathcal{H} \text{ is a finite restriction subclass of } \mathcal{F} \\ \text{which is } \ell\text{-irreducible, and } \mathrm{sfat}_2(\mathcal{H}) = b} \right\}.$$

The algorithm `FilterStep` is presented in Algorithm 1. For an input positive integer $r_{\max}$ and

Algorithm 1: `FilterStep`

Input: A class $\mathcal{F}$ with $d := \mathrm{sfat}_2(\mathcal{F})$, and a sequence $(\ell_{r,t})_{r,t \geq 0}$ of positive integers that is non-decreasing in r, a parameter $r_{\max}$.

1. For each $t \in \{0, 1, \ldots, d\}$, set $\mathscr{L}_t \leftarrow \emptyset$.
2. For $0 \leq t \leq d$ and $0 \leq r \leq r_{\max}$, define $\mathscr{I}_{r,t} := \mathscr{I}_{\ell_{r,t}, d-t}(\mathcal{F})$. Also set $\mathscr{I}_{r_{\max}+1,t} := \emptyset$ for $0 \leq t \leq d$.
3. For $t \in \{0, 1, \ldots, d\}$:
   (a) For $r \in \{r_{\max}, r_{\max} - 1, \ldots, 0\}$:
      i. For each $\mathcal{H} \in \mathscr{I}_{r,t} \backslash \mathscr{I}_{r+1,t}$: *(Since the sequence $\ell_{r,t}$ is non-decreasing in r, we have $\mathscr{I}_{r+1,t} \subset \mathscr{I}_{r,t}$ for all r, t. Note that this step makes sense since $\mathscr{I}_{r,t}$ is countable; an arbitrary enumeration of $\mathscr{I}_{r,t}$ may be used.)*
         A. If there is some $\mathcal{L} \in \mathscr{L}_{d-t}$ and $\mathbf{A} \subset \mathcal{X} \times [K]$ with $|\mathbf{A}| \leq \ell_{r,t} - 1$ so that $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d - t$ and for all $(x, y) \in \mathbf{A}$, $\mathrm{SOA}_{\mathcal{L}}(x) = \mathrm{SOA}_{\mathcal{H}}(x) = y$, then set $\mathcal{L}_{\mathrm{rep}}(\mathcal{H}) \leftarrow \mathcal{L}$.
         B. Else, add $\mathcal{H}$ to $\mathscr{L}_{d-t}$, and set $\mathcal{L}_{\mathrm{rep}}(\mathcal{H}) \leftarrow \mathcal{H}$.
4. Output the sets $\mathscr{L}_t$, $0 \leq t \leq d$, as well as the mapping $\mathcal{L}_{\mathrm{rep}}(\cdot)$.

a sequence $(\ell_{r,t})_{r,t}$ defined for $0 \le r \le r_{\max}, 0 \le t \le d$, the algorithm defines a mapping $\mathcal{L}_{\text{rep}}(\cdot)$, which maps each $\mathcal{H} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$, for $0 \le r \le r_{\max}$ and $0 \le t \le d$, into some "filtered set" $\mathscr{L}_{d-t}$. For $\mathcal{H} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$, the class $\mathcal{L}_{\text{rep}}(\mathcal{H})$ should be interpreted as a representative of $\mathcal{H}$ which approximates it well, in the sense of the following lemma:

Lemma 14 *Fix inputs $\mathcal{F}, (\ell_{r,t})_{r,t\ge 0}, r_{\max}$ to* `FilterStep`*. For any $0 \le r \le r_{\max}, 0 \le t \le d$, and any $\mathcal{H} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$, we have that $\| \text{SOA}_{\mathcal{H}} - \text{SOA}_{\mathcal{L}_{\text{rep}}(\mathcal{H})} \|_\infty \le 1$.*

The algorithm `FilterStep` is designed so that its output sets $\mathscr{L}_{d-t}$, $0 \le t \le d$, satisfy the following sparsity-type property:

Lemma 15 *Fix inputs $\mathcal{F}, (\ell_{r,t})_{r,t\ge 0}, r_{\max}$ to* `FilterStep`*. For any $0 \le t \le d$ and $0 \le r \le r_{\max}$, and any $\mathbf{A} \subset \mathcal{X} \times [K]$ with $|\mathbf{A}| \le \ell_{r,t} - 1$ so that $\text{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d - t$, there is at most one element $\mathcal{L} \in \mathscr{L}_{d-t} \cap \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ so that for all $(x, y) \in \mathbf{A}$, $\text{SOA}_{\mathcal{L}}(x) = y$.*

5.2. Reducing trees and `SOAFilter`

In this section we describe the algorithm `SOAFilter` in full; before doing so, we introduce the notion of *reducing tree* in the following two definitions:

Definition 16 (Augmented tree) *For $d \ge 1, K \in \mathbb{N}$, an* augmented K-ary $\mathcal{X}$-valued tree *of depth d is defined exactly the same as a K-ary $\mathcal{X}$-valued tree (Definition 5), with the exception that there is a unique value of $k_1 \in [K]$ so that the sequence (k_1) lies in the domain of $\mathbf{x}_2$ (in particular, requirement 2 in Definition 5 is dropped for $t = 1$). Moreover, the only node associated with a sequence of length 1 is the node associated with (k_1). We will say that the augmented tree $\mathbf{x}$ is* rooted by the pair $(\mathbf{x}_1, k_1)$.

One should think of an augmented $\mathcal{X}$-labeled tree $\mathbf{x}$ of depth d which is rooted by the pair (x, k) as an $\mathcal{X}$-labeled tree $\mathbf{x}'$ of depth $d - 1$ for which we created a new root labeled by x and attached to it a single child (labeled by k), which is the root of the tree $\mathbf{x}'$. (Note that we have $\mathbf{x}_1 = x$ here.)

Definition 17 (Reducing tree) *Suppose $\mathcal{H} \subset [K]^{\mathcal{X}}$, and let $d := \text{sfat}_2(\mathcal{H})$. Fix an increasing sequence $(\ell_t)_{t\ge 0}$ of positive integers. Given a point $(x, y) \in \mathcal{X} \times [K]$ so that $\text{sfat}_2(\mathcal{H}|_{(x,y)}) < \text{sfat}_2(\mathcal{H})$, we say that an augmented K-ary $\mathcal{X}$-labeled tree $\mathbf{x}$ rooted by the pair (x, y) is a* reducing tree for the pair (x, y) and the sequence $(\ell_t)_{t\ge 0}$ *if any leaf v of the tree satisfies:*

- *$\mathcal{H}|_{\mathbf{A}(v)}$ is either empty or is ℓ_t-irreducible, where $t := d - \text{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$.*
- *$\text{depth}(v) \le \sum_{t'=0}^{t-1} \ell_{t'}$. Moreover, for any $1 \le \tilde{t} < t$, there is some node v' which is an ancestor of v so that $\text{sfat}_2(\mathcal{H}|_{\mathbf{A}(v')}) \le d - \tilde{t}$ and $\text{depth}(v') \le \sum_{t'=0}^{\tilde{t}-1} \ell_{t'}$.*

Lemma 36 in the appendix shows that reducing trees exist.

The algorithm `SOAFilter` is presented in Algorithm 2. It takes as input some hypothesis $\hat{g} : \mathcal{X} \to [K]$ and a class $\mathcal{F} \subset [K]^{\mathcal{X}}$, as well as parameters $\tau_{\max}, r_{\max} \in \mathbb{N}$. Its output is a set $\mathscr{R}_{\hat{g}}$, consisting of sub-classes of $\mathcal{F}$. The set $\mathscr{R}_{\hat{g}}$ should be interpreted as a set of "representatives" of $\hat{g}$ in the sense that for $\mathcal{L} \in \mathscr{R}_{\hat{g}}$, under appropriate conditions, we will have that $\text{SOA}_{\mathcal{L}}$ is a good ℓ_∞-approximation of $\hat{g}$ (i.e., $\| \text{SOA}_{\mathcal{L}} - \hat{g} \|_\infty$ is small); see Lemma 18 below.

The algorithm `SOAFilter` proceeds as follows. It first runs the algorithm `FilterStep` for the class $\mathcal{F}$, which produces "filtered sets" $\mathscr{L}_{d-t}, 0 \le t \le d$, of sub-classes of $\mathcal{F}$; each element

of $\mathscr{R}_{\hat{g}}$ will belong to some set $\mathscr{L}_{d-t}$. `SOAFilter` then tries to find finite sets $\mathbf{A} \subset \mathcal{X} \times [K]$ so that both (a) $\| \mathrm{SOA}_{\mathcal{F}|_{\mathbf{A}}} - \hat{g} \|_\infty$ is small and (b) so that for some $\mathcal{L}$ in one of the "filtered sets" $\mathscr{L}_{d-t}$ produced by `FilterStep`, it holds that $\mathrm{SOA}_{\mathcal{L}}(x) = y$ for each $(x, y) \in \mathbf{A}$; such sets $\mathcal{L}$ will be added to $\mathscr{R}_{\hat{g}}$ (step 4(a)ii). The sets $\mathbf{A}$ are built up gradually as follows: if some set $\mathbf{A}$ in the process of being built up is so that $\| \mathrm{SOA}_{\mathcal{F}|_{\mathbf{A}}} - \hat{g} \|_\infty$ is large, then we may choose some $x_{\mathbf{A}} \in \mathcal{X}$ so that $| \mathrm{SOA}_{\mathcal{H}}(x_{\mathbf{A}}) - \hat{g}(x_{\mathbf{A}}) |$ is large (step 4(a)iii). For y not too far from $\hat{g}(x_{\mathbf{A}})$, it will follow that we can construct a reducing tree with respect to the class $\mathcal{F}|_{\mathbf{A}}$ at the point $(x_{\mathbf{A}}, y)$ (step 4(a)ivA). For some of the leaves v of this reducing tree, we will then add $\mathbf{A}(v)$ to $\mathbf{A}$ to create a new set $\mathbf{A}'$ (one for each such leaf v), and continue to process each of these new sets $\mathbf{A}'$ (step 4(a)ivB). Intuitively, adding $\mathbf{A}(v)$ to $\mathbf{A}$ "restricts" the class of functions $\mathcal{F}|_{\mathbf{A}}$ under consideration so that all functions in it (and therefore its SOA hypothesis $\mathrm{SOA}_{\mathcal{F}|_{\mathbf{A}}}$) well-approximates $\hat{g}(x_{\mathbf{A}})$ at $x_{\mathbf{A}}$. Since for all leaves v of the reducing tree we must have that $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}(v) \cup \mathbf{A}}) < \mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}})$, this process must eventually terminate. We will show that some sequence of restrictions, corresponding to a choice of leaf of the reducing tree created at each step, will create some set $\mathbf{A}$ with our desired properties (a) and (b) above. All rigorous details of the algorithm are presented in Algorithm 2. Lemma 18 provides the main guarantee for `SOAFilter`.

Lemma 18 ("Strong stability") *Fix any positive integer $\bar{\ell}$. Suppose that $\mathcal{G} \subset \mathcal{F}$ is nonempty, $\hat{g} \in [K]^{\mathcal{X}}$, that $\| \mathrm{SOA}_{\mathcal{G}} - \hat{g} \|_\infty \leq \chi$ for some $\chi > 0$, and that $\mathcal{G}$ is $(\bar{\ell} \cdot (d+3)^d)$-irreducible. Then there is some $\bar{\ell}$-irreducible $\mathcal{L}^\star \subset \mathcal{F}$, depending only on $\mathcal{G}$, so that $\| \mathrm{SOA}_{\mathcal{L}^\star} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq (2 + 2\chi)(d+1) + 1$ and so that $\mathcal{L}^\star \in \mathscr{R}_{\hat{g}}$, where $\mathscr{R}_{\hat{g}}$ is the output of `SOAFilter` when given as inputs $\mathcal{F}$, $\hat{g}$, $r_{\max} = (d+1)$, $\tau_{\max} = (2 + 2\chi)(d+1)$ and the sequence $\ell_{r,t} := \bar{\ell} \cdot (r+2)^t$ for $0 \leq r \leq (d+1)$, $0 \leq t \leq d$.*

Moreover, all $\mathcal{L} \in \mathscr{R}_{\hat{g}}$ satisfy $\| \mathrm{SOA}_{\mathcal{L}} - \hat{g} \|_\infty \leq (2 + 2\chi)(d+1)$ and are $\bar{\ell}$-irreducible.

We provide a brief sketch of the proof of Lemma 18; the full proof is given in the appendix. The final statement of the lemma follows from step 5 of `SOAFilter`. To prove the remainder of the lemma, for $0 \leq \tau \leq (2 + 2\chi)(d+1)$ and $2 \leq r \leq (d+1)$, define $\mu(r, \tau) := \max_{(\mathcal{H}, \ell) \in \mathscr{G}_{r,\tau}} \{\mathrm{sfat}_2(\mathcal{H})\}$, where

$$\mathscr{G}_{r,\tau} := \left\{ (\mathcal{H}, \ell_{r,t}) : \begin{array}{c} \mathcal{H} \subset \mathcal{F} \text{ is } \ell_{r,t}\text{-irreducible and a finite restriction subclass of } \mathcal{F}, \\ \text{where } t = d - \mathrm{sfat}_2(\mathcal{H}), \text{ and } \| \mathrm{SOA}_{\mathcal{H}} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq \tau. \end{array} \right\}.$$

Since $\mathcal{G}$ is $\ell_{(d+1),d}$-irreducible, and for all t, r we have $\ell_{r,t} \leq \ell_{(d+1),d}$, we have that $(\mathcal{G}, \ell_{r,t}) \in \mathscr{G}_{r,\tau}$ for $t = d - \mathrm{sfat}_2(\mathcal{G})$ and all $0 \leq r \leq (d+1), 0 \leq \tau \leq (2 + 2\chi)(d+1)$, i.e., $\mathscr{G}_{r,\tau}$ is nonempty and so $\mu(r, \tau)$ is well-defined. It is straightforward to show, using that μ is non-decreasing in τ and non-increasing in r, that we can find some $r^\star, \tau^\star$ so that $\mu(r^\star, \tau^\star) = \mu(r^\star - 1, \tau^\star + 2 + 2\chi)$. Informally, this property of $r^\star, \tau^\star$ provides a source of "stability" which may be exploited to find some $\mathcal{L}^\star$ and show that it satisfies the claimed properties in Lemma 18.

We next explain how $\mathcal{L}^\star$ is defined: choose some $(\mathcal{H}^\star, \ell^\star)$ which achieves the maximum in (29) for $r = r^\star$, $\tau = \tau^\star$; letting $t^\star = d - \mathrm{sfat}_2(\mathcal{H}^\star)$, we have $\ell^\star = \ell_{r^\star, t^\star}$. Let $\mathcal{L}_{\mathrm{rep}}(\cdot)$ be the mapping defined as the output of `FilterStep` with the input class $\mathcal{F}$, the sequence $(\ell_{r,t})_{0 \leq r \leq r_{\max}, 0 \leq t \leq d}$, and $r_{\max} = d + 1$ (these are the parameters used in Step 1 of `SOAFilter`). Now set $\mathcal{L}^\star = \mathcal{L}_{\mathrm{rep}}(\mathcal{H}^\star) \in \mathscr{L}_{d-t^\star} \cap \mathscr{I}_{\ell_{r^\star,t^\star}, d-t^\star}(\mathcal{F})$; this is well-defined since $\mathcal{H}^\star \in \mathscr{I}_{\ell_{r^\star,t^\star}, d-t^\star}(\mathcal{F})$. It can be shown that $\mathcal{L}^\star$ satisfies the claimed properties of Lemma 18; full details are given in the appendix.

Finally, in Lemma 42 we show that $|\mathscr{R}_{\hat{g}}| \leq K^{\bar{\ell} \cdot (d+4)^d}$ for the parameter settings in Lemma 18.

Algorithm 2: `SOAFilter`

Input: Class $\mathcal{F} \subset [K]^{\mathcal{X}}$, $d := \mathrm{sfat}_2(\mathcal{F})$, sequence $(\ell_{r,t})_{r,t\geq 0}$, $r_{\max} \in \mathbb{N}$, tolerance parameter $\tau_{\max} \in \mathbb{N}$, $\chi \in \mathbb{N}$, $\hat{g} \in [K]^{\mathcal{X}}$. It is assumed that $r_{\max}, \tau_{\max}$ are multiples of $d+1$; let $r_0 := r_{\max}/(d+1), \tau_0 := \tau_{\max}/(d+1)$. Initialize $\mathscr{R}_{\hat{g}} \leftarrow \emptyset$.

1. Run the algorithm `FilterStep` (Algorithm 1) with $\mathcal{F}$, $(\ell_{r,t})_{r,t\geq 0}$, and $r_{\max}$ as input, and let the output sets be denoted $(\mathscr{L}_t)_{0\leq t\leq d}$.
2. For each $0 \leq s \leq d, 0 \leq j \leq d$, set $\mathscr{Q}_{j,s} \leftarrow \emptyset$. *($\mathscr{Q}_{j,s}$ will be a collection of finite subsets $\mathbf{A} \subset \mathcal{X} \times [K]$ defined for each index pair s, j.)*
3. Set $\mathscr{Q}_{j,0} \leftarrow \{\emptyset\}$ for each j (i.e., $\mathscr{Q}_{j,0}$ has a single element, which is the empty set).
4. For $j \in \{0, 1, \ldots, d\}$: let $r \leftarrow r_{\max} - jr_0 - 1$, $\tau \leftarrow j\tau_0 + 2 + \chi$:
   (a) For $s \in \{0, 1, \ldots, d\}$:
   - For each $\mathbf{A} \in \mathscr{Q}_{j,s}$, letting $\mathcal{H} := \mathcal{F}|_{\mathbf{A}}$:
     i. If $\mathcal{H}$ is empty, continue on with the next $\mathbf{A} \in \mathscr{Q}_{j,s}$.
     ii. If $\| \mathrm{SOA}_{\mathcal{H}} - \hat{g}\|_\infty \leq \tau$:
        – If there is some $\mathcal{L} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F}) \cap \mathscr{L}_{d-t}$ so that for all $(x,y) \in \mathbf{A}$, $\mathrm{SOA}_{\mathcal{L}}(x) = y$, then add any such $\mathcal{L}$ to $\mathscr{R}_{\hat{g}}$.
        – Continue (i.e., go to step 4(a)i with the next $\mathbf{A} \in \mathscr{Q}_{j,s}$).
     iii. Else, we have $\| \mathrm{SOA}_{\mathcal{H}} - \hat{g}\|_\infty > \tau$; then choose some $x_{\mathbf{A}} \in \mathcal{X}$ so that $|\mathrm{SOA}_{\mathcal{H}}(x_{\mathbf{A}}) - \hat{g}(x_{\mathbf{A}})| \geq \tau + 1$.
     iv. Let $k \leftarrow \hat{g}(x_{\mathbf{A}})$. For $y \in \{k-\tau+1 \vee 0, k-\tau+2 \vee 0, \ldots, k+\tau-1 \wedge K\}$:
        A. Let $t_{\mathbf{A}} := d - \mathrm{sfat}_2(\mathcal{H})$, and let $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$ be a reducing tree with respect to $\mathcal{H}$ for the point $(x_{\mathbf{A}}, y)$ and the sequence $(\ell_{r,t+t_{\mathbf{A}}})_{0\leq t\leq d-t_{\mathbf{A}}}$, as constructed per Lemma 36. *(Note that the reducing tree is well-defined since $|k - \mathrm{SOA}_{\mathcal{H}}(x_{\mathbf{A}})| \geq \tau + 1$ and so any y with $|y-k| \leq \tau - 1$ must satisfy $\mathrm{sfat}_2(\mathcal{H}|_{(x_{\mathbf{A}},y)}) < \mathrm{sfat}_2(\mathcal{H})$.)*
        B. For each leaf v of the tree $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$, if it is the case that (a) $\mathcal{F}|_{\mathbf{A}\cup\mathbf{A}(v)}$ is nonempty, and (b) for each $(x,y) \in \mathbf{A}(v)$, $|\hat{g}(x) - y| \leq \tau - 1$, then add $\mathbf{A} \cup \mathbf{A}(v)$ to $\mathscr{Q}_{j,s+1}$.
5. Remove all $\mathcal{L} \in \mathscr{R}_{\hat{g}}$ from $\mathscr{R}_{\hat{g}}$ with $\| \mathrm{SOA}_{\mathcal{L}} - \hat{g}\|_\infty > (2+2\chi)(d+1)$, then output $\mathscr{R}_{\hat{g}}$.

6. Putting it all together with `RegLearn`: on the proof of Theorem 1

Theorem 1 may be obtained as a reasonably straightforward consequence of the results presented in the previous sections; the full algorithm (`RegLearn`; Algorithm 4) is presented in the appendix. For positive integers n_0, m, we will draw $n := n_0 m$ samples (x,y) from some distribution P on $\mathcal{X} \times [K]$, and partition them into m groups of n_0 samples. For $1 \leq j \leq m$, the jth group of n_0 samples will be fed to the algorithm `ReduceTreeReg`, which outputs some $\{\hat{g}_1^{(j)}, \ldots, \hat{g}_{M_j}^{(j)}\}$ of candidate hypotheses, satisfying the weak stability guarrantee of Lemma 13. Then each of $\hat{g}_1^{(j)}, \ldots, \hat{g}_{M_j}^{(j)}$ will be fed to `SOAFilter`, which produces an output set $\mathscr{R}_{\hat{g}_i^{(j)}}$ for each $1 \leq i \leq M_j$, consisting of hypotheses all of which have low population error. The combination of Lemma 13 and the strong stability property of Lemma 18 gives that there is some hypothesis $h^\star : \mathcal{X} \to [K]$, depending only on $\mathcal{F}, P$, so that with probability $1/O(d)$ over the n_0 samples, $h^\star \in \hat{\mathscr{R}}^{(j)} := \bigcup_{i=1}^{M_j} \mathscr{R}_{\hat{g}_i^{(j)}}$. We will also be able to bound $|\mathscr{R}^{(j)}|$ by $K^{2^{\tilde{O}(d)}}$. Then we will apply Proposition 7 with m users whose sets

are $\hat{\mathscr{R}}^{(1)}, \ldots, \hat{\mathscr{R}}^{(m)}$. By choosing the number of groups m to be large enough, we may ensure that some $h^\star$ occurs in a number of groups greater than the additive error in Proposition 7, which ensures that the private sparse selection algorithm outputs some such $h^\star$ with high probability. Full details of the proof are presented in Appendix E.

7. Conclusion and future work

In this paper we showed that the condition $\liminf_{\eta \downarrow 0} \eta \cdot \mathrm{sfat}_\eta(\mathcal{H}) = 0$ is sufficient for the class $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$ to be privately learnable. A natural question is whether this growth condition can be relaxed; it seems that new techniques will be required even to prove that all classes $\mathcal{H}$ with $\eta \cdot \mathrm{sfat}_\eta(\mathcal{H}) \leq 1$ for all $\eta > 0$ are privately learnable, if this is even true (such classes are all online learnable since $\mathrm{sfat}_\eta(\mathcal{H})$ is necessarily finite). An example of a natural hypothesis class for which our growth condition is not satisfied is infinite-dimensional ℓ_2 regression: in particular, set $\mathcal{X} = \ell_2^\infty = \{(x_1, x_2, \ldots) : \ x_i \in \mathbb{R}, \sum_{i=1}^\infty x_i^2 \leq 1\}$ and $\mathcal{H} = \ell_2^\infty = \{(w_1, w_2, \ldots) : \ w_i \in \mathbb{R}, \sum_{i=1}^\infty w_i^2 \leq 1\}$, and then for $h = (w_1, w_2, \ldots)$ and $x = (x_1, x_2, \ldots)$, define $h(x) := \langle w, x \rangle = \sum_{i=1}^\infty w_i x_i$. It can be shown that $\mathrm{sfat}_\eta(\mathcal{H}) \asymp 1/\eta^2 \gg 1/\eta$ as $\eta \to 0$.

Another interesting question is whether the sample complexity bound of Theorem 1 can be improved to one that is polynomial in $\mathrm{sfat}_\eta(\mathcal{H})$; for the setting of binary classification, it *is* possible to obtain sample complexity bounds polynomial in the appropriate complexity parameter for online learnability, namely the Littlestone dimension (Ghazi et al., 2020a).

Acknowledgments

I am grateful to Sasha Rakhlin and Roi Livni for helpful suggestions.

References

Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam D. Smith, and Salil P. Vadhan. Differentially private simple linear regression. *CoRR*, abs/2007.05157, 2020. URL https://arxiv.org/abs/2007.05157.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615631, July 1997.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 852860, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, USA, 1st edition, 2009.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463482, March 2003. ISSN 1532-4435.

Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434 – 452, 1996.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, page 464473, USA, 2014. IEEE Computer Society. ISBN 9781479965175.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40328-6.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94:401–437, 2014.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *JMLR*, 20(146):1–33, 2019.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

Garrett Bernstein and Daniel R. Sheldon. Differentially private bayesian linear regression. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 523–533, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f90f2aca5c640289d0a29417bcb63a37-Abstract.html.

Mark Bun. A computational separation between private learning and online learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649, 2015. doi: 10.1109/FOCS.2015.45.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *ITCS*, page 369380, 2016. ISBN 9781450340571.

Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing - STOC 2018*, pages 74–86, Los Angeles, CA, USA, 2018. ACM Press. ISBN 978-1-4503-5559-9. doi: 10.1145/3188745.3188946. URL http://dl.acm.org/citation.cfm?doid=3188745.3188946.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *Proceedings of the 61st Annual IEEE Symposium of Foundations of Computer Science (FOCS '20)*, 2020.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *CoRR*, abs/1902.04495, 2019.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 289–296. Curran Associates, Inc., 2009.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12(null):10691109, July 2011. ISSN 1532-4435.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013. ISSN 1551-305X, 1551-3068. doi: 10.1561/0400000042. URL http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1000–1019, Barcelona, Spain, 13–15 Jun 2014. PMLR.

Dylan J. Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 26262637, Red Hook, NY, USA, 2018. Curran Associates Inc.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper private pac learning. *arXiv:2012.03893*, 2020a.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Young Hun Jung, Baekjin Kim, and Ambuj Tewari. On the equivalence between online and private learnability beyond binary classification. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *COLT*, pages 2263–2285, 2020.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Rashkodnikova, and Adam Smith. What can we learn privately? In *FOCS*, pages 531–540, 2008.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464 – 497, 1994.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *FOCS*, pages 68–77, 1987.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55, 2003.

Seth Neel, Aaron Roth, and Zhiwei Steven Wu. How to use heuristics for differential privacy. In *FOCS*, pages 72–93, 2019.

Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, and Salil Vadhan. Bridging the gap between computer science and legal approaches to privacy. *Harvard Journal of Law & Technology*, 31:687–780, 2016 2018.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvri, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1232–1264, Barcelona, Spain, 13–15 Jun 2014a. PMLR.

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction. 2014b.

Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1704–1722, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(6):155–186, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161:111–153, 2015b.

Aaron Roth and Michael Kearns. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

Salil Vadhan. The Complexity of Differential Privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, Cham, 2017.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 93–103. AUAI Press, 2018.

Appendix A. Additional preliminaries

In this section we introduce some additional preliminaries which will be useful in our proofs.

A.1. Fat-shattering dimension and uniform convergence

In this section we overview some uniform convergence properties of real-valued classes and their discretizations. For a class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ and $\eta > 0$, the *η-fat shattering dimension* of $\mathcal{H}$ is defined as the largest positive integer d so that there are d points $x_1, \ldots, x_d \in \mathcal{X}$ and real numbers $s_1, \ldots, s_d \in \mathbb{R}$ so that for each $b = (b_1, \ldots, b_d) \in \{0,1\}^d$, there is a function $h \in \mathcal{H}$ so that, for $1 \le i \le d$, $h(x_i) \ge s_i + \eta$ if $b_i = 1$, and $h(x_i) \le s_i - \eta$ if $b_i = 0$.

We will use the following result showing that finiteness of the fat-shattering dimension of $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$ implies that it exhibits uniform convergence.

Theorem 19 (Uniform convergence; Mendelson and Vershynin (2003)) *There are constants $C_0 \ge 1$ and $0 < c_0 \le 1$ so that the following holds. For any $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, any distribution Q on $\mathcal{X} \times [-1,1]$, and any $\gamma \in (0,1)$, it holds that*

$$\Pr_{S_n \sim Q^n}\left[\sup_{h\in\mathcal{H}}\left|\mathrm{err}_Q(h) - \mathrm{err}_{\hat{Q}_{S_n}}(h)\right| > C_0 \cdot \left(\inf_{\eta \ge 0}\left\{\eta + \frac{1}{\sqrt{n}}\int_\eta^1 \sqrt{\mathrm{fat}_{c_0\eta'}(\mathcal{H})\log(1/\eta')}\, d\eta'\right\} + \sqrt{\frac{\log(1/\gamma)}{n}}\right)\right] \le \gamma. \tag{1}$$

The specific form of Theorem 19 may be derived from (Rakhlin and Sridharan, 2014b, Corollary 12.8) (which is a corollary of (Mendelson and Vershynin, 2003, Theorem 1)) by applying the symmetrization lemma together with McDiarmid's inequality (see the proof of Theorem 8 in Bartlett and Mendelson (2003)). By upper bounding the integral in (1) by $\sqrt{\mathrm{fat}_{c_0\eta}(\mathcal{H})\log(1/\eta)}$ for some choice of $\eta \in (0,1)$, we obtain the following consequence, which only depends on the fat-shattering dimension of $\mathcal{H}$ at a single scale $c_0\eta$, yet may be weaker than Theorem 19.

Corollary 20 (Uniform convergence, simplified) *There are constants $C_0 \ge 1$ and $0 < c_0 \le 1$ so that the following holds. For any $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, any distribution Q on $\mathcal{X} \times [-1,1]$, and any $\gamma \in (0,1/2), \eta \in (0,1/2)$, it holds that, for any*

$$n \ge C_0 \cdot \frac{\mathrm{fat}_{c_0\eta}(\mathcal{H})\log(1/\eta) + \log(1/\gamma)}{\eta^2},$$

we have

$$\Pr_{S_n \sim Q^n}\left[\sup_{h\in\mathcal{H}}\left|\mathrm{err}_Q(h) - \mathrm{err}_{\hat{Q}_{S_n}}(h)\right| > \eta\right] \le \gamma. \tag{2}$$

A.2. Uniform convergence for discretized classes

Recall that we defined discretized classes and distributions in Section 2.1. In this section we state (straightforward) consequences of Corollary 20 for such discretized classes.

For $y, y' \in [-1,1]$, note that

$$\frac{\lceil 2/\eta \rceil \cdot |y - y'|}{2} - 1 \le \left| \lfloor y \rfloor_\eta - \lfloor y' \rfloor_\eta \right| \le \frac{\lceil 2/\eta \rceil \cdot |y - y'|}{2} + 1,$$

Therefore, for $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, a distribution Q on $\mathcal{X} \times [-1,1]$, and any $h \in \mathcal{H}$, we have that

$$\frac{\lceil 2/\eta \rceil \cdot \mathrm{err}_Q(h)}{2} - 1 \le \mathrm{err}_{\lfloor Q \rfloor_\eta}(\lfloor h \rfloor_\eta) \le \frac{\lceil 2/\eta \rceil \cdot \mathrm{err}_Q(h)}{2} + 1. \tag{3}$$

Using (3), we have the following corollary of Corollary 20 showing a uniform convergence result for the discretized class corresponding to a class of finite fat-shattering dimension.

Corollary 21 *There are constants $C_0 \ge 1$ and $0 < c_0 \le 1$ so that the following holds. For any $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, any distribution Q on $\mathcal{X} \times [-1,1]$, and any $\gamma \in (0,1/2), \alpha \in (0,1/2)$, it holds that, for any*

$$n \ge C_0 \cdot \frac{\mathrm{fat}_{c_0\alpha}(\mathcal{H}) \log(1/\alpha) + \log(1/\gamma)}{\alpha^2}, \tag{4}$$

we have

$$\Pr_{S_n \sim Q^n}\left[\sup_{h \in \mathcal{H}} \left| \mathrm{err}_{\lfloor Q \rfloor_\alpha}(\lfloor h \rfloor_\alpha) - \mathrm{err}_{\lfloor \hat{Q}_{S_n} \rfloor_\alpha}(\lfloor h \rfloor_\alpha) \right| > 3 \right] \le \gamma. \tag{5}$$

Proof [Proof of Corollary 21] Suppose first that $\mathrm{err}_{\lfloor Q \rfloor_\alpha}(\lfloor h \rfloor_\alpha) - \mathrm{err}_{\lfloor \hat{Q}_{S_n} \rfloor_\alpha}(\lfloor h \rfloor_\alpha) > 0$. Then, with probability at least $1 - \gamma$ over $S_n \sim Q^n$, as long as C_0 and c_0 in (4) are sufficiently large and small, respectively,

$$\begin{aligned}
&\mathrm{err}_{\lfloor Q \rfloor_\alpha}(\lfloor h \rfloor_\alpha) - \mathrm{err}_{\lfloor \hat{Q}_{S_n} \rfloor_\alpha}(\lfloor h \rfloor_\alpha) \\
&\le \left(\frac{\lceil 2/\alpha \rceil \cdot \mathrm{err}_Q(h)}{2} + 1 \right) - \left(\frac{\lceil 2/\alpha \rceil \cdot \mathrm{err}_{\hat{Q}_{S_n}}(h)}{2} - 1 \right) && (6) \\
&= \frac{\lceil 2/\alpha \rceil}{2} \cdot (\mathrm{err}_Q(h) - \mathrm{err}_{\hat{Q}_{S_n}}(h)) + 2 \\
&\le \frac{\lceil 2/\alpha \rceil}{2} \cdot 2/\lceil 2/\alpha \rceil + 2 && (7) \\
&= 3, && (8)
\end{aligned}$$

where (6) follows from (3), and (7) follows from Corollary 20 with $\eta = 2/\lceil 2/\alpha \rceil = \Theta(\alpha)$ (and holds with probability at least $1-\gamma$ over $S_n \sim Q^n$). The case that $\mathrm{err}_{\lfloor Q \rfloor_\alpha}(\lfloor h \rfloor_\alpha) - \mathrm{err}_{\lfloor \hat{Q}_{S_n} \rfloor_\alpha}(\lfloor h \rfloor_\alpha) < 0$ is handled similarly. ■

The following result, also a consequence of Corollary 20, is similar to Corollary 21, but it states the sample complexity bound in terms of the quantity $\mathrm{sfat}_2(\mathcal{F})$ of a discretized class $\mathcal{F}$, at the expense of having a larger constant in (9) (not explicitly computed here; compare to (5)). Strictly speaking, Corollary 21 is not necessary for our purposes, but we use it to improve certain constants in our bounds.

Corollary 22 *There are constant $C_0, C_1 \geq 1$ so that the following holds. For any $K \in \mathbb{N}$, $\mathcal{F} \subset [K]^{\mathcal{X}}$, any distribution P on $\mathcal{X} \times [K]$, and any $\gamma \in (0, 1/2)$, it holds that, for any*

$$n \geq C_0 K^2 \cdot (\mathrm{fat}_2(\mathcal{F}) \log(K) + \log(1/\gamma)),$$

we have

$$\Pr_{S_n \sim P^n}\left[\sup_{f \in \mathcal{F}} \left|\mathrm{err}_P(f) - \mathrm{err}_{\hat{P}_{S_n}}(f)\right| > C_1\right] \leq \gamma. \tag{9}$$

Proof Define the class $\tilde{\mathcal{F}} \subset [-1,1]^{\mathcal{X}}$ as follows: for each $f \in \mathcal{F}$, there is a function $\tilde{f} \in \tilde{\mathcal{F}}$, defined as $\tilde{f}(x) := \frac{2}{K} f(x) - 1$. Note that $\mathrm{fat}_2(\mathcal{F}) = \mathrm{fat}_{1/K}(\tilde{\mathcal{F}})$. Let c_0, C_0 be the constants of Corollary 20. Using Corollary 20 with $\eta = \frac{1}{c_0 K}$, we have that for $n \geq \frac{K^2 C_0}{c_0^2} \cdot (\mathrm{fat}_2(\mathcal{F}) \log(C_0 K) + \log(1/\gamma))$, it holds that for any distribution Q on $\mathcal{X} \times [-1,1]$,

$$\Pr_{S_n \sim Q^n}\left[\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left|\mathrm{err}_Q(\tilde{f}) - \mathrm{err}_{\hat{Q}_{S_n}}(\tilde{f})\right| > \frac{1}{c_0 K}\right] \leq \gamma.$$

The claimed statement (9) follows by setting $C_1 = 1/(2c_0)$ and increasing C_0 by a sufficiently large amount. ■

We may upper bound the fat-shattering dimension and the sequential fat-shattering dimension of $\lfloor \mathcal{H} \rfloor_\eta$ in terms of the corresponding quantities for $\mathcal{H}$:

Lemma 23 *Suppose $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, and $\eta > 0$. Then it holds that $\mathrm{fat}_2(\lfloor \mathcal{H} \rfloor_\eta) \leq \mathrm{fat}_\eta(\mathcal{H})$, and $\mathrm{sfat}_2(\lfloor \mathcal{H} \rfloor_\eta) \leq \mathrm{sfat}_\eta(\mathcal{H})$.*

Proof This follows from the fact that for any $\lfloor h \rfloor_\eta \in \lfloor \mathcal{H} \rfloor_\eta$ and any $s \in \mathbb{R}$, if it holds that $|\lfloor h \rfloor_\eta(x) - s| \geq 1$, then since

$$\left|\left(\frac{2(\lfloor h \rfloor_\eta(x) - 1)}{\lceil 2/\eta \rceil} - 1\right) - h(x)\right| \leq \eta/2$$

for all $x \in \mathcal{X}$, and

$$\left|\left(\frac{2(\lfloor h \rfloor_\eta(x) - 1)}{\lceil 2/\eta \rceil} - 1\right) - \left(\frac{2(s-1)}{\lceil 2/\eta \rceil} - 1\right)\right| \geq \frac{2}{\lceil 2/\eta \rceil} \geq \eta,$$

we must have that

$$\left|\left(\frac{2(s-1)}{\lceil 2/\eta \rceil} - 1\right) - h(x)\right| \geq \eta/2.$$

■

A.3. Attaching a tree via a node

The following definition will be useful when arguing about trees in the context of irrecucibility:

Definition 24 (Attaching a tree via a node) *Suppose that $\mathbf{x}, \mathbf{x}'$ are K-ary $\mathcal{X}$-valued trees of depths d and $d' \geq 1$, respectively, and that v is a leaf of $\mathbf{x}$, corresponding to some tuple $(\bar{k}_1, \dots, \bar{k}_{t_0}) \in [K]^t$ (in particular, the depth of v is t_0). We say that the tree $\mathbf{x}''$ is obtained by* attaching the tree $\mathbf{x}'$ to $\mathbf{x}$ via the leaf v, *where $\mathbf{x}''$ is the depth-$(d' + t_0)$ tree defined as follows: for all $1 \leq t \leq d' + t_0$, and $k_1, \dots, k_{t-1} \in [K]$,*

$$\mathbf{x}''_t(k_1, \dots, k_{t-1}) = \begin{cases} \mathbf{x}_t(k_1, \dots, k_{t-1}) & : t \leq t_0 \text{ or } (k_1, \dots, k_{t-1}) \neq (\bar{k}_1, \dots, \bar{k}_{t-1}) \\ \mathbf{x}'_{t-t_0}(k_{t_0+1}, k_{t_0+2}, \dots, k_{t-1}) & : t > t_0 \text{ and } (k_1, \dots, k_{t-1}) = (\bar{k}_1, \dots, \bar{k}_{t-1}). \end{cases}$$

(If, in either case above, either $\mathbf{x}_t(k_1, \dots, k_{t-1})$ or $\mathbf{x}'_{t-t_0}(k_{t_0+1}, \dots, k_{t-1})$ is not defined, then $\mathbf{x}''_t(k_1, \dots, k_{t-1})$ is not defined, i.e., $(k_1, \dots, k_{t-1})$ is not in the domain of $\mathbf{x}''_t$.)

In words, $\mathbf{x}''$ is obtained as follows: the node v is given the label of $\mathbf{x}'_1$, and the sub-tree of $\mathbf{x}''$ rooted at v is identical to $\mathbf{x}'$ (and otherwise is identical to $\mathbf{x}$).

A.4. Laplace distribution

For a positive real number $b > 0$, write $\mathrm{Lap}(b)$ to denote the random variable $X \in \mathbb{R}$ with probability density function $\Pr[X = x] = \frac{1}{2b} \exp(-|x|/b)$. A straightforward computation gives that for any $t > 0$, $\Pr[|X| \geq t \cdot b] = \exp(-t)$.

Appendix B. Proofs for Section 3: irreducibility

This section presents basic properties of the notion of irreducibility from Definition 9. Some of the results are analogous to those in the setting for classification (Ghazi et al., 2020a); this is indicated where it is the case.

B.1. Basic properties of irreducibility

Lemma 10 *Suppose $\mathcal{G} \subset [K]^{\mathcal{X}}$ is irreducible. Then there are at most 2 values of $k \in [K]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$, and if there are 2 values, they differ by 1.*

Proof Let $d := \mathrm{sfat}_2(\mathcal{G})$, and suppose without loss of generality that $k > k'$. Suppose for the purpose of contradiction that for some $k, k' \in [K]$ with $|k - k'| \geq 2$, we have $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G}|_{(x,k')}) = \mathrm{sfat}_2(\mathcal{G}) = d$. Let $\mathbf{x}, \mathbf{y}$ be complete binary trees of depth d shattered by $\mathcal{G}|_{(x,k)}, \mathcal{G}|_{(x,k')}$, respectively, witnessed by trees $\mathbf{s}, \mathbf{t}$, respectively. We construct a tree $\mathbf{z}$ of depth $d + 1$ shattered by $\mathcal{G}$, as follows: for any $k_2, \dots, k_{d+1} \in \{1, 2\}$, set $\mathbf{z}_{t+1}(1, k_2, \dots, k_t) = \mathbf{x}_t(k_2, \dots, k_t)$, $\mathbf{z}_{t+1}(2, k_2, \dots, k_t) = \mathbf{y}_t(k_2, \dots, k_t)$ for $1 \leq t \leq d$ and $\mathbf{z}_1 = x$. (In words, we are setting $\mathbf{x}, \mathbf{y}$ to be the left and right subtrees of a node labeled by x.) We claim that $\mathbf{z}$ is 2-shattered by $\mathcal{G}$: indeed, a witness $\mathbf{r}$ may be defined as follows: define $\mathbf{r}_{t+1}(1, k_2, \dots, k_t) = \mathbf{s}_t(k_2, \dots, k_t)$, $\mathbf{r}_{t+1}(2, k_2, \dots, k_t) = \mathbf{t}_t(k_2, \dots, k_t)$ for $1 \leq t \leq d$, and $\mathbf{r}_1 = \frac{k+k'}{2}$. (In words, $\mathbf{r}$ is the tree rooted by a node labeled by $\frac{k+k'}{2}$, whose left and right subtrees are given by $\mathbf{s}, \mathbf{t}$, respectively.) That $\mathbf{r}$ witnesses the shattering follows from the fact that $\mathbf{s}, \mathbf{t}$ are witnesses to the shattering of $\mathcal{G}|_{(x,k)}, \mathcal{G}|_{(x,k')}$ by $\mathbf{x}, \mathbf{y}$, respectively, and the fact that for any $f \in \mathcal{G}|_{(x,k)}$, $f' \in \mathcal{G}|_{(x,k')}$, we have that $f(\mathbf{z}_1) - \mathbf{r}_1 \geq 1$ and $-(f'(\mathbf{z}_1) - \mathbf{r}_1) \geq 1$. ■

Lemma 25 *Suppose $\mathcal{G} \subset [K]^{\mathcal{X}}$ has $\mathrm{sfat}_2(\mathcal{G}) = 0$. Then $\mathcal{G}$ is ℓ-irreducible for all $\ell \in \mathbb{N}$.*

Proof Let $\mathbf{x}$ be a K-ary $\mathcal{X}$-valued tree of depth ℓ. Since $\bigcup_{\text{leaves } v \text{ of } \mathbf{x}} \mathcal{G}|_{\mathbf{A}(v)} = \mathcal{G}$, the tree $\mathbf{x}$ must have some leaf v so that $\mathcal{G}|_{\mathbf{A}(v)}$ is nonempty. For such v, we must have that $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(v)}) \geq 0$, and since $\mathcal{G}|_{\mathbf{A}(v)} \subset \mathcal{F}$, we have $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(v)}) = 0$, as desired. ■

Lemma 26 *Suppose $\mathcal{G} \subset [K]^{\mathcal{X}}$ is ℓ-irreducible for $\ell \geq 1$. Then for any $x \in \mathcal{X}$, there is some $k \in [K]$ so that $\mathcal{G}|_{(x,k)}$ is $(\ell-1)$-irreducible and $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$.*

Proof The statement of the lemma follows immediately from Definition 9 if $\ell = 1$, so we may assume from here on that $\ell \geq 2$.

Fix any $x \in \mathcal{X}$. Our goal is to show that there is some $k \in [K]$ so that the following holds: for any K-ary $\mathcal{X}$-valued tree $\mathbf{x}$, of depth $\ell - 1$, $\mathbf{x}$ has some leaf v so that $\mathrm{sfat}_2(\mathcal{G}|_{\{(x,k)\} \cup \mathbf{A}(v)}) = \mathrm{sfat}_2(\mathcal{G})$. We now consider two cases:

Case 1. There is a unique $k' \in [K]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k')}) = \mathrm{sfat}_2(\mathcal{G})$. In this case, we set $k = k'$. Now consider any K-ary $\mathcal{X}$-valued tree $\mathbf{x}$ of depth $\ell - 1$. Let $\tilde{\mathbf{x}}$ be the tree of depth ℓ whose root is given by x and so that each child of the root is a copy of the tree $\mathbf{x}$; formally, for $k_1, \ldots, k_\ell \in [K]$, $\tilde{\mathbf{x}}_{t+1}(k_1, \ldots, k_t) = \mathbf{x}_t(k_2, \ldots, k_t)$ for $1 \leq t \leq \ell - 1$, and $\tilde{\mathbf{x}}_1 = x$. The ℓ-irreducibility of $\mathcal{G}$ guarantees the existence of some tuple $k_1, \ldots, k_\ell$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k_1),(\tilde{\mathbf{x}}_2(k_1),k_2),\ldots,(\tilde{\mathbf{x}}_\ell(k_{1:\ell-1}),k_\ell)}) = \mathrm{sfat}_2(\mathcal{G})$. Since for all $k' \neq k$, we have $\mathrm{sfat}_2(\mathcal{G}|_{(x,k')}) < \mathrm{sfat}_2(\mathcal{G})$, it holds that $k_1 = k$. Letting v be the leaf of $\mathbf{x}$ associated to the tupe $(k_2, \ldots, k_\ell)$, we see that $\mathrm{sfat}_2(\mathcal{G}|_{\{(x,k)\} \cup \mathbf{A}(v)}) = \mathrm{sfat}_2(\mathcal{G})$, as desired.

Case 2. For some $k_0 \in [K]$, it holds that $\mathrm{sfat}_2(\mathcal{G}|_{(x,k_0)}) = \mathrm{sfat}_2(\mathcal{G}|_{(x,k_0+1)}) = \mathrm{sfat}_2(\mathcal{G})$, and for all $k' \neq k_0$, $\mathrm{sfat}_2(\mathcal{G}|_{(x,k')}) < \mathrm{sfat}_2(\mathcal{G})$ (see Lemma 10). Suppose for the purpose of contradiction that there did not exist a choice of $k \in \{k_0, k_0+1\}$ so that $\mathcal{G}|_{(x,k)}$ is $(\ell-1)$-irreducible. Then for each $k \in \{k_0, k_0+1\}$, there is some tree $\mathbf{x}^{(k)}$ of depth $\ell - 1$ so that for any choice of $k_2, \ldots, k_\ell \in [K]$ we have $\mathrm{sfat}_2(\mathcal{G}|_{(x,k),(\mathbf{x}_1^{(k)},k_2),\ldots,(\mathbf{x}_{\ell-1}^{(k)}(k_{2:\ell-1}),k_{\ell-1})}) < \mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$.

Now let $\tilde{\mathbf{x}}$ be the tree of depth ℓ whose root is given by x, so that the k'-th child of the root, for $k' \neq k_0 + 1$, is a copy of the tree $x^{(k_0)}$, and so that the (k_0+1)-th child of the root is a copy of the tree $x^{(k_0+1)}$. (The k'-th children of the root for $k' \in \{k_0, k_0+1\}$ can in fact be arbitrary.) Formally, for $k_1, \ldots, k_\ell \in [K]$, we have $\tilde{\mathbf{x}}_1 = x$ and

$$\tilde{\mathbf{x}}_{t+1}(k_1, \ldots, k_t) = \begin{cases} \mathbf{x}^{(k_0)}(k_2, \ldots, k_t) & : \quad k_1 \neq k_0 + 1 \\ \mathbf{x}^{(k_0+1)}(k_2, \ldots, k_t) & : \quad k_1 = k_0 + 1. \end{cases}$$

Now consider any sequence $(k_1, \ldots, k_\ell)$, and let its associated leaf in $\tilde{\mathbf{x}}$ be denoted v. If $k_1 \notin \{k_0, k_0+1\}$, then $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(v)}) \leq \mathrm{sfat}_2(\mathcal{G}|_{(x,k_1)}) < \mathrm{sfat}_2(\mathcal{G})$. If $k_1 \in \{k_0, k_0+1\}$, then

$$\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(v)}) = \mathrm{sfat}_2(\mathcal{G}|_{(x,k_1),(\mathbf{x}_1^{(k_1)},k_2),\ldots,(\mathbf{x}_{\ell-1}^{(k_1)}(k_2:k_{\ell-1}),k_\ell)}) < \mathrm{sfat}_2(\mathcal{G}|_{(x,k_1)}) = \mathrm{sfat}_2(\mathcal{G}).$$

This contradicts the ℓ-irreducibility of $\mathcal{G}$, completing the proof. ■

The following lemma is analogous to (Ghazi et al., 2020a, Lemma 4.2):

Lemma 27 *Suppose $\mathcal{H} \subset \mathcal{G} \subset [K]^{\mathcal{X}}$, and that $\mathrm{sfat}_2(\mathcal{G}) = \mathrm{sfat}_2(\mathcal{H})$. If $\mathcal{H}$ is ℓ-irreducible, then so is $\mathcal{G}$.*

Proof The ℓ-irreducibility of $\mathcal{H}$ implies that for any K-ary $\mathcal{X}$-valued tree $\mathbf{x}$ of depth ℓ, there is some choice of $k_1, \dots, k_\ell \in [K]$ so that

$$\mathrm{sfat}_2(\mathcal{G}|_{(\mathbf{x}_1,k_1),\dots,(\mathbf{x}_\ell(k_{1:\ell-1}),k_\ell)}) \geq \mathrm{sfat}_2(\mathcal{H}|_{(\mathbf{x}_1,k_1),\dots,(\mathbf{x}_\ell(k_{1:\ell-1}),k_\ell)}) = \mathrm{sfat}_2(\mathcal{H}) = \mathrm{sfat}_2(\mathcal{G}).$$

But since $\mathcal{G}|_{(\mathbf{x}_1,k_1),\dots,(\mathbf{x}_\ell(k_{1:\ell-1}),k_\ell)} \subset \mathcal{G}$, the inequality above must be an equality, and this ensures that $\mathcal{G}$ is ℓ-irreducible. ■

B.2. Properties of SOA hypotheses

Lemma 12 *Suppose $\mathcal{H} \subset \mathcal{G}$, $\mathrm{sfat}_2(\mathcal{H}) = \mathrm{sfat}_2(\mathcal{G})$, and that $\mathcal{H}$ is irreducible. Then for all $x \in \mathcal{X}$, $|\mathrm{SOA}_{\mathcal{H}}(x) - \mathrm{SOA}_{\mathcal{G}}(x)| \leq 1$.*

Proof Fix any $x \in \mathcal{X}$, and let $k := \mathrm{SOA}_{\mathcal{H}}(x)$. Then $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) \geq \mathrm{sfat}_2(\mathcal{H}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{H}) = \mathrm{sfat}_2(\mathcal{G})$, and so $\mathrm{sfat}_2(\mathcal{G}|_{(x,k)}) = \mathrm{sfat}_2(\mathcal{G})$. By Lemma 10 and Definition 11, we have that $\mathrm{SOA}_{\mathcal{G}}(x) \in \{k-1, k, k+1\}$, as desired. ■

Lemma 28 *Suppose $\mathcal{G} \subset [K]^{\mathcal{X}}$ is ℓ-irreducible. Consider any $\ell' \leq \ell$, and any set $\mathbf{A} \subset \mathcal{X} \times [K]$ of size $|\mathbf{A}| \leq \ell'$, so that each $(x, y) \in \mathbf{A}$ satisfies $y = \mathrm{SOA}_{\mathcal{G}}(x)$. Then $\mathcal{G}' := \mathcal{G}|_{\mathbf{A}}$ is $(\ell - \ell')$-irreducible and satisfies $\mathrm{sfat}_2(\mathcal{G}') = \mathrm{sfat}_2(\mathcal{G})$.*

Proof We first prove the statement for the case $\ell' = 1$. Consider some $(x, y) \in \mathcal{X} \times [K]$, so that $y = \mathrm{SOA}_{\mathcal{G}}(x)$. By Definition 11, for $\mathcal{G}' := \mathcal{G}|_{(x,y)}$, we have $\mathrm{sfat}_2(\mathcal{G}') = \mathrm{sfat}_2(\mathcal{G})$. By Lemma 26, there is some $y' \in [K]$ so that $\mathrm{sfat}_2(\mathcal{G}|_{(x,y')}) = \mathrm{sfat}_2(\mathcal{G})$ and so that $\mathcal{G}|_{(x,y')}$ is $(\ell - 1)$-irreducible. By Definition 11 we must have $y \in \{y' - 1, y', y' + 1\}$ and $\mathcal{G}|_{(x,y)}$ is $(\ell - 1)$-irreducible as well.

We now prove the statement for general ℓ' by induction. Suppose the statement holds for some value $\ell' < \ell$. Consider some set $\mathbf{A} \subset \mathcal{X} \times [K]$ of size $|\mathbf{A}| = \ell' + 1$, and write $\mathbf{A} = \tilde{\mathbf{A}} \cup \{(x, y)\}$, for $|\tilde{\mathbf{A}}| = \ell'$ and some $(x, y) \in \mathcal{X} \times [K]$. By the inductive hypothesis we have that $\mathcal{G}|_{\tilde{\mathbf{A}}}$ is $(\ell - \ell')$-irreducible and satisfies $\mathrm{sfat}_2(\mathcal{G}|_{\tilde{\mathbf{A}}}) = \mathrm{sfat}_2(\mathcal{G})$. By the case $\ell' = 1$ proven above we have that $(\mathcal{G}|_{\tilde{\mathbf{A}}})|_{(x,y)} = \mathcal{G}|_{\mathbf{A}}$ is $(\ell - \ell' - 1)$-irreducible and satisfies $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}}) = \mathrm{sfat}_2(\mathcal{G}|_{\tilde{\mathbf{A}}}) = \mathrm{sfat}_2(\mathcal{G})$, as desired. ■

The below lemma is analogous to (Ghazi et al., 2020a, lemma 4.4).

Lemma 29 *For a class $\mathcal{F} \subset [K]^{\mathcal{X}}$ with $\mathrm{sfat}_2(\mathcal{F}) = d$, set*

$$\tilde{\mathcal{F}}_{d+1} := \{\mathrm{SOA}_{\mathcal{G}} \ : \ \mathcal{G} \subset \mathcal{F}, \ \mathcal{G} \text{ is nonempty and } (d+1)\text{-irreducible}\}.$$

Then $\mathrm{sfat}_2(\tilde{\mathcal{F}}_{d+1}) = d$ as well.

Proof Note that $\mathcal{F} \subset \tilde{\mathcal{F}}_{d+1}$, since for any $f \in \mathcal{F}$, $\{f\}$ is ℓ-irreducible for all $\ell \in \mathbb{N}$, and $\mathrm{SOA}_{\{f\}} = f$. Thus $\mathrm{sfat}_2(\tilde{\mathcal{F}}_{d+1}) \geq d$. To see the upper bound on $\mathrm{sfat}_2(\tilde{\mathcal{F}}_{d+1})$, suppose for the purpose of contradiction that $\tilde{\mathcal{F}}_{d+1}$ shatters an $\mathcal{X}$-valued binary tree $\mathbf{x}$ of depth $d + 1$. Let $\mathbf{s}$ be a witness tree to this shattering. We will show that $\mathcal{F}$ also shatters $\mathbf{x}$ (witnessed by $\mathbf{s}$), which leads to the desired contradiction.

Fix any sequence $(k_1, \ldots, k_{d+1}) \in \{1,2\}^{d+1}$. Since $\mathbf{x}$ is shattered by $\tilde{\mathcal{F}}_{d+1}$, there must be some $\mathcal{G} \subset \mathcal{F}$ that is $(d+1)$-irreducible so that for $1 \le t \le d+1$,

$$(3 - 2k_t) \cdot (\mathrm{SOA}_{\mathcal{G}}(\mathbf{x}_t(k_{1:t-1})) - \mathbf{s}_t(k_{1:t-1})) \ge 1.$$

For $1 \le t \le d+1$, set $y_t := \mathrm{SOA}_{\mathcal{G}}(\mathbf{x}_t(k_{1:t-1}))$. Since $\mathcal{G}$ is $(d+1)$-irreducible, by Lemma 28, we have that

$$\mathrm{sfat}_2(\mathcal{G}|_{(\mathbf{x}_1,y_1),(\mathbf{x}_2(k_1),y_2),\ldots,(\mathbf{x}_{d+1}(k_{1:d}),y_{d+1})}) = \mathrm{sfat}_2(\mathcal{G}) \ge 0.$$

Thus there must be some $f \in \mathcal{G} \subset \mathcal{F}$ so that for $1 \le t \le d+1$, $f(\mathbf{x}_t(k_{1:t-1})) = y_t$. Since the above argument holds for any choice of $(k_1, \ldots, k_{d+1}) \in \{1,2\}^{d+1}$, it follows that $\mathbf{x}$ is shattered by $\mathcal{F}$, witnessed by $\mathbf{s}$. ■

Appendix C. Proofs for the `ReduceTreeReg` algorithm (Section 4)

In this section we introduce the `ReduceTreeReg` algorithm reference in Section 4 and state its main guarantee of weak stability reference in Lemma 13 (the informal version of Lemmas 34 and Lemma 35). The algorithm and its analysis is very similar to that in Ghazi et al. (2020a); we provide all proofs for completeness, but indicate the corresponding results in Ghazi et al. (2020a) where appropriate.

Suppose $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ (e.g., $\mathcal{F} \subset [K]^{\mathcal{X}}$); for each $\alpha > 0$, and a distribution P on $\mathcal{X} \times \mathbb{R}$, define

$$\mathcal{F}_{P,\alpha} := \{f \in \mathcal{F} : \mathrm{err}_P(f) \le \alpha\}.$$

For a dataset $S_n \in (\mathcal{X} \times \mathbb{R})^n$, note that, under the event $\sup_{f \in \mathcal{F}} \left|\mathrm{err}_P(f) - \mathrm{err}_{\hat{P}_{S_n}}(f)\right| \le \alpha_0$, for each $\alpha \in [0,1]$ it holds that

$$\mathcal{F}_{\hat{P}_{S_n},\alpha-2\alpha_0} \subset \mathcal{F}_{P,\alpha-\alpha_0} \subset \mathcal{F}_{\hat{P}_{S_n},\alpha}. \tag{10}$$

The below lemma is analogous to Lemma 4.7 of Ghazi et al. (2020a); the proof is almost identical to that in Ghazi et al. (2020a), but we provide it for completeness.

Lemma 30 *Fix some $\ell, \ell' \in \mathbb{N}$ with $\ell > \ell'$ and hypothesis classes $\mathcal{H} \subset \mathcal{G} \subset [K]^{\mathcal{X}}$. Suppose we are given $S^\star \in (\mathcal{X} \times [K])^{\ell-\ell'}$ so that $\mathcal{H}|_{S^\star}$ is ℓ-irreducible, and that*

$$\mathrm{sfat}_2(\mathcal{G}|_{S^\star}) = \mathrm{sfat}_2(\mathcal{H}|_{S^\star}) =: q^\star \ge 0. \tag{11}$$

Suppose that $\mathbf{x}$ is a K-ary $\mathcal{X}$-valued tree of depth at most $\ell - \ell'$, and that for all leaves v of $\mathbf{x}$, $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(v)}) \le q^\star$. Then there is some leaf $\hat{v}$ of $\mathbf{x}$ so that $\| \mathrm{SOA}_{\mathcal{J}|_{S^\star}} - \mathrm{SOA}_{\mathcal{J}'|_{\mathbf{A}(\hat{v})}} \|_\infty \le 4$ for all hypothesis classes $\mathcal{J}', \mathcal{J}$ satisfying $\mathcal{H} \subset \mathcal{J}' \subset \mathcal{G}$ and $\mathcal{H} \subset \mathcal{J} \subset \mathcal{G}$.

Moveover, the leaf $\hat{v}$ satisfies:

1. $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(\hat{v})}) = \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\hat{v})}) = q^\star$.
2. *$\mathcal{H}|_{\mathbf{A}(\hat{v})}$ is ℓ'-irreducible.*

Proof The fact that $\mathcal{H}|_{S^\star}$ is ℓ-irreducible together with (11) and Lemma 27 gives that $\mathcal{G}|_{S^\star}$ and $\mathcal{J}|_{S^\star}$ are ℓ-irreducible for any $\mathcal{J}$ satisfying $\mathcal{H} \subset \mathcal{J} \subset \mathcal{G}$.

We now define a leaf $\hat{v}$ of $\mathbf{x}$ as follows: first choose $k_1 := \mathrm{SOA}_{\mathcal{H}|_{S^\star}}(\mathbf{x}_1)$, then for $t > 2$, if the node corresponding to the sequence $(k_1, \dots, k_{t-1})$ is not a leaf of $\mathbf{x}$, set $k_t := \mathrm{SOA}_{\mathcal{H}|_{S^\star}}(\mathbf{x}_t(k_{1:t-1}))$. This process will stop (i.e., the node corresponding to $(k_1, \dots, k_t)$ will be a leaf for some t) after at most $\ell - \ell'$ steps (since $\mathrm{depth}(\mathbf{x}) \le \ell - \ell'$), and we let the resulting leaf be $\hat{v}$. Since $|\mathbf{A}(\hat{v})| \le \mathrm{depth}(\mathbf{x}) \le \ell - \ell'$ and for each $(x, y) \in \mathbf{A}(\hat{v})$ we have $y = \mathrm{SOA}_{\mathcal{H}|_{S^\star}}(x)$, by Lemma 28, it holds that $\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}$ is ℓ'-irreducible and satisfies $\mathrm{sfat}_2(\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}) = \mathrm{sfat}_2(\mathcal{H}|_{S^\star}) = q^\star$.

Next, using the assumption that $\mathrm{sfat}_2(\mathcal{H}|_{S^\star}) = q^\star \ge \mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(\hat{v})})$ (as $\hat{v}$ is a leaf of $\mathbf{x}$) together with the ℓ'-irreducibility of $\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}$, we see that for any K-ary $\mathcal{X}$-valued tree $\mathbf{y}$ of depth at most ℓ', there is some leaf u of $\mathbf{y}$ so that

$$\begin{aligned}
\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\hat{v}) \cup \mathbf{A}(u)}) &\ge \mathrm{sfat}_2(\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v}) \cup \mathbf{A}(u)}) && (12)\\
&= \mathrm{sfat}_2(\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}) \\
&= \mathrm{sfat}_2(\mathcal{H}|_{S^\star}) && (13)\\
&\ge \mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(\hat{v})}) \ge \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\hat{v})}). && (14)
\end{aligned}$$

Since $\mathcal{H}|_{\mathbf{A}(\hat{v}) \cup \mathbf{A}(u)} \subset \mathcal{H}|_{\mathbf{A}(\hat{v})}$, it follows that the inequalities in (12) and (14) are equalities. For any $x \in \mathcal{X}$, interpret it as a depth-0 tree $\mathbf{y}$ whose root node is labeled by x, set $k(x) \in [K]$ to be the value ensuring that (12) through (14) holds. It then follows from Lemma 10 that

$$|\,\mathrm{SOA}_{\mathcal{H}|_{\mathbf{A}(\hat{v})}}(x) - k(x)| \le 1 \tag{15}$$

for all $x \in \mathcal{X}$.

From equalities (12) through (13), we have that for all $x \in \mathcal{X}$ (again letting the tree $\mathbf{y}$ be the depth-0 tree whose root is labeled by x), $\mathrm{sfat}_2(\mathcal{H}|_{S^\star}) = \mathrm{sfat}_2(\mathcal{H}|_{S^\star \cup \{(x, k(x))\}})$. Thus

$$|\,\mathrm{SOA}_{\mathcal{H}|_{S^\star}}(x) - k(x)| \le 1 \tag{16}$$

for all $x \in \mathcal{X}$.

Since $\mathcal{H}|_{S^\star}$ is irreducible, by Lemma 12, we have that for all $x \in \mathcal{X}$ and $\mathcal{J}$ satisfying $\mathcal{H} \subset \mathcal{J} \subset \mathcal{G}$,

$$|\,\mathrm{SOA}_{\mathcal{H}|_{S^\star}}(x) - \mathrm{SOA}_{\mathcal{J}|_{S^\star}}(x)| \le 1. \tag{17}$$

From (15), (16), (17) and the triangle inequality we see that $\|\,\mathrm{SOA}_{\mathcal{J}|_{S^\star}} - \mathrm{SOA}_{\mathcal{H}|_{\mathbf{A}(\hat{v})}}\|_\infty \le 3$. This establishes the desired closeness of SOA hypotheses for $\mathcal{J}' = \mathcal{H}$. Before establishing this for all $\mathcal{H}'$ satisfying $\mathcal{H} \subset \mathcal{J}' \subset \mathcal{G}$, we first show items 1 and 2.

Using (13) and (14) (which, as we argued above, are all equalities) gives that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\hat{v})}) = \mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}(\hat{v})}) = q^\star$, establishing item 1. Item 2 is a consequence of the fact that $\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}$ is ℓ'-irreducible, $\mathrm{sfat}_2(\mathcal{H}|_{S^\star \cup \mathbf{A}(\hat{v})}) = \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\hat{v})})$ (by (12) through (14)), and Lemma 27.

Items 1 and 2 together with Lemma 12 imply that for any hypothesis class $\mathcal{J}'$ satisfying $\mathcal{H} \subset \mathcal{J}' \subset \mathcal{G}$, we have that

$$\|\,\mathrm{SOA}_{\mathcal{J}'|_{\mathbf{A}(\hat{v})}} - \mathrm{SOA}_{\mathcal{H}'|_{\mathbf{A}(\hat{v})}}\|_\infty \le 1. \tag{18}$$

Then (15), (16), (17), and (18) together with the triangle inequality give that $\|\,\mathrm{SOA}_{\mathcal{J}'|_{\mathbf{A}(\hat{v})}} - \mathrm{SOA}_{\mathcal{J}|_{S^\star}}\|_\infty \le 4$ for all $\mathcal{J}', \mathcal{J}$ satisfying $\mathcal{H} \subset \mathcal{J}' \subset \mathcal{G}$, $\mathcal{H} \subset \mathcal{J} \subset \mathcal{G}$. ■

Algorithm 3: `ReduceTreeReg`

Input: Parameters $n, \ell' \in \mathbb{N}$, $\alpha_\Delta, \alpha_1 \in \mathbb{R}_+$. Distribution $\hat{P}_{S_n}$ over $\mathcal{X}$. Hypothesis class $\mathcal{F}$, with $d := \mathrm{sfat}_2(\mathcal{F})$.

1. Initialize a counter $t = 1$ (t counts the depth of the tree constructed at each step of the algorithm).
2. For $1 < t \le d+1$, set $\alpha_t := \alpha_1 - (t-1)\cdot\alpha_\Delta$.
3. For $1 \le t \le d$, set $\ell_t := \ell' \cdot 2^t$.
4. Initialize $\hat{\mathbf{x}}^{(0)} = \{v_0\}$ to be a tree with a single (unlabeled) leaf v_0. (In general $\hat{\mathbf{x}}^{(t)}$ will be the tree produced by the algorithm after step t is completed.)
5. Initialize $\hat{\mathcal{L}}_1 = \{v_0\}$. (In general $\hat{\mathcal{L}}_t$ will be the set of leaves of the tree before step t is started.)
6. For $t \in \{1, 2, \ldots, d\}$:
   (a) For each leaf $v \in \hat{\mathcal{L}}_t$ and $\alpha \ge 0$, set $\hat{\mathcal{G}}(\alpha, v) := \mathcal{F}_{\hat{P}_{S_n},\alpha}|_{\mathbf{A}(v)}$. (Note that since the only way the tree changes from round to round is by adding children to existing nodes, $\mathbf{A}(v)$ will never change for a node v that already exists.)
   (b) Let $\hat{w}_t^\star := \max_{v\in\hat{\mathcal{L}}_t} \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v))$ be the maximum sequential fat-shattering dimension of any of the classes $\hat{\mathcal{G}}(\alpha_t, v)$.
   Also let $\hat{\mathcal{L}}'_t := \{v \in \hat{\mathcal{L}}_t : \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) = \hat{w}_t^\star\}$.
   (c) If $\hat{w}_t^\star < 0$, halt and output ERROR. *(We show that this never occurs under appropriate assumptions in Lemma 31.)*
   (d) If there is some $v \in \hat{\mathcal{L}}'_t$ so that $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) = \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \ge 0$ and $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is ℓ_t-irreducible, then break out of the loop and go to step 7.
   (e) Else, for each node $v \in \hat{\mathcal{L}}'_t$:
      i. If $\hat{\mathcal{G}}(\alpha_t, v)$ is empty or $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) < \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v))$, move on to the next v.
      ii. Else, we must have that $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is not ℓ_t-irreducible. Let ℓ_v be chosen as small as possible so that $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is not ℓ_v-irreducible; then $\ell_v \le \ell_t$. Then there is some K-ary $\mathcal{X}$-valued tree $\mathbf{x}$ of depth ℓ_v, so that for any choice of $k_1, \ldots, k_{\ell_v} \in [K]$, we have
      $$\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)|_{(\mathbf{x}_1,k_1),\ldots,(\mathbf{x}_{\ell_v}(k_{1:\ell_v-1}),k_{\ell_v})}) < \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)). \tag{19}$$
      iii. Attach the tree $\mathbf{x}$ to $\hat{\mathbf{x}}^{(t-1)}$ via the leaf v (per Definition 24).
   (f) Let the current tree (with the additions of the previous step) be denoted by $\hat{\mathbf{x}}^{(t)}$, and let $\hat{\mathcal{L}}_{t+1}$ be the list of the leaves of $\hat{\mathbf{x}}^{(t)}$, i.e., the nodes which have not (yet) been assigned labels or children.
7. Let t_{final} be the final value of t the algorithm *completed* the loop of step 6e for before breaking out of the above loop (i.e., if the break at step 6d was taken at step t, then $t_{\mathrm{final}} = t-1$; if the break was never taken, then $t_{\mathrm{final}} = d$). Let $\hat{w}^\star_{t_{\mathrm{final}}+1}$ and $\hat{\mathcal{L}}'_{t_{\mathrm{final}}+1}$ be defined as in Step 6b.
8. Output the set $\hat{\mathcal{L}}' := \hat{\mathcal{L}}'_{t_{\mathrm{final}}+1}$ of leaves of the tree $\hat{\mathbf{x}}^{(t_{\mathrm{final}})}$, and the tree $\hat{\mathbf{x}} := \hat{\mathbf{x}}^{(t_{\mathrm{final}})}$. Finally, output the set
$$\hat{\mathcal{S}} := \{\mathrm{SOA}_{\hat{\mathcal{G}}(\alpha_{t_{\mathrm{final}}+1} - 2\alpha_\Delta/3, v)} : v \in \hat{\mathcal{L}}' \text{ and } \hat{\mathcal{G}}(\alpha_{t_{\mathrm{final}}+1} - 2\alpha_\Delta/3, v) \text{ is } \ell'\text{-irreducible \& nonempty}\}. \tag{20}$$

C.1. `ReduceTreeReg` algorithm

Throughout this section we fix a positive integer K, a distribution P on $\mathcal{X} \times [K]$, a function class $\mathcal{F} \subset [K]^{\mathcal{X}}$, and write $d := \text{sfat}_2(\mathcal{F})$. The algorithm `ReduceTreeReg` takes as input some parameters $k' \in \mathbb{N}$, $\alpha_1, \alpha_\Delta > 0$, as well as some dataset $S_n \in (\mathcal{X} \times [K])^n$ consisting of n samples $(x, k) \in \mathcal{X} \times [K]$, which is accessed through its empirical distribution $\hat{P}_{S_n}$. Given these parameters, define the event E_{good} to be

$$E_{\text{good}} := \left\{ \sup_{f \in \mathcal{F}} \left| \text{err}_P(f) - \text{err}_{\hat{P}_{S_n}}(f) \right| \le \frac{\alpha_\Delta}{6} \right\}. \tag{21}$$

Though the algorithm `ReduceTreeReg` is well-defined regardless of whether E_{good} holds, several of the lemmas in this section regarding correctness of `ReduceTreeReg` will rely on E_{good} holding; in Section E we will show that when the dataset S_n is drawn according to an appropriate distribution, E_{good} will hold with high probability with respect to this draw.

The below lemma is analogous to Lemma 5.1 of Ghazi et al. (2020a).

Lemma 31 *Suppose the inputs $S_n, \alpha_1, \alpha_\Delta$ of `ReduceTreeReg` are chosen so that $\mathcal{F}_{\hat{P}_{S_n}, \alpha_1 - (d+1)\cdot\alpha_\Delta}$ is nonempty. Then `ReduceTreeReg` never halts and outputs ERROR at step 6c. Moreover, the set $\hat{\mathcal{L}}'$ output by `ReduceTreeReg` satisfies the following property: letting $t = t_{\text{final}} + 1 \in [d+1]$, there is some leaf $v \in \hat{\mathcal{L}}'$ so that $\text{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) = \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \ge 0$ and $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is ℓ_t-irreducible.*

Proof If, for some t, the algorithm `ReduceTreeReg` breaks at step 6d, then the inclusion of the lemma is immediate: the condition to break in step 6d gies that for some $v \in \hat{\mathcal{L}}'_t = \hat{\mathcal{L}}'_{t_{\text{final}}+1}$, we have that $\text{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) = \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \ge 0$ and $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is ℓ_t-irreducible.

Next we show that the algorithm never halts and outputs ERROR at step 6c. Note that for each $1 \le t \le d+1$, the tree $\hat{\mathbf{x}}^{(t-1)}$ has the property that each non-leaf node has exactly K children, one corresponding to each label in $[K]$ (this is by Definition 5); thus, we have that, for each t, and each $\alpha \ge 0$,

$$\mathcal{F}_{\hat{P}_{S_n}, \alpha} = \bigcup_{v \in \hat{\mathcal{L}}_t} \mathcal{F}_{\hat{P}_{S_n}, \alpha}|_{\mathbf{A}(v)} = \bigcup_{v \in \hat{\mathcal{L}}_t} \hat{\mathcal{G}}(\alpha, v). \tag{22}$$

Since $\mathcal{F}_{\hat{P}_{S_n}, \alpha_1 - (d+1)\cdot\alpha_\Delta}$ is nonempty (by assumption), $\mathcal{F}_{\hat{P}_{S_n}, \alpha_t} \supset \mathcal{F}_{\hat{P}_{S_n}, \alpha_1 - (d+1)\cdot\alpha_\Delta}$ is nonempty for $1 \le t \le d+1$. Thus there is some $v \in \hat{\mathcal{L}}_t$ so that $\hat{\mathcal{G}}(\alpha_t, v)$ is nonempty, i.e., $\hat{w}^\star_t = \max_{v \in \hat{\mathcal{L}}_t} \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \ge 0$.

Otherwise, the algorithm performs a total of d iterations. We claim that $\hat{w}^\star_{d+1} = 0$. We first show that for all $t \ge 1$, $\hat{w}^\star_{t+1} < \hat{w}^\star_t$. To see this, note that each leaf v in $\hat{\mathcal{L}}_{t+1}$ belongs to one of the following categories:

- $v \in \hat{\mathcal{L}}_t \backslash \hat{\mathcal{L}}'_t$. (This includes the case that $\hat{\mathcal{G}}(\alpha_t, v)$ is empty.) In this case, we have

$$\text{sfat}_2(\hat{\mathcal{G}}(\alpha_{t+1}, v)) \le \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) < \hat{w}^\star_t.$$

- $v \in \hat{\mathcal{L}}'_t$ and $\text{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) < \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v))$. Using that $\alpha_{t+1} = \alpha_t - \alpha_\Delta$, we obtain

$$\text{sfat}_2(\hat{\mathcal{G}}(\alpha_{t+1}, v)) = \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) < \text{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \le \hat{w}^\star_t.$$

- v corresponds to some leaf u of some K-ary $\mathcal{X}$-valued tree $\mathbf{x}$ which is attached to $\hat{\mathbf{x}}^{(t-1)}$ via some leaf v_0 of $\hat{\mathbf{x}}^{(t-1)}$ (as constructed in steps 6(e)ii and 6(e)iii of the algorithm). Then $\mathbf{A}(v) = \mathbf{A}(v_0) \cup \mathbf{A}(u)$, and so

$$\begin{aligned}\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_{t+1}, v)) &\le \mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n}, \alpha_t - \alpha_\Delta}|_{\mathbf{A}(v)}) \\ &= \mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n}, \alpha_t - \alpha_\Delta}|_{\mathbf{A}(v_0) \cup \mathbf{A}(u)}) \\ &< \mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n}, \alpha_t - \alpha_\Delta}|_{\mathbf{A}(v_0)}) \\ &\le \hat{w}^\star_t,\end{aligned}$$

  where the strict inequality follows from (19) (the set $\{(\mathbf{x}_1, k_1), \ldots, (\mathbf{x}_{\ell_v}(k_{1:\ell_v - 1}), k_{\ell_v})\}$ is exactly $\mathbf{A}(u)$), and the last inequality follows from the fact that $v_0 \in \hat{\mathcal{L}}_t$.

Thus all leaves v in $\hat{\mathcal{L}}_{t+1}$ satisfy $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_{t+1}, v)) < \hat{w}^\star_t$, i.e., $\hat{w}^\star_{t+1} < \hat{w}^\star_t$. Since $\hat{w}^\star_1 \le d$ as $\hat{\mathcal{G}}(\alpha_t, v) \subset \mathcal{F}$, we obtain that $\hat{w}^\star_{d+1} \le 0$. We have already shown that $\hat{w}^\star_{d+1} \ge 0$, and so $\hat{w}^\star_{d+1} = 0$.

By assumption, $\mathcal{F}_{\hat{P}_{S_n}, \alpha_{d+1} - \alpha_\Delta} = \mathcal{F}_{\hat{P}_{S_n}, \alpha_1 - (d+1)\cdot\alpha_\Delta}$ is nonempty, and therefore, by (22), and therefore, for some leaf $v \in \hat{\mathcal{L}}'_{d+1} = \hat{\mathcal{L}}'$, we have $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_{d+1}, v)) = \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_{d+1} - \alpha_\Delta, v)) = 0$. Moreover, $\hat{\mathcal{G}}(\alpha_{d+1} - \alpha_\Delta, v)$ is ℓ_{d+1}-irreducible since a class with sequential fat-shattering dimension 0 is ℓ-irreducible for all $\ell \in \mathbb{N}$ (Lemma 25). ■

The below lemma is analogous to Lemma 5.2 of Ghazi et al. (2020a).

Lemma 32 *For all t the tree $\hat{\mathbf{x}}^{(t)}$ of Algorithm 3 has depth at most $\ell_{t+1} - \ell'$. In particular, the tree $\hat{\mathbf{x}}$ has depth at most $\ell_{t_{\mathrm{final}}+1} - \ell'$.*

Proof We prove by induction that the depth of $\hat{\mathbf{x}}^{(t)}$, denoted $\mathrm{depth}(\hat{\mathbf{x}}^{(t)})$, satisfies $\mathrm{depth}(\hat{\mathbf{x}}^{(t)}) \le \ell_{t+1} - \ell' = \ell' \cdot 2^{t+1} - \ell'$. For the base case, note that $\mathrm{depth}(\hat{\mathbf{x}}^{(0)}) = 0 < 2\ell' - \ell' = \ell' \cdot 2^t - \ell'$. For any $t > 0$, The only step of `ReduceTreeReg` at which $\hat{\mathbf{x}}^{(t-1)}$ is modified (to produce $\hat{\mathbf{x}}^{(t)}$) is step 6(e)ii, when some trees of depth at most ℓ_t are attached to $\hat{\mathbf{x}}^{(t-1)}$ via some leaves. Thus we have

$$\mathrm{depth}(\hat{\mathbf{x}}^{(t)}) \le \mathrm{depth}(\hat{\mathbf{x}}^{(t-1)}) + \ell_t \le \ell_t - \ell' + \ell_t = \ell_{t+1} - \ell'.$$

■

For each $\alpha > 0$ and $t \in [d+1]$, define the set:

$$\mathcal{M}_{\alpha,t} := \left\{ S \in (\mathcal{X} \times [K])^{\le (\ell_t - \ell')} : \begin{array}{c} \mathcal{F}_{P, \alpha - \alpha_\Delta/3}|_S \text{ is } \ell_t\text{-irreducible and nonempty,} \\ \text{and } \mathrm{sfat}_2(\mathcal{F}_{P,\alpha-\alpha_\Delta/3}|_S) = \mathrm{sfat}_2(\mathcal{F}_{P,\alpha+\alpha_\Delta/3}|_S) \end{array} \right\}. \tag{23}$$

Notice that $\mathcal{M}_{\alpha,t}$ depends on $\mathcal{F}, P$. The below lemma is analogous to Lemma 5.3 of Ghazi et al. (2020a).

Lemma 33 *Suppose that E_{good} holds. Then for $t = t_{\mathrm{final}} + 1$, the set $\mathcal{M}_{\alpha_t - \alpha_\Delta/2, t}$ is nonempty.*

Proof Set $t = t_{\mathrm{final}} + 1$. Let v be a node in the set $\hat{\mathcal{L}}'$ (so that v is a leaf of $\hat{\mathbf{x}}^{(t_{\mathrm{final}})} = \hat{\mathbf{x}}^{(t-1)}$) produced by `ReduceTreeReg` as guaranteed by Lemma 31, i.e., so that $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) = \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \ge 0$ and so that $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is ℓ_t-irreducible. Since the event E_{good} holds,

$$\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v) = \mathcal{F}_{\hat{P}_{S_n}, \alpha_t - \alpha_\Delta}|_{\mathbf{A}(v)} \subset \mathcal{F}_{P, \alpha_t - 5\alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{P, \alpha_t - \alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{\hat{P}_{S_n}, \alpha_t}|_{\mathbf{A}(v)} = \hat{\mathcal{G}}(\alpha_t, v).$$

It follows from Lemma 27 that $\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v)}$ is ℓ_t-irreducible and that $\mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}|_{\mathbf{A}(v)}) = \mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v)}) \geq 0$. Since the depth of the tree $\hat{\mathbf{x}}^{(t-1)} = \hat{\mathbf{x}}^{(t_{\mathrm{final}})}$ is at most $\ell_t - \ell'$ (Lemma 32), it follows that the number of tuples in $\mathbf{A}(v)$ is at most $\ell_t - \ell'$; thus $\mathbf{A}(v) \in \mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$. ■

For any $\alpha > 0, t \in [d+1]$ for which $\mathcal{M}_{\alpha,t}$ is nonempty, define:

$$S^\star_{\alpha,t} \in \arg\max_{S\in\mathcal{M}_{\alpha,t}} \{\mathrm{sfat}_2(\mathcal{F}_{P,\alpha}|_S)\}, \qquad q^\star_{\alpha,t} := \max_{S\in\mathcal{M}_{\alpha,t}} \{\mathrm{sfat}_2(\mathcal{F}_{P,\alpha}|_S)\} \geq 0. \tag{24}$$

Also set

$$\sigma^\star_{\alpha,t} := \mathrm{SOA}_{\mathcal{F}_{P,\alpha}|_{S^\star_{\alpha,t}}}. \tag{25}$$

The below lemma is analogous to Lemma 5.4 of Ghazi et al. (2020a).

Lemma 34 ("Weak stability") *Suppose that E_{good} holds and $\mathcal{F}_{\hat{P}_{S_n},\alpha_1-d\cdot\alpha_\Delta} = \mathcal{F}_{\hat{P}_{S_n},\alpha_{d+1}}$ is nonempty. Then the following holds: for $t = t_{\mathrm{final}} + 1 \in [d+1]$ and some leaf $\hat{v} \in \hat{\mathcal{L}}'$, we have $\|\sigma^\star_{\alpha_t-\alpha_\Delta/2,t} - \mathrm{SOA}_{\hat{\mathcal{G}}(\alpha_t-2\alpha_\Delta/3,\hat{v})}\|_\infty \leq 5$. (In particular, for this t, $\sigma^\star_{\alpha_t-\alpha_\Delta/2,t}$ is well-defined, i.e., $\mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$ is nonempty.)*

Moreover, $\hat{\mathcal{G}}(\alpha_t - 2\alpha_\Delta/3, \hat{v})$ is ℓ'-irreducible and nonempty, and $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - 2\alpha_\Delta/3, \hat{v})) = q^\star_{\alpha_t-\alpha_\Delta/2,t} \geq 0$.

Proof By Lemma 31, for $t := t_{\mathrm{final}}+1 \in [d+1]$, there is some leaf $v' \in \hat{\mathcal{L}}'$ so that $\mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)) = \mathrm{sfat}_2(\hat{\mathcal{G}}(\alpha_t, v)) \geq 0$ and $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v)$ is ℓ_t-irreducible. Since the event E_{good} holds, for each node v of the tree $\hat{\mathbf{x}}$ output by `ReduceTreeReg`, we have that

$$\begin{aligned}\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v)} \subset \mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{\hat{P}_{S_n},\alpha_t-4\alpha_\Delta/6}|_{\mathbf{A}(v)} \\ \subset \mathcal{F}_{P,\alpha_t-3\alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{\hat{P}_{S_n},\alpha_t-2\alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}|_{\mathbf{A}(v)} \subset \mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(v)}.\end{aligned} \tag{26}$$

Now we apply Lemma 30 with $\mathcal{J} = \mathcal{J}' = \mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}, \mathcal{H} = \mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}, \mathcal{G} = \mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}, \ell = \ell_t, \ell' = \ell'$, $\mathbf{x}$ equal to the tree $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(t_{\mathrm{final}})}$ output by `ReduceTreeReg`, and $S^\star = S^\star_{\alpha_t-\alpha_\Delta/2,t}$. Since $t = t_{\mathrm{final}} + 1$, Lemma 33 guarantees that $S^\star_{\alpha_t-\alpha_\Delta/2,t}$ is well-defined (i.e., $\mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$ is nonempty). We check that the preconditions of Lemma 30 hold: First, note that (11) holds by definition of $\mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$ in (23) and since $S^\star \in \mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$. Moreover, $\mathcal{H}|_{S^\star} = \mathcal{F}_{P,\alpha_t-\alpha_\Delta/2-\alpha_\Delta/3}|_{S^\star}$ is ℓ_t-irreducible, again by (23) and since $S^\star \in \mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$. By definition of $q^\star_{\alpha,t}$ in (24), we have

$$q^\star_{\alpha_t-\alpha_\Delta/2,t} = \mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{S^\star}) = \mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}|_{S^\star}).$$

Lemma 32 establishes that the depth of $\hat{\mathbf{x}}$ is at most $\ell_t - \ell'$, so $|\mathbf{A}(v')| \leq \ell_t - \ell'$. Next, from the guarantee on v' in Lemma 31 (i.e., that $\hat{\mathcal{G}}(\alpha_t - \alpha_\Delta, v') = \mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')}$ is ℓ_t-irreducible), the fact that $\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')} \subset \mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v')}$ (by (26)), and Lemma 27, we have that $\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v')}$ is ℓ_t-irreducible. (To apply Lemma 27 here, we need that $\mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(v')}) = \mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')})$, which follows from $\mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')}) = \mathrm{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')})$ and (26).) Since also $\mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}) = \mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/6})$, we have that $\mathbf{A}(v') \in \mathcal{M}_{\alpha_t-\alpha_\Delta/2,t}$, so the definition of $q^\star_{\alpha,t}$ gives

$$q^\star_{\alpha_t-\alpha_\Delta/2,t} \geq \mathrm{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{\mathbf{A}(v')}).$$

Moreover, for any other leaf u of the tree $\hat{\mathbf{x}}$, we have, by definition of $\hat{\mathcal{L}}' = \hat{\mathcal{L}}'_{t_{\text{final}}+1}$,

$$\text{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}|_{\mathbf{A}(u)}) \le \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(u)}) \le \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(v')}) = \text{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{\mathbf{A}(v')}) \le q^\star_{\alpha_t-\alpha_\Delta/2,t},$$

(The first inequality above holds due to (26), the second inequality is due to the fact that $v' \in \hat{\mathcal{L}}'_{t_{\text{final}}+1}$ (see step 6b of `ReduceTreeReg`), and the equality holds due to (26) and $\text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-\alpha_\Delta}|_{\mathbf{A}(v')}) = \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(v')})$.) This completes the verification that all hypotheses of Lemma 30 hold. Then Lemma 30 with $\mathcal{J}' = \mathcal{J} = \mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}$, we get that for some leaf $\hat{v}$ of $\hat{\mathbf{x}}$, we have

$$\|\,\text{SOA}_{\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{S^\star}} - \text{SOA}_{\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{\mathbf{A}(\hat{v})}}\,\| = \|\sigma^\star_{\alpha_t-\alpha_\Delta/2,t} - \text{SOA}_{\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{\mathbf{A}(\hat{v})}}\,\| \le 4.$$

Moreover, item 1 of Lemma 30 gives that $\text{sfat}_2(\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}) = \text{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}) = q^\star_{\alpha_t-\alpha_\Delta/2,t}$, and item 2 gives that $\mathcal{F}_{P,\alpha_t-5\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}$ is ℓ'-irreducible. From (26), it follows that $\text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-4\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}) = \text{sfat}_2(\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}|_{\mathbf{A}(\hat{v})}) = \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-2\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}) = q^\star_{\alpha_t-\alpha_\Delta/2,t} \ge 0$, and that $\mathcal{F}_{\hat{P}_{S_n},\alpha_t-4\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})} = \hat{\mathcal{G}}(\alpha_t - 2\alpha_\Delta/3, \hat{v})$ is ℓ'-irreducible (from Lemma 27). Then by (26) and Lemma 12, we have

$$\begin{aligned}&\|\sigma^\star_{\alpha_t-\alpha_\Delta/2,t} - \text{SOA}_{\hat{\mathcal{G}}(\alpha_t-2\alpha_\Delta/3,\hat{v})}\,\|_\infty \\ &\le \|\sigma^\star_{\alpha_t-\alpha_\Delta/2,t} - \text{SOA}_{\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2|_{\mathbf{A}(\hat{v})}}}\,\|_\infty + \|\,\text{SOA}_{\mathcal{F}_{P,\alpha_t-\alpha_\Delta/2}}|_{\mathbf{A}(\hat{v})} - \text{SOA}_{\hat{\mathcal{G}}(\alpha_t-2\alpha_\Delta/3,\hat{v})}\,\|_\infty \le 4+1 = 5.\end{aligned}$$

Finally, we check that $\hat{v} \in \hat{\mathcal{L}}' = \hat{\mathcal{L}}'_t = \hat{\mathcal{L}}'_{t_{\text{final}}+1}$, i.e., all leaves u of the tree $\hat{\mathbf{x}}$ satisfy $\text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(u)}) \le \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(\hat{v})})$. This is a consequence of the fact that for all such u,

$$\text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(\hat{v})}) \ge \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t-2\alpha_\Delta/6}|_{\mathbf{A}(\hat{v})}) = q^\star_{\alpha_t-\alpha_\Delta/2,t} \ge \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(v')}) \ge \text{sfat}_2(\mathcal{F}_{\hat{P}_{S_n},\alpha_t}|_{\mathbf{A}(u)}),$$

since $v' \in \hat{\mathcal{L}}'$ (by definition). ■

Lemma 35 *The set $\hat{\mathcal{S}}$ output by `ReduceTreeReg` has size $|\hat{\mathcal{S}}| \le K^{\ell' \cdot 2^{d+1}}$.*

Proof We show that for $t \in [d]$, the tree $\hat{\mathbf{x}}^{(t)}$ has at most $\prod_{t'=1}^{t} K^{\ell_{t'}}$ leaves. This statement is a simple consequence of the fact that $\mathbf{x}^{(0)}$ has a single leaf, and the tree $\hat{\mathbf{x}}^{(t)}$ is formed by attaching a trees of depth at most ℓ_t to some of the leaves of $\hat{\mathbf{x}}^{(t-1)}$. Thus the number of leaves of $\hat{\mathbf{x}}^{(t)}$ is at most

$$\prod_{t'=1}^{d} K^{\ell_{t'}} = K^{\ell_1+\cdots+\ell_d} \le K^{\ell' \cdot 2^{d+1}}.$$

■

Appendix D. Proofs for Section 5: the algorithm `SOAFilter`

In this section we give proofs for all results in Section 5, and state several additional lemmas which will be useful in our proofs. Throughout we suppose that we are given a hypothesis class $\mathcal{F} \subset [K]^{\mathcal{X}}$ and write $d := \text{sfat}_2(\mathcal{F})$.

D.1. Existence of reducing trees

Recall the definition of *reducing tree* from Definition 17. Lemma 36 shows that such trees exist.

Lemma 36 *For any class $\mathcal{H} \subset [K]^{\mathcal{X}}$ with $d := \mathrm{sfat}_2(\mathcal{H})$, any sequence $(\ell_t)_{t \geq 0}$ of positive integers, and any $(x, y) \in \mathcal{X} \times [K]$ for which $\mathrm{sfat}_2(\mathcal{H}|_{(x,y)}) < \mathrm{sfat}_2(\mathcal{H})$, there is a reducing tree $\mathbf{x}$ (of depth at least 1) for the pair (x, y), the sequence (ℓ_t), and the class $\mathcal{H}$.*

Moreover, $\mathbf{x}$ may be chosen so that for each $1 \leq t \leq d$, $\mathbf{x}$ has at most $K^{\sum_{t'=0}^{t-1} \ell_{t'}}$ leaves v so that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) = d - t$.

Proof We define a sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ of augmented $\mathcal{X}$-labeled trees. We begin by defining the tree $\mathbf{x}^{(0)}$, which is of depth 1 and consists of a root, labeled by x, together with a single child (which is its only leaf), for which the edge to the root is labeled by y. Now, suppose we are given the tree $\mathbf{x}^{(s)}$, for some $s \geq 0$. To define the tree $\mathbf{x}^{(s+1)}$, we begin with the tree $\mathbf{x}^{(s)}$, and then add some subtrees below some of the leaves of $\mathbf{x}^{(s)}$; we will say that each node of $\mathbf{x}^{(s)}$ *corresponds* to its copy in this copy of $\mathbf{x}^{(s)}$ in $\mathbf{x}^{(s+1)}$, as well as to its copies in $\mathbf{x}^{(s+2)}, \mathbf{x}^{(s+3)}, \ldots$. In particular, for each leaf v of $\mathbf{x}^{(s)}$:

- If $\mathcal{H}|_{\mathbf{A}(v)}$ is empty or ℓ_t-irreducible, where $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$, we move onto the next leaf.
- Otherwise, by the definition of irreducibility, there is some K-ary $\mathcal{X}$-valued tree $\mathbf{x}'$ of depth at most ℓ_t (again, with $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$) so that for each leaf v' of $\mathbf{x}'$, it holds that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v) \cup \mathbf{A}(v')}) < \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. Then we attach $\mathbf{x}'$ to $\mathbf{x}$ via the leaf v, i.e., we label the leaf v with $\mathbf{x}'_1$ and add a copy of the tree $\mathbf{x}'$ to $\mathbf{x}$ rooted at the leaf v (Definition 24).

We claim that $\mathbf{x}^{(d)} = \mathbf{x}^{(d-1)}$, namely that for any leaf v of $\mathbf{x}^{(d-1)}$, we have that $\mathcal{H}|_{\mathbf{A}(v)}$ is ℓ_t-irreducible, where $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. To do this, we introduce the following notation: for $s \geq 0$, let $\mathcal{B}^{(s)}$ denote the set of leaves of $\mathbf{x}^{(s)}$ so that $\mathcal{H}|_{\mathbf{A}(v)}$ is not empty or ℓ_t-irreducible for $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. We now prove the following claim:

Claim 37 *For $0 \leq s \leq d - 1$, for each leaf $v \in \mathcal{B}^{(s)}$, $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) \leq d - s - 1$.*

Proof [Proof of Claim 37] We use induction on s. The base case $s = 0$ is immediate since the tree $\mathbf{x}^{(0)}$ has a single leaf v which satisfies $\mathbf{A}(v) = \{(x, y)\}$, and $\mathrm{sfat}_2(\mathcal{H}|_{(x,y)}) < \mathrm{sfat}_2(\mathcal{H}) = d$ is assumed.

To establish the inductive step, note that any leaf $v \in \mathcal{B}^{(s+1)}$ does not correspond to a leaf v' of $\mathbf{x}^{(s)}$. Rather, there is some leaf $\tilde{v}$ of $\mathbf{x}^{(s)}$ and some tree $\mathbf{x}'$, as well as some leaf $\tilde{v}'$ of $\mathbf{x}'$ so that v is the leaf $\tilde{v}'$ attached to $\mathbf{x}^{(s)}$ via $\tilde{v}$. In particular, we have $\mathbf{A}(v) = \mathbf{A}(\tilde{v}) \cup \mathbf{A}(\tilde{v}')$ and $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\tilde{v}) \cup \mathbf{A}(\tilde{v}')}) < \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\tilde{v})})$. By the inductive hypothesis, $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\tilde{v})}) \leq d - s - 1$, and so $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) \leq d - s - 2$, completing the inductive step. ■

We now set $\mathbf{x} = \mathbf{x}^{(d-1)}$. It follows from Claim 37 that for all leaves v of $\mathbf{x}$, either $v \notin \mathcal{B}^{(s)}$, in which case $\mathcal{H}|_{\mathbf{A}(v)}$ is empty or ℓ_t-irreducible for $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$, or $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) \leq 0$, i.e., $\mathcal{H}|_{\mathbf{A}(v)}$ is empty or ℓ-irreducible for all $\ell \in \mathbb{N}$ (Lemma 25).

To establish that $\mathbf{x}$ is a reducing tree, we need to establish the second item in Definition 17 regarding $\mathrm{depth}(v)$ for leaves v of $\mathbf{x}$. To do so, we establish the following claim:

Claim 38 *Fix any $0 \leq s \leq d - 1$. For each leaf v of $\mathbf{x}^{(s)}$, letting $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$, we have that $\mathrm{depth}(v) \leq \sum_{t'=0}^{t-1} \ell_{t'}$.*

Proof We establish the claim using induction on s. For the base case $s = 0$, the only leaf v of $\mathbf{x}^{(0)}$ satisfies $\mathrm{depth}(v) = 1$, which is bounded above by $\sum_{t'=0}^{t-1} \ell_{t'}$ (Note that we have $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) \geq 1$ here.)

To establish the inductive step, consider any leaf v of $\mathbf{x}^{(s+1)}$ for some $0 \leq s \leq d-2$, and let $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. If v corresponds to some leaf v' of $\mathbf{x}^{(s)}$ then certainly $\mathrm{depth}(v) \leq \sum_{t'=0}^{t-1} \ell_{t'}$, by the inductive hypothesis. Otherwise (as in the proof of Claim 37), there is some leaf $\tilde{v}$ of $\mathbf{x}^{(s)}$, some tree $\mathbf{x}'$ of depth at most $\ell_{\tilde{t}}$ (where $\tilde{t} := d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\tilde{v})})$), as well as some leaf $\tilde{v}'$ of $\mathbf{x}'$, so that v is the leaf $\tilde{v}'$ attached to $\mathbf{x}^{(s)}$ via $\tilde{v}$. Moreover, it holds that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) = d - t < \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(\tilde{v})}) = d - \tilde{t}$, i.e., $t > \tilde{t}$. It follows that

$$\mathrm{depth}(v) \leq \mathrm{depth}(\tilde{v}) + \ell_{\tilde{t}} \leq \sum_{t'=0}^{\tilde{t}-1} \ell_{t'} + \ell_{\tilde{t}} \leq \sum_{t'=0}^{t-1} \ell_{t'}, \tag{27}$$

as desired. ■

Applying Claim 38 for $s = d-1$, we get that for each leaf v of $\mathbf{x}$, $\mathrm{depth}(v) \leq \sum_{t'=0}^{t-1} \ell_{t'}$ for $t = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. Moreover, again fixing a leaf v of $\mathbf{x}$, let s_v denote the minimum value of $s' \geq 0$ so that v corresponds to a leaf v' in $\mathbf{x}^{(s')}$. For each $0 \leq s' < s_v$, there is a unique leaf $w_{s'}$ of $\mathbf{x}^{(s')}$ (in fact, $w_{s'} \in \mathcal{B}^{(s')}$) so that $w_{s'}$ is an ancestor of the leaf v' in $\mathbf{x}^{(s')}$. Also let $w_{s_v} = v$. For any given $1 \leq \tilde{t} < t$, choose $s' \leq s_v$ as small as possible so that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(w_{s'})}) \leq d - \tilde{t}$. We must have $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(w_{s'-1})}) > d - \tilde{t}$, and so, letting $\hat{t} := d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(w_{s'-1})})$, (similarly to (27)) it follows that

$$\mathrm{depth}(w_{s'}) \leq \mathrm{depth}(w_{s'-1}) + \ell_{\hat{t}} \leq \sum_{t'=0}^{\hat{t}-1} \ell_{t'} + \ell_{\hat{t}} \leq \sum_{t'=0}^{\tilde{t}-1} \ell_{t'},$$

which completes the verification that $\mathbf{x}$ is a reducing tree.

To establish the last claim of the lemma, note that Claim 38 with $s = d-1$ implies that to specify a leaf v of $\mathbf{x}$ with $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) = d - t$, we need to specify a sequence of at most $\sum_{t'=0}^{t-1} \ell_{t'}$ integers in $[K]$ (as the tree $\mathbf{x}$ is K-ary). Moreover, the set of such sequences, taken over all leaves v with $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) = d - t$, must be prefix-free (as a leaf cannot be an ancestor of another leaf). Thus the number of leaves v with $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)}) = d - t$ is at most $K^{\sum_{t'=0}^{t-1} \ell_{t'}}$. ■

D.2. Proofs for the `FilterStep` algorithm

Lemma 39 *Suppose $\mathcal{F} \subset [K]^{\mathcal{X}}$ and $\mathbf{A} \subset \mathcal{X} \times [K]$ is a subset of $\mathcal{X} \times [K]$ of size at most $\ell - 1$, for some positive integer ℓ. Suppose $\mathcal{G}, \mathcal{G}' \subset \mathcal{F}$ are ℓ-irreducible and satisfy, for each $(x, y) \in \mathbf{A}$, $\mathrm{SOA}_{\mathcal{G}}(x) = \mathrm{SOA}_{\mathcal{G}'}(x) = y$. If also $\mathrm{sfat}_2(\mathcal{G}) = \mathrm{sfat}_2(\mathcal{G}') = \mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}})$, then*

$$\|\mathrm{SOA}_{\mathcal{G}'} - \mathrm{SOA}_{\mathcal{G}}\|_{\infty} \leq 1. \tag{28}$$

Proof By Lemma 28 applied to the classes $\mathcal{G}, \mathcal{G}'$, it holds that $\mathcal{G}|_{\mathbf{A}}$ and $\mathcal{G}'|_{\mathbf{A}}$ are 1-irreducible and satisfy $\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A}}) = \mathrm{sfat}_2(\mathcal{G}'|_{\mathbf{A}}) = \mathrm{sfat}_2(\mathcal{G}) = \mathrm{sfat}_2(\mathcal{G}') = \mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}})$.

If there were some $x \in \mathcal{X}$ together with $k, k' \in [K]$ so that $|k - k'| \geq 2$ so that

$$\mathrm{sfat}_2(\mathcal{G}|_{\mathbf{A} \cup \{(x,k)\}}) = \mathrm{sfat}_2(\mathcal{G}), \qquad \mathrm{sfat}_2(\mathcal{G}'|_{\mathbf{A} \cup \{(x,k')\}}) = \mathrm{sfat}_2(\mathcal{G}'),$$

and since $\mathcal{G}, \mathcal{G}' \subset \mathcal{F}$, we would have that

$$\text{sfat}_2(\mathcal{F}|_{\mathbf{A}\cup\{(x,k)\}}) = \text{sfat}_2(\mathcal{F}|_{\mathbf{A}\cup\{(x,k')\}}) = \text{sfat}_2(\mathcal{F}|_{\mathbf{A}}),$$

which is a contradiction to Lemma 10. ■

Lemma 14 uses Lemma 39 to show that any class $\mathcal{H}$ belonging to one of the sets $\mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ constructed in `FilterStep` is close in ℓ_∞ norm to its representative $\mathcal{L}_{\text{rep}}(\mathcal{H})$.

Lemma 14 *Fix inputs $\mathcal{F}, (\ell_{r,t})_{r,t\geq 0}, r_{\max}$ to `FilterStep`. For any $0 \leq r \leq r_{\max}, 0 \leq t \leq d$, and any $\mathcal{H} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$, we have that $\| \text{SOA}_{\mathcal{H}} - \text{SOA}_{\mathcal{L}_{\text{rep}}(\mathcal{H})} \|_\infty \leq 1$.*

Proof [Proof of Lemma 14] Fix some $\mathcal{H} \in \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ so that either $r = r_{\max}$ or $\mathcal{H} \notin \mathscr{I}_{\ell_{r+1,t},d-t}(\mathcal{F})$, and recall that $d - t = \text{sfat}_2(\mathcal{H})$. If, in the iteration of the for loop in step 3(a)i when the given $\mathcal{H}$ is considered (which corresponds to the value r), the branch in step 3(a)iB is taken, then we have $\| \text{SOA}_{\mathcal{H}} - \text{SOA}_{\mathcal{L}_{\text{rep}}(\mathcal{H})} \|_\infty = 0$. The nontrivial case is that the branch in step 3(a)iA is taken: in this case, choose $\mathcal{L} \in \mathscr{L}_{d-t}$ and $\mathbf{A} \subset \mathcal{X} \times [K]$ so that $|\mathbf{A}| \leq \ell_{r,t} - 1$, $\text{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d - t$ and so that for all $(x, y) \in \mathbf{A}$, $\text{SOA}_{\mathcal{L}}(x) = \text{SOA}_{\mathcal{H}}(x) = y$.

Certainly $\mathcal{H}$ is $\ell_{r,t}$-irreducible. The same holds for $\mathcal{L}$, since the only classes that have been added to $\mathscr{L}_{d-t}$ at the time when $\mathcal{H}$ is reached in step 3(a)i must belong to $\mathscr{I}_{\ell_{r',t},d-t}(\mathcal{F})$ for some $r' \geq r$, and for all $r' \geq r$, we have $\ell_{r',t} \geq \ell_{r,t}$. We also have $\text{sfat}_2(\mathcal{L}) = \text{sfat}_2(\mathcal{H}) = d - t$ since this is the case for all elements of $\mathscr{L}_{d-t}$.

By Lemma 39 with $\mathcal{G} = \mathcal{H}, \mathcal{G}' = \mathcal{L}, \ell = \ell_{r,t}$, it follows that since $\ell_{r,t} - 1 \geq |\mathbf{A}|$, we have that

$$\| \text{SOA}_{\mathcal{H}} - \text{SOA}_{\mathcal{L}} \|_\infty \leq 1,$$

as desired. ■

Lemma 15 *Fix inputs $\mathcal{F}, (\ell_{r,t})_{r,t\geq 0}, r_{\max}$ to `FilterStep`. For any $0 \leq t \leq d$ and $0 \leq r \leq r_{\max}$, and any $\mathbf{A} \subset \mathcal{X} \times [K]$ with $|\mathbf{A}| \leq \ell_{r,t} - 1$ so that $\text{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d - t$, there is at most one element $\mathcal{L} \in \mathscr{L}_{d-t} \cap \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ so that for all $(x, y) \in \mathbf{A}$, $\text{SOA}_{\mathcal{L}}(x) = y$.*

Proof [Proof of Lemma 15] Suppose for the purpose of contradiction there were two distinct $\mathcal{L}, \mathcal{L}' \in \mathscr{L}_{d-t} \cap \mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ so that for all $(x, y) \in \mathbf{A}$, $\text{SOA}_{\mathcal{L}}(x) = \text{SOA}_{\mathcal{L}'}(x) = y$. By construction all elements of $\mathscr{L}_{d-t}$ are elements of $\mathscr{I}_{\ell_{r,t},d-t}(\mathcal{F})$ for some r. Suppose (without loss of generality) that $\mathcal{L}'$ is considered after $\mathcal{L}$ in the for loop in step 3(a)i of `FilterStep`. Since $|\mathbf{A}| \leq \ell_{r',t} - 1$ for all $r' \geq r$, when $\mathcal{L}'$ is considered in the for loop in step 3(a)i of `FilterStep`, we would not add $\mathcal{L}'$ to $\mathscr{L}_{d-t}$ and could instead set $\mathcal{L}_{\text{rep}}(\mathcal{L}') \leftarrow \mathcal{L}$. ■

D.3. Proofs for the `SOAFilter` algorithm

Lemma 40 *Fix $\mathcal{F} \subset [K]^{\mathcal{X}}$, and in the context of the algorithm `SOAFilter`, consider any $0 \leq j \leq d$ and $1 \leq s \leq d$, and set $r = r_{\max} - jr_0 - 1$. For any $\mathbf{A} \in \mathscr{Q}_{j,s}$, letting $t := d - \text{sfat}_2(\mathcal{F}|_{\mathbf{A}})$, it holds that $\mathcal{F}|_{\mathbf{A}}$ is $\ell_{r,t}$-irreducible.*

Proof Given $\mathbf{A} \in \mathscr{Q}_{j,s}$, let $\mathcal{H} := \mathcal{F}|_{\mathbf{A}}$. There is some $\mathbf{A}' \in \mathscr{Q}_{j,s-1}$ so that, letting $\mathcal{H}' := \mathcal{F}|_{\mathbf{A}'}$, there is some $y \in [K]$ and leaf v of the tree $\mathbf{x}^{(\mathcal{H}',(x_{\mathbf{A}'},y))}$ so that $\mathbf{A} = \mathbf{A}' \cup \mathbf{A}(v)$ (see step 4(a)ivB of `SOAFilter`). Let $t' := d - \mathrm{sfat}_2(\mathcal{H}')$. Since the tree $\mathbf{x}^{(\mathcal{H}',(x_{\mathbf{A}'},y))}$ is a reducing tree with respect to $\mathcal{H}'$ for the pair $(x_{\mathbf{A}'}, y)$ and the sequence $(\ell_{r,t+t'})_{0\le t\le d-t'}$, we have that $\mathcal{H}'|_{\mathbf{A}(v)} = \mathcal{F}|_{\mathbf{A}}$ is $\ell_{r,(\mathrm{sfat}_2(\mathcal{H}')-\mathrm{sfat}_2(\mathcal{H}))+t'}$-irreducible, i.e., $\ell_{r,t}$-irreducible (see Definition 17). ■

Lemma 41 *Fix $\mathcal{F} \subset [K]^{\mathcal{X}}$, and in the context of the algorithm `SOAFilter` consider any $0 \le j \le d$ and $1 \le s \le d$, and let $r = r_{\max} - jr_0 - 1$. Then the following statements hold:*

1. *For any $\mathbf{A} \in \mathscr{Q}_{j,s}$, let $t := d - \mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}})$; then $|\mathbf{A}| \le \sum_{t'=0}^{t-1} \ell_{r,t'}$.*
2. *For any $\mathbf{A} \in \mathscr{Q}_{j,s}$, let $\mathcal{H} := \mathcal{F}|_{\mathbf{A}}$, $t := d - \mathrm{sfat}_2(\mathcal{H})$, and consider any of the reducing trees $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$ constructed in step 4(a)ivA of `SOAFilter`, and any leaf v of $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$. Then for any $t < \tilde{t} \le d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$, there is some node v' of $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$ which is an ancestor of v (or is v itself) and so that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v')}) \le d - \tilde{t}$ and $|\mathbf{A} \cup \mathbf{A}(v')| \le \sum_{t'=0}^{\tilde{t}-1} \ell_{r,t'}$.*

Proof Fix any j, let $r = r_{max} - jr_0 - 1$, and write $\mathscr{Q}_j := \bigcup_{0\le s\le d} \mathscr{Q}_{j,s}$. We begin with the proof of item 1, which we establish via induction on t; the base case $t = 0$ is immediate since the only element $\mathbf{A} \in \mathscr{Q}_j$ with $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d$ is $\mathbf{A} = \emptyset$. Suppose the statement of the lemma holds for all $\mathbf{A} \in \mathscr{Q}_j$ with $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}}) > d - t_0$, for any $t_0 \ge 0$. Now, for any $0 \le s \le d$, fix any $\mathbf{A} \in \mathscr{Q}_{j,s}$ with $\mathrm{sfat}_2(\mathcal{F}|_{\mathbf{A}}) = d - t_0$. By construction of $\mathscr{Q}_{j,s}$, there is some $\mathbf{A}' \in \mathscr{Q}_{j,s-1}$, together with some $(x_{\mathbf{A}'}, y) \in \mathcal{X} \times [K]$, so that the following holds. Let us set $\mathcal{H} := \mathcal{F}|_{\mathbf{A}}$, $\mathcal{H}' := \mathcal{F}|_{\mathbf{A}'}$ and $t_0' := d - \mathrm{sfat}_2(\mathcal{H}') < t_0$; then for some leaf v of the reducing tree $\mathbf{x}^{(\mathcal{H}',(x_{\mathbf{A}'},y))}$ (which is defined with respect to the sequence $(\ell_{r,t_0'+t'})_{0\le t'\le d-t_0'}$), we have that $\mathcal{H} = \mathcal{H}'|_{\mathbf{A}(v)}$. By definition of a reducing tree, we have that

$$|\mathbf{A}(v)| \le \sum_{q=0}^{(d-t_0')-(d-t_0)-1} \ell_{r,q+t_0'} = \sum_{t'=t_0'}^{t_0-1} \ell_{r,t'}.$$

By the inductive hypothesis, it holds that $|\mathbf{A}'| \le \sum_{t'=0}^{t_0'-1} \ell_{r,t'}$. Then

$$|\mathbf{A}' \cup \mathbf{A}(v)| \le \sum_{t'=0}^{t_0-1} \ell_{r,t'},$$

which establishes part 1.

Next we establish part 2. Fix $\mathbf{A}, \mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}, v$ as in the statement of the lemma, and consider any $t < \tilde{t} \le d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$. By the definition of a reducing tree there is some node v' of $\mathbf{x}^{(\mathcal{H},(x_{\mathbf{A}},y))}$ which is an ancestor of v so that $\mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v')}) \le d - \tilde{t}$ and so that $\mathrm{depth}(v') \le \sum_{t'=0}^{\tilde{t}-t-1} \ell_{r,t'+t}$. (If $\tilde{t} = d - \mathrm{sfat}_2(\mathcal{H}|_{\mathbf{A}(v)})$ we may just choose $v' = v$.) Using part 1, we obtain that

$$|\mathbf{A} \cup \mathbf{A}(v')| \le \sum_{t'=0}^{t-1} \ell_{r,t'} + \sum_{t'=0}^{\tilde{t}-t-1} \ell_{r,t'+t} = \sum_{t'=0}^{\tilde{t}-1} \ell_{r,t'}.$$

■

Finally we are ready to establish the main strong stability result of `SOAFilter`.

Lemma 18 *Fix any positive integer $\bar{\ell}$. Suppose that $\mathcal{G} \subset \mathcal{F}$ is nonempty, $\hat{g} \in [K]^{\mathcal{X}}$, that $\| \mathrm{SOA}_{\mathcal{G}} - \hat{g} \|_\infty \leq \chi$ for some $\chi > 0$, and that $\mathcal{G}$ is $(\bar{\ell} \cdot (d+3)^d)$-irreducible. Then there is some $\bar{\ell}$-irreducible $\mathcal{L}^\star \subset \mathcal{F}$, depending only on $\mathcal{G}$, so that $\| \mathrm{SOA}_{\mathcal{L}^\star} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq (2+2\chi)(d+1)+1$ and so that $\mathcal{L}^\star \in \mathscr{R}_{\hat{g}}$, where $\mathscr{R}_{\hat{g}}$ is the output of `SOAFilter` when given as inputs $\mathcal{F}$, $\hat{g}$, $r_{\max} = (d+1)$, $\tau_{\max} = (2+2\chi)(d+1)$ and the sequence $\ell_{r,t} := \bar{\ell} \cdot (r+2)^t$ for $0 \leq r \leq (d+1)$, $0 \leq t \leq d$.*

Moreover, all $\mathcal{L} \in \mathscr{R}_{\hat{g}}$ satisfy $\| \mathrm{SOA}_{\mathcal{L}} - \hat{g} \|_\infty \leq (2+2\chi)(d+1)$ and are $\bar{\ell}$-irreducible.

Proof The final statement of the lemma follows from step 5 of `SOAFilter`.

We proceed to prove the remainder of the lemma. For $0 \leq \tau \leq (2+2\chi)(d+1)$ and $2 \leq r \leq (d+1)$, define

$$\mu(r,\tau) := \max_{(\mathcal{H},\ell) \in \mathscr{G}_{r,\tau}} \left\{ \mathrm{sfat}_2(\mathcal{H}) \right\}, \tag{29}$$

where

$$\mathscr{G}_{r,\tau} := \left\{ (\mathcal{H}, \ell_{r,t}) : \begin{array}{c} \mathcal{H} \subset \mathcal{F} \text{ is } \ell_{r,t}\text{-irreducible and a finite restriction subclass of } \mathcal{F}, \\ \text{where } t = d - \mathrm{sfat}_2(\mathcal{H}), \text{ and } \| \mathrm{SOA}_{\mathcal{H}} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq \tau. \end{array} \right\}. \tag{30}$$

Since $\mathcal{G}$ is $\ell_{(d+1),d}$-irreducible, and for all t, r we have $\ell_{r,t} \leq \ell_{(d+1),d}$, we have that $(\mathcal{G}, \ell_{r,t}) \in \mathscr{G}_{r,\tau}$ for $t = d - \mathrm{sfat}_2(\mathcal{G})$ and all $0 \leq r \leq (d+1), 0 \leq \tau \leq (2+2\chi)(d+1)$, i.e., $\mathscr{G}_{r,\tau}$ is nonempty and so $\mu(r,\tau)$ is well-defined. Thus, for all r, τ in this range, it holds that for fixed r, $\tau \mapsto \mu(r,\tau)$ is a non-decreasing function of τ, and for fixed τ, $r \mapsto \mu(r,\tau)$ is a non-increasing function of r (since for any t, $r \mapsto \ell_{r,t}$ is an increasing function). By Lemma 43, there is some $r^\star, \tau^\star$ with $r^\star = (d+1) - j^\star$, $\tau^\star = (2+2\chi)j^\star$ for some $0 \leq j^\star \leq d$, so that $\mu(r^\star, \tau^\star) = \mu(r^\star - 1, \tau^\star + 2 + 2\chi)$.

Now choose some $(\mathcal{H}^\star, \ell^\star)$ which achieves the maximum in (29) for $r = r^\star$, $\tau = \tau^\star$; letting $t^\star = d - \mathrm{sfat}_2(\mathcal{H}^\star)$, we have that $\ell^\star = \ell_{r^\star, t^\star}$. Let $\mathcal{L}_{\mathrm{rep}}(\cdot)$ be the mapping defined as the output of `FilterStep` with the input class $\mathcal{F}$, the sequence $(\ell_{r,t})_{0 \leq r \leq r_{\max}, 0 \leq t \leq d}$, and $r_{\max} = d+1$ (these are exactly the parameters used in Step 1 of `SOAFilter`). Now set $\mathcal{L}^\star = \mathcal{L}_{\mathrm{rep}}(\mathcal{H}^\star) \in \mathscr{L}_{d-t^\star} \cap \mathscr{I}_{\ell_{r^\star,t^\star}, d-t^\star}(\mathcal{F})$; note that this is well-defined since $\mathcal{H}^\star \in \mathscr{I}_{\ell_{r^\star,t^\star}, d-t^\star}(\mathcal{F})$.

By definition of $\mathcal{H}^\star$ we have that

$$\| \mathrm{SOA}_{\mathcal{H}^\star} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq \tau^\star.$$

By Lemma 14, the fact that $\| \mathrm{SOA}_{\mathcal{G}} - \hat{g} \|_\infty \leq \chi$ (by assumption), and the triangle inequality, it follows that

$$\| \mathrm{SOA}_{\mathcal{L}^\star} - \hat{g} \|_\infty \leq \tau^\star + 1 + \chi. \tag{31}$$

Next consider the execution of `SOAFilter` (Algorithm 2) in the iteration of the for loop in line 4 corresponding to $j = j^\star$ (and with input $\hat{g}$ and $\ell_{r,t}, r_{\max}, \tau_{\max}$ as in the lemma statement; note that we have $\tau_0 = 2 + 2\chi, r_0 = 1$ in the context of `SOAFilter`). In particular, in this iteration of the loop we have $\tau = \tau^\star + 2 + \chi, r = r^\star - 1$. We define a particular sequence $\hat{\mathbf{A}}_0 \in \mathscr{Q}_{j^\star,0}, \hat{\mathbf{A}}_1 \in \mathscr{Q}_{j^\star,1}, \ldots, \hat{\mathbf{A}}_{\hat{s}} \in \mathscr{Q}_{j^\star,\hat{s}}$, for some $\hat{s} \leq d+1$ (to be defined below). First set $\hat{\mathbf{A}}_0 = \emptyset$. Given the choice of $\hat{\mathbf{A}}_s \in \mathscr{Q}_{j^\star,s}$, for any $s \geq 0$, define $\hat{\mathbf{A}}_{s+1}$ as follows: consider the iteration of the for loop over $\mathscr{Q}_{j^\star,s}$ (i.e., the bullet point in step 4a) for which $\mathbf{A} = \hat{\mathbf{A}}_s \in \mathscr{Q}_{j^\star,s}$. If it holds that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_s}} - \hat{g} \|_\infty \leq \tau^\star + 2 + \chi$, meaning that the branch in step 4(a)ii is taken, then set $\hat{s} = s$ (in which case $\hat{\mathbf{A}}_{s+1}$ is not defined). Otherwise, on step 4(a)iv on the iteration of the for loop corresponding to $\mathbf{A} = \hat{\mathbf{A}}_s$, choose $y = \mathrm{SOA}_{\mathcal{L}^\star}(x_{\hat{\mathbf{A}}_s})$ (which is of distance at most $\tau^\star + 1 + \chi = (\tau^\star + 2 + \chi) - 1$ from $\hat{g}(x_{\hat{\mathbf{A}}_s})$). Then let v be the unique leaf of the reducing tree $\mathbf{x}^{(\mathcal{F}|_{\hat{\mathbf{A}}_s}, (x_{\hat{\mathbf{A}}_s}, y))}$

corresponding to $\mathrm{SOA}_{\mathcal{L}^\star}$ in the sense that for all $(x', y') \in \mathbf{A}(v)$, $\mathrm{SOA}_{\mathcal{L}^\star}(x') = y'$. Now set $\hat{\mathbf{A}}_{s+1} := \hat{\mathbf{A}}_s \cup \mathbf{A}(v) \in \mathscr{Q}_{j^\star, s+1}$ (again we use that for each such pair (x', y'), $|\hat{g}(x') - y'| \leq (\tau^\star + 2 + \chi) - 1$). Notice that the definition of $\hat{\mathbf{A}}_{s+1}$ from $\hat{\mathbf{A}}_s$ above relies on the fact that $\mathcal{F}|_{\hat{\mathbf{A}}_s}$ is nonempty for each s; we will establish that this is case below, which will show that the $\hat{\mathbf{A}}_s$ are well-defined for $0 \leq s \leq \hat{s}$. Finally, if there is no $0 \leq s \leq d$ so that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_s}} - \hat{g}\|_\infty \leq \tau^\star + 2 + \chi$, then define $\hat{s} = d + 1$ (we will show that this will not be the case).

We claim that (a) for each $0 \leq s \leq \hat{s}$, all $\hat{\mathbf{A}}_s \in \mathscr{Q}_{j^\star, s}$ are well-defined, (b) $\hat{s} \leq d$, and (c) $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}) = d - t^\star$. (Recall that $d - t^\star = \mathrm{sfat}_2(\mathcal{L}^\star) = \mathrm{sfat}_2(\mathcal{H}^\star) = \mu(r^\star, \tau^\star) = \mu(r^\star - 1, \tau^\star + 4)$.) We show this in several steps:

- We begin by showing that for all $s \leq \min\{\hat{s}, d\}$, it holds that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s}) \geq \mathrm{sfat}_2(\mathcal{L}^\star)$. This immediately implies that $\hat{\mathbf{A}}_s$ is well-defined for all $0 \leq s \leq \min\{\hat{s}, d\}$, since the fact that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s}) \geq 0$ implies that $\mathcal{F}|_{\hat{\mathbf{A}}_s}$ is nonempty. Suppose that this is not the case; then choose $s < \hat{s}$ as large as possible so that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s}) \geq \mathrm{sfat}_2(\mathcal{L}^\star)$ (in particular, $\hat{\mathbf{A}}_s$ is well-defined and $\mathcal{F}|_{\hat{\mathbf{A}}_s}$ is nonempty). Let $y = \mathrm{SOA}_{\mathcal{L}^\star}(x_{\hat{\mathbf{A}}_s})$. Let v be the unique leaf of the tree $\mathbf{x}^{(\mathcal{F}|_{\hat{\mathbf{A}}_s}, (x_{\hat{\mathbf{A}}_s}, y))}$ corresponding to $\mathrm{SOA}_{\mathcal{L}^\star}$ in the sense that for all $(x', y') \in \mathbf{A}(v)$, we have $\mathrm{SOA}_{\mathcal{L}^\star}(x') = y'$. By definition of s and of $\hat{\mathbf{A}}_{s+1}$ we must have that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s \cup \mathbf{A}(v)}) < \mathrm{sfat}_2(\mathcal{L}^\star) \leq \mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s})$. By part 2 of Lemma 41 with $\tilde{t} = t^\star + 1 = \mathrm{sfat}_2(\mathcal{L}^\star) + 1$ and $\mathbf{A} = \hat{\mathbf{A}}_s$, there is some node v' of the tree $\mathbf{x}^{(\mathcal{F}|_{\hat{\mathbf{A}}_s}, (x_{\hat{\mathbf{A}}_s}, y))}$ which is an ancestor of v and satisfies $\mathrm{sfat}_2(\mathcal{L}^\star|_{\hat{\mathbf{A}} \cup \mathbf{A}(v')}) \leq \mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}} \cup \mathbf{A}(v')}) < \mathrm{sfat}_2(\mathcal{L}^\star)$ as well as $|\hat{\mathbf{A}}_s \cup \mathbf{A}(v')| \leq \sum_{t'=0}^{t^\star} \ell_{r^\star - 1, t'}$. Now notice that for each pair $(x', y') \in \hat{\mathbf{A}}_s \cup \mathbf{A}(v')$, we have that $\mathrm{SOA}_{\mathcal{L}^\star}(x') = y'$ by construction. But since $\mathcal{L}^\star$ is $\ell_{r^\star, t^\star}$-irreducible, this is a contradiction in light of Lemma 28 and the fact that

$$\sum_{t'=0}^{t^\star} \ell_{r^\star - 1, t'} \leq \ell_{r^\star, t^\star}$$

for all possible $r^\star \geq 1, t^\star \geq 0$ for our choice of $\ell_{r,t} = \bar{\ell} \cdot (r+2)^t$.
- Next we show that $\hat{s} \leq d$ (which implies that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}} - \hat{g}\|_\infty \leq \tau^\star + 2 + \chi$). To do this we note that since the tree $\mathbf{x}^{(\mathcal{F}|_{\hat{\mathbf{A}}_s}, (x_{\hat{\mathbf{A}}_s}, y))}$ used to define $\hat{\mathbf{A}}_{s+1}$ from $\hat{\mathbf{A}}_s$ is a reducing tree for the class $\mathcal{F}|_{\hat{\mathbf{A}}_s}$, we must have that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{s+1}}) < \mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s})$, and so for $s \leq \hat{s}$, $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_s}) \leq d - s$. If it is not the case that $\hat{s} \leq d$ (i.e., $\hat{s} = d+1$), then by the previous item for $s = d$, we have that $0 \geq \mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_d}) \geq \mathrm{sfat}_2(\mathcal{L}^\star)$, which implies that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_d}) = \mathrm{sfat}_2(\mathcal{L}^\star) = 0$ since $\mathcal{L}^\star$ is nonempty. In particular, by Lemma 25, $\mathcal{L}^\star, \mathcal{F}|_{\hat{\mathbf{A}}_d}$ are ℓ-irreducible for all $\ell \in \mathbb{N}$. By Lemma 39 with $\mathbf{A} = \hat{\mathbf{A}}_d$, $\mathcal{G} = \mathcal{L}^\star, \mathcal{G}' = \mathcal{F}|_{\hat{\mathbf{A}}_d}$, since for all $(x', y') \in \hat{\mathbf{A}}_d$, we have $\mathrm{SOA}_{\mathcal{L}^\star}(x') = \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_d}}(x') = y'$, it follows that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_d}} - \mathrm{SOA}_{\mathcal{L}^\star} \|_\infty \leq 1$. Together with the triangle inequality and (31), this gives $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_d}} - \hat{g}\|_\infty \leq \tau^\star + 2 + \chi$. But this means that in step 4(a)ii of `SOAFilter`, it holds that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_d}} - \hat{g}\|_\infty \leq \tau = \tau^\star + 2 + \chi$, and thus the branch in that step is taken, i.e., we set $\hat{s} = d$. This shows it cannot be the case that $\hat{s} = d + 1$, as desired.
- Finally we show that $\mathrm{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}) \leq \mu(r^\star - 1, \tau^\star + 2 + 2\chi)$. By definition of $\mu(\cdot, \cdot)$ it suffices to show that $\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}} \in \mathscr{G}_{r^\star - 1, \tau^\star + 2 + 2\chi}$. By the definition of $\hat{s}$ and the fact that $\hat{s} \leq d$, we have that $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}} - \hat{g}\|_\infty \leq \tau^\star + 2 + \chi$, and thus $\| \mathrm{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}} - \mathrm{SOA}_{\mathcal{G}} \|_\infty \leq \tau^\star + 2 + 2\chi$.

By Lemma 40, we have that $\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}$ is $\ell_{r^\star-1,t}$-irreducible for $t = d - \text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}})$. Hence $\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}} \in \mathscr{G}_{r^\star-1,\tau^\star+2+2\chi}$, and thus $\mu(r^\star-1,\tau^\star+2+2\chi) \geq \text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}})$.

- From the first and second items above it follows that $\text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}) \geq \text{sfat}_2(\mathcal{L}^\star)$, and the third item above shows that $\text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}) \leq \mu(r^\star-1,\tau^\star+2+2\chi) = \mu(r^\star,\tau^\star) = \text{sfat}_2(\mathcal{L}^\star)$. Thus $\text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}) = \text{sfat}_2(\mathcal{L}^\star) = d - t^\star$.

Part 1 of Lemma 41 gives that $|\hat{\mathbf{A}}_{\hat{s}}| \leq \sum_{t'=0}^{t^\star-1} \ell_{r^\star-1,t'} < \ell_{r^\star-1,t^\star} < \ell_{r^\star,t^\star}$. By Lemma 15, there is at most one choice of $\mathcal{L} \in \mathscr{L}_{d-t^\star}$ so that $\mathcal{L}$ is $\ell_{r^\star,t^\star}$-irreducible and for each $(x,y) \in \hat{\mathbf{A}}_{\hat{s}}$, $\text{SOA}_{\mathcal{L}}(x) = y$. Notice that $\mathcal{L}^\star$ is one such choice of $\mathcal{L}$. Thus $\mathcal{L}^\star$ must be added to $\mathscr{R}_{\hat{g}}$ in step 4(a)ii of `SOAFilter` when $\hat{\mathbf{A}}_{\hat{s}}$ is considered in the for loop.

By Lemma 39 with $\mathbf{A} = \hat{\mathbf{A}}_{\hat{s}}$, $\mathcal{G} = \mathcal{L}^\star, \mathcal{G}' = \mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}$, since $|\hat{\mathbf{A}}_{\hat{s}}| < \ell_{r^\star-1,t^\star}$, $\text{sfat}_2(\mathcal{L}^\star) = \text{sfat}_2(\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}})$, $\mathcal{L}^\star$ and $\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}$ are both $\ell_{r^\star-1,t^\star}$-irreducible, and for all $(x,y) \in \hat{\mathbf{A}}_{\hat{s}}$, $\text{SOA}_{\mathcal{L}^\star}(x) = \text{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}}(x) = y$, we have that $\|\text{SOA}_{\mathcal{L}^\star} - \text{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}}\|_\infty \leq 1$. Together with $\|\text{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}} - \text{SOA}_{\mathcal{G}}\|_\infty \leq \tau^\star + 2 + 2\chi$ and $\tau^\star \leq (2+2\chi)d$, we get that $\|\text{SOA}_{\mathcal{L}^\star} - \text{SOA}_{\mathcal{G}}\|_\infty \leq (2+2\chi)(d+1)+1$. Moreover, since $\|\text{SOA}_{\mathcal{F}|_{\hat{\mathbf{A}}_{\hat{s}}}} - \hat{g}\|_\infty \leq \tau^\star + 2 + \chi \leq (2+2\chi)(d+1) - 1$, we have that $\text{SOA}_{\mathcal{L}^\star}$ is not eliminated from $\mathscr{R}_{\hat{g}}$ in step 5 of `SOAFilter`. ■

Lemma 42 *In the algorithm `SOAFilter`, we have the following upper bound on the size of the output set $\mathscr{R}_{\hat{g}}$:*

$$|\mathscr{R}_{\hat{g}}| \leq \sum_{r=0}^{r_{\max}} K^{\sum_{t'=0}^{d-1} \ell_{r,t'}}.$$

In particular, for the choice $r_{\max} = (d+1)$ and $\ell_{r,t} = \bar{\ell} \cdot (r+2)^t$ (for any $\bar{\ell} \in \mathbb{N}$), we get

$$|\mathscr{R}_{\hat{g}}| \leq K^{\bar{\ell} \cdot (d+4)^d}.$$

Proof Fix any $0 \leq j \leq d$ considered in the for loop on step 4 of `SOAFilter`. Let $\tau = j\tau_0 + 3, r = r_{\max} - jr_0 - 1$. For accounting purposes, we define the following tree T whose non-leaf nodes are labeled by elements of $\mathcal{X}$ (the tree T does *not* satisfy the requirements of Definitions 5 or 16). The root of the tree T is labeled by the point $x_\emptyset$ defined in step 4(a)iii of `SOAFilter` corresponding to $\emptyset \in \mathscr{Q}_{j,0}$ (in the event that this step is never reached, then at most a single element is added to $\mathscr{R}_{\hat{g}}$ in `SOAFilter` for the value of j under consideration). We will call some of the nodes of T *special*; the root is special. Each special node of T is labeled by some $x_{\mathbf{A}}$ corresponding to the execution of step 4(a)iii in `SOAFilter` for some $0 \leq s \leq d$ and $\mathbf{A} \in \mathscr{Q}_{j,s}$. For each special node u we define its descendents inductively as follows. The (immediate) children of u in T are defined as follows: u has at most $2\tau - 1 \wedge K$ children, corresponding to each of the elements y of $\{k - \tau + 1 \vee 0, \ldots, k + \tau - 1 \wedge K\}$, where $k = \hat{g}(x_{\mathbf{A}})$. Each such child corresponding to some such y is labeled by the unique child of the root of the reducing tree $\mathbf{x}^{(\mathcal{F}|_{\mathbf{A}},(x_{\mathbf{A}},y))}$. Then we append the reducing tree $\mathbf{x}^{(\mathcal{F}|_{\mathbf{A}},(x_{\mathbf{A}},y))}$ (except its root) to T via this child. The leaves of a reducing tree are not labeled by elements of $\mathcal{X}$, but we label some leaves v of $\mathbf{x}^{(\mathcal{F}|_{\mathbf{A}},(x_{\mathbf{A}},y))}$ as follows. For any leaf v of $\mathbf{x}^{(\mathcal{F}|_{\mathbf{A}},(x_{\mathbf{A}},y))}$, if $\mathbf{A} \cup \mathbf{A}(v)$ is not added to $\mathscr{Q}_{j,s+1}$ in step 4(a)ivB, then v (viewed as a node of T) is defined to be a leaf of T, in which case we do not assign it a label. Otherwise, we have that $\mathbf{A}' := \mathbf{A} \cup \mathbf{A}(v) \in \mathscr{Q}_{j,s+1}$; if either of the branches in steps 4(a)i or 4(a)ii are taken when $\mathbf{A}'$ is considered in the for loop (for the value $s+1$), then v has no children in the tree T (i.e., is a leaf of T) and again is assigned no label. Otherwise, v is labeled by the element $x_{\mathbf{A}'}$ defined in step 4(a)iii,

in which case we say that v is special and we repeat the process described above with $\mathbf{A}'$ replacing $\mathbf{A}$. Notice that the construction of T maintains the following property: for each special node v of T which is labeled by $x_{\mathbf{A}}$, the ancestor set of v in T is exactly $\mathbf{A}$.

By construction of T, each element added to $\mathscr{R}_{\hat{g}}$ in step 4(a)ii of `SOAFilter` for the value of j under consideration corresponds to a distinct leaf of the tree T, whose ancestor set is given by some $\mathbf{A} \in \mathscr{Q}_{j,s}$ for some $0 \le s \le d$. So it suffices to bound the number of such leaves of T. Note that each node of T has at most K children; indeed, the special nodes of T have at most $2\tau - 1 \wedge K \le K$ children, and the remaining nodes are identified with nodes of various K-ary reducing trees. Moreover, the depth (i.e., distance to the root) of any leaf of T whose ancestor set is given by some $\mathbf{A} \in \mathscr{Q}_{j,s}$ for some $s \le d$ is at most $|\mathbf{A}| \le \sum_{t'=0}^{d-1} \ell_{r,t'}$, by part 1 of Lemma 41. Thus the number of leaves of T is at most $K^{\sum_{t'=0}^{d-1} \ell_{r,t'}}$. Hence

$$|\mathscr{R}_{\hat{g}}| \le \sum_{r=0}^{r_{\max}} K^{\sum_{t'=0}^{d-1} \ell_{r,t'}},$$

and for the choice $\ell_{r,t} = \bar{\ell} \cdot (r+2)^t$ and $r_{\max} = (d+1)$, this number is at most

$$\sum_{r=0}^{d+1} K^{\bar{\ell} \cdot (r+2)^d} \le K^{\bar{\ell} \cdot (d+4)^d}.$$

■

Lemma 43 *Fix positive integers A, B, d, and let $\mu : \{0,1,\dots,A(d+1)\} \times \{0,1,\dots,B(d+1)\} \to \mathbb{Z}$ be a function so that $0 \le \mu(a,b) \le d$ for all a, b and so that for each fixed b, $a \mapsto \mu(a,b)$ is non-decreasing and for each fixed a, $b \mapsto \mu(a,b)$ is non-decreasing. Then there is some $0 \le i \le d$ so that the pair $(a,b) := (Ai, Bi)$ satisfies $\mu(a,b) = \mu(a+A, b+B)$.*

Proof Consider the $d+1$ pairs $(0,0), (A,B), (2A,2B), \dots, (A(d+1), B(d+1))$. If for each $0 \le i \le d+1$, $\mu(Ai,Bi) \ne \mu(A(i+1), B(i+1))$, then we have $0 \le \mu(0,0) < \mu(A,B) < \cdots < \mu(A(d+1), B(d+1)) \le d$, which is impossible since $\mu(a,b)$ is an integer for all a, b in the domain of μ. ■

Appendix E. `RegLearn`: Private learning algorithm for regression

In this section we combine the procedures described in the previous sections to produce an algorithm for privately learning a real-valued hypothesis class. At a high level, our algorithm `RegLearn` (Algorithm 4) proceeds as follows: given a class $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$ and samples from a distribution Q supported on $\mathcal{X} \times [-1,1]$, it first discretizes $\mathcal{H}$ as described in Section 2.1: to avoid confusion with notation in other sections, we denote the discretization parameter as $\bar{\eta} > 0$. In particular, we set $\mathcal{F} := \lfloor \mathcal{H} \rfloor_{\bar{\eta}} \subset [K]^{\mathcal{X}}$ (with $K = \lceil 2/\bar{\eta} \rceil$) and $P := \lfloor Q \rfloor_{\bar{\eta}}$, so that P is a distribution over $\mathcal{X} \times [K]$. We then use Algorithm 3 applied to the class $\mathcal{F}$ to learn a hypothesis $\hat{g} \in [K]^{\mathcal{X}}$ with low population error with respect to $\lfloor Q \rfloor_{\bar{\eta}}$ and which satisfies the "weak stability" guarantee of Lemma 34. Using Algorithm 2 we then produce a set of hypotheses $\mathscr{R}_{\hat{g}}$, satisfying the "strong stability" guarantee of Lemma 18. Repeating this procedure sufficiently many times using independent datasets drawn

Algorithm 4: `RegLearn`

Input: Parameters $\varepsilon, \delta, \bar{\eta}, \beta \in (0,1)$, irreducibility parameter $\bar{\ell} \in \mathbb{N}$, i.i.d. samples $(x,y) \in \mathcal{X} \times [-1,1]$ from a distribution Q, hypothesis class $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$.

1. Set $\mathcal{F} := \lfloor \mathcal{H} \rfloor_{\bar{\eta}}$, and write $K := \lceil 2/\bar{\eta} \rceil$, so that $\mathcal{F} \subset [K]^{\mathcal{X}}$.
   Set $m \leftarrow \frac{C\bar{\ell}(2\operatorname{sfat}_2(\mathcal{F})+6)^{\operatorname{sfat}_2(\mathcal{F})+4}\log^2\left(\frac{1}{\varepsilon\delta\beta\bar{\eta}}\right)}{\varepsilon\bar{\eta}^2}$, $n_0 \leftarrow C_0 \cdot \frac{\operatorname{fat}_{c_0\bar{\eta}}(\mathcal{H})\log(1/\bar{\eta})+\log(4m/\beta)}{\bar{\eta}^2}$, $n \leftarrow n_0 m$, $\alpha_\Delta \leftarrow 18$, where C_0, c_0 are the constants of Corollary 21, and $C > 0$ is a sufficiently large constant.
   Also set $\ell' \leftarrow \max\left\{\bar{\ell} \cdot (d+3)^d, C_0 K^2 (d \log K + 1)\right\}$, where C_0 is the constant of Corollary 22.
2. Let $n_1 = \frac{C_0 \cdot \operatorname{fat}_{c_0\bar{\eta}}(\mathcal{H})\log(1/\bar{\eta})+\log(8/\beta)}{\varepsilon\bar{\eta}^2}$, where C_0, c_0 are the constants of Corollary 21. Set $T_{n_1} \sim Q^{n_1}$ to be an independent sample from the distribution Q of size n_1. Set
$$\hat{\eta} := \inf_{f \in \mathcal{F}}\left\{\operatorname{err}_{\lfloor \hat{Q}_{T_{n_1}} \rfloor_{\bar{\eta}}}(f)\right\} + \operatorname{Lap}\left(\frac{2K}{\varepsilon n_1}\right).$$
   to be the sum of the smallest achievable empirical error on T_{n_1} and a Laplace random variable with scale $2K/(\varepsilon n_1)$. *($\hat{\eta}$ is a private estimate of the optimal error achievable by a classifier in $\mathcal{F}$, which is neededd to apply* `ReduceTreeReg`*.)*
   Then set $\alpha_1 := \hat{\eta} + \alpha_\Delta/2 + d \cdot \alpha_\Delta$.
3. For $1 \le j \le m$:
   (a) Let $S_{n_0} \sim Q^{n_0}$ be an independent sample from the distribution Q.
   (b) Run the algorithm `ReduceTreeReg` with the class $\mathcal{F}$, distribution $\lfloor \hat{Q}_{S_{n_0}} \rfloor_{\bar{\eta}}$, $n = n_0$ and the parameters $\alpha_1, \alpha_\Delta, \ell'$ defined in steps 1 and 2.
   Let its output set $\hat{\mathcal{S}}$ (defined in (20)) be denoted by $\hat{\mathscr{S}}^{(j)}$.
4. For $1 \le j \le m$:
   (a) Set $\hat{\mathscr{R}}^{(j)} \leftarrow \emptyset$. *($\hat{\mathscr{R}}^{(j)}$ will hold hypotheses of the form $g : \mathcal{X} \to [K]$.)*
   (b) For each hypothesis $\hat{g} \in \hat{\mathscr{S}}^{(j)}$, apply the algorithm `SOAFilter` to the hypothesis $\hat{g} : \mathcal{X} \to [K]$, with the other inputs as follows: the hypothesis class is $\mathcal{F}$, the sequence $\ell_{r,t}$ is given by $\bar{\ell} \cdot (r+2)^t$, parameters $\tau_{\max} = 12 \cdot (\operatorname{sfat}_2(\mathcal{F})+1)$, $r_{\max} = \operatorname{sfat}_2(\mathcal{F})+1$.
   (c) Denote the output set of `SOAFilter` by $\hat{\mathscr{R}}_{\hat{g}}$; for each $\mathcal{L} \in \hat{\mathscr{R}}_{\hat{g}}$, add $\operatorname{SOA}_{\mathcal{L}}$ to the set $\hat{\mathscr{R}}^{(j)}$.
5. Run the (ε, δ)-differentially private (m, s)-sparse selection protocol of Proposition 7 with sparsity $s = K^{C\bar{\ell}(2\cdot\operatorname{sfat}_2(\mathcal{F})+6)^{\operatorname{sfat}_2(\mathcal{F})+2}K^2 \cdot \operatorname{sfat}_2(\mathcal{F})\log K}$ on the sets $\hat{\mathscr{R}}^{(1)}, \ldots, \hat{\mathscr{R}}^{(m)}$; the universe $\mathcal{U}$ for the sparse selection protocol is equal to the set of all $\operatorname{SOA}_{\mathcal{L}}$, for $\mathcal{L} \subset \mathcal{F}$ irreducible. Denote its output by $\operatorname{SOA}_{\hat{\mathcal{L}}} : \mathcal{X} \to [K]$, for some $\hat{\mathcal{L}} \subset \mathcal{F}$. Output the class $\hat{\mathcal{L}}$, as well as the function $\hat{h} : \mathcal{X} \to [-1,1]$, defined by
$$\hat{h}(x) := -1 + \frac{2}{K} \cdot (\operatorname{SOA}_{\hat{\mathcal{L}}}(x) - 1).$$

from the distribution Q and using the sparse selection procedure of Proposition 7, we may finally produce a regressor in $[-1,1]^{\mathcal{X}}$ which is differentially private.

The below theorem states the main guarantee for the algorithm `RegLearn`:

Theorem 44 *There are constants $c_0 \leq 1, C \geq 1, C_1 \geq 1$ so that the following holds.*[6] *Suppose we are given $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$, as well as $\varepsilon, \delta, \bar{\eta}, \beta \in (0,1)$ and $\bar{\ell} \in \mathbb{N}$. For*

$$n = C \cdot \frac{\bar{\ell} \cdot \mathrm{fat}_{c_0\bar{\eta}}(\mathcal{H}) \cdot (2 \cdot \mathrm{sfat}_{\bar{\eta}}(\mathcal{H}) + 6)^{\mathrm{sfat}_{\bar{\eta}}(\mathcal{H})+5} \log^3\left(\frac{\mathrm{sfat}_{\bar{\eta}}(\mathcal{H}) \cdot \bar{\ell}}{\varepsilon\delta\beta\bar{\eta}}\right)}{\varepsilon\bar{\eta}^4},$$

if the algorithm `RegLearn` (Algorithm 4) takes as input n i.i.d. samples $(x_1,y_1),\ldots,(x_n,y_n)$ from any distribution Q on $\mathcal{X} \times [-1,1]$, then it is (ε,δ)-differentially private and its output hypothesis $\hat{h}$ satisfies

$$\Pr_{(x_1,y_1),\ldots,(x_n,y_n)}\left[\mathrm{err}_Q(\hat{h}) \leq \inf_{h\in\mathcal{H}}\{\mathrm{err}_Q(h)\} + 30(\mathrm{sfat}_{\bar{\eta}}(\mathcal{H})+2)\cdot\bar{\eta} + 2C_1\bar{\eta}\right] \geq 1-\beta.$$

Moreover, under the same $(1-\beta)$-probability event, the class $\hat{\mathcal{L}} \subset [\lceil 2/\bar{\eta}\rceil]^{\mathcal{X}}$ output by `RegLearn` is $\bar{\ell}$-irreducible.

Proof In the proof we will often refer to the values $n_0, m, \bar{\eta}, \alpha_\Delta, \alpha_1, K, \mathcal{F}$ which are set in steps 1 through 2 of `RegLearn`. Throughout the proof we will write $d := \mathrm{sfat}_2(\mathcal{F}) \leq \mathrm{sfat}_{\bar{\eta}}(\mathcal{H})$ (Lemma 23). Since our choice of n_0 satisfies

$$n_0 \geq C_0 \cdot \frac{\mathrm{fat}_{c_0\bar{\eta}}(\mathcal{H})\log(1/\bar{\eta}) + \log(4m/\beta)}{\bar{\eta}^2},$$

where c_0, C_0 are the constants of Corollary 21, then by Corollary 21 and the choice of $\alpha_\Delta = 18$, we have that

$$\Pr_{S_{n_0}\sim Q^{n_0}}\left[\begin{smallmatrix}E_{\mathrm{good}} \text{ holds for the dataset } S_{n_0}\\ \text{and the distribution } \lfloor Q\rfloor_{\bar{\eta}}\end{smallmatrix}\right] = \Pr_{S_{n_0}\sim Q^{n_0}}\left[\sup_{f\in\mathcal{F}}\left|\mathrm{err}_{\lfloor Q\rfloor_{\bar{\eta}}}(f) - \mathrm{err}_{\lfloor \hat{Q}_{S_{n_0}}\rfloor_{\bar{\eta}}}(f)\right| \leq \alpha_\Delta/6\right] \geq 1 - \frac{\beta}{4m}.$$

(Recall the definition of E_{good} in (21).)

For $1 \leq j \leq m$, let $S_{n_0}^{(j)} := \{(x_1^{(j)},y_1^{(j)}),\ldots,(x_{n_0}^{(j)},y_{n_0}^{(j)})\}$ be the dataset of size n_0 drawn i.i.d. from Q in the jth iteration of step 3 of `RegLearn`. For convenience of notation let $\hat{Q}^{(j)} := \hat{Q}_{S_{n_0}^{(j)}} = \frac{1}{n_0}\sum_{i=1}^{n_0}\delta_{(x_i^{(j)},y_i^{(j)})}$ denote the empirical measure over $S_{n_0}^{(j)}$. Then by the union bound the probability that E_{good} holds for each of the datasets $S_{n_0}^{(j)}$ is at least $1-\beta/4$, i.e.,

$$\Pr\left[\forall j \in [m] : \sup_{f\in\mathcal{F}}\left|\mathrm{err}_{\lfloor Q\rfloor_{\bar{\eta}}}(f) - \mathrm{err}_{\lfloor \hat{Q}^{(j)}\rfloor_{\bar{\eta}}}(f)\right| \leq \alpha_\Delta/6\right] \geq 1 - \frac{\beta}{4}. \tag{32}$$

Let E_0 be the event inside the probability above, namely that E_{good} holds for each $S_{n_0}^{(j)}$.

The bulk of the proof of Theorem 44 is to show the following claims:

The first, Claim 45, shows that α_1 in step 2 in of `RegLearn` is differentially private and is with high probability an upper bound on the optimal error with respect to the true distribution Q:

6. In particular, c_0 is the corresponding constant of Corollary 21 and C_1 is the corresponding constant of Corollary 22.

Claim 45 (Privacy and accuracy of α_1) *The value α_1 produced in step 2 of* `RegLearn` *is $(\varepsilon, 0)$-differentially private as a function of the dataset T_{n_1} (and thus the entire dataset of n samples used by* `RegLearn`*). Moreover, α_1, satisfies the following:*

$$\Pr\left[\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\alpha_\Delta \geq \alpha_1 - d\cdot\alpha_\Delta \geq \inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\alpha_\Delta/6\right]\geq 1-\beta/4. \quad (33)$$

Claim 46 *There is an event E_1 that occurs with probability at least $1-\beta/2$ over the randomness of the dataset and the algorithm, so that under $E_0\cap E_1$,* `RegLearn` *outputs a class $\hat{\mathcal{L}}\subset\mathcal{F}$ which is $\bar\ell$-irreducible and satisfies $\hat{\mathcal{L}}\in\mathscr{R}^{(j)}$ for some $1\leq j\leq m$.*

Claim 47 *Let C_0, C_1 be the constants of Corollary 22. Suppose $\ell'\geq C_0K^2(d\log(K)+1)$. Under the event $E_1\cap E_0$, the output $\hat h$ of* `RegLearn` *satisfies*

$$\mathrm{err}_Q(\hat h)\leq \inf_{h\in\mathcal{H}}\left\{\mathrm{err}_Q(h)\right\}+30(d+2)\bar\eta+2C_1\bar\eta. \quad (34)$$

Assuming Claims 45, 46 and 47, we complete the proof of Theorem 44. By Claim 47, under the event $E_0\cap E_1$ (which holds with probability at least $1-\beta$), we have that the output hypothesis $\hat h:\mathcal{X}\to[-1,1]$ of `RegLearn` satisfies (34). Moreover, by Claim 46, under $E_0\cap E_1$, the class $\hat{\mathcal{L}}$ output by `RegLearn` is $\bar\ell$-irreducible.

Next we argue that the outputs $(\hat{\mathcal{L}},\hat h)$ of `RegLearn` are (ε,δ)-differentially private as a function of its input dataset (which consists of the disjoint union of the datasets $T_{n_1}, S_{n_0}^{(1)},\ldots,S_{n_0}^{(m)}$, which we denote as R). Let us consider two neighboring datasets R, R'. If they differ in a sample corresponding to T_{n_1}, then we have that for any event E, $\Pr_R[(\hat{\mathcal{L}},\hat h)\in E]\leq e^\varepsilon\cdot\Pr_{R'}[(\hat{\mathcal{L}},\hat h)\in E]$ by the $(\varepsilon,0)$-differential privacy of α_1 (Claim 45) and the post-processing lemma for differential privacy (Dwork and Roth, 2013, Proposition 2.1) (since for fixed $S_{n_0}^{(1)},\ldots,S_{n_0}^{(m)}$, $(\hat{\mathcal{L}},\hat h)$ are randomized functions of α_1). Otherwise, R, R' differ in a sample corresponding to one of $S_{n_0}^{(1)},\ldots,S_{n_0}^{(m)}$. Then the (ε,δ)-differential privacy guarantee of the sparse selection protocol of Proposition 7 guarantees that for any fixed α_1, for any event E, $\Pr_R[(\hat{\mathcal{L}},\hat h)\in E]\leq e^\varepsilon\cdot\Pr_{R'}[(\hat{\mathcal{L}},\hat h)\in E]+\delta$. This establishes that $(\hat{\mathcal{L}},\hat h)$ are differentially private as a function of R.

Summarizing, letting $d=\mathrm{sfat}_2(\mathcal{F})\leq\mathrm{sfat}_{\bar\eta}(\mathcal{H})$ and $d':=\mathrm{fat}_{c_0\bar\eta}(\mathcal{H})$ (where c_0 is the constant of Corollary 21), the sample complexity of `RegLearn` is

$$\begin{aligned} n_0\cdot m+n_1 &\leq C\cdot\frac{\bar\ell(2d+6)^{d+4}\log^2\left(\frac{1}{\varepsilon\delta\beta\bar\eta}\right)\cdot(d'\log(1/\bar\eta)+\log(m/\beta))}{\varepsilon\bar\eta^4}\\ &\leq C'\cdot\frac{\bar\ell d'(2d+6)^{d+5}\log^3\left(\frac{d\bar\ell}{\varepsilon\delta\beta\bar\eta}\right)}{\varepsilon\bar\eta^4}, \end{aligned}$$

where C, C' are sufficiently large constants.

It only remains to prove Claims 45, 46, and 47, which we do so below.

Proof [Proof of Claim 45] Let $C_0\geq 1, c_0\leq 1$ be the constants of Corollary 21; then by Corollary 21, as long as

$$n_1\geq C_0\cdot\frac{\mathrm{fat}_{c_0\bar\eta}(\mathcal{H})\log(1/\bar\eta)+\log(8/\beta)}{\bar\eta^2}, \quad (35)$$

we have

$$\Pr_{T_{n_1}\sim Q^{n_1}}\left[\sup_{f\in\mathcal{F}}\left|\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)-\mathrm{err}_{\lfloor \hat{Q}_{T_{n_1}}\rfloor_{\bar\eta}}(f)\right|>\frac{\alpha_\Delta}{6}\right]\le\beta/8.$$

Let Y denote the random variable drawn according to $\mathrm{Lap}(2K/(\varepsilon n_1))$ in step 2 of `RegLearn`. Then $\Pr[|Y|>2Kt/(\varepsilon n_1)]=\exp(-t)$ for all $t>0$, and in particular, as long as

$$n_1\ge C_1\cdot\frac{\log(1/\beta)}{\varepsilon\bar\eta} \tag{36}$$

for a sufficiently large constant C_1, it holds that $\Pr\left[|Y|>\frac{\alpha_\Delta}{6}\right]\le\beta/8$.

Under the event that both $\sup_{f\in\mathcal{F}}\left|\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)-\mathrm{err}_{\lfloor \hat{Q}_{T_{n_1}}\rfloor_{\bar\eta}}(f)\right|\le\alpha_\Delta/6$ and $|Y|\le\alpha_\Delta/6$, which holds with probability at least $1-\beta/4$, we get that

$$\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\frac{5\alpha_\Delta}{6}\ge\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor \hat{Q}_{T_{n_1}}\rfloor_{\bar\eta}}(f)\right\}+Y+\frac{\alpha_\Delta}{2}\ge\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\frac{\alpha_\Delta}{6}.$$

Note that the choice of n_1 in step 1 ensures that both (35) and (36) hold (as long as the constant C is sufficiently large). Recalling that $\hat\eta=\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor \hat{Q}_{T_{n_1}}\rfloor_{\bar\eta}}(f)\right\}+Y$ and $\alpha_1-d\cdot\alpha_\Delta=\hat\eta+\alpha_\Delta/2$, we get that (33) holds.

To see the differential privacy of α_1, note that the function that maps $T_{n_1}=\{(x_1,y_1),\ldots,(x_{n_1},y_{n_1})\}$ to $\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor \hat{Q}_{T_{n_1}}\rfloor_{\bar\eta}}(f)\right\}$ has sensitivity at most K/n_1, since $|f(x)-y|\le K$ for each $(x,y)\in\mathcal{X}\times[K]$ and $f\in\mathcal{F}$. Since $Y\sim\mathrm{Lap}((K/n_1)\cdot(2/\varepsilon))$, we get that α_1 is $(\varepsilon/2,0)$-differentially private as a function of the dataset T_{n_1}. ■

Proof [Proof of Claim 46] Recall that $\mathcal{F}=\lfloor\mathcal{H}\rfloor_{\bar\eta}$ and $P=\lfloor Q\rfloor_{\bar\eta}$, as well as $d=\mathrm{sfat}_2(\mathcal{F})\le\mathrm{sfat}_{\bar\eta}(\mathcal{H})$ (Lemma 23). For $\alpha>0$, $t\in[d+1]$, recall the definition of $\mathcal{M}_{\alpha,t}$ in (23) (defined with respect to $\mathcal{F}$ and P), and for those α,t for which $\mathcal{M}_{\alpha,t}$ is nonempty, the definition of $\sigma^\star_{\alpha,t}$ in (25). By definition of $\hat{\mathscr{S}}^{(j)}$ (see (20) and step 3 of `RegLearn`) and Lemma 34, as long as $\mathcal{F}_{\lfloor\hat{Q}^{(j)}\rfloor_{\bar\eta},\alpha_{d+1}}$ is nonempty, then under the event E_0, each $\mathscr{S}^{(j)}$ contains at least one hypothesis of the form $\mathrm{SOA}_{\hat{\mathcal{G}}}$, where $\|\sigma^\star_{\alpha_t-\alpha_\Delta/2,t}-\mathrm{SOA}_{\hat{\mathcal{G}}}\|_\infty\le5$. By the pigeonhole principle, some t satisfies this property for at least $\lceil m/(d+1)\rceil$ sets $\mathscr{S}^{(j)}$; let us denote this t by $t^\star$. We must also verify that $\mathcal{F}_{\lceil\hat{Q}^{(j)}\rceil_{\bar\eta},\alpha_{d+1}}$ is nonempty; to do so, let $E_{1,0}$ be the event that

$$\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\alpha_\Delta\ge\alpha_1-d\cdot\alpha_\Delta\ge\inf_{f\in\mathcal{F}}\left\{\mathrm{err}_{\lfloor Q\rfloor_{\bar\eta}}(f)\right\}+\alpha_\Delta/6. \tag{37}$$

By Claim 45, the probability that $E_{1,0}$ holds (over the choices of the algorithm `RegLearn`) is at least $1-\beta/4$. Then noting that $\alpha_{d+1}=\alpha_1-d\cdot\alpha_\Delta$ and using (32), we get that $\mathcal{F}_{\lfloor Q^{(j)}\rfloor_{\bar\eta},\alpha_{d+1}}$ is nonempty under the event $E_0\cap E_{1,0}$.

Since $\ell_t\ge\ell'$ for all $t\ge1$ (step 3 of `ReduceTreeReg`), it holds from (25) and (23) that $\sigma^\star_{\alpha_{t^\star}-\alpha_\Delta/2,t^\star}$ is of the form $\mathrm{SOA}_{\mathcal{G}}$ for some $\mathcal{G}\subset\mathcal{F}$ which is $\ell_{t^\star}$-irreducible, and thus ℓ'-irreducible. By Lemma 18 with $\chi=5$, as long as $\ell'\ge\bar\ell\cdot(d+3)^d$, there is some $\mathcal{L}^\star\subset\mathcal{F}$ which is $\bar\ell$-irreducible, depending only on $\mathcal{G}$, so that for any $\hat g$ satisfying $\|\hat g-\sigma^\star_{\alpha_{t^\star}-\alpha_\Delta/2,t^\star}\|_\infty\le5$, $\mathcal{L}^\star\in\mathscr{R}_{\hat g}$, where $\mathscr{R}_{\hat g}$ is as in step 4c of `RegLearn`. Thus, among the sets $\mathscr{R}^{(1)},\ldots,\mathscr{R}^{(m)}$, there are at least $\lceil m/(d+1)\rceil$ of them containing $\mathcal{L}^\star$.

By Lemma 35, we have that for each $1 \le j \le m$, $|\hat{\mathcal{S}}^{(j)}| \le K^{\ell' \cdot 2^{d+1}}$. By Lemma 42, each element $\hat{g} \in \mathscr{S}^{(j)}$ gives rise to $|\mathscr{R}_{\hat{g}}| \le K^{\bar{\ell} \cdot (d+4)^d}$ elements of $\mathscr{R}_{\hat{g}}$, all of which are added to $\hat{\mathscr{R}}^{(j)}$. Thus, recalling the definition of ℓ' in step 1 of `RegLearn`, we have that

$$|\hat{\mathscr{R}}^{(j)}| \le K^{\ell' \cdot 2^{d+1} + \bar{\ell} \cdot (d+4)^d} \le K^{\bar{\ell} \cdot (2d+6)^{d+2} + CK^2 d \log K} \le K^{C\bar{\ell}(2d+6)^{d+2} K^2 d \log K},$$

where $C > 0$ is a sufficiently large constant.

Now choose $\nu > 0$ so that the $(m, K^{C\bar{\ell}(2d+6)^{d+2}K^2 d \log K})$-sparse selection protocol of Proposition 7 (with universe $\mathcal{U}$ given by the family of all $\mathrm{SOA}_{\mathcal{G}}$, where $\mathcal{G} \subset \mathcal{F}$ is a finite restriction subclass of $\mathcal{F}$; this family must include all elements of $\mathscr{R}^{(j)}$, $1 \le j \le m$), has error at most ν on some event $E_{1,1}$ with probability at least $1 - \beta/4$. By (Ghazi et al., 2020b, Lemma 36), we may choose $\nu = \frac{C}{\varepsilon} \log\left(\frac{mK^{C\bar{\ell}(2d+6)^{d+2}K^2 d \log K}}{\varepsilon\delta\beta}\right)$ for a sufficiently large constant C.

Now set $E_1 = E_{1,0} \cap E_{1,1}$. Then under the event $E_0 \cap E_1$, as long as $\nu < \lceil m/(d+1) \rceil$, the hypothesis $\hat{\mathcal{L}}$ output by the sparse selection protocol belongs to $\hat{\mathscr{R}}^{(j)}$ for some $1 \le j \le m$. That $\hat{\mathcal{L}}$ is $\bar{\ell}$-irreducible follows from the fact $\mathscr{R}^{(j)}$ is the union of output sets $\mathscr{R}_{\hat{g}}$ of `SOAFilter`, for various functions $g : \mathcal{X} \to [K]$, and $\mathscr{R}_{\hat{g}}$ consists of $\bar{\ell}$-irreducible classes (Lemma 18).

To ensure $\nu < \lceil m/(d+1) \rceil$, it suffices to have, for C' a sufficiently large constant,

$$m > \frac{C'd}{\varepsilon} \cdot \left(\log(m) + \log\left(\frac{1}{\varepsilon\delta\beta}\right) + \bar{\ell}(2d+6)^{d+3} K^2 \log^2 K\right),$$

for which it in turn suffices that

$$m \ge \frac{C''\bar{\ell}(2d+6)^{d+4} \log^2\left(\frac{1}{\varepsilon\delta\beta\bar{\eta}}\right)}{\varepsilon\bar{\eta}^2},$$

where we have used that $K = \lceil 2/\bar{\eta} \rceil$, and C'' is a sufficiently large constant.

■

Proof [Proof of Claim 47] By Claim 46, under the event $E_1 \cap E_0$, `RegLearn` outputs a class $\hat{\mathcal{L}} \in \mathscr{R}^{(j)}$ for some $1 \le j \le m$, which is $\bar{\ell}$-irreducible. For the remainder of the proof we assume that $E_1 \cap E_0$ holds and fix such a j. By Lemma 18 (with $\chi = 5$), there is some $\hat{g} \in \mathscr{S}^{(j)}$ so that $\|\mathrm{SOA}_{\hat{\mathcal{L}}} - \hat{g}\|_\infty \le 12(d+1)$. Set $\hat{P}^{(j)} := \lfloor \hat{Q}^{(j)} \rfloor_{\bar{\eta}}$. By definition, each element $\hat{g} \in \hat{\mathscr{S}}^{(j)}$ is of the form $\mathrm{SOA}_{\mathcal{F}_{\hat{P}^{(j)}, \alpha_t - 2\alpha_\Delta/3}|_{\mathbf{A}(v)}}$ for some $1 \le t \le d$ and some node v of the tree $\hat{\mathbf{x}}$ output by `ReduceTreeReg` for which $\mathcal{F}_{\hat{P}^{(j)}, \alpha_t - 2\alpha_\Delta/3}|_{\mathbf{A}(v)}$ is nonempty and ℓ'-irreducible (see (20)). Fix any such element, and write $\mathcal{J} := \mathcal{F}_{\hat{P}^{(j)}, \alpha_t - 2\alpha_\Delta/3}|_{\mathbf{A}(v)}$. By definition we have that each $f \in \mathcal{J} \subset \mathcal{F}_{\hat{P}^{(j)}, \alpha_t - 2\alpha_\Delta/3}$ satisfies, under the event $E_1 \cap E_0$,

$$\mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \le \mathrm{err}_{\hat{P}^{(j)}}(f) + \alpha_\Delta/6 \le \alpha_t - \alpha_\Delta/2 \le \alpha_1 - \alpha_\Delta/2 \le \inf_{f \in \mathcal{F}} \left\{\mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f)\right\} + (d+1)\alpha_\Delta, \tag{38}$$

where the first inequality holds under E_0 (see (32)) and the final inequality follows from (37), which holds under $E_1 \cap E_0$ (in particular, it holds under the event $E_{1,0}$ defined in the proof of Claim 45, which is included in E_1).

Recall the definition of finite restriction subclasses of $\mathcal{F}$ from Section 3. Since $\mathcal{X}$ is countable, the set of all finite restriction subclasses of $\mathcal{X}$ is countable; thus the set of all finite unions of finite restriction subclasses of $\mathcal{F}$ is countable as well. Define

$$\tilde{\mathcal{F}} = \mathcal{F} \cup \left\{ \mathrm{SOA}_{\mathcal{G}} : \substack{\mathcal{G} \subset \mathcal{F},\, \mathcal{G} \text{ is nonempty, } (d+1)\text{-irreducible,} \\ \text{and a finite union of finite restriction subclasses of } \mathcal{F}} \right\}.$$

Then $\tilde{\mathcal{F}}$ is countable, and Lemma 29 gives that $\mathrm{fat}_2(\tilde{\mathcal{F}}) \le \mathrm{sfat}_2(\tilde{\mathcal{F}}) = d$.

Let C_0, C_1 be the constants of Corollary 22, and choose $n_2 \ge C_0 K^2 \cdot (d \log(K) + 1)$ (recall $K = \lceil 2/\bar{\eta} \rceil$). By Corollary 22 applied to the class $\tilde{\mathcal{F}}$, we have:

$$\Pr_{S_{n_2} \sim Q^{n_2}} \left[\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left| \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\tilde{f}) - \mathrm{err}_{\lfloor \hat{Q}_{S_{n_2}} \rfloor_{\bar{\eta}}}(\tilde{f}) \right| > C_1 \right] \le 1/2.$$

Choose some dataset $S_{n_2} \in (\mathcal{X} \times [-1,1])^{n_2}$ so that $\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left| \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\tilde{f}) - \mathrm{err}_{\lfloor \hat{Q}_{S_{n_2}} \rfloor_{\bar{\eta}}}(\tilde{f}) \right| \le C_1$ holds, and write $\hat{P} := \lfloor \hat{Q}_{S_{n_2}} \rfloor_{\bar{\eta}}$ as the discretization of the empirical distribution $\hat{Q}_{S_{n_2}}$. Then by (38), each $f \in \mathcal{J}$ satisfies

$$\mathrm{err}_{\hat{P}}(f) \le \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)\alpha_\Delta + C_1. \tag{39}$$

We next claim that $\mathrm{err}_{\hat{P}}(\mathrm{SOA}_{\mathcal{J}}) \le \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)\alpha_\Delta + C_1$. Suppose for the purpose of contradiction that this is not the case. Let us write $S_{n_2} = \{(x_1, y_1), \ldots, (x_{n_2}, y_{n_2})\}$. For $1 \le i \le n_2$, write $\tilde{y}_i := \mathrm{SOA}_{\mathcal{J}}(x_i)$. Since $\mathcal{J}$ is ℓ'-irreducible and the definition of ℓ' in step 1 of `RegLearn` ensures $\ell' \ge n_2$, it holds that

$$\mathrm{sfat}_2(\mathcal{J}|_{\{(x_1,\tilde{y}_1),\ldots,(x_{n_2},\tilde{y}_{n_2})\}}) = \mathrm{sfat}_2(\mathcal{J}) \ge 0.$$

Thus there is some $f \in \mathcal{J}$ so that $f(x_i) = \tilde{y}_i = \mathrm{SOA}_{\mathcal{J}}(\tilde{x}_i)$ for $1 \le i \le n_2$. Thus $\mathrm{err}_{\hat{P}}(\mathrm{SOA}_{\mathcal{J}}) = \mathrm{err}_{\hat{P}}(f) > \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)\alpha_\Delta + C_1$, which contradicts (39). Since $\mathrm{SOA}_{\mathcal{J}} \in \tilde{\mathcal{F}}$ (as $\ell' \ge n_2 \ge d+1$), it follows from the choice of S_{n_2} that

$$\mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\mathrm{SOA}_{\mathcal{J}}) \le \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)\alpha_\Delta + 2C_1. \tag{40}$$

Recalling that $\| \mathrm{SOA}_{\hat{\mathcal{L}}} - \mathrm{SOA}_{\mathcal{J}} \|_\infty \le 12(d+1)$ and using (40), we get that

$$\mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\mathrm{SOA}_{\hat{\mathcal{L}}}) \le \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\mathrm{SOA}_{\mathcal{J}}) + \| \mathrm{SOA}_{\hat{\mathcal{L}}} - \mathrm{SOA}_{\mathcal{J}} \|_\infty \le \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)(\alpha_\Delta + 12) + 2C_1.$$

Finally, using (3) with $\eta = \bar{\eta}$ and the definition of $\hat{h} : \mathcal{X} \to [-1,1]$ in step 5 of `RegLearn` (which implies that $\mathrm{SOA}_{\hat{\mathcal{L}}} = \lfloor \hat{h} \rfloor_{\bar{\eta}}$), we get

$$\begin{aligned}
\mathrm{err}_Q(\hat{h}) &\le \frac{2(1 + \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(\mathrm{SOA}_{\hat{\mathcal{L}}}))}{\lceil 2/\bar{\eta} \rceil} \\
&\le \frac{2(1 + \inf_{f \in \mathcal{F}} \left\{ \mathrm{err}_{\lfloor Q \rfloor_{\bar{\eta}}}(f) \right\} + (d+1)(\alpha_\Delta + 12) + 2C_1)}{\lceil 2/\bar{\eta} \rceil} \\
&\le \frac{2(1 + \frac{\lceil 2/\bar{\eta} \rceil}{2} \cdot \inf_{h \in \mathcal{H}} \left\{ \mathrm{err}_Q(h) \right\} + 1 + (d+1)(\alpha_\Delta + 12) + 2C_1)}{\lceil 2/\bar{\eta} \rceil} \\
&\le \inf_{h \in \mathcal{H}} \left\{ \mathrm{err}_Q(h) \right\} + \bar{\eta} \cdot (d+2)(\alpha_\Delta + 12) + 2C_1 \bar{\eta} \\
&= \inf_{h \in \mathcal{H}} \left\{ \mathrm{err}_Q(h) \right\} + 30(d+2)\bar{\eta} + 2C_1 \bar{\eta},
\end{aligned}$$

where the last line follows from the choice of $\alpha_\Delta = 18$. ■

■