Minimizing a Submodular Function from Samples

Eric Balkanski
Harvard University
ericbalkanski@g.harvard.edu

Yaron Singer
Harvard University
yaron@seas.harvard.edu

Abstract

In this paper we consider the problem of minimizing a submodular function from training data. Submodular functions can be efficiently minimized and are consequently heavily applied in machine learning. There are many cases, however, in which we do not know the function we aim to optimize, but rather have access to training data that is used to learn it. In this paper we consider the question of whether submodular functions can be minimized when given access to its training data. We show that even learnable submodular functions cannot be minimized within any non-trivial approximation when given access to polynomially-many samples. Specifically, we show that there is a class of submodular functions with range in $[0, 1]$ such that, despite being PAC-learnable and minimizable in polynomial-time, no algorithm can obtain an approximation strictly better than $1/2 - o(1)$ using polynomially-many samples drawn from any distribution. Furthermore, we show that this bound is tight via a trivial algorithm that obtains an approximation of $1/2$.

1 Introduction

For well over a decade now, submodular minimization has been heavily studied in machine learning (e.g. [SK10, JB11, JLB11, NB12, EN15, DTK16]). This focus can be largely attributed to the fact that if a set function $f : 2^N \to \mathbb{R}$ is submodular, meaning it has the following property of diminishing returns: $f(S \cup \{a\}) - f(S) \geq f(T \cup \{a\}) - f(T)$ for all $S \subseteq T \subseteq N$ and $a \notin T$, then it can be optimized efficiently: its minimizer can be found in time that is polynomial in the size of the ground set N [GLS81, IFF01]. In many cases, however, we do not know the submodular function, and instead learn it from data (e.g. [BH11, IJB13, FKV13, FK14, Bal15, BVW16]). The question we address in this paper is whether submodular functions can be (approximately) minimized when the function is not known but can be learned from training data.

An intuitive approach for optimization from training data is to learn a surrogate function from training data that predicts the behavior of the submodular function well, and then find the minimizer of the surrogate learned and use that as a proxy for the true minimizer we seek. The problem however, is that this approach does not generally guarantee that the resulting solution is close to the true minimum of the function. One pitfall is that the surrogate may be non-submodular, and despite approximating the true submodular function arbitrarily well, the surrogate can be intractable to minimize. Alternatively, it may be that the surrogate is submodular, but its minimum is arbitrarily far from the minimum of the true function we aim to optimize (see examples in Appendix A).

Since optimizing a surrogate function learned from data may generally result in poor approximations, one may seek learning algorithms that are guaranteed to produce surrogates whose optima well-approximate the true optima and are tractable to compute. More generally, however, it is possible that there is some other approach for optimizing the function from the training samples, without learning a model. Therefore, at a high level, the question is whether a reasonable number of training samples suffices to minimize a submodular function. We can formalize this as *optimization from samples*.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Optimization from samples. We will say that a class of functions $\mathcal{F} = \{f : 2^N \to [0,1]\}$ is *α-optimizable from samples over distribution $\mathcal{D}$* if for every $f \in \mathcal{F}$ and $\delta \in (0,1)$, when given $\text{poly}(|N|)$ i.i.d. samples $\{(S_i, f(S_i))\}_{i=1}^m$ where $S_i \sim \mathcal{D}$, with probability at least $1-\delta$ over the samples one can construct an algorithm that returns a solution $S \subseteq N$ s.t,

$$f(S) - \min_{T \subseteq N} f(T) \leq \alpha.$$

This framework was recently introduced in [BRS17] for the problem of submodular maximization where the standard notion of approximation is multiplicative. For submodular minimization, since the optimum may have zero value, the suitable measure is that of additive approximations for $[0,1]$-bounded functions, and the goal is to obtain a solution which is a $o(1)$ additive approximation to the minimum (see e.g. [CLSW16, EN15, SK10]). The question is then:

Can submodular functions be minimized from samples?

Since submodular functions can be minimized in polynomial-time, it is tempting to conjecture that when the function is learnable it also has desirable approximation guarantees from samples, especially in light of positive results in related settings of *submodular maximization*:

- **Constrained maximization.** For functions that can be maximized in polynomial time under a cardinality constraint, like modular and unit-demand functions, there are polynomial time algorithms that obtain an arbitrarily good approximation using polynomially-many samples [BRS16, BRS17]. For general monotone submodular functions which are NP-hard to maximize under cardinality constraints, there is no algorithm that can obtain a reasonable approximation from polynomially-many samples [BRS17]. For the problem of unconstrained minimization, submodular functions can be optimized in polynomial time;
- **Unconstrained maximization.** For unconstrained maximization of general submodular functions, the problem is NP-hard to maximize (e.g. `MAX-CUT`) and one seeks constant factor approximations. For this problem, there is an extremely simple algorithm that uses no queries and obtains a good approximation: choose elements uniformly at random with probability $1/2$ each. This algorithm achieves a constant factor approximation of $1/4$ for general submodular functions. For *symmetric* submodular functions (i.e. $f(S) = f(N \setminus S)$), this algorithm is a $1/2$-approximation which is optimal, since no algorithm can obtain an approximation ratio strictly better than $1/2$ using polynomially-many *value* queries, even for symmetric submodular functions [FMV11]. For unconstrained symmetric submodular *minimization*, there is an appealing analogue: the empty set and the ground set N are guaranteed to be minimizers of the function (see Section 2). This algorithm, of course, uses no queries either. The parallel between these two problems seems quite intuitive, and it is tempting to conjecture that like for unconstrained submodular maximization, there are optimization from samples algorithms for general unconstrained submodular minimization with good approximation guarantees.

Main result. Somewhat counter-intuitively, we show that despite being computationally tractable to optimize, submodular functions cannot be minimized from samples to within a desirable guarantee, even when these functions are learnable. In particular, we show that there is no algorithm for minimizing a submodular function from polynomially-many samples drawn from any distribution that obtains an additive approximation of $1/2 - o(1)$, even when the function is `PAC`-learnable. Furthermore, we show that this bound is tight: the algorithm which returns the empty set or ground set each with probability $1/2$ achieves at least a $1/2$ approximation. Notice that this also implies that in general, there is no learning algorithm that can produce a surrogate whose minima is close to the minima of the function we aim to optimize, as otherwise this would contradict our main result.

Technical overview. At a high level, hardness results in optimization from samples are shown by constructing a family of functions, where the values of functions in the family are likely to be indistinguishable for the samples, while having very different optimizers. The main technical difficulty is to construct a family of functions that concurrently satisfy these two properties (indistinguishability and different optimizers), and that are also `PAC`-learnable. En route to our main construction, we first construct a family of functions that are completely indistinguishable given samples drawn from the uniform distribution, in which case we obtain a $1/2 - o(1)$ impossibility result (Section 2). The

general result that holds for any distribution requires heavier machinery to argue about more general families of functions where some subset of functions can be distinguished from others given samples. Instead of satisfying the two desired properties for all functions in a fixed family, we show that these properties hold for all functions in a randomized subfamily (Section 3.2). We then develop an efficient learning algorithm for the family of functions constructed for the main hardness result (Section 3.3). This algorithm builds multiple linear regression predictors and a classifier to direct a fresh set to the appropriate linear predictor. The learning of the classifier and the linear predictors relies on multiple observations about the specific structure of this class of functions.

1.1 Related work

The problem of optimization from samples was introduced in the context of constrained submodular maximization [BRS17, BRS16]. In general, for maximizing a submodular function under a cardinality constraint, no algorithm can obtain a constant factor approximation guarantee from any samples. As discussed above, for special classes of submodular functions that can be optimized in polynomial time under a cardinality constraint, and for unconstrained maximization, there are desirable optimization from samples guarantees. It is thus somewhat surprising that submodular minimization, which is an unconstrained optimization problem that is optimizable in polynomial time in the value query model, is hard to optimize from samples. From a technical perspective the constructions are quite different. In maximization, the functions constructed in [BRS17, BRS16] are monotone so the ground set would be an optimal solution if the problem was unconstrained. Instead, we need to construct novel non-monotone functions. In convex optimization, recent work shows a tight $1/2$-inapproximability for convex minimization from samples [BS17]. Although there is a conceptual connection between that paper and this one, from a technical perspective these papers are orthogonal. The discrete analogue of the family of convex functions constructed in that paper is not (even approximately) a family of submodular functions, and the constructions are significantly different.

2 Warm up: the Uniform Distribution

As a warm up to our main impossibility result, we sketch a tight lower bound for the special case in which the samples are drawn from the uniform distribution. At a high level, the idea is to construct a function which considers some special subset of "good" elements that make its value drops when a set contains *all* such "good" elements. When samples are drawn from the uniform distribution and "good" elements are sufficiently rare, there is a relatively simple construction that obfuscates which elements the function considers "good", which then leads to the inapproximability.

2.1 Hardness for uniform distribution

We construct a family of functions $\mathcal{F}$ where $f_i \in \mathcal{F}$ is defined in terms of a set $G_i \subset N$ of size $\sqrt{n}$. For each such function we call G_i the set of *good* elements, and $B_i = N \setminus G_i$ its *bad* elements. We denote the number of good and bad elements in a set S by g_S and b_S, dropping the subscripts (S and i) when clear from context, so $g = |G_i \cap S|$ and $b = |B_i \cap S|$. The function f_i is defined as follows:

$$f_i(S) := \begin{cases} \frac{1}{2} + \frac{1}{2n} \cdot (g + b) & \text{if } g < \sqrt{n} \\ \frac{1}{2n} \cdot b & \text{if } g = \sqrt{n} \end{cases}$$

It is easy to verify that these functions are submodular with range in $[0, 1]$ (see illustration in Figure 1a). Given samples drawn uniformly at random (u.a.r.), it is impossible to distinguish good and bad elements since with high probability (w.h.p.) $g < \sqrt{n}$ for all samples. Informally, this implies that a good learner for $\mathcal{F}$ over the uniform distribution $\mathcal{D}$ is $f'(S) = 1/2 + |S|/(2n)$.

Intuitively, $\mathcal{F}$ is not $1/2 - o(1)$ optimizable from samples because if an algorithm cannot learn the set of good elements G_i, then it cannot find S such $f_i(S) < 1/2 - o(1)$ whereas the optimal solution $S_i^\star = G_i$ has value $f_i(G_i) = 0$.

Theorem 1. *Submodular functions $f : 2^N \to [0, 1]$ are not $1/2 - o(1)$ optimizable from samples drawn from the uniform distribution for the problem of submodular minimization.*

Proof. The details for the derivation of concentration bounds are in Appendix B. Consider f_k drawn u.a.r. from $\mathcal{F}$ and let $f^\star = f_k$ and $G^\star = G_k$. Since the samples are all drawn from the uniform distribution, by standard application of the Chernoff bound we have that every set S_i in the sample respects $|S_i| \leq 3n/4$, w.p. $1 - e^{-\Omega(n)}$. For sets $S_1, \ldots, S_m$, all of size at most $3n/4$, when f_j is drawn u.a.r. from $\mathcal{F}$ we get that $|S_i \cap G_j| < \sqrt{n}$, w.p. $1 - e^{-\Omega(n^{1/2})}$ for all $i \in [m]$, again by Chernoff, and since $m = \text{poly}(n)$. Notice that this implies that w.p. $1 - e^{-\Omega(n^{1/2})}$ for all $i \in [m]$:

$$f_j(S_i) = \frac{1}{2} + \frac{|S_i|}{2n}$$

Now, let $\mathcal{F}'$ be the collection of all functions f_j for which $f_j(S_i) = 1/2 + |S_i|/(2n)$ on all sets $\{S_i\}_{i=1}^m$. The argument above implies that $|\mathcal{F}'| = (1 - e^{-\Omega(n^{1/2})})|\mathcal{F}|$. Thus, since $f^\star$ is drawn u.a.r. from $\mathcal{F}$ we have that $f^\star \in \mathcal{F}'$ w.p. $1 - e^{-\Omega(n^{1/2})}$, and we condition on this event.

Let S be the (possibly randomized) solution returned by the algorithm. Observe that S is *independent* of $f^\star \in \mathcal{F}'$. In other words, the algorithm cannot learn any information about which function in $\mathcal{F}'$ generates the samples. By Chernoff, if we fix S and choose f u.a.r. from $\mathcal{F}$, then, w.p. $1 - e^{-\Omega(n^{1/6})}$:

$$f(S) \geq \frac{1}{2} - o(1)$$

Thus, since $|\mathcal{F}'| = (1 - e^{-\Omega(n^{1/2})})|\mathcal{F}|$, w.p. $1 - e^{-\Omega(n^{1/6})}$ over the choice of $f^\star$, we have that $f^\star(S) \geq 1/2 - o(1)$ and we condition on this event. Conditioning on all events, the value of the set S returned by the algorithm is $1/2 - o(1)$ whereas the optimal solution is $f^\star(G^\star) = 0$. Since all the events we conditioned on occur with exponentially high probability, this concludes the proof. □

2.2 A tight upper bound

We now show that the result above is tight. In particular, by randomizing between the empty set and the ground set we get a solution whose value is at most $1/2$. In the case of symmetric submodular functions, the empty set and the ground set are minima. Notice, that this does not require any samples.

Proposition 2. *The algorithm which returns the empty set $\emptyset$ or the ground N with probability $1/2$ each is a $1/2$ additive approximation for the problem of unconstrained submodular minimization.*

Proof. Let $S \subseteq N$, observe that

$$f(N \setminus S) - f(\emptyset) = f_\emptyset(N \setminus S) \geq f_S(N \setminus S) = f(N) - f(S)$$

where the inequality is by submodularity. Thus, we obtain

$$\frac{1}{2}(f(N) + f(\emptyset)) \leq \frac{1}{2}f(S) + \frac{1}{2}f(N \setminus S) \leq f(S) + \frac{1}{2}.$$

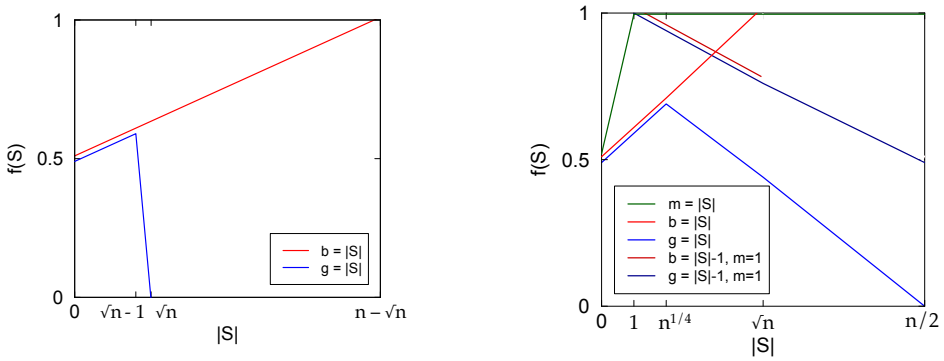
In particular, this holds for $S \in \text{argmin}_{T \subseteq N} f(T)$. □

Since $f(N) + f(\emptyset)) \leq f(S) + f(N \setminus S)$ for all $S \subseteq N$, the ground set N and the empty set $\emptyset$ are minima if f is symmetric (i.e. $f(S) = f(N \setminus S)$ for all $S \subseteq N$).

3 General Distribution

In this section, we show our main result, namely that there exists a family of submodular functions such that, despite being PAC-learnable for all distributions, no algorithm can obtain an approximation better than $1/2 - o(1)$ for the problem of unconstrained minimization.

The functions in this section build upon the previous construction, though are inevitably more involved in order to achieve learnability and inapproximability on *any* distribution. The functions constructed for the uniform distribution do not yield inapproximability for general distributions due to the fact that the indistinguishability between two functions no longer holds when sets S of large size are sampled with non-negligible probability. Intuitively, in the previous construction, once a set is sufficiently large the good elements of the function can be distinguished from the bad ones. The main idea to get around this issue is to introduce *masking* elements M. We construct functions such that, for sets S of large size, good and bad elements are indistinguishable if S contains *at least one* masking element.



(a) Uniform distribution

(b) General distribution

Figure 1: An illustration of the value of a set S of good (blue), bad (red), and masking (green) elements as a function of $|S|$ for the functions constructed. For the general distribution case, we also illustrate the value of a set S of good (dark blue) and bad (dark red) elements when S also contains at least one masking element.

The construction. Each function $f_i \in \mathcal{F}$ is defined in terms of a partition P_i of the ground set into *good*, *bad*, and *masking* elements. The partitions we consider are $P_i = (G_i, B_i, M_i)$ with $|G_i| = n/2$, $|B_i| = \sqrt{n}$, and $|M_i| = n/2 - \sqrt{n}$. Again, when clear from context, we drop indices of i and S and the number of good, bad, and masking elements in a set S are denoted by g, b, and m. For such a given partition P_i, the function f_i is defined as follows (see illustration in Figure 1b):

$$f_i(S) = \frac{1}{2} + \begin{cases} \frac{1}{2\sqrt{n}} \cdot (b+g) & \textbf{Region } \mathcal{X}: \text{ if } m = 0 \text{ and } g < n^{1/4} \\ \frac{1}{2\sqrt{n}} \cdot \left(b + n^{1/4}\right) - \frac{1}{n} \cdot \left(g - n^{1/4}\right) & \textbf{Region } \mathcal{Y}: \text{ if } m = 0 \text{ and } g \geq n^{1/4} \\ \frac{1}{2} - \frac{1}{n} \cdot (b+g) & \textbf{Region } \mathcal{Z}: \text{ otherwise} \end{cases}$$

3.1 Submodularity

In the appendix, we prove that the functions f_i constructed as above are indeed submodular (Lemma 10). By rescaling f_i with an additive term of $n^{1/4}/(2\sqrt{n}) = 1/(2n^{1/4})$, it can be easily verified that its range is in $[0, 1]$. We use the non-normalized definition as above for ease of notation.

3.2 Inapproximability

We now show that $\mathcal{F}$ cannot be minimized within a $1/2 - o(1)$ approximation given samples from any distribution. We first define $\mathcal{F}^M$, which is a randomized subfamily of $\mathcal{F}$. We then give a general lemma that shows that if two conditions of indistinguishability and gap are satisfied then we obtain inapproximability. We then show that these two conditions are satisfied for the subfamily $\mathcal{F}^M$.

A randomization over masking elements. Instead of considering a function f drawn u.a.r. from $\mathcal{F}$ as in the uniform case, we consider functions f in a *randomized subfamily* of functions $\mathcal{F}^M \subseteq \mathcal{F}$ to obtain the indistinguishability and gap conditions. Given the family of functions $\mathcal{F}$, let M be a *uniformly random subset* of size $n/2 - \sqrt{n}$ and define $\mathcal{F}^M \subset \mathcal{F}$:

$$\mathcal{F}^M := \{f_i \in \mathcal{F} : (G_i, B_i, M)\}.$$

Since masking elements are distinguishable from good and bad elements, they need to be the same set of elements for each function in family $\mathcal{F}^M$ to obtain indistinguishability of functions in $\mathcal{F}^M$.

The inapproximability lemma. In addition to this randomized subfamily of functions, another main conceptual departure of the following inapproximability lemma from the uniform case is that no assumption can be made about the samples, such as their size, since the distribution is arbitrary. The ordering of the quantifiers for the two conditions of this lemma is crucial.

Lemma 3. *Let $\mathcal{F}$ be a family of functions and $\mathcal{F}' = \{f_1, \ldots, f_p\} \subseteq \mathcal{F}$ be a subfamily of functions drawn from some distribution. Assume the following two conditions hold:*

1. ***Indistinguishability.*** *For all $S \subseteq N$, w.p. $1 - e^{-\Omega(n^{1/4})}$ over $\mathcal{F}'$: for every $f_i, f_j \in \mathcal{F}'$,*
$$f_i(S) = f_j(S);$$
2. ***α-gap.*** *Let $S_i^\star$ be a minimizer of f_i, then w.p. 1 over $\mathcal{F}'$: for all $S \subseteq N$,*
$$\mathop{\mathbb{E}}_{f_i \sim U(\mathcal{F}')} \left[f_i(S) - f_i\left(S_i^\star\right)\right] \geq \alpha;$$

Then, $\mathcal{F}$ is not α-minimizable from strictly less than $e^{\Omega(n^{1/4})}$ samples over any distribution $\mathcal{D}$.

The proof is deferred to the appendix, but at a high level the main ideas can be summarized as follows. We use a probabilistic argument to switch from the randomization over $\mathcal{F}'$ to the randomization over $S \sim \mathcal{D}$ and show that there exists a deterministic $F \subseteq \mathcal{F}$ such that $f_i(S) = f_j(S)$ for all $f_i, f_j \in F$ w.h.p. over $S \sim \mathcal{D}$. By a union bound this holds for all samples S. Thus, for such a family of functions $F = \{f_1, \ldots, f_p\}$, the choices of an algorithm that is given samples from f_i for $i \in [m]$ are *independent* of i. By the α-gap condition, this implies that there exists at least one $f_i \in F$ for which a solution S returned by the algorithm is at least α away from $f_i(S_i^\star)$.

Indistinguishability and gap of $\mathcal{F}$. We now show the indistinguishability and gap conditions, with $\alpha = 1/2 - o(1)$, which immediately imply a $1/2 - o(1)$ inapproximability by Lemma 3. For the indistinguishability, it suffices to show that good and bad elements are indistinguishable since the masking elements are identical for all functions in $\mathcal{F}^M$. Good and bad elements are indistinguishable since, w.h.p., a set S is not in region $\mathcal{Y}$, which is the only region distinguishing good and bad elements.

Lemma 4. *For all $S \subseteq N$ s.t. $|S| < n^{1/4}$: For all $f_i \in \mathcal{F}^M$,*
$$f_i(S) = \frac{1}{2} + \begin{cases} \frac{1}{2\sqrt{n}} \cdot (b+g) & \text{if } m = 0 \text{ (Region } \mathcal{X}\text{)} \\ \frac{1}{2} - \frac{1}{n} \cdot (b+g) & \text{otherwise (Region } \mathcal{Z}\text{)} \end{cases}$$
and for all $S \subseteq N$ such that $|S| \geq n^{1/4}$, with probability $1 - e^{-\Omega(n^{1/4})}$ over $\mathcal{F}^M$: For all $f_i \in \mathcal{F}^M$,
$$f_i(S) = 1 - \frac{1}{n} \cdot (b+g) \qquad \text{(Region } \mathcal{Z}\text{)}$$

Proof. Let $S \subseteq N$. If $|S| < n^{1/4}$, then the proof follows immediately from the definition of f_i. If $|S| \geq n^{1/4}$, then, the number of masking elements m in S is $m = |M \cap S|$ for all $f_i \in \mathcal{F}^M$. We then get $m \geq 1$, for all $f_i \in \mathcal{F}^M$, with probability $1 - e^{-\Omega(n^{1/4})}$ over $\mathcal{F}^M$ by Chernoff bound. The proof then follows again immediately from the definition of f_i. □

Next, we show the gap. The gap is since the good elements can be any subset of $N \setminus M$.

Lemma 5. *Let $S_i^\star$ be a minimizer of f_i. With probability 1 over $\mathcal{F}^M$, for all $S \subseteq N$,*
$$\mathop{\mathbb{E}}_{f_i \sim U(\mathcal{F}^M)} [f_i(S)] \geq \frac{1}{2} - o(1).$$

Proof. Let $S \subseteq N$ and $f_i \sim U(\mathcal{F}^M)$. Note that the order of the quantifiers in the statement of the lemma implies that S can be dependent on M, but that it is independent of i. There are three cases. If $m \geq 1$, then S is in region $\mathcal{Z}$ and $f_i(S) \geq 1/2$. If $m = 0$ and $|S| \leq n^{7/8}$, then S is in region $\mathcal{X}$ or $\mathcal{Y}$ and $f_i(S) \geq 1/2 - n^{7/8}/n = \frac{1}{2} - o(1)$. Otherwise, $m = 0$ and $|S| \geq n^{7/8}$. Since S is independent of i, by Chernoff bound, we get
$$(1 - o(1)) \cdot |S| \leq \frac{n/2 + \sqrt{n}}{\sqrt{n}} \cdot b, \frac{n/2 + \sqrt{n}}{n/2} \cdot g \leq (1 + o(1)) \cdot |S|$$
with probability $1 - e^{-\Omega(n^{1/4})}$. Thus S is in region $\mathcal{Y}$ and
$$f_i(S) \geq \frac{1}{2} + (1 - o(1)) \frac{1}{2\sqrt{n}} \cdot \frac{\sqrt{n}}{n/2 + \sqrt{n}} \cdot |S| - (1 + o(1)) \frac{1}{n} \cdot \frac{n/2}{n/2 + \sqrt{n}} \cdot |S| \geq \frac{1}{2} - o(1).$$
Thus, we obtain $\mathbb{E}_{f_i \sim U(\mathcal{F}^M)} [f_i(S)] \geq \frac{1}{2} - o(1)$. □

Combining the above three lemmas, we obtain the inapproximability result.

Lemma 6. *The problem of submodular minimization cannot be approximated with a* $1/2 - o(1)$ *additive approximation given* $\mathrm{poly}(n)$ *samples from any distribution* $\mathcal{D}$.

Proof. For any set $S \subseteq N$, observe that the number $g + b$ of elements in S that are either good or bad is the same for any two functions $f_i, f_j \in \mathcal{F}^M$ and for any $\mathcal{F}^M$. Thus, by Lemma 4, we obtain the indistinguishability condition. Next, the optimal solution $S_i^\star = G_i$ of f_i has value $f_i(G_i) = o(1)$, so by Lemma 5, we obtain the α-gap condition with $\alpha = 1/2 - o(1)$. Thus $\mathcal{F}$ is not $1/2 - o(1)$ minimizable from samples from any distribution $\mathcal{D}$ by Lemma 3. The class of functions $\mathcal{F}$ is a class of submodular functions by Lemma 10 (in Appendix C). □

3.3 Learnability of $\mathcal{F}$

We now show that every function in $\mathcal{F}$ is efficiently learnable from samples drawn from any distribution $\mathcal{D}$. Specifically, we show that for any $\epsilon, \delta \in (0, 1)$ the functions are $(\epsilon, \delta) - \texttt{PAC}$ learnable with the absolute loss function (or any Lipschitz loss function) using $\mathrm{poly}(1/\epsilon, 1/\delta, n)$ samples and running time. At a high level, since each function f_i is piecewise-linear over three different regions $\mathcal{X}_i, \mathcal{Y}_i$, and $\mathcal{Z}_i$, the main idea is to exploit this structure by first training a classifier to distinguish between regions and then apply linear regression in different regions.

The learning algorithm. Since every function $f \in \mathcal{F}$ is piecewise linear over three different regions, there are three different linear functions $f_{\mathcal{X}}, f_{\mathcal{Y}}, f_{\mathcal{Z}}$ s.t. for every $S \subseteq N$ its value $f(S)$ can be expressed as $f_{\mathcal{R}}(S)$ for some region $\mathcal{R} \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$. The learning algorithm produces a predictor $\tilde{f}$ by using a multi-label classifier and a set of linear predictors $\{f_{\tilde{\mathcal{X}}}, f_{\tilde{\mathcal{Y}}}\} \cup \{\cup_{i \in \tilde{M}} f_{\tilde{\mathcal{Z}}_i}\}$. The multi-label classifier creates a mapping from sets to regions, $g : 2^N \to \{\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}\} \cup \{\cup_{i \in \tilde{M}} \tilde{\mathcal{Z}}_i\}$, s.t. $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are approximated by $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}, \cup_{i \in \tilde{M}} \tilde{\mathcal{Z}}_i$. Given a sample $S \sim \mathcal{D}$, using the algorithm then retuns $\tilde{f}(S) = f_{g(S)}(S)$. We give a formal description below (detailed description is in Appendix D).

Algorithm 1 A learning algorithm for $f \in \mathcal{F}$ which combines classification and linear regression.

Input: samples $\mathcal{S} = \{(S_j, f(S_j))\}_{j \in [m]}$

$(\tilde{\mathcal{Z}}, \tilde{M}) \leftarrow (\emptyset, \emptyset)$

for $i = 1$ **to** n **do**

$\tilde{\mathcal{Z}}_i \leftarrow \{S : a_i \in S, S \notin \tilde{\mathcal{Z}}\}$

$f_{\tilde{\mathcal{Z}}_i} \leftarrow \mathrm{ERM}^{\mathrm{reg}}(\{(S_j, f(S_j)) : S_j \in \tilde{\mathcal{Z}}_i\})$ linear regression

if $\sum_{(S_j, f(S_j)) : S_j \in \tilde{\mathcal{Z}}_i} |f_{\tilde{\mathcal{Z}}_i}(S_j) - f(S_j)| = 0$ **then**

$\tilde{\mathcal{Z}} \leftarrow \tilde{\mathcal{Z}} \cup \tilde{\mathcal{Z}}_i, \tilde{M} \leftarrow \tilde{M} \cup \{a_i\}$

$C \leftarrow \mathrm{ERM}^{\mathrm{cla}}(\{(S_j, f(S_j)) : S_j \notin \tilde{\mathcal{Z}}, j \le m/2\})$ train a classifier for regions $\mathcal{X}, \mathcal{Y}$

$(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) \leftarrow (\{S : S \notin \tilde{\mathcal{Z}}, C(S) = 1\}, \{S : S \notin \tilde{\mathcal{Z}}, C(S) = -1\})$

return $\tilde{f} \leftarrow S \mapsto \begin{cases} |S|/(2\sqrt{n}) & \text{if } S \in \tilde{\mathcal{X}} \\ f_{\tilde{\mathcal{Y}}}(S) = \mathrm{ERM}^{\mathrm{reg}}(\{(S_j, f(S_j)) : S_j \in \tilde{\mathcal{Y}}, j > m/2\}) & \text{if } S \in \tilde{\mathcal{Y}} \\ f_{\tilde{\mathcal{Z}}_i}(S) : i = \min(\{i' : a_{i'} \in S \cap \tilde{M}\}) & \text{if } S \in \tilde{\mathcal{Z}} \end{cases}$

Overview of analysis of the learning algorithm. There are two main challenges in training the algorithm. The first is that the region $\mathcal{X}$, $\mathcal{Y}$, or $\mathcal{Z}$ that a sample $(S_j, f(S_j))$ belongs to is not known. Thus, even before being able to train a classifier which learns the regions $\tilde{\mathcal{X}}$, $\tilde{\mathcal{Y}}$, $\tilde{\mathcal{Z}}$ using the samples, we need to learn the region a sample S_j belongs to using $f(S_j)$. The second is that the samples $\mathcal{S}_{\mathcal{R}}$ used for training a linear regression predictor $f_{\mathcal{R}}$ over region $\mathcal{R}$ need to be carefully selected so that $\mathcal{S}_{\mathcal{R}}$ is a collection of i.i.d. samples from the distribution $S \sim \mathcal{D}$ conditioned on $S \in \mathcal{R}$ (Lemma 20).

We first discuss the challenge of labeling samples with the region they belong to. Observe that for a fixed masking element $a_i \in M$, $f \in \mathcal{F}$ is linear over all sets S containing a_i since these sets are all in region $\mathcal{Z}$. Thus, there must exist a linear regression predictor $f_{\tilde{\mathcal{Z}}_i} = \mathrm{ERM}^{\mathrm{reg}}(\cdot)$ with zero empirical loss over all samples S_j containing a_i if $a_i \in M$ (and thus $S_j \in \mathcal{Z}$). $\mathrm{ERM}^{\mathrm{reg}}(\cdot)$ minimizes the empirical loss on the input samples over the class of linear regression predictors with bounded norm

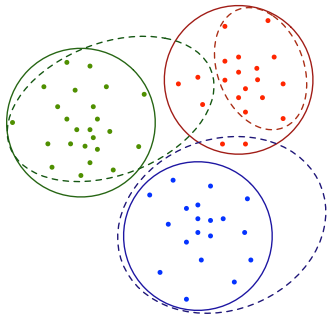


Figure 2: An illustration of the regions. The dots represent the samples, the corresponding full circles represent the regions $\mathcal{X}$ (red), $\mathcal{Y}$ (blue), and $\mathcal{Z}$ (green). The ellipsoids represent the regions $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}, \tilde{\mathcal{Z}}$ learned by the classifier. Notice that $\tilde{\mathcal{Z}}$ has no false negatives.

(Lemma 19). If $f_{\tilde{\mathcal{Z}}_i}$ has zero empirical loss, we directly classify any set S containing a_i as being in $\tilde{\mathcal{Z}}$. Next, for a sample $(S_j, f(S_j))$ not in $\tilde{\mathcal{Z}}$, we can label these samples since $S_j \in \mathcal{X}$ if and only if $f(S_j) = |S_j|/(2\sqrt{n})$. With these labeled samples $\mathcal{S}'$, we train a binary classifier $C = \text{ERM}^{\text{cla}}(\mathcal{S}')$ that indicates if S s.t. $S \notin \tilde{\mathcal{Z}}$ is in region $\tilde{\mathcal{X}}$ or $\tilde{\mathcal{Y}}$. $\text{ERM}^{\text{cla}}(\mathcal{S}')$ minimizes the empirical loss on labeled samples $\mathcal{S}'$ over the class of halfspaces $\mathbf{w} \in \mathbb{R}^n$ (Lemma 23).

Regarding the second challenge, we cannot use all samples S_j s.t. $S_j \in \tilde{\mathcal{Y}}$ to train a linear predictor $f_{\tilde{\mathcal{Y}}}$ for region $\tilde{\mathcal{Y}}$ since these same samples were used to define $\tilde{\mathcal{Y}}$, so they are not a collection of i.i.d. samples from the distribution $S \sim \mathcal{D}$ conditioned on $S \in \tilde{\mathcal{Y}}$. To get around this issue, we partition the samples into two distinct collections, one to train the classifier C and one to train $f_{\tilde{\mathcal{Y}}}$ (Lemma 24). Next, given $T \in \tilde{\mathcal{Z}}$, we predict $f_{\tilde{\mathcal{Z}}_i}(T)$ where i is s.t. $a_i \in T \cap \tilde{M}$ (breaking ties lexicographically) which performs well since $\tilde{f}_{\tilde{\mathcal{Z}}_i}$ has zero empirical error for $a_i \in \tilde{M}$ (Lemma 22). Since we break ties lexicographically, $\tilde{f}_{\tilde{\mathcal{Z}}_i}$ must be trained over samples S_j such that $a_i \in S_j$ and $a_{i'} \notin S_j$ for i' s.t. $i' < i$ and $a_{i'} \in \tilde{M}$ to obtain i.i.d. samples from the same distribution as $T \sim \mathcal{D}$ conditioned on T being directed to $\tilde{f}_{\tilde{\mathcal{Z}}_i}$ (Lemma 21).

The analysis of the learning algorithm leads to the following main learning result.

Lemma 7. *Let $\tilde{f}$ be the predictor returned by Algorithm 1, then w.p. $1 - \delta$ over $m \in O(n^3 + n^2(\log(2n/\delta))/\epsilon^2)$ samples $\mathcal{S}$ drawn i.i.d. from any distribution $\mathcal{D}$, $\mathbb{E}_{S\sim\mathcal{D}}[|\tilde{f}(S) - f(S)|] \leq \epsilon$.*

3.4 Main Result

We conclude this section with our main result which combines Lemmas 6 and 7.

Theorem 8. *There exists a family of $[0, 1]$-bounded submodular functions $\mathcal{F}$ that is efficiently PAC-learnable and that cannot be optimized from polynomially many samples drawn from any distribution $\mathcal{D}$ within a $1/2 - o(1)$ additive approximation for unconstrained submodular minimization.*

4 Discussion

In this paper, we studied the problem of submodular minimization from samples. Our main result is an impossibility, showing that even for learnable submodular functions it is impossible to find a non-trivial approximation to the minimizer with polynomially-many samples, drawn from any distribution. In particular, this implies that minimizing a general submodular function learned from data cannot yield desirable guarantees. In general, it seems that the intersection between learning and optimization is elusive, and a great deal still remains to be explored.

References

[Bal15] Maria-Florina Balcan. Learning submodular functions with applications to multi-agent systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, page 3, 2015.

[BH11] Maria-Florina Balcan and Nicholas JA Harvey. Learning submodular functions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 793–802. ACM, 2011.

[BRS16] Eric Balkanski, Aviad Rubinstein, and Yaron Singer. The power of optimization from samples. In *Advances in Neural Information Processing Systems*, pages 4017–4025, 2016.

[BRS17] Eric Balkanski, Aviad Rubinstein, and Yaron Singer. The limitations of optimization from samples. *Proceedings of the Forty-Ninth Annual ACM on Symposium on Theory of Computing*, 2017.

[BS17] Eric Balkanski and Yaron Singer. The sample complexity of optimizing a convex function. In *COLT*, 2017.

[BVW16] Maria-Florina Balcan, Ellen Vitercik, and Colin White. Learning combinatorial functions from pairwise comparisons. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 310–335, 2016.

[CLSW16] Deeparnab Chakrabarty, Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. Subquadratic submodular function minimization. *arXiv preprint arXiv:1610.09800*, 2016.

[DTK16] Josip Djolonga, Sebastian Tschiatschek, and Andreas Krause. Variational inference in mixed probabilistic submodular models. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1759–1767, 2016.

[EN15] Alina Ene and Huy L. Nguyen. Random coordinate descent methods for minimizing decomposable submodular functions. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 787–795, 2015.

[FK14] Vitaly Feldman and Pravesh Kothari. Learning coverage functions and private release of marginals. In *COLT*, pages 679–702, 2014.

[FKV13] Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. Representation, approximation and learning of submodular functions using low-rank decision trees. In *COLT*, pages 711–740, 2013.

[FMV11] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM J. Comput.*, 40(4):1133–1153, 2011.

[GLS81] Martin Grotschel, Laszlo Lovasz, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[IFF01] Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *J. ACM*, 48(4):761–777, 2001.

[IJB13] Rishabh K. Iyer, Stefanie Jegelka, and Jeff A. Bilmes. Curvature and optimal algorithms for learning and minimizing submodular functions. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2742–2750, 2013.

[JB11] Stefanie Jegelka and Jeff Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1897–1904. IEEE, 2011.

[JLB11] Stefanie Jegelka, Hui Lin, and Jeff A Bilmes. On fast approximate submodular minimization. In *Advances in Neural Information Processing Systems*, pages 460–468, 2011.

[NB12] Mukund Narasimhan and Jeff A Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. *arXiv preprint arXiv:1207.1404*, 2012.

[SK10] Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2010.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

A Optimization from Samples via Learning then Optimization

In this section, we discuss the issues with the approach which consists of first learning a surrogate of the true function and then optimizing this surrogate. A first issue is that the surrogate may be non-submodular, and even though we might be able to approximate the true function everywhere, the surrogate may be intractable to optimize (Section A.1). A second issue is that the surrogate is submodular and approximates the true function on all the samples but that its minimum is far from the minimum of the true function (Section A.2).

A.1 The surrogate is a good approximation of the true function but is not submodular

Consider the following submodular function defined over a partition of the ground set N into a set of good elements G and a set of bad elements B, each of size $n/2$:

$$f(S) = \frac{1}{2} + \frac{1}{n}\left(|S \cap B| - |S \cap G|\right).$$

Let $\tilde{f}$ be a surrogate for f defined as follows:

$$\tilde{f}(S) = \frac{1}{2} + \begin{cases} \frac{1}{n}\left(|S \cap B| - |S \cap G|\right) & \text{if } ||S \cap B| - |S \cap G|| \geq \frac{1}{2}\epsilon n \\ 0 & \text{otherwise} \end{cases}$$

It is easy to verify that the surrogate $\tilde{f}$ ϵ-approximates f on all sets. However $\tilde{f}$ is intractable to optimize. Intuitively, by concentration bounds, a query S of the algorithm has value $\tilde{f}(S) = 1/2$ for an exponentially high fraction of the sets S. Thus, with polynomially many adaptive queries $(S, f(S))$ of the choice of the algorithm, the algorithm will probably not be able to distinguish $\tilde{f}(S)$ from the constant function $1/2$ everywhere. Since the optimal solution is $S^\star = G$ and has value 0, no algorithm can then do better than the trivial $1/2$-approximation to minimize $\tilde{f}$.

A.2 The surrogate is submodular but its minimum is far from the true minimum

We illustrate the second issue with samples from the uniform distribution. Similarly, consider the following submodular function defined over a partition of the ground set N into a set of good elements G and a set of bad elements B, each of size $n/2$

$$f(S) := \begin{cases} \frac{1}{2} + \frac{1}{2n} \cdot \left(|S \cap G| + |S \cap B|\right) & \text{if } |S \cap G| < n/2 \\ \frac{1}{2n} \cdot |S \cap B| & \text{if } |S \cap G| = n/2 \end{cases}$$

This function is similar to the functions constructed in Section 2. Let $\tilde{f}$ be a surrogate for f where G and B are interchanged:

$$\tilde{f}(S) := \begin{cases} \frac{1}{2} + \frac{1}{2n} \cdot \left(|S \cap B| + |S \cap G|\right) & \text{if } |S \cap B| < n/2 \\ \frac{1}{2n} \cdot |S \cap G| & \text{if } |S \cap B| = n/2 \end{cases}$$

It is easy to verify that given polynomially many samples S from the uniform distribution, $\tilde{f}(S) = f(S)$ with high probability, so $\tilde{f}$ is consistent with all the samples. However its minimum B, which is such that $\tilde{f}(B) = 0$ is a bad solution for the true underlying function since $f(B) = 3/4$.

B Concentration bounds

Lemma 9 (Chernoff Bound)**.** *Let $X_1, \ldots, X_n$ be independent indicator random variables. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. For $0 < \delta < 1$,*

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}.$$

B.1 Concentration bounds from Section 2

- Every set S_i in the sample respects $|S_i| \leq 3n/4$, w.p. $1-e^{-\Omega(n)}$. Consider a set $S_i \sim U$. Let X_j be an indicator variable indicating if $j \in S_i$. By Chernoff bound with $|S| = \sum_{j=1}^{n} X_j$, $\mu = n/2$, $\delta = 1/2$, $\Pr[|S_i| - n/2 \geq n/4] \leq 2e^{-n/24}$. By a union bound over all polynomially many samples, the claim holds with probability $1 - e^{-\Omega(n)}$.
- For sets $S_1, \ldots, S_m$, $m = \text{poly}(n)$, all of size at most $3n/4$, when f_j is drawn u.a.r. from $\mathcal{F}$ we get that $|S_i \cap G_j| < \sqrt{n}$, w.p. $1 - e^{-\Omega(n^{1/2})}$. Consider a set T of size $n/4$. Let $X_1, \ldots, X_{n/4}$ be indicator variables indicating if elements in T are also in G_j. By Chernoff bound with $|T \cap G_j| = \sum_{i=1}^{n/4} X_i$, $\mu = n/4 \cdot \sqrt{n}/n = \sqrt{n}/4$, $\delta = 1/2$, $\Pr[|T \cap G_j| - \sqrt{n}/4 \geq \sqrt{n}/8] \leq 2e^{-n^{1/2}/48}$. Thus, with probability $1 - e^{-\Omega(n^{1/2})}$, $|T \cap G_j| > 0$, and for a set S of size at most $3n/4$, $|(N \setminus S) \cap G_j| > 0$ and $G_j \not\subseteq S$. Thus, by a union bound, it holds with probability $1 - e^{-\Omega(n^{1/2})}$ over f_j that for all samples S, $G_j \not\subseteq S$.
- Fix S and choose f u.a.r. from $\mathcal{F}$, then, w.p. $1 - e^{-\Omega(n^{1/6})}$:

$$f(S) \geq \frac{1}{2} - o(1).$$

  If $|S| \leq n - n^{2/3}$, let $T \subseteq N \setminus S$ such that $|T| = n^{2/3}$. Let $X_1, \ldots, X_{n^{2/3}}$ be indicator variables indicating if elements in T are also in G. By Chernoff bound with $|T \cap G| = \sum_{i=1}^{n^{2/3}} X_i$, $\mu = n^{2/3} \cdot \sqrt{n}/n = n^{1/6}$, $\delta = 1/2$, $\Pr[|T \cap G| - n^{1/6} \geq n^{1/6}/2] \leq 2e^{-n^{1/6}/12}$. Thus, with probability $1 - e^{-\Omega(n^{1/6})}$, $|T \cap G| > 0$ and $|S \cap G| < \sqrt{n}$. Thus, $f(S) \geq 1/2$. If $|S| > n - n^{2/3}$, then $f(S) \geq (n - 2n^{2/3})/(2n) = n(1 - o(1))/(2n(1 - o(1))) = (1 - o(1))/2$.

B.2 Concentration bounds from Section 3

- For all $S \subseteq N$ such that $|S| \geq n^{1/4}$, with probability $1 - e^{-\Omega(n^{1/4})}$ over $\mathcal{F}^M$, we have $|M \cap S| \geq 1$. Let X_i, for $i \in S$, indicate if $i \in M$. With $\mu = |S|(n/2 - \sqrt{n})/n$ and $\delta = 1/2$,

$$\Pr\left[\left||M \cap S| - |S| \cdot \frac{n/2 - \sqrt{n}}{n}\right| \geq |S| \cdot \frac{n/2 - \sqrt{n}}{2n}\right] \leq 2e^{-n^{1/4} \cdot \frac{n/2 - \sqrt{n}}{12n}} = e^{-\Omega(n^{1/4})}.$$

  Thus, $|M \cap S| \geq 1$ with probability at least $1 - e^{-\Omega(n^{1/4})}$.
- Let $S \subseteq N$ and $f_i \in \mathcal{F}^M$ such that $m = 0$, $|S| \geq n^{3/4}$, and S independent of i, then

$$(1 - o(1)) \cdot |S| \leq \frac{n/2 + \sqrt{n}}{\sqrt{n}} \cdot b, \frac{n/2 + \sqrt{n}}{n/2} \cdot g \leq (1 + o(1)) \cdot |S|.$$

  Let X_j, for $j \in S$, indicate if $j \in B_i$. With $\mu = |S|(\sqrt{n}/(n/2 + \sqrt{n}))$ and $\delta = n^{-1/16}$,

$$\Pr\left[\left|g - \frac{\sqrt{n}}{n/2 + \sqrt{n}} \cdot |S|\right| \geq n^{-1/16} \cdot \frac{\sqrt{n}}{n/2 + \sqrt{n}} \cdot |S|\right] \leq 2e^{-n^{-1/8} \cdot \frac{\sqrt{n}}{3(n/2 + \sqrt{n})} \cdot n^{7/8}} = e^{-\Omega(n^{1/4})}$$

  Similarly,

$$\Pr\left[\left|g - \frac{n/2}{n/2 + \sqrt{n}} \cdot |S|\right| \geq n^{-1/16} \cdot \frac{n/2}{n/2 + \sqrt{n}} \cdot |S|\right] \leq 2e^{-n^{-1/8} \cdot \frac{n/2}{3(n/2 + \sqrt{n})} \cdot n^{7/8}} = e^{-\Omega(n^{3/4})}$$

  and we obtain the desired result.

C Missing analysis from Sections 3.1 and 3.2

Lemma 10. *For all $f \in \mathcal{F}$, the function f is submodular.*

Proof. We show that the marginal contribution $f_S(a) := f(S \cup \{a\}) - f(S)$ of an element $a \in N$ is decreasing.

$$a \in B, f_S(a) = \begin{cases} \frac{1}{2\sqrt{n}} & \text{if } m = 0 \\ -\frac{2}{n} & \text{otherwise} \end{cases}$$

$$a \in G, f_S(a) = \begin{cases} \frac{1}{2\sqrt{n}} & \text{if } m = 0 \text{ and } g \leq n^{1/4} \\ -\frac{2}{n} & \text{otherwise} \end{cases}$$

$$\begin{aligned} a \in M \text{ and } m = 0, f_S(a) &= \frac{1}{2} - \frac{b}{n} - \frac{b}{2\sqrt{n}} - \frac{g}{n} - \begin{cases} \frac{g}{2\sqrt{n}} & \text{if } g < n^{1/4} \\ \frac{n^{1/4}}{2\sqrt{n}} - \frac{g-n^{1/4}}{n} & \text{if } g \geq n^{1/4} \end{cases} \\ &= \frac{1}{2} - \frac{b}{n} - \frac{b}{2\sqrt{n}} - \begin{cases} \frac{g}{n} + \frac{g}{2\sqrt{n}} & \text{if } g < n^{1/4} \\ \frac{g-n^{1/4}}{n} + \frac{n^{1/4}}{n} + \frac{n^{1/4}}{2\sqrt{n}} - \frac{g-n^{1/4}}{n} & \text{if } g \geq n^{1/4} \end{cases} \\ &= \frac{1}{2} - \frac{b}{n} - \frac{b}{2\sqrt{n}} - \begin{cases} \frac{g}{n} + \frac{g}{2\sqrt{n}} & \text{if } g < n^{1/4} \\ \frac{n^{1/4}}{n} + \frac{n^{1/4}}{2\sqrt{n}} & \text{if } g \geq n^{1/4} \end{cases} \end{aligned}$$

$$a \in M \text{ and } m > 0, f_S(a) = 0$$

For $a \in G$ or $a \in B$ it is immediate that these marginal contributions are decreasing. For $a \in M$, note that

$$\frac{1}{2} - \frac{b}{n} - \frac{b}{2\sqrt{n}} - \begin{cases} \frac{g}{n} + \frac{g}{2\sqrt{n}} & \text{if } g < n^{1/4} \\ \frac{n^{1/4}}{n} + \frac{n^{1/4}}{2\sqrt{n}} & \text{if } g \geq n^{1/4} \end{cases} \geq \frac{1}{2} - \frac{\sqrt{n}}{n} - \frac{\sqrt{n}}{2\sqrt{n}} - \frac{n^{1/4}}{n} + \frac{n^{1/4}}{2\sqrt{n}} \geq 0$$

so $f_S(a)$ is also decreasing. □

Lemma 3. *Let $\mathcal{F}$ be a family of functions and $\mathcal{F}' = \{f_1, \ldots, f_p\} \subseteq \mathcal{F}$ be a subfamily of functions drawn from some distribution. Assume the following two conditions hold:*

1. ***Indistinguishability.*** *For all $S \subseteq N$, w.p. $1 - e^{-\Omega(n^{1/4})}$ over $\mathcal{F}'$: for every $f_i, f_j \in \mathcal{F}'$,*
$$f_i(S) = f_j(S);$$
2. ***α-gap.*** *Let $S_i^\star$ be a minimizer of f_i, then w.p. 1 over $\mathcal{F}'$: for all $S \subseteq N$,*
$$\mathop{\mathbb{E}}_{f_i \sim U(\mathcal{F}')} \left[f_i(S) - f_i\left(S_i^\star\right) \right] \geq \alpha;$$

Then, $\mathcal{F}$ is not α-minimizable from strictly less than $e^{\Omega(n^{1/4})}$ samples over any distribution $\mathcal{D}$.

Proof. We first claim that for any distribution $\mathcal{D}$, there exists a family of functions $F \subseteq \mathcal{F}$ such that with probability $1 - e^{-\Omega(n^{1/4})}$ over $S \sim \mathcal{D}$, $f_i(S) = f_j(S)$ for all $f_i, f_j \in F$. Let $I(\mathcal{F}', S)$ be the event that $f_i(S) = f_j(S)$ for all $f_i, f_j \in \mathcal{F}'$. Then,

$$\begin{aligned} \sum_{F \subseteq \mathcal{F}} \Pr[\mathcal{F}' = F] \Pr_{S \sim \mathcal{D}}[I(F, S)] &= \sum_{F \subseteq \mathcal{F}} \Pr[\mathcal{F}' = F] \sum_{S \subseteq N} \mathbb{1}_{I(F,S)} \Pr[S \sim \mathcal{D}] \\ &= \sum_{S \subseteq N} \Pr[S \sim \mathcal{D}] \sum_{F \subseteq \mathcal{F}} \Pr[\mathcal{F}' = F] \mathbb{1}_{I(F,S)} \\ &= \sum_{S \subseteq N} \Pr[S \sim \mathcal{D}] \Pr_{\mathcal{F}'}[I(\mathcal{F}', S)] \\ &\geq \min_{S \subseteq N} \Pr_{\mathcal{F}'}[I(\mathcal{F}', S)]. \end{aligned}$$

Thus, there exists some $F = \{f_1, \ldots, f_p\}$ such that

$$\Pr_{S \sim \mathcal{D}}[I(F, S)] \geq \min_{S \subseteq N} \Pr_{\mathcal{F}'}[I(\mathcal{F}', S)].$$

Since $\min_{S \subseteq N} \Pr_{\mathcal{F}'}[I(\mathcal{F}', S)] = 1 - e^{-\Omega(n^{1/4})}$. Then, by a union bound over the samples, $f_i(S) = f_j(S)$ for all $f_i, f_j \in F$ and for all samples S with probability $1 - e^{-\Omega(n^{1/4})}$, and we assume this is the case, as well as the gap condition, for the remaining of the proof.

It follows that the choices of the algorithm given samples from f_i, $i \in [p]$, are independent of i. Pick $i \in [p]$ uniformly at random and consider the (possibly randomized) vector S returned by the algorithm. Since S is independent of i and by the α-gap condition,

$$\mathop{\mathbb{E}}_{f_i \sim U(\mathcal{F}')}[f_i(S) - f_i(S_i^\star)] \geq \alpha.$$

Thus, there exists at least one $f_i \in F$ such that for f_i, the algorithm is at least an additive factor α away from $f_i(S_i^\star)$. □

D Missing analysis from Section 3.3

We begin by reviewing known results for classification and linear regression using the VC-dimension and the Rademacher complexity (Section D.1) and then give the missing analysis for the learning algorithm (Section D.2).

D.1 VC-dimension and Rademacher complexities essentials

We review learning results needed for the analysis. These results use the VC-dimension and the Rademacher complexity, two of the most common tools to bound the generalization error of a learning algorithm. We formally define the VC-dimension and the Rademacher complexity using definitions from [SSBD14]. We begin with the VC-dimension, which is for classes of binary functions. We first define the concepts of restriction to a set and of shattering, which are useful to define the VC-dimension.

Definition 11. *(Restriction of $\mathcal{H}$ to A). Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $A = \{a_1, \ldots, a_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to A is the set of functions from A to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,*

$$\mathcal{H}_A = \{(h(a_1), \ldots, h(a_m)) \,:\, h \in \mathcal{H}\},$$

where we represent each function from A to $\{0,1\}$ as a vector in $\{0,1\}^{|A|}$.

Definition 12. *(Shattering). A hypothesis class $\mathcal{H}$ shatters a finite set $A \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to A is the set of all functions from A to $\{0,1\}$. That is, $|\mathcal{H}_A| = 2^{|A|}$.*

Definition 13. *(VC-dimension). The VC-dimension of a hypothesis class $\mathcal{H}$ is the maximal size of a set $S \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.*

Next, we define the Rademacher complexity, which is for more complex classes of functions than binary functions, such as real-valued functions.

Definition 14. *(Rademacher complexity). Let σ be distributed i.i.d. according to $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = 1/2$. The Rademacher complexity $R(A)$ of set of vectors $A \subset \mathbb{R}^m$ is defined as*

$$R(A) := \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \right].$$

This first result bounds the generalization error of a class of binary functions in terms of the VC-dimension of these classifiers.

Theorem 15 ([SSBD14])**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{-1,1\}$ and $f : \mathcal{X} \mapsto \{-1,1\}$ be some "correct" function. Assume that the VC-dimension of $\mathcal{H}$ is d. Then, there is an absolute constant C such that with $m \geq C(d + \log(1/\delta))/\epsilon^2$ i.i.d. samples $\mathbf{x}^1, \ldots, \mathbf{x}^m \sim \mathcal{D}$,*

$$\left| \Pr_{S \sim \mathcal{D}}[h(S) \neq f(S)] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h(S_i) \neq f(S_i)} \right| \leq \epsilon$$

for all $h \in \mathcal{H}$, with probability $1 - \delta$ over the samples.

We use the class of halfspaces for classification, for which we know the VC-dimension.

Theorem 16 ([SSBD14]). *Let $b \in \mathbb{R}$. The class of functions $\{\mathbf{x} \mapsto \text{sign}(\mathbf{w}^\intercal \mathbf{x}) + b : \mathbf{w} \in \mathbb{R}^n\}$ has VC dimension n.*

The following result combines the generalization error of a class of functions in terms of its Rademacher complexity with the Rademacher complexity of linear functions over a ρ-Lipschitz loss function.

Theorem 17 ([SSBD14]). *Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X}$ such that with probability 1 over $\mathbf{x} \sim \mathcal{D}$ we have that $\|\mathbf{x}\|_\infty \leq R$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 \leq B\}$ and let $\ell(\mathbf{w}, (\mathbf{x}, y)) = \phi(\mathbf{w}^\intercal \mathbf{x}, y)$ such that for all $y \in \mathcal{Y}$, $a \mapsto \phi(a, y)$ is an ρ-Lipschitz function and such that $\max_{a \in [-BR, BR]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample of size m,*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} [|\ell(\mathbf{w}, (\mathbf{x}, y))|] \leq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} |\ell(\mathbf{w}, (\mathbf{x}, y))| + 2\rho BR \sqrt{\frac{2 \log(2d)}{m}} + c \sqrt{\frac{2 \log(2/\delta)}{m}}$$

for all $\mathbf{w} \in \mathcal{H}$.

D.2 Missing analysis for the learning algorithm

We formally define PAC learnability with absolute loss.

Definition 18 (PAC learnability with absolute loss). *A class of functions $\mathcal{F}$ is PAC learnable if there exist a function $m_{\mathcal{F}} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$, and every function $f \in \mathcal{F}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. samples $(S, f(S))$ with $S \sim \mathcal{D}$, the algorithm returns a function $\tilde{f}$ such that, with probability at least $1 - \delta$ (over the choice of the m training examples),*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}} \left[\left| \tilde{f}(S) - f(S) \right| \right] \leq \epsilon.$$

For the remaining of this section, let $f \in \mathcal{F}$ be the function defined over partition (G, B, M) for which the learning algorithm is given samples $\mathcal{S}$. Let $\mathbf{x}_S$ denote the $0 - 1$ vector corresponding to the set S, i.e., $x_i = \mathbb{1}_{i \in S}$. We define the following subcollection of samples, $\mathcal{S}_{\tilde{\mathcal{Z}}_i} := \{(S, f(S)) \in \mathcal{S} : S \in \tilde{\mathcal{Z}}_i\}$, $\mathcal{S}^{\leq}_{\tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}} := \{(S_j, f(S_j)) \in \mathcal{S} : S_j \notin \tilde{\mathcal{Z}}, j \leq m/2\}$, $\mathcal{S}^{>}_{\tilde{\mathcal{Y}}} := \{(S_j, f(S_j)) \in \mathcal{S} : S_j \in \tilde{\mathcal{Y}}, j > m/2\}$. Let $\mathcal{R}$ be a region of sets, we define $\mathcal{D}_{\mathcal{R}}$ to be the distribution $S \sim \mathcal{D}$ conditioned on $S \in \mathcal{R}$. The linear regression predictors $f_{\tilde{\mathcal{Z}}_i}$ and $f_{\tilde{\mathcal{Y}}}$ and the classifier C are formally defined as follows.

$$\begin{aligned}
\tilde{\mathbf{w}}_i &:= \mathop{\text{argmin}}_{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\|_1 \leq 1} \sum_{(S_j, f(S_j)) \in \mathcal{S}_{\tilde{\mathcal{Z}}_i}} |\mathbf{w}^\intercal \mathbf{x}_S + 1 - f(S)| \\
\tilde{\mathbf{w}}_C &:= \mathop{\text{argmin}}_{\mathbf{w} \in \mathbb{R}^n} \sum_{(S_j, f(S_j)) \in \mathcal{S}^{\leq}_{\tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}}} \mathbb{1}_{\text{sign}\left(\mathbf{w}^\intercal \mathbf{x}_S + n^{1/4}\right) = \text{sign}(f(S) - |S|/(2\sqrt{n}))} \\
\tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}} &:= \mathop{\text{argmin}}_{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\|_1 \leq 1} \sum_{(S_j, f(S_j)) \in \mathcal{S}^{>}_{\tilde{\mathcal{Y}}}} |\mathbf{w}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)|
\end{aligned}$$

where $b_{\mathcal{Y}}$ is the constant $b_{\mathcal{Y}} = 1/2 + 1/(2n^{1/4}) + 1/n^{3/4}$. Then,

$$\begin{aligned}
f_{\tilde{\mathcal{Z}}_i}(S) &:= \tilde{\mathbf{w}}_i^\intercal \cdot \mathbf{x}_S + 1 \\
C(S) &:= \text{sign}\left(\tilde{\mathbf{w}}_C^\intercal \mathbf{x}_S + n^{1/4}\right) \\
f_{\tilde{\mathcal{Y}}}(S) &:= \tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}}^\intercal \cdot \mathbf{x}_S + b_{\mathcal{Y}}
\end{aligned}$$

We show that there are no false negatives for $\tilde{Z}$, which is important for the existence of a linear classifier C with zero empirical error on samples $\mathcal{S}^{\leq}_{\tilde{X} \cup \tilde{Y}}$.

Lemma 19. *If $S \notin \tilde{\mathcal{Z}}$, then $S \in \mathcal{X} \cup \mathcal{Y}$.*

Proof. The proof is by contrapositive. If $S \notin \mathcal{X} \cup \mathcal{Y}$, then there exists $i \in M$ such that $i \in S$. Then note that $f(S)$ is an affine function over all S such that $i \in S$. So it must be the case that the empirical minimizer $\tilde{\mathbf{w}}_i$ computed has zero empirical loss. Thus $i \in \tilde{M}$. □

We give a general lemma for the expected error of a linear regression predictor.

Lemma 20. *Assume $\mathcal{S}'$ is a collection of $m \geq \epsilon^{-2}(\log(2n) + \log(2/\delta))/2$ i.i.d. samples from some distribution $\mathcal{D}'$. Then, with probability at least $1 - \delta$,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}'} [|\mathbf{w}^\intercal \mathbf{x}_S + b - f(S)|] \leq \frac{1}{|\mathcal{S}'|} \sum_{S \in \mathcal{S}'} |\mathbf{w}^\intercal \mathbf{x}_S + b - f(S)| + \epsilon$$

for all $\mathbf{w} \in \mathbb{R}^n$ such that $||\mathbf{w}||_1 \leq 1$.

Proof. Consider the setting of Theorem 17 with the absolute loss, so $\ell(\mathbf{w}, (\mathbf{x}_S, f(S)) = \phi(\mathbf{w}^\intercal \mathbf{x}_S, f(S)) = |\mathbf{w}^\intercal \mathbf{x}_S + b - f(S)|$ for some $b \in \mathbb{R}$. Notice that the absolute loss is 1-Lipschitz, so $\rho = 1$. We also have $d = n$ and $\|\mathbf{x}_S\|_\infty \leq 1 = R$ for all S. We consider $\|\mathbf{w}\|_1 \leq 1 = B$. Such B and R imply that $c = 2$ satisfies the condition of Theorem 17. Thus, if $\mathcal{S}'$ is a set of $m \geq \epsilon^{-2}(\log(2n) + \log(2/\delta))/2$ i.i.d. samples from a distribution $\mathcal{D}'$, then with probability at least $1 - \delta$,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}'} [|\tilde{\mathbf{w}}^\intercal \mathbf{x}_S + b - f(S)|] \leq \frac{1}{|\mathcal{S}'|} \sum_{S \in \mathcal{S}'} \left|\tilde{\mathbf{w}}_{\mathcal{Y}}^\intercal \mathbf{x}_S + b - f(S)\right| + \epsilon$$

for all $\mathbf{w} \in \mathbb{R}^n$ such that $||\mathbf{w}||_1 \leq 1$. □

We show that if $S \in \tilde{Z}_i$ with $i \in \tilde{M}$, then the learner $\tilde{f}$ directs S to $f_{\tilde{\mathcal{Z}}_i}$.

Lemma 21. *Assume $S \in \tilde{\mathcal{Z}}_i$, with $i \in \tilde{M}$. Then, $\tilde{f}(S) = f_{\tilde{\mathcal{Z}}_i}(S)$*

Proof. We argue that $i = \min(\{i' : i \in S \cap \tilde{M}\})$. Assume by contradiction that there exists $i' < i$ such that $i' \in S \cap \tilde{M}$. Then it must be the case that $S \in \tilde{\mathcal{Z}}$ before iteration i of the algorithm. But then, S would not have been considered for $\tilde{\mathcal{Z}}_i$. □

The following lemma bounds the error of linear regression predictors $f_{\tilde{\mathcal{Z}}_i}$ with $i \in \tilde{M}$.

Lemma 22. *Assume $|\mathcal{S}_{\tilde{\mathcal{Z}}_i}| \geq \epsilon^{-2}(\log(2n) + \log(2/\delta))/2$ and $i \in \tilde{M}$. Then with probability $1 - \delta$ over $\mathcal{S}_{\tilde{\mathcal{Z}}_i}$,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\tilde{\mathcal{Z}}_i}} \left[\left|\tilde{f}(S) - f(S)\right|\right] = \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\tilde{\mathcal{Z}}_i}} [|\tilde{\mathbf{w}}_i^\intercal \mathbf{x}_S + 1 - f(S)|] \leq \epsilon.$$

Proof. By Lemma 21,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\tilde{\mathcal{Z}}_i}} \left[\left|\tilde{f}(S) - f(S)\right|\right] = \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\tilde{\mathcal{Z}}_i}} [|\tilde{\mathbf{w}}_i^\intercal \mathbf{x}_S + 1 - f(S)|]$$

Let $i \in \tilde{M}$, then, the empirical loss of $\tilde{\mathbf{w}}_i$ is zero, i.e., $\sum_{S \in \mathcal{S}_i} |\tilde{\mathbf{w}}_i^\intercal \mathbf{x}_S + 1 - f(S)| = 0$ by Algorithm 1 since $i \in \tilde{M}$. The collection of samples $\mathcal{S}_{\tilde{\mathcal{Z}}_i}$ consists of m i.i.d. samples S from $\mathcal{D}_{\tilde{\mathcal{Z}}_i}$, so by Lemma 20,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\tilde{\mathcal{Z}}_i}} [|\tilde{\mathbf{w}}_i^\intercal \mathbf{x}_S + 1 - f(S)|] \leq \epsilon.$$

□

Next, we bound the error of classifier C.

Lemma 23. *Assume $|\mathcal{S}^{\leq}_{\tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}}| \geq C(n + \log(1/\delta))/\epsilon^2$. Then, with probability $1 - \delta$ over $\mathcal{S}^{\leq}_{\tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}}$,*

$$\Pr_{S \sim \mathcal{D}_{\tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}}} \left[\text{sign}\left((\tilde{\mathbf{w}}_C)^\intercal \mathbf{x}_S + n^{1/4}\right) = \text{sign}\left(S \in \mathcal{X}\right)\right] \leq \epsilon.$$

Proof. Note that the support of $\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}$ does not contain any set $S \in \mathcal{Z}$ by Lemma 19. Thus we only consider $S \in \mathcal{X} \cup \mathcal{Y}$ for the remaining of the analysis. Consider the classifier

$$h^\star(S) = \text{sign}\left((\mathbf{w}^\star)^\intercal \mathbf{x}_S + n^{1/4}\right)$$

where $w_i^\star = -1$ if $i \in G$ and $w_i^\star = 0$ otherwise. Then $h^\star(S) = 1$ if $S \in \mathcal{X}$ and $h^\star(S) = -1$ if $S \in \mathcal{Y}$. Since $\mathbf{w}^\star$ has zero empirical error over $\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}$, it must also be the case for the empirical risk minimizer $\tilde{\mathbf{w}}_C$,

$$\sum_{(S,f(S))\in\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}} \mathbb{1}_{\text{sign}\left((\tilde{\mathbf{w}}_C)^\intercal \mathbf{x}_S + n^{1/4}\right)=\text{sign}(S\in\mathcal{X})} = 0.$$

By Theorem 16, the class of functions $\{S \mapsto \text{sign}(\mathbf{w}^\intercal \mathbf{x} + n^{1/4}) : \mathbf{w} \in \mathbb{R}^n\}$ has VC dimension n. Since $\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}$ is a collection of i.i.d. samples S from $\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}$, we conclude by Theorem 15 that with probability $1-\delta$ over $\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}$,

$$\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}}\left[\text{sign}\left((\tilde{\mathbf{w}}_C)^\intercal \mathbf{x}_S + n^{1/4}\right) = \text{sign}\left(S \in \mathcal{X}\right)\right] \leq \epsilon$$

with $|\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}| \geq C(n + \log(1/\delta))/\epsilon^2$.

□

We bound the error of the linear regression predictor $\tilde{\mathbf{w}}_{\tilde{Y}}$. Note the additional $\frac{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}|}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}$ term that is due the predictor being trained over $\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}$ which might contain samples in $\mathcal{X}$ which have been misclassified in $\tilde{\mathcal{Y}}$.

Lemma 24. *Assume $|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}| \geq \epsilon^{-2}(\log(2n) + \log(2/\delta))/2$. Then with probability $1-\delta$ over $\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}$,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}\left[\left|\tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right|\right] \leq \epsilon + \frac{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}|}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}.$$

Proof. Consider the affine function $(\mathbf{w}_{\mathcal{Y}}, b)$ over all sets in region $\mathcal{Y}$. This function has empirical loss:

$$\begin{aligned}
&\frac{1}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}\sum_{(S,f(S))\in\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}}\left|\mathbf{w}_{\mathcal{Y}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right| \\
=&\frac{1}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}\left(\sum_{(S,f(S))\in\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}:S\in\mathcal{X}}\left|\mathbf{w}_{\mathcal{Y}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right| + \sum_{(S,f(S))\in\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}:S\in\mathcal{Y}}\left|\mathbf{w}_{\mathcal{Y}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right|\right) \\
\leq&\frac{1}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}\left(|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}| + 0\right) \\
=&\frac{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}|}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}
\end{aligned}$$

Since $\tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}}$ is the empirical loss minimizer, it has smaller empirical loss than $\mathbf{w}_{\mathcal{Y}}$, so $\sum_{(S,f(S))\in\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}}\left|\tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right| \leq \frac{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}|}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}$.

Since $\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}$ consists of m i.i.d. samples from $\mathcal{D}_{\tilde{\mathcal{Y}}}$ (in particular, none of these samples were used to train classifier C) and by Theorem 20, we conclude that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}\left[\left|\tilde{\mathbf{w}}_{\tilde{\mathcal{Y}}}^\intercal \mathbf{x}_S + b_{\mathcal{Y}} - f(S)\right|\right] \leq \epsilon + \frac{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}\cap\mathcal{X}|}{|\mathcal{S}^{>}_{\tilde{\mathcal{Y}}}|}.$$

□

We show a general lemma to analyze the cases where the linear predictors are trained over a small number of samples.

Lemma 25. *Let $\mathcal{S}$ be the collection of m i.i.d. samples drawn from $\mathcal{D}$. Assume $m \geq \frac{12}{\epsilon}\log(2/\delta)$ and let $\mathcal{S}' \subseteq \mathcal{S}$, then it is either the case that*

$$\Pr_{S\sim\mathcal{D}}[S \in \mathcal{S}'] \leq \epsilon$$

or

$$|\mathcal{S} \cap \mathcal{S}'| \geq \frac{\epsilon m}{2}$$

with probability at least $1-\delta$.

Proof. By Chernoff bound with $\mu = \epsilon m$ and $m \geq \frac{12}{\epsilon}\log(2/\delta)$:

$$\Pr_{\mathcal{S}}[|\mathcal{S} \cap \mathcal{S}'| - \epsilon m| \geq \epsilon m/2] \leq 2e^{-\epsilon m/12} \leq \delta$$

□

We bound the additional $\frac{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}|}{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|}$ term from Lemma 24 in the case where the number of samples $\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}$ is large enough.

Lemma 26. *Assume $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}| \geq 12n^2\log(2n/\delta)/\epsilon^2$, then*

$$\frac{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}|}{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|} \leq \frac{3}{2}\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}] + \frac{3\epsilon}{2n}$$

with probability $1-\delta/n$

Proof. If $\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}] \geq \epsilon/n$, then by the Chernoff bound:

$$\begin{aligned}\Pr\left[\left||\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}| - |\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]\right| \geq \frac{1}{2}\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]\,|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\right] &\leq 2e^{-|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S\in\mathcal{X}]/12} \\ &\leq 2e^{-|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\epsilon/(12n)}\end{aligned}$$

and $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}| \leq \frac{3}{2}|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]$ with probability $1-\delta/n$ since $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}| \geq 12n^2\log(2n/\delta)/\epsilon^2$. Otherwise, by another Chernoff bound:

$$\begin{aligned}&\Pr\left[\left||\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}| - |\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]\right| \geq \frac{\epsilon}{2n}\cdot|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\right] \\ =&\Pr\left[\left||\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}| - |\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]\right| \geq \frac{1}{2n}\frac{\epsilon}{\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S\in\mathcal{X}]}\cdot|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}]\right] \\ \leq&2e^{-|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S\in\mathcal{X}](\frac{1}{2n}\frac{\epsilon}{\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S\in\mathcal{X}]})^2/3} \\ \leq&2e^{-|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\epsilon^2/(12n^2)}\end{aligned}$$

and since $\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}] \leq \epsilon/n$, $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}| \leq 3|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|\epsilon/(2n)$ with probability $1-\delta/n$ since $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}| \geq 12n^2\log(2n/\delta)/\epsilon^2$ □

We are now ready to put all the pieces together and show the main result for the learning algorithm.

Lemma 7. *Let $\tilde{f}$ be the predictor returned by Algorithm 1, then w.p. $1-\delta$ over $m \in O(n^3 + n^2(\log(2n/\delta))/\epsilon^2)$ samples $\mathcal{S}$ drawn i.i.d. from any distribution $\mathcal{D}$, $\mathbb{E}_{S\sim\mathcal{D}}[|\tilde{f}(S) - f(S)|] \leq \epsilon$.*

Proof. Let $\ell(S) = \left|\tilde{f}(S) - f(S)\right|$ be the loss by the algorithm. We divide the analysis in three cases dependent on which region $S \sim \mathcal{D}$ falls into.

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}}[\ell(S)] = \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}}]\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[\ell(S)] + \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}]\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[\ell(S)] + \sum_{i\in\tilde{M}}\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Z}}_i]\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Z}}_i}}[\ell(S)]$$

We analyze each of these three cases separately. Note that since $f(S) \in [0,1]$ and since all the linear regression predictors $\mathbf{w}$ have bounded norm $\|\mathbf{w}\|_1 \leq 1$, $\ell(S) \leq 2$.

If $S \in \tilde{\mathcal{X}}$. Then, $S \notin \mathcal{Z}$ by Lemma 19. Thus it is either the case that $S \in \mathcal{X}$ or $S \in \mathcal{Y}$. We get

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[\ell(S)] &= \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{X}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\mathcal{X}\cap\tilde{\mathcal{X}}}}[\ell(S)] + \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{Y}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cap\mathcal{Y}}}[\ell(S)] \\ &\leq 1 \cdot \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\mathcal{X}\cap\tilde{\mathcal{X}}}}\left[\frac{|S|}{2\sqrt{n}} - \frac{|S|}{2\sqrt{n}}\right] + \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{Y}] \cdot 2 \\ &\leq 2 \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{Y}]\end{aligned}$$

If $S \in \tilde{\mathcal{Y}}$. Then, if $\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \leq \epsilon/(2n)$,

$$\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[\ell(S)] \leq \frac{\epsilon}{n}$$

Otherwise, by Lemma 25, $|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}| \geq \epsilon m/(4n) \geq (n/\epsilon)^2 \log(2n/\delta)$, so by Lemma 24, with probability at least $1 - \delta/n$,

$$\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[\ell(S)] \leq \frac{\epsilon}{n} + \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \left(\frac{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}|}{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|}\right)$$

If $S \in \tilde{\mathcal{Z}}_i$. Then if $\Pr_{S\sim\mathcal{D}}[S \in \mathcal{Z}_i] \leq \epsilon/(4n)$,

$$\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Z}}_i] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i}[\ell(S)] \leq \frac{\epsilon}{2n}.$$

Otherwise, by Lemma 25 $|\mathcal{S}_i| \geq \epsilon m/(8n) \geq (2n/\epsilon)^2 \log(2n/\delta)$, so by Lemma 22, with probability at least $1 - \delta/n$,

$$\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Z}}_i] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i}[\ell(S)] \leq \frac{\epsilon}{2n}.$$

Combining the three cases, and by a union bound over the at most n events with probability at least $1 - \delta/n$, we have that with probaiblity at least $1 - \delta$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{S\sim\mathcal{D}}[\ell(S)] \\ &= \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[\ell(S)] + \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[\ell(S)] + \sum_{i\in\tilde{M}} \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Z}}_i] \mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\tilde{\mathcal{Z}}_i}}[\ell(S)] \\ &\leq 2 \cdot \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}}] \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{Y}] + \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \left[\frac{\epsilon}{n} + \frac{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}} \cap \mathcal{X}|}{|\mathcal{S}^{\geq}_{\tilde{\mathcal{Y}}}|}\right] + n \cdot \frac{\epsilon}{2n} \\ &\leq 2 \cdot \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}}] \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}}}[S \in \mathcal{Y}] + \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{Y}}] \left[\frac{\epsilon}{n} + \frac{3}{2} \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{Y}}}}[S \in \mathcal{X}] + \frac{3\epsilon}{2n}\right] + \frac{\epsilon}{2} \\ &\leq 2 \cdot \Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}] \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}}[S \in (\tilde{\mathcal{X}} \cap \mathcal{Y}) \cup (\tilde{\mathcal{Y}} \cap \mathcal{X})] + \frac{3\epsilon}{4}\end{aligned}$$

where the third inequality is by Lemma 26. If $\Pr_{S\sim D}[S \in \tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}] \leq \frac{\epsilon}{n}$,

$$\Pr_{S\sim\mathcal{D}}[S \in \tilde{\mathcal{X}} \cup \tilde{\mathcal{Y}}] \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}}[S \in (\tilde{\mathcal{X}} \cap \mathcal{Y}) \cup (\tilde{\mathcal{Y}} \cap \mathcal{X})] \leq \frac{\epsilon}{n}.$$

Otherwise, by Lemma 25, $|\mathcal{S}^{\leq}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}| \geq \epsilon m/2n \geq c_1(n^3 + n^2 \log(n/\delta))/\epsilon^2$ with probability $1 - \delta/n$, so by Lemma 23,

$$\Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}}[S \in (\tilde{\mathcal{X}} \cap \mathcal{Y}) \cup (\tilde{\mathcal{Y}} \cap \mathcal{X})] = \Pr_{S\sim\mathcal{D}_{\tilde{\mathcal{X}}\cup\tilde{\mathcal{Y}}}}[\mathrm{sign}\left((\tilde{\mathbf{w}}_C)^\intercal \mathbf{x}_S + n^\epsilon\right) \neq \mathrm{sign}\left(S \in \mathcal{X}\right)] \leq \frac{\epsilon}{n}.$$

and we conlude that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}}\left[\left|\tilde{f}(S) - f(S)\right|\right] \leq \epsilon$$

with probability $1 - \delta$. □