Teacher Improves Learning by Selecting a Training Subset

Yuzhe Ma **Robert Nowak** **Philippe Rigollet*** **Xuezhou Zhang** **Xiaojin Zhu**
University of Wisconsin-Madison *Massachusetts Institute of Technology

Abstract

We call a learner super-teachable if a teacher can trim down an iid training set while making the learner learn even better. We provide sharp super-teaching guarantees on two learners: the maximum likelihood estimator for the mean of a Gaussian, and the large margin classifier in 1D. For general learners, we provide a mixed-integer nonlinear programming-based algorithm to find a super teaching set. Empirical experiments show that our algorithm is able to find good super-teaching sets for both regression and classification problems.

1 Introduction

Consider the following question: a learner receives an iid training set S drawn from a distribution parametrized by θ^*. There is a teacher who knows θ^*. Can the teacher select a subset from S so the learner estimates θ^* better from the subset than from S?

This question is distinct from training set reduction (see e.g. [19, 43, 42]) in that the teacher can use the knowledge of θ^* to carefully design the subset. It is, in fact, a coding problem: Can the teacher approximately encode θ^* using items in S for a known decoder, which is the learner? As such, the question is not a machine learning task but rather a machine teaching one [47, 20, 45].

This question is relevant for several nascent applications. One application is in understanding blackbox models such as deep nets. Often observation to a blackbox model is limited to its predicted label $y = \theta^*(x)$ given input x. One way to interpret a blackbox model is to locally train an interpretable model with data points S labeled by the blackbox model around the region of interest [35]. We, however, ask for more: to reduce the size of the training set S for the local learner *while* making the learner approximate the blackbox better. The reduced training set itself also serves as representative examples of local model behavior. Another application is in education. Imagine a teacher who has a teaching goal θ^*. This is a reasonable assumption in practice: e.g. a geology teacher has the knowledge of the actual decision boundaries between rock categories. However, the teacher is constrained to teach with a given textbook (or a set of courseware) S. To the extent that the student is quantified mathematically, the teacher wants to select pages in the textbook with the guarantee that the student learns better from those pages than from gulping the whole book.

But is the question possible? The following example says yes. Consider learning a threshold classifier on the interval $[-1, 1]$, with true threshold at $\theta^* = 0$. Let S have n items drawn uniformly from the interval and labeled according to θ^*. Let the learner be a hard margin SVM, which places the estimated threshold in the middle of the inner-most pair in S with different labels: $\hat{\theta}_S = (x_- + x_+)/2$ where x_- is the largest negative training item and x_+ the smallest positive training item in S. It is well known that $|\hat{\theta}_S - \theta^*|$ converges at a rate of $1/n$: the intuition being that the average space between adjacent items is $O(1/n)$.

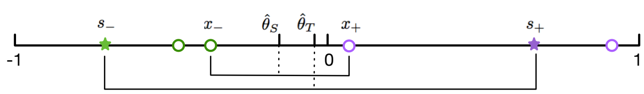


Figure 1: The original training set S with $n = 6$ items (circles and stars; green=negative, purple=positive), and the most-symmetric training set (stars) the teacher selects.

The teacher knows everything but cannot tell θ^* directly to the learner. Instead, it can select the *most-symmetric pair* in S about θ^* and give them to the learner as a two-item training set. We will prove later that the risk on the most symmetric pair is $O(1/n^2)$, that is, learning from the selected subset surpasses learning from S. Thus we observe something interesting: the teacher can turn a larger training set S into a smaller and better subset for the midpoint classifier. We call this phenomenon **super-teaching**.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

2 Formal Definition of Super Teaching

Let $\mathbb{Z}$ be the data space: for unsupervised learning $\mathbb{Z} = \mathbb{X}$, while for supervised learning $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$. Let $p_{\mathbb{Z}}$ be the underlying distribution over $\mathbb{Z}$. We take a function view of the learner: a learner A is a function $A : \cup_{n=0}^{\infty} \mathbb{Z}^n \mapsto \Theta$, where Θ is the learner's hypothesis space. The notation $\cup_{n=0}^{\infty} \mathbb{Z}^n$ defines the "set of (potentially non-iid) training sets", namely multisets of any size whose elements are in $\mathbb{Z}$. Given any training set $T \in \cup_{n=0}^{\infty} \mathbb{Z}^n$, we assume A returns a unique hypothesis $A(T) \triangleq \hat{\theta}_T \in \Theta$. The learner's risk $R(\theta)$ for $\theta \in \Theta$ is defined as:

$$R(\theta) = \mathbf{E}_{p_{\mathbb{Z}}} \left[\ell(\theta(x), y)\right], \text{ or } R(\theta) = \|\theta - \theta^*\|_2. \tag{1}$$

The former is for prediction tasks where $\ell()$ is a loss function and $\theta(x)$ denotes the prediction on x made by model θ; the latter is for parameter estimation where we assume a realizable model $p_{\mathbb{Z}} = p_{\theta^*}$ for some $\theta^* \in \Theta$.

We now introduce a clairvoyant teacher B who has full knowledge of $p_{\mathbb{Z}}, A, R$. The teacher is also given an iid training set $S = \{z_1, \ldots, z_n\} \sim p_{\mathbb{Z}}$. If the teacher teaches S to A, the learner will incur a risk $R(A(S)) \triangleq R(\hat{\theta}_S)$. The teacher's goal is to judiciously select a subset $B(S) \subset S$ to act as a "super teaching set" for the learner so that $R(\hat{\theta}_{B(S)}) < R(\hat{\theta}_S)$. Of course, to do so the teacher must utilize her knowledge of the learning task, thus the subset is actually a function $B(S, p_{\mathbb{Z}}, A, R)$. In particular, the teacher knows $p_{\mathbb{Z}}$ already, and this sets our problem apart from machine learning. For readability we suppress these extra parameters in the rest of the paper. We formally define super teaching as follows.

Definition 1 (Super Teaching). *B is a super teacher for learner A if $\forall \delta > 0, \exists N$ such that $\forall n \geq N$*

$$\mathbf{P}_S \left[R(\hat{\theta}_{B(S)}) \leq c_n R(\hat{\theta}_S)\right] > 1 - \delta, \tag{2}$$

where $S \overset{iid}{\sim} p_{\mathbb{Z}}^n, B(S) \subset S$, and $c_n \leq 1$ is a sequence we call super teaching ratio.

Obviously, $c_n = 1$ can be trivially achieved by letting $B(S) = S$ so we are interested in small c_n. There are two fundamental questions: (1) Do super teachers provably exist? (2) How to compute a super teaching set $B(S)$ in practice?

We answer the first question positively by exhibiting super teaching on two learners: maximum likelihood estimator for the mean of a Gaussian in section 3, and 1D large margin classifier in section 4. Guarantees on super teaching for general learners remain future work. Nonetheless, empirically we can find a super teaching set for many general learners: We formulate the second question as mixed-integer nonlinear programming in section 5. Empirical experiments in section 6 demonstrates that one can find a good $B(S)$ effectively.

3 Analysis on Super Teaching for the MLE of Gaussian mean

In this section, we present our first theoretical result on super teaching, when the learner A_{MLE} is the maximum likelihood estimator (MLE) for the mean of a Gaussian. Let $\mathbb{Z} = \mathbb{X} = \mathbb{R}$, $\Theta = \mathbb{R}$, $p_{\mathbb{Z}}(x) = \mathcal{N}(\theta^*, 1)$. Given a sample S of size n drawn from $p_{\mathbb{Z}}$, the learner computes the MLE for the mean: $\hat{\theta}_S = A_{MLE}(S) = \frac{1}{n} \sum_{i=1}^n x_i$. We define the risk as $R(\hat{\theta}_S) = |\hat{\theta}_S - \theta^*|$. The teacher we consider is the optimal k-subset teacher B_k, which uses the best subset of size k to teach:

$$B_k(S) \in \text{argmin}_{T \subset S, |T| = k} R(\hat{\theta}_T). \tag{3}$$

To build intuition, it is well-known that the risk of A_{MLE} under S is $O(1/\sqrt{n})$ because the variance under n items shrinks like $1/n$. Now consider $k = 1$. Since the teacher B_1 knows θ^*, under our setting the best teaching strategy is for her to select the item in S closest to θ^*, which forms the singleton teaching set $B_1(S)$. One can show that with large probability this closest item is $O(1/n)$ away from θ^* (the central part of a Gaussian density is essentially uniform). Therefore, we already see a super teaching ratio of $c_n = n^{-\frac{1}{2}}$. More generally, our main result below shows that B_k achieves a super teaching ratio $c_n = O(n^{-k+\frac{1}{2}})$:

Theorem 1. *Let B_k be the optimal k-subset teacher. $\forall \epsilon \in (0, \frac{2k-1}{4}), \forall \delta \in (0, 1)$, $\exists N(k, \epsilon, \delta)$ such that $\forall n \geq N(k, \epsilon, \delta)$, $\mathbf{P}\left[R(\hat{\theta}_{B_k(S)}) \leq c_n R(\hat{\theta}_S)\right] > 1 - \delta$, where $c_n = \frac{k^{k-\epsilon}}{\sqrt{k}} n^{-k+\frac{1}{2}+2\epsilon}$.*

Toward proving the theorem, [1] we first recall the standard rate $R(\hat{\theta}_S) \approx n^{-\frac{1}{2}}$ if A_{MLE} learns from the whole training set S:

Proposition 2. *Let S be an n-item iid sample drawn from $\mathcal{N}(\theta^*, 1)$. $\forall \epsilon > 0$, $\forall \delta \in (0, 1)$, $\exists N_1(\epsilon, \delta)$ such that $\forall n \geq N_1$,*

$$\mathbf{P}\left[n^{-\frac{1}{2}-\epsilon} < R(\hat{\theta}_S) < n^{-\frac{1}{2}+\epsilon}\right] > 1 - \delta. \tag{4}$$

Proof. $R(\hat{\theta}_S) = |\hat{\theta}_S - \theta^*|$ and $\hat{\theta}_S - \theta^* \sim \mathcal{N}(0, n^{-1}) = \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}}$. Let $\alpha = n^{-\frac{1}{2}-\epsilon}$ and $\beta = n^{-\frac{1}{2}+\epsilon}$. We have

$$\begin{aligned} \mathbf{P}\left[R(\hat{\theta}_S) \leq \alpha\right] &= 2\int_0^{\alpha} \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} dx \\ &< 2\int_0^{\alpha} \sqrt{\frac{n}{2\pi}} dx = 2\alpha\sqrt{\frac{n}{2\pi}} = \sqrt{\frac{2}{\pi}} n^{-\epsilon}, \end{aligned} \tag{5}$$

[1] Remark: we introduced an auxiliary variable ϵ which controls the implicit tradeoff between c_n, how much super teaching helps, and N, how soon super teaching takes effect. When $\epsilon \to 0$ the teaching ratio c_n approaches $O(n^{-k+\frac{1}{2}})$, but as we will see $N(k, \epsilon, \delta) \to \infty$. Similarly, k also affects the tradeoff: the teaching ratio is smaller as we enlarge k, but $N(k, \epsilon, \delta)$ increases.

$$\mathbf{P}\left[R(\hat{\theta}_S) \geq \beta\right] = 2\int_{\beta}^{\infty} \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} dx$$
$$< 2\int_{\beta}^{\infty} \frac{x}{\beta}\sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} dx = \int_{\beta^2}^{\infty} \frac{1}{\beta}\sqrt{\frac{n}{2\pi}} e^{-\frac{ny}{2}} dy \quad (6)$$
$$= \frac{1}{\beta}\sqrt{\frac{2}{n\pi}} e^{-\frac{n\beta^2}{2}} < \frac{1}{\beta}\sqrt{\frac{2}{n\pi}} = \sqrt{\frac{2}{\pi}} n^{-\epsilon}.$$

Thus $\mathbf{P}\left[\alpha < R(\hat{\theta}_S) < \beta\right] = 1 - \mathbf{P}\left[R(\hat{\theta}_S) \leq \alpha\right] - \mathbf{P}\left[R(\hat{\theta}_S) \geq \beta\right] > 1 - 2\sqrt{\frac{2}{\pi}} n^{-\epsilon}$. Let $N_1(\epsilon, \delta) = (\frac{1}{\delta}\sqrt{\frac{8}{\pi}})^{\frac{1}{\epsilon}}$, then $\forall n \geq N_1$, $\mathbf{P}\left[\alpha < R(\hat{\theta}_S) < \beta\right] > 1 - \delta$. □

We now work out the risk of A_{MLE} if it learns from the optimal k-subset teacher B_k. Theorem 4 says that this risk is very small and sharply concentrated around $R(\hat{\theta}_{B_k(S)}) \approx n^{-k}$. To prove Theorem 4, we first give the following lemma.

Lemma 3. *Denote $C_k^n = \binom{n}{k}$. Let the index set $I = \{1, 2, ...n\}$ where $n \geq 4k$. Consider all subsets of size k, then there are at most $4^k C_k^{2k} C_{2k-1}^n$ ordered pairs of subsets that are overlapping but not identical.*

Proof. Let I_1 and I_2 be two subsets of size k and they overlap on t indexes. Then the total number of distinct indexes that appear in $I_1 \cup I_2$ is $2k - t$. There are C_{2k-t}^n ways of choosing such $2k - t$ indexes. Next we determine which t indexes are overlapping ones. We have C_t^{2k-t} ways of choosing such t indexes. Finally we have C_{k-t}^{2k-2t} ways of selecting half of the non-overlapping indexes and attribute them to I_1. Thus in total we have $O_t = C_{2k-t}^n C_t^{2k-t} C_{k-t}^{2k-2t}$ ordered pairs of subsets that overlap on t indexes. By our assumption $n \geq 4k$ we have $C_{2k-t}^n \leq C_{2k-1}^n$. Also note that $C_t^{2k-t} < C_t^{2k}$ and $C_{k-t}^{2k-2t} < C_k^{2k}$, thus $O_t < C_{2k-1}^n C_t^{2k} C_k^{2k}$. Therefore the total number of ordered pairs of subsets that are overlapping but not identical is

$$O = \sum_{t=1}^{k-1} O_t < \sum_{t=1}^{k-1} C_{2k-1}^n C_t^{2k} C_k^{2k}$$
$$< \sum_{t=0}^{2k} C_{2k-1}^n C_t^{2k} C_k^{2k} = 4^k C_k^{2k} C_{2k-1}^n. \quad (7)$$

□

Now we prove the risk of the optimal k-subset teacher.

Theorem 4. *Let B_k be the optimal k-subset teacher. Let S be an n-item iid sample drawn from $\mathcal{N}(\theta^*, 1)$. $\forall \epsilon \in (0, k), \forall \delta \in (0, 1)$, $\exists N_2(k, \epsilon, \delta)$ such that $\forall n \geq N_2$,*

$$\mathbf{P}\left[\frac{1}{\sqrt{k}}(\frac{k}{n})^{k+\epsilon} < R(\hat{\theta}_{B_k(S)}) < \frac{1}{\sqrt{k}}(\frac{k}{n})^{k-\epsilon}\right] > 1 - \delta. \quad (8)$$

Proof. Let $I \subseteq \{1, 2, ..., n\}$ and $|I| = k$, define $\gamma_I = \frac{1}{\sqrt{k}}\sum_{i \in I}(x_i - \theta^*)$. Let S_I denote the subset indexed by I. Note that $\hat{\theta}_{S_I} = \frac{1}{k}\sum_{i\in I} x_i$ and $R(\hat{\theta}_{S_I}) = |\hat{\theta}_{S_I} - \theta^*| = |\frac{1}{k}\sum_{i\in I} x_i - \theta^*| = \frac{1}{\sqrt{k}}|\gamma_I|$. Also note that $R(\hat{\theta}_{B_k(S)}) = \inf_I R(\hat{\theta}_{S_I}) = \frac{1}{\sqrt{k}} \inf_I |\gamma_I|$. Thus to prove Theorem 4 it suffices to prove

$$\mathbf{P}\left[(\frac{k}{n})^{k+\epsilon} < \inf_I |\gamma_I| < (\frac{k}{n})^{k-\epsilon}\right] \to 1. \quad (9)$$

Let $\alpha = (\frac{k}{n})^{k+\epsilon}$ and $\beta = (\frac{k}{n})^{k-\epsilon}$. We first prove the lower bound. Note that γ_I has the same distribution for all I. Thus by the union bound,

$$\mathbf{P}\left[\inf_I |\gamma_I| \leq \alpha\right] = \mathbf{P}\left[\exists I : |\gamma_I| \leq \alpha\right] \leq C_k^n \mathbf{P}\left[|\gamma_{I_1}| \leq \alpha\right], \quad (10)$$

where $I_1 = \{1, 2, ..., k\}$. Since $\gamma_{I_1} \sim \mathcal{N}(0, 1)$, we have

$$\mathbf{P}\left[|\gamma_{I_1}| \leq \alpha\right] = \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < \sqrt{\frac{2}{\pi}}\alpha. \quad (11)$$

Note that $C_k^n \leq (\frac{en}{k})^k$. Thus,

$$\mathbf{P}\left[\inf_I |\gamma_I| \leq \alpha\right] < (\frac{en}{k})^k \sqrt{\frac{2}{\pi}}\alpha = \sqrt{\frac{2}{\pi}} e^k (\frac{k}{n})^{\epsilon} \to 0. \quad (12)$$

Thus $\exists N_2'(k, \epsilon, \delta)$ such that $\forall n \geq N_2'$,

$$\mathbf{P}\left[\inf_I |\gamma_I| \leq \alpha\right] < \frac{\delta}{2}. \quad (13)$$

To show the upper bound, we define $t_I = \mathbb{1}\left[|\gamma_I| < \beta\right]$, where $\mathbb{1}\left[\right]$ is the indicator function. Let $T = \sum_I t_I$. Then it suffices to show $\lim_{n\to\infty} \mathbf{P}\left[T = 0\right] = 0$. Note that

$$\mathbf{P}\left[T = 0\right] = \mathbf{P}\left[T - \mathbf{E}\left[T\right] = -\mathbf{E}\left[T\right]\right]$$
$$\leq \mathbf{P}\left[(T - \mathbf{E}\left[T\right])^2 \geq (\mathbf{E}\left[T\right])^2\right] \leq \frac{\mathbf{V}\left[T\right]}{(\mathbf{E}\left[T\right])^2}, \quad (14)$$

where the last inequality follows from the Markov inequality. Now we lower bound $\mathbf{E}\left[T\right]$.

$$\mathbf{E}\left[T\right] = \mathbf{E}\left[\sum_I t_I\right] = \sum_I \mathbf{E}\left[t_I\right] = C_k^n \mathbf{E}\left[t_{I_1}\right]$$
$$= C_k^n \mathbf{P}\left[|\gamma_{I_1}| < \beta\right] = C_k^n \int_{-\beta}^{\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx. \quad (15)$$

Note that $\epsilon < k$, thus $\beta < 1$. For $x \in (-\beta, \beta)$, $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} > \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi e}}$. Also note that $C_k^n > (\frac{n}{k})^k$, thus

$$\mathbf{E}\left[T\right] > (\frac{n}{k})^k \frac{1}{\sqrt{2\pi e}} 2\beta = \sqrt{\frac{2}{\pi e}} (\frac{n}{k})^{\epsilon}. \quad (16)$$

Now we upper bound $\mathbf{V}[T]$.

$$\mathbf{V}[T] = \sum_{I,I'} \mathbf{Cov}[t_I, t_{I'}] = \sum_{I,I', |I\cap I'|\geq 1} \mathbf{Cov}[t_I, t_{I'}]. \tag{17}$$

Note that for Bernoulli random variable t_I, $\mathbf{V}[t_I] \leq \mathbf{E}[t_I]$. Thus if $I = I'$, then

$$\begin{aligned}\mathbf{Cov}[t_I, t_{I'}] &= \mathbf{V}[t_I] \leq \mathbf{E}[t_I] = \mathbf{P}[|\gamma_I| < \beta] \\ &= \int_{-\beta}^{\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < \frac{1}{\sqrt{2\pi}} 2\beta = \sqrt{\frac{2}{\pi}} (\frac{k}{n})^{k-\epsilon}.\end{aligned} \tag{18}$$

Otherwise $1 \leq |I \cap I'| \leq k-1$, that is, I and I' overlap but not identical, then

$$\begin{aligned}\mathbf{Cov}[t_I, t_{I'}] &= \mathbf{E}[t_I t_{I'}] - \mathbf{E}[t_I]\mathbf{E}[t_{I'}] \leq \mathbf{E}[t_I t_{I'}] \\ &= \mathbf{P}[|\gamma_I| < \beta, |\gamma_{I'}| < \beta].\end{aligned} \tag{19}$$

Note that γ_I and $\gamma_{I'}$ are jointly Gaussian with covariance

$$\begin{aligned}\mathbf{Cov}[\gamma_I, \gamma_{I'}] &= \frac{1}{k} \sum_{i\in I, i'\in I'} \mathbf{Cov}[x_i - \theta^*, x_{i'} - \theta^*] \\ &= \frac{1}{k} \sum_{i\in I, i'\in I', i=i'} 1 = \frac{|I\cap I'|}{k} \triangleq \rho,\end{aligned} \tag{20}$$

where $\frac{1}{k} \leq \rho \leq \frac{k-1}{k}$. The joint PDF of two standard normal distributions x, y with covariance ρ is

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}}. \tag{21}$$

Note that $f(x,y) \leq \frac{1}{2\pi\sqrt{1-\rho^2}}$, thus

$$\begin{aligned}\mathbf{P}[|\gamma_I| < \beta, |\gamma_{I'}| < \beta] &\leq \iint_{|x|<\beta, |y|<\beta} \frac{1}{2\pi\sqrt{1-\rho^2}} dxdy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} (2\beta)^2 = \frac{2}{\pi\sqrt{1-\rho^2}} \beta^2.\end{aligned} \tag{22}$$

Since $\frac{2}{\pi\sqrt{1-\rho^2}} \leq \frac{2}{\pi\sqrt{1-(\frac{k-1}{k})^2}} \leq \frac{2}{\pi\sqrt{\frac{k}{k^2}}} = \frac{2\sqrt{k}}{\pi}$, thus

$$\mathbf{P}[|\gamma_I| < \beta, |\gamma_{I'}| < \beta] \leq \frac{2\sqrt{k}\beta^2}{\pi} = \frac{2\sqrt{k}}{\pi} (\frac{k}{n})^{2k-2\epsilon}. \tag{23}$$

According to Lemma 3, there are at most $4^k C_k^{2k} C_{2k-1}^n$ pairs of I and I' such that $1 \leq |I \cap I'| \leq k-1$. Thus,

$$\begin{aligned}\mathbf{V}[T] &= \sum_I \mathbf{V}[t_I] + \sum_{I\neq I', |I\cap I'|\geq 1} \mathbf{Cov}[t_I, t_{I'}] \\ &\leq C_k^n \sqrt{\frac{2}{\pi}} (\frac{k}{n})^{k-\epsilon} + 4^k C_k^{2k} C_{2k-1}^n \frac{2\sqrt{k}}{\pi} (\frac{k}{n})^{2k-2\epsilon} \\ &\leq \sqrt{\frac{2}{\pi}} (\frac{en}{k})^k (\frac{k}{n})^{k-\epsilon} + 4^k C_k^{2k} (\frac{en}{2k-1})^{2k-1} \frac{2\sqrt{k}}{\pi} (\frac{k}{n})^{2k-2\epsilon} \\ &= \sqrt{\frac{2}{\pi}} e^k (\frac{n}{k})^{\epsilon} + \frac{4\sqrt{k}}{\pi} C_k^{2k} (\frac{2ek}{2k-1})^{2k-1} (\frac{n}{k})^{2\epsilon-1}.\end{aligned} \tag{24}$$

Now plug (24) and (16) into (14), we have

$$\mathbf{P}[T=0] \leq a_1(k)(\frac{n}{k})^{-\epsilon} + a_2(k)(\frac{n}{k})^{-1} \to 0, \tag{25}$$

where $a_1 = \sqrt{\frac{\pi}{2}} e^{k+1}$ and $a_2(k) = 2\sqrt{k} e C_k^{2k} (\frac{2ek}{2k-1})^{2k-1}$. Thus $\exists N_2''(k,\epsilon,\delta)$ such that $\forall n \geq N_2''$,

$$\mathbf{P}\left[\inf_I |\gamma_I| \geq \beta\right] < \frac{\delta}{2}. \tag{26}$$

Let $N_2(k,\epsilon,\delta) = \max\{N_2'(k,\epsilon,\delta), N_2''(k,\epsilon,\delta)\}$, combining (13) and (26) concludes the proof. $\square$

Now we can conclude super-teaching by comparing Theorem 4 and Proposition 2:

Proof of Theorem 1. Let $\alpha = \frac{1}{\sqrt{k}} (\frac{k}{n})^{k-\epsilon}$ and $\beta = n^{-\frac{1}{2}-\epsilon}$. By Proposition 2, $\forall \epsilon \in (0, \frac{2k-1}{4}), \forall \delta \in (0,1)$, $\exists N_1(\epsilon, \frac{\delta}{2})$ such that $\forall n \geq N_1$, $\mathbf{P}\left[R(\hat{\theta}_S) > \beta\right] > 1 - \frac{\delta}{2}$. By Theorem 4, $\exists N_2(k, \epsilon, \frac{\delta}{2})$ such that $\forall n \geq N_2$, $\mathbf{P}\left[R(\hat{\theta}_{B_k(S)}) < \alpha\right] > 1 - \frac{\delta}{2}$. Let $c_n = \frac{k^{k-\epsilon}}{\sqrt{k}} n^{-k+\frac{1}{2}+2\epsilon}$. Since $\epsilon < \frac{2k-1}{4}$, c_n is a decreasing sequence in n with $\lim_{n\to\infty} c_n = 0$. Let $N_3(k,\epsilon)$ be the first integer such that $c_{N_3} \leq 1$. Let $N(k,\epsilon,\delta) = \max\{N_1(\epsilon, \frac{\delta}{2}), N_2(k,\epsilon,\frac{\delta}{2}), N_3(k,\epsilon)\}$. By a union bound $\forall n \geq N(k,\epsilon,\delta)$, $\mathbf{P}\left[R(\hat{\theta}_{B_k(S)}) < \alpha, R(\hat{\theta}_S) \geq \beta\right] > 1-\delta$. Since $\frac{\alpha}{\beta} = c_n$, we have $\mathbf{P}\left[R(\hat{\theta}_{B_k(S)}) \leq c_n R(\hat{\theta}_S)\right] > 1 - \delta$, where $c_n \leq c_{N_3} \leq 1$. $\square$

4 Analysis on Super Teaching for 1D Large Margin Classifier

We present our second theoretical result, this time on teaching a 1D large margin classifier. Let $\mathbb{X} = [-1,1]$, $\mathbb{Y} = \{-1,1\}$, $\Theta = [-1,1]$, $\theta^* = 0$, $p_{\mathbb{Z}}(x,y) = p_{\mathbb{Z}}(x) p_{\mathbb{Z}}(y \mid x)$ where $p_{\mathbb{Z}}(x) = U(\mathbb{X})$ and $p_{\mathbb{Z}}(y = 1 \mid x) = \mathbb{1}[x \geq \theta^*]$. Let $x_- \triangleq \max_{i:y_i=-1} x_i$ and $x_+ \triangleq \min_{i:y_i=+1} x_i$ be the inner-most pair of opposite labels in S if they exist. We formally define the large margin classifier $A_{lm}(S)$ as

$$\hat{\theta}_S = A_{lm}(S) = \begin{cases} (x_- + x_+)/2 & \text{if } x_-, x_+ \text{ exist} \\ -1 & \text{if } S \text{ all positive} \\ 1 & \text{if } S \text{ all negative.} \end{cases} \tag{27}$$

The risk is defined as $R(\hat{\theta}_S) = |\hat{\theta}_S - \theta^*| = |\hat{\theta}_S|$. The teacher we consider is the most symmetric teacher, who selects the most symmetric pair about θ^* in S and gives it to the learner. We define the most-symmetric teacher B_{ms}:

$$B_{ms}(S) = \begin{cases} \{(s_-, -1), (s_+, 1)\} & \text{if } s_-, s_+ \text{ exist,} \\ \{(x_1, y_1)\} & \text{otherwise.} \end{cases} \tag{28}$$

where $(s_-, s_+) \in \text{argmin}_{(x,-1),(x',1)\in S} |\frac{x+x'}{2} - \theta^*|$. $\square$

Our main result shows that learning from the whole set S achieves the well-known $O(1/n)$ risk, but surprisingly B_{ms} achieves $O(1/n^2)$ risk, therefore it is an approximately $c_n = O(n^{-1})$ super teaching ratio.

Theorem 5. *Let S be an n-item iid sample drawn from $p_{\mathbb{Z}}$. Then $\forall \delta \in (0, 1)$, $\exists N(\delta)$ such that $\forall n \geq N$, $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \leq c_n R(\hat{\theta}_S)\right] > 1 - \delta$, where $c_n = \frac{32}{n\delta} \ln \frac{6}{\delta}$.*

Before proving Theorem 5, we first show that B_{ms} is an optimal teacher for the large margin classifier.

Proposition 6. *B_{ms} is an optimal teacher for the large margin classifier $\hat{\theta}_S$.*

Proof. We show $R(\hat{\theta}_{B_{ms}(S)}) \leq R(\hat{\theta}_{B(S)})$ for any B and any S.

If $|B_{ms}(S)| = 1$, then S is either all positive or all negative. In both cases $R(\hat{\theta}_{B(S)}) = 1$ for any B by definition. Thus $R(\hat{\theta}_{B_{ms}(S)}) \leq R(\hat{\theta}_{B(S)})$.

Otherwise $|B_{ms}(S)| = 2$, then if $B(S)$ is all positive or all negative, we have $R(\hat{\theta}_{B(S)}) = 1$ and thus $R(\hat{\theta}_{B_{ms}(S)}) \leq R(\hat{\theta}_{B(S)})$. Otherwise let x_-^B, x_+^B be the inner most pair of $B(S)$. Since $x_-^B, x_+^B \in S$, then by definition of B_{ms}, $R(\hat{\theta}_{B_{ms}(S)}) = |\frac{s_- + s_+}{2} - \theta^*| \leq |\frac{x_-^B + x_+^B}{2} - \theta^*| = R(\hat{\theta}_{B(S)})$. $\square$

Now we show that learning on the whole S incurs $O(n^{-1})$ risk. First, we give the following lemma for the exact tail probability of $R(\hat{\theta}_S)$.

Lemma 7. *For the large margin classifier $\hat{\theta}_S$, we have*

$$\mathbf{P}\left[R(\hat{\theta}_S) > \epsilon\right] = \begin{cases} (1-\epsilon)^n + (\epsilon)^n & 0 < \epsilon \leq \dfrac{1}{2} \\ (\dfrac{1}{2})^{n-1} & \dfrac{1}{2} < \epsilon < 1 \\ 0 & \epsilon = 1. \end{cases} \tag{29}$$

The proof for Lemma 7 is in the appendix.

Now we show that $R(\hat{\theta}_S)$ is $O(n^{-1})$.

Theorem 8. *Let S be an n-item iid sample drawn from $p_{\mathbb{Z}}$. Then $\forall \delta \in (0, 1)$ and $\forall n \geq 2$,*

$$\mathbf{P}\left[R(\hat{\theta}_S) > \frac{\delta}{n}\right] > 1 - \delta. \tag{30}$$

Proof. According to Lemma 7, for $\epsilon \leq \frac{1}{2}$, we have

$$\mathbf{P}\left[R(\hat{\theta}_S) > \epsilon\right] > (1-\epsilon)^n > 1 - n\epsilon. \tag{31}$$

Note that $n \geq 2$, thus $\frac{\delta}{n} \leq \frac{1}{2}$. Let $\epsilon = \frac{\delta}{n}$ in (31) we have

$$\mathbf{P}\left[R(\hat{\theta}_S) > \frac{\delta}{n}\right] > 1 - n\frac{\delta}{n} = 1 - \delta. \tag{32}$$

$\square$

Now we work out the risk of the most symmetric teacher B_{ms}. To bound the risk of B_{ms} we need the following key lemma, which shows that the sample complexity with the teacher is $O(\epsilon^{-1/2})$.

Lemma 9. *Let $n = 4m$, where m is an integer. Let S be an n-item iid sample drawn from $p_{\mathbb{Z}}$. $\forall \epsilon > 0, \forall \delta \in (0, 1)$, $\exists \mathbb{M}(\epsilon, \delta) = \max\{\frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta}, (\frac{1}{\epsilon} \ln \frac{3}{\delta})^{\frac{1}{2}}\}$ such that $\forall m \geq \mathbb{M}(\epsilon, \delta)$, $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \leq \epsilon\right] > 1 - \delta$.*

Proof. We give a proof sketch and the details are in the appendix. Let $S_1 = \{x \mid (x, 1) \in S\}$ and $S_2 = \{x \mid (x, -1) \in S\}$ respectively. Then we have $|S_1| + |S_2| = 4m$. Define event $E_1 : \{|S_1| \geq m \wedge |S_2| \geq m\}$. Given that $m \geq \frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta}$, one can show $P(E_1) > 1 - \frac{\delta}{3}$. Since $|S_1| + |S_2| = 4m$, either $|S_1| \geq 2m$ or $|S_2| \geq 2m$. Without loss of generality we assume $|S_1| \geq 2m$. We then divide the interval [0, 1] equally into $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor$ segments. The length of each segment is $\frac{1}{N} = O(\frac{1}{m^2})$ as Figure 2 shows.

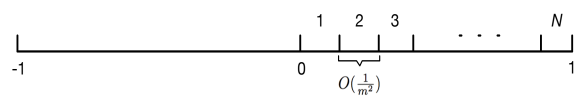


Figure 2: segments

Let N_o be the number of segments that are occupied by the points in S_1. Note that N_o is a random variable. Let E_2 be the event that $N_o \geq m$. Then one can show $P(E_2) > 1 - \frac{\delta}{3}$. By union bound, we have $P(E_1, E_2) > 1 - \frac{2\delta}{3}$. Let E_3 be the following event: there exist a point x_2 in S_2 such that $-x_2$, the flipped point, lies in the same segment as some point x_1 in S_1. One can show that $P(E_3 \mid E_1, E_2) > 1 - \frac{\delta}{3}$. Thus $P(E_3) \geq P(E_1, E_2, E_3) = P(E_3 \mid E_1, E_2) P(E_1, E_2) \geq (1 - \frac{\delta}{3})(1 - \frac{2\delta}{3}) > 1 - \delta$. If E_3 happens, then $|x_1 + x_2| = |x_1 - (-x_2)| \leq \frac{1}{N}$. Note that $m \geq (\frac{1}{\epsilon} \ln \frac{3}{\delta})^{\frac{1}{2}}$ and $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor \geq \frac{m^2}{2} (\ln \frac{3}{\delta})^{-1}$, thus $\frac{1}{N} \leq \frac{2}{m^2} \ln \frac{3}{\delta} \leq 2\epsilon$. Therefore $R(\hat{\theta}_{B_{ms}(S)}) = |\frac{s_- + s_+}{2}| \leq |\frac{x_1 + x_2}{2}| \leq \epsilon$. $\square$

Rewriting ϵ in Lemma 9 as a function of n, we have the following theorem.

Theorem 10. *Let S be an n-item iid sample dawn from $p_{\mathbb{Z}}$, then $\exists N_1(\delta) = \frac{12e}{\ln 4 - 1} \ln \frac{3}{\delta}$ such that $\forall n \geq N_1$,*

$$\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \leq \frac{16}{n^2} \ln \frac{3}{\delta}\right] > 1 - \delta. \tag{33}$$

Proof. Note that if $n \geq N_1(\delta) = \frac{12e}{\ln 4 - 1} \ln \frac{3}{\delta}$, then $m = \frac{n}{4} \geq \frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta}$, thus the minimum ϵ that satisfies $m \geq \mathbb{M}(\epsilon, \delta)$ is $\frac{1}{m^2} \ln \frac{3}{\delta} = \frac{16}{n^2} \ln \frac{3}{\delta}$. $\square$

Now we can conclude super teaching:

Proof of Theorem 5. According to Theorem 10, $\exists N_1(\frac{\delta}{2})$ such that $\forall n \geq N_1$, $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \leq \frac{16}{n^2}\ln\frac{6}{\delta}\right] > 1 - \frac{\delta}{2}$. Note that $N_1 \geq 2$, thus according to Theorem 8, $\forall n \geq N_1$, $\mathbf{P}\left[R(\hat{\theta}_S) > \frac{\delta}{2n}\right] > 1 - \frac{\delta}{2}$. Let $c_n = \frac{32}{n\delta}\ln\frac{6}{\delta}$ and $N_2(\delta) = \frac{32}{\delta}\ln\frac{6}{\delta}$ so that $c_{N_2} = 1$. Let $N(\delta) = \max\{N_1(\delta), N_2(\delta)\}$. By union bound, $\forall n \geq N$, with probability at least $1-\delta$, we have both $R(\hat{\theta}_S) > \frac{\delta}{2n}$ and $R(\hat{\theta}_{B_{ms}(S)}) \leq \frac{16}{n^2}\ln\frac{6}{\delta}$, which gives $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \leq c_n R(\hat{\theta}_S)\right] > 1 - \delta$, where $c_n \leq c_{N_2} = 1$. $\square$

5 An MINLP Algorithm for Super Teaching

Although the problem of proving super teaching ratios for a specific learner is interesting, we now focus on an algorithm to find a super teaching set for general learners *given* a training set S. That is, we find a subset $B(S) \subset S$ so that $R(\hat{\theta}_{B(S)}) < R(\hat{\theta}_S)$. We start by formulating super teaching as a subset selection problem. To this end, we introduce binary indicator variables $b_1, \ldots, b_n$ where $b_i = 1$ means $z_i \in S$ is included in the subset. We consider learners A that can be defined via convex empirical risk minimization:

$$A(S) \triangleq \text{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} \tilde{\ell}(\theta, z_i) + \frac{\lambda}{2}\|\theta\|^2. \tag{34}$$

For simplicity we assume there is a unique global minimum which is returned by argmin. Note that we use $\tilde{\ell}$ in (34) to denote the (surrogate) convex loss used by A in performing empirical risk minimization. For example, $\tilde{\ell}$ may be the negative log likelihood for logistic regression. $\tilde{\ell}$ is potentially different from ℓ (e.g. the 0-1 loss) used by the teacher to define the teaching risk R in (1).

We formulate super teaching as the following bilevel combinatorial optimization problem:

$$\min_{b\in\{0,1\}^n, \hat{\theta}\in\Theta} R(\hat{\theta}) \tag{35}$$

$$\text{s.t.} \quad \hat{\theta} = \text{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} b_i \tilde{\ell}(\theta, z_i) + \frac{\lambda}{2}\|\theta\|^2. \tag{36}$$

Under mild conditions, we may replace the lower level optimization problem (i.e. the machine learning problem (36)) by its first order optimality (KKT) conditions:

$$\min_{b\in\{0,1\}^n, \hat{\theta}\in\Theta} R(\hat{\theta}) \tag{37}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} b_i \nabla_\theta \tilde{\ell}(\hat{\theta}, z_i) + \lambda\hat{\theta} = 0.$$

This reduces the bilevel problem but the constraint is nonlinear in general, leading to a mixed-integer nonlinear program (MINLP), for which effective solvers exist. We use the MINLP solver in NEOS [15].

6 Simulations

We now apply the framework in section 5 to logistic regression and ridge regression, and show that the solver indeed selects a super-teaching subset that is far better than the original training set S.

6.1 Teaching Logistic Regression A_{lr}

Let $\mathbb{X} = \mathbb{R}^d$, $\Theta = \mathbb{R}^d$, $\theta^* = (\frac{1}{\sqrt{d}}, ..., \frac{1}{\sqrt{d}})$, $p_{\mathbb{Z}}(x) = \mathcal{N}(0, I)$. Let $p_{\mathbb{Z}}(y \mid x) = \mathbb{1}\left[x^\top\theta^* > 0\right]$, which is deterministic given x. Logistic regression estimates $\hat{\theta}_S = A_{lr}(S)$ with (34), where $\lambda = 0.1$ and $\tilde{\ell}(z_i) = \log(1 + \exp(-y_i x_i^\top \theta))$. In contrast, The teacher's risk is defined to be the expected 0-1 loss: $R(\hat{\theta}) = \mathbf{E}_{p_{\mathbb{Z}}}\left[\mathbb{1}\left[\hat{\theta}(x) \neq y\right]\right]$, where $\hat{\theta}(x)$ is the label of x predicted by $\hat{\theta}$. Since $p_{\mathbb{Z}}$ is symmetric about the origin, the risk can be rewritten in terms of the angle between $\hat{\theta}$ and θ^*: $R(\hat{\theta}) = \arccos(\frac{\hat{\theta}^\top\theta^*}{||\hat{\theta}||\cdot||\theta^*||})/\pi$. Instantiating (37) we have

$$\min_{b\in\{0,1\}^n, \hat{\theta}\in\mathbb{R}^d} \quad \arccos(\frac{\hat{\theta}^\top\theta^*}{||\hat{\theta}||\cdot||\theta^*||})/\pi \tag{38}$$

$$\text{s.t.} \quad \lambda\hat{\theta} - \sum_{i=1}^{n} \frac{b_i y_i x_i}{1 + \exp(y_i x_i^\top \hat{\theta})} = 0.$$

We run experiments to study the effectiveness and scalability of the NEOS MINLP solver on (38), specifically with respect to the training set size $n = |S|$ and dimension d.

In the first set of experiments we fix $d = 2$ and vary $n = 16, 64, 256$ and 1024. For each n we run 10 trials. In each trial we draw an n-item iid sample $S \sim p_{\mathbb{Z}}$ and call the solver on (38). The solver's solution to $b_1 \ldots b_n$ indicates the super teaching set $B(S)$. We then compute an empirical version of the super teaching ratio:

$$\hat{c}_n = R(\hat{\theta}_{B(S)})/R(\hat{\theta}_S).$$

	Logistic Regression			Ridge Regression		
$n = \lvert S\rvert$	$\hat{c}_n$	$\lvert B(S)\rvert/n$	time (s)	$\hat{c}_n$	$\lvert B(S)\rvert/n$	time (s)
16	8.5e-4	0.50	3.4e-1	7.8e-3	0.50	6.3e-1
64	1.3e-3	0.69	3.5e+0	7.5e-3	0.70	5.8e+0
256	6.3e-3	0.67	6.0e+1	5.6e-3	0.84	1.4e+2
1024	1.3e-2	0.86	1.4e+3	4.1e-3	0.92	3.3e+3

Table 1: Super teaching as n changes.

In the left half of Table 1 we report the median of the following quantities over 10 trials: $\hat{c}_n$, the fraction of the training items selected for super teaching $|B(S)|/n$, and the NEOS server running time.

The main result is that $\hat{c}_n \ll 1$ for all n, which means the solver indeed selects a super-teaching set $B(S)$ that is far

	Logistic Regression			Ridge Regression		
d	$\hat{c}_n$	$\lvert B(S)\rvert/n$	time (s)	$\hat{c}_n$	$\lvert B(S)\rvert/n$	time (s)
2	3.1e-3	0.67	5.4e-1	3.3e-3	0.55	6.6e+0
4	2.4e-3	0.44	8.5e+1	7.2e-3	0.53	5.8e+1
8	1.8e-1	0.39	4.1e+0	1.5e-1	0.47	6.0e+0
16	5.6e-1	0.42	5.1e+0	4.3e-1	0.59	9.3e+0
32	8.2e-1	0.58	1.0e+1	6.4e-1	0.86	3.0e+0

Table 2: Super teaching as d changes.

better than the original iid training set S. Therefore, MINLP is a valid algorithm for finding a super teaching set.

Second, we note that the solver tends to select a large subset since the median $|B(S)|/n \geq 1/2$. This is interesting as it is known that when S is dense, one can select extremely sparse super teaching sets, as small as a few items, to teach effectively [28]. Understanding the different regimes remains future work.

Finally, the running time grows fast with n. For example, when $n = 1024$ it takes around half an hour to solve (38). Future work needs to address this bottleneck in applying MINLP to large problems.

In the second set of experiments we fix $n = 32$ and vary $d = 2, 4, 8, 16, 32$. The left half of Table 2 shows the results. The empirical teaching ratio $\hat{c}_n$ is still below 1 in all cases, showing super teaching. But as the dimension of the problem increases $\hat{c}_n$ deteriorates toward 1. Nonetheless, even when $d = n$ we still see a median super teaching ratio of 0.82; the corresponding super teaching set $B(S)$ has only 58% training items than the dimension. It is interesting that the MINLP algorithm intentionally created a "high dimensional" learning problem (as in higher dimension d than selected training items $|B(S)|$) to achieve better teaching, knowing that the learner A_{lr} is regularized. The running time does not change dramatically.

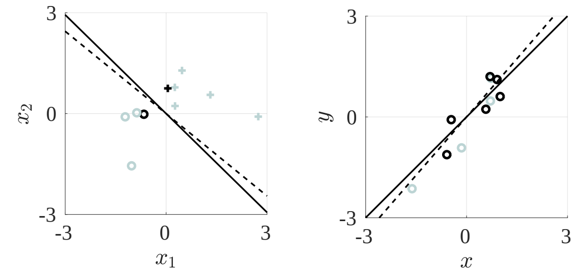
6.2 Teaching Ridge Regression A_{rr}

Let $\mathbb{X} = \mathbb{R}^d$, $\Theta = \mathbb{R}^d$, $\theta^* = (\frac{1}{\sqrt{d}}, ..., \frac{1}{\sqrt{d}})$, $p_{\mathbb{Z}}(x) = \mathcal{N}(0, I)$, $p_{\mathbb{Z}}(y \mid x) = \mathcal{N}(y; x^\top \theta^*, 0.1)$. Let the teaching risk be the parameter difference: $R(\hat{\theta}) = \|\hat{\theta} - \theta^*\|$. Given a sample S with n iid items drawn from $p_{\mathbb{Z}}$, ridge regression estimates $\hat{\theta}_S = A_{rr}(S)$ with $\lambda = 0.1$ and $\tilde{\ell}(z_i) = (x_i^\top \hat{\theta} - y_i)^2$. The corresponding MINLP is:

$$\min_{b \in \{0,1\}^n, \hat{\theta} \in \mathbb{R}^d} \quad ||\hat{\theta} - \theta^*|| \tag{39}$$

$$\text{s.t.} \quad \lambda \hat{\theta} + 2 \sum_{i=1}^{n} b_i (x_i^\top \hat{\theta} - y_i) x_i = 0.$$

We run the same set of experiments. Tables 1 and 2 show the results, which are qualitatively similar to teaching logistic regression. Again, we see the empirical super teaching ratio $\hat{c}_n \ll 1$, indicating the presence of super teaching.



(a) logistic regression (b) ridge regression

Figure 3: Typical trials from the MINLP algorithm

Finally, Figure 3 visualizes one typical trial each for teaching logistic regression and ridge regression. S consists of both dark and light points, while the dark ones representing $B(S)$ optimized by MINLP. The dashed line shows $\hat{\theta}_S$, while the solid lines shows $\hat{\theta}_{B(S)}$. The ground truth ($x_1 + x_2 = 0$ in logistic regression, $y = x$ in ridge regression) essentially overlaps with the solid lines. Specifically, the super taught models $\hat{\theta}_{B(S)}$ have negligible risks of 2.5e-4 and 3.3e-3, whereas models $\hat{\theta}_S$ trained from the whole iid sample S incur much larger risks of 0.03 and 0.16, respectively.

7 Related Work

There has been several research threads in different communities aimed at reducing a data set while maintaining its utility. The first thread is training set reduction [19, 43, 42], which during training time prunes items in S in an attempt to improve the learned model. The second thread is coresets [22, 6], a summary of S such that models learned on the summary are provably competitive with models learned on the full data set S. But as they do not know the target model $p_{\mathbb{Z}}$ or θ^*, these methods cannot truly achieve super teaching. The third thread is curriculum learning [11] which showed that smart initialization is useful for nonconvex optimization. In contrast, our teacher can directly encode the true model and therefore obtain faster rates. The final thread is sample compression [17], where a compression function chooses a subset $T \subset S$ and a reconstruction function to form a hypothesis. Our present work has some similarity with compression, which allows increased accuracy since compression bounds can be used as regularization [26].

The theoretical study of machine teaching has focused on the teaching dimension, i.e. the minimum training set size needed to exactly teach a target concept θ^* [20, 38, 46, 18, 29, 45, 16, 44, 48, 9, 3, 4, 21, 31, 8, 7, 25, 5, 36, 23, 10]. Most of the prior work assumed a synthetic teaching setting where S is the whole item space, which is often unrealistic. Liu *et al.* considered approximate teaching in the finite S setting [30], though their analysis focused on a specific SGD learner. Our super teaching setting applies to arbitrary learners, and we allow approximate teaching – namely we do not require the teacher to teach exactly the target model, which is infeasible in our pool-based teaching setting with a finite S.

Machine teaching applications include education [14, 33, 39, 27, 13, 34], computer security [2, 1, 32], and interactive machine learning [40, 12, 24]. By establishing the existence of super-teaching, the present paper can guide the process of finding a more effective training set for these applications.

8 Discussions and Conclusion

We presented super-teaching: when the teacher already knows the target model, she can often choose from a given training set a smaller subset that trains a learner better. We proved this for two learners, and provided an empirical algorithm based on mixed integer nonlinear programming to find a super teaching set.

However, much needs to be done on the theory of super teaching. We give two counterexamples to illustrate that not all learners are super-teachable.

Example 1 (MLE of interval)**.** *Let $\mathbb{X} = [0, \theta^*]$, where $\theta^* \in \mathbb{R}^+$. $p_{\mathbb{Z}}(x) = U(\mathbb{X})$. Given a n-item training set S, the MLE for θ^* is $\hat{\theta}_S = A_{int}(S) = \max_{i=1:n} x_i$. The risk is defined as $R(\hat{\theta}_S) = |\hat{\theta}_S - \theta^*|$. We show A_{int} is not super-teachable. $\hat{\theta}_{B(S)} = \max_{x_i \in B(S)} x_i \le \max_{x_i \in S} x_i = \hat{\theta}_S$. Since $\hat{\theta}_S \le \theta^*$, $R(\hat{\theta}_{B(S)}) = |\hat{\theta}_{B(S)} - \theta^*| \ge |\hat{\theta}_S - \theta^*| = R(\hat{\theta}_S)$.*

We can generalize this to a classification setting, and show that neither the least nor the greatest consistent hypothesis is not super-teachable:

Example 2 (Consistent learners)**.** *Let $\mathbb{X} = [x_{\min}, x_{\max}] \subset \mathbb{Z}$ be an interval over the integer grid. The hypothesis space is $\Theta = \{[a, b] \subseteq \mathbb{X} : y = 1 \text{ in } [a, b] \text{ and } -1 \text{ outside}\}$. $\theta^* = [a^*, b^*] \in \Theta$. $p_{\mathbb{Z}}$ is uniform on $\mathbb{X}$ and noiseless y labeled according to θ^*. The risk $R(\hat{\theta}_S)$ is the size of the symmetric difference between the two intervals $\hat{\theta}_S$ and θ^*, normalized by $x_{\max} - x_{\min}$. Given a sample S, the least consistent learner A_{lc} learns the tightest interval over positive items in S: $\hat{\theta}_S^{lc} = A_{lc}(S) \triangleq \left[\min_{\substack{i=1:n\\y_i=1}} x_i, \max_{\substack{i=1:n\\y_i=1}} x_i\right]$. $\hat{\theta}_S^{lc} = \emptyset$ if S does not contain positive items. The greatest consistent learner A_{gc} extends the hypothesis interval in both directions as much as possible before hitting negative points in S. If S has no positive we define $\hat{\theta}_S^{gc} = \emptyset$, too.*

Proposition 11. *Neither A_{lc} nor A_{gc} is super-teachable.*

Proof. We first show A_{lc} is not super-teachable. Note that A_{lc} learns the tightest interval consistent with S, thus we always have $\hat{\theta}_S^{lc} \subseteq \theta^*$. Now we show that $\hat{\theta}_{B(S)}^{lc} \subseteq \hat{\theta}_S^{lc}$ is always true so that $R(\hat{\theta}_S^{lc}) \le R(\hat{\theta}_{B(S)}^{lc})$ follows.

If $\theta^* = \emptyset$, then trivially $\hat{\theta}_{B(S)}^{lc} = \hat{\theta}_S^{lc} = \emptyset$.

Now assume $\theta^* \neq \emptyset$. If $\exists (x, 1) \in B(S)$, let $[a_1, b_1] = \hat{\theta}_{B(S)}^{lc}$. Note that $\hat{\theta}_S^{lc} \neq \emptyset$ because $B(S) \subseteq S$ and thus S has at least one positive point. Let $\hat{\theta}_S^{lc} = [a_2, b_2]$. Now $a_1 = \min\{x \mid (x, 1) \in B(S)\} \ge \min\{x \mid (x, 1) \in S\} = a_2$, and $b_1 = \max\{x \mid (x, 1) \in B(S)\} \le \max\{x \mid (x, 1) \in S\} = b_2$. Thus we have $\hat{\theta}_{B(S)}^{lc} \subseteq \hat{\theta}_S^{lc}$. If $\nexists (x, 1) \in B(S)$, $\hat{\theta}_{B(S)}^{lc} = \emptyset$ and $\hat{\theta}_{B(S)}^{lc} \subseteq \hat{\theta}_S^{lc}$ is always true.

Thus $\hat{\theta}_{B(S)}^{lc} \subseteq \hat{\theta}_S^{lc} \subseteq \theta^*$ for any B and any S.

The proof for A_{gc} is similar by showing $\theta^* \subseteq \hat{\theta}_S^{gc} \subseteq \hat{\theta}_{B(S)}^{gc}$. □

This leads to an open question: which family of learners are super teachable? We offer a conjecture here: we speculate that MLEs (and the derived MAP estimates or regularized empirical risk minimizers) which satisfy the asymptotic normality conditions [41] are super teachable. This conjecture is motivated by its similarity to the proof in section 3. Also note that the two counterexamples are classic examples of MLE that do *not* satisfy the asymptotic normality conditions.

Another open question concerns the optimal super-teaching subset size k for a given training set of size n. For example, our result on teaching the MLE of Gaussian mean indicates that the rate improves as k grows. However, our analysis only applies to a fixed k. Further research is needed to identify the optimal k.

Acknowledgments: R.N. acknowledges support by NSF IIS-1447449 and CCF-1740707. P.R. is supported in part by grants NSF DMS-1712596, NSF DMS-TRIPODS-1740751, DARPA W911NF-16-1-0551, ONR N00014-17-1-2147 and a grant from the MIT NEC Corporation. X.Z. is supported in part by NSF CCF-1704117, IIS-1623605, CMMI-1561512, DGE-1545481, and CCF-1423237.

References

[1] S. Alfeld, X. Zhu, and P. Barford. Data poisoning attacks against autoregressive models. *AAAI*, 2016.

[2] S. Alfeld, X. Zhu, and P. Barford. Explicit defense actions against test-set attacks. In *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[3] D. Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

[4] D. Angluin and M. Krikis. Teachers, learners and black boxes. *COLT*, 1997.

[5] D. Angluin and M. Krikis. Learning from different teachers. *Machine Learning*, 51(2):137–163, 2003.

[6] O. Bachem, M. Lucic, and A. Krause. Practical Coreset Constructions for Machine Learning. *ArXiv e-prints*, Mar. 2017.

[7] F. J. Balbach. Measuring teachability using variants of the teaching dimension. *Theor. Comput. Sci.*, 397(1-3):94–113, 2008.

[8] F. J. Balbach and T. Zeugmann. Teaching randomized learners. *COLT*, pages 229–243, 2006.

[9] F. J. Balbach and T. Zeugmann. Recent developments in algorithmic teaching. In *Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, pages 1–18, 2009.

[10] S. Ben-David and N. Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998.

[11] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In L. Bottou and M. Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, Montreal, June 2009. Omnipress.

[12] M. Cakmak and M. Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI*, 2012.

[13] M. Cakmak and A. Thomaz. Mixed-initiative active learning. *ICML Workshop on Combining Learning Strategies to Reduce Label Cost*, 2011.

[14] B. Clement, P.-Y. Oudeyer, and M. Lopes. A comparison of automatic teaching strategies for heterogeneous student populations. In *Educational Data Mining (EDM)*, 2016.

[15] J. Czyzyk, M. P. Mesnier, and J. J. Moré. The NEOS server. *IEEE Computational Science and Engineering*, 5(3):68–75, 1998.

[16] T. Doliwa, G. Fan, H. U. Simon, and S. Zilles. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15:3107–3131, 2014.

[17] S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[18] Z. Gao, C. Ries, H. U. Simon, and S. Zilles. Preference-based teaching. *Journal of Machine Learning Research*, 18(31):1–32, 2017.

[19] S. Garcia, J. Derrac, J. Cano, and F. Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012.

[20] S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and Systems Sciences*, 50(1):20–31, 1995.

[21] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *Journal of Computer and Systems Sciences*, 52(2):255–267, 1996.

[22] S. Har-Peled. *Geometric approximation algorithms*, volume 173. American mathematical society Boston, 2011.

[23] T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the eighth Annual Conference on Computational Learning Theory (COLT)*, pages 108–117, 1995.

[24] F. Khan, X. Zhu, and B. Mutlu. How do humans teach: On curriculum learning and teaching dimension. *NIPS*, 2011.

[25] H. Kobayashi and A. Shinohara. Complexity of teaching by a restricted number of examples. *COLT*, pages 293–302, 2009.

[26] A. Kontorovich, S. Sabato, and R. Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. *arXiv preprint arXiv:1705.08184*, 2017.

[27] R. Lindsey, M. Mozer, W. J. Huggins, and H. Pashler. Optimizing instructional policies. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2778–2786. 2013.

[28] J. Liu and X. Zhu. The teaching dimension of linear learners. *Journal of Machine Learning Research*, 17(162):1–25, 2016.

[29] J. Liu, X. Zhu, and H. G. Ohannessian. The teaching dimension of linear learners. In *The 33rd International Conference on Machine Learning (ICML)*, 2016.

[30] W. Liu, B. Dai, J. M. Rehg, and L. Song. Iterative machine teaching. In *ICML*, 2017.

[31] H. D. Mathias. A model of interactive teaching. *J. Comput. Syst. Sci.*, 54(3):487–501, 1997.

[32] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. *AAAI*, 2015.

[33] K. Patil, X. Zhu, L. Kopec, and B. C. Love. Optimal teaching for limited-capacity human learners. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[34] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by POMDP planning. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, AIED'11, pages 280–287, Berlin, Heidelberg, 2011. Springer-Verlag.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[36] R. L. Rivest and Y. L. Yin. Being taught can be faster than asking questions. *COLT*, 1995.

[37] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[38] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337—-348, 1991.

[39] A. Singla, I. Bogunovic, G. Bartok, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. In *ICML*, pages 154–162, 2014.

[40] J. Suh, X. Zhu, and S. Amershi. The label complexity of mixed-initiative classifier training. *International Conference on Machine Learning (ICML)*, 2016.

[41] H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[42] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, 2000.

[43] X. Zeng and X.-w. Chen. SMO-based pruning methods for sparse least squares support vector machines. *IEEE transactions on Neural Networks*, 16(6):1541–1546, 2005.

[44] X. Zhu. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1905–1913, 2013.

[45] X. Zhu. Machine teaching: an inverse problem to machine learning and an approach toward optimal education. *AAAI*, 2015.

[46] X. Zhu, J. Liu, and M. Lopes. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *The 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[47] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty. An Overview of Machine Teaching. *ArXiv e-prints*, Jan. 2018. https://arxiv.org/abs/1801.05927.

[48] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.

Supplemental Material

Lemma 7. *For the large margin classifier $\hat{\theta}_S$, we have*

$$\mathbf{P}\left[R(\hat{\theta}_S) > \epsilon\right] = \begin{cases} (1-\epsilon)^n + (\epsilon)^n & 0 < \epsilon \le \frac{1}{2} \\ (\frac{1}{2})^{n-1} & \frac{1}{2} < \epsilon < 1 \\ 0 & \epsilon = 1. \end{cases} \tag{29}$$

Proof. The risk is $R(\hat{\theta}_S) = |\hat{\theta}_S|$. Define event $E : \{\exists(x,-1) \in S \wedge \exists(x,+1) \in S\}$. $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right]$ can be decomposed into two components depending on if E happens as (40) shows.

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right] = \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] + \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right]. \tag{40}$$

$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right] = \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E^c\right]\mathbf{P}\left[E^c\right]$. Note that $\mathbf{P}\left[E^c\right] = (\frac{1}{2})^{n-1}$, $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E^c\right]$ is 0 if $\epsilon < 1$ and 1 if $\epsilon = 1$ because $\hat{\theta}_S = \pm 1$ always holds given E^c happens. Thus,

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right] = \begin{cases} 0 & \text{if } \epsilon < 1 \\ (\frac{1}{2})^{n-1} & \text{if } \epsilon = 1. \end{cases} \tag{41}$$

Now we compute $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right]$. Let n_+ be the number of positive points in S. Define $E_i : \{n_+ = i\}$. Note $E_i \cap E_j = \emptyset$ if $i \ne j$ and $E = \cup_{i=1}^{n-1} E_i$, thus

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E_i\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right]\mathbf{P}\left[E_i\right]. \tag{42}$$

$\mathbf{P}\left[E_i\right] = C_i^n(\frac{1}{2})^n$. Note that $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right] = \mathbf{P}\left[|\frac{x_- + x_+}{2}| \le \epsilon \mid E_i\right]$. To compute it, we first compute $F_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = \mathbf{P}\left[-x_- \le \epsilon_1, x_+ \le \epsilon_2 \mid E_i\right]$. Given E_i happens, $\mathbf{P}\left[-x_- \le \epsilon_1 \mid E_i\right] = 1 - (1-\epsilon_1)^{n-i}$ and $\mathbf{P}\left[x_+ \le \epsilon_2 \mid E_i\right] = 1 - (1-\epsilon_2)^i$. Also since $-x_- \le \epsilon_1$ and $x_+ \le \epsilon_2$ are independent given E_i happens, thus

$$F_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = \mathbf{P}\left[-x_- \le \epsilon_1, x_+ \le \epsilon_2 \mid E_i\right] = [1-(1-\epsilon_1)^{n-i}][1-(1-\epsilon_2)^i]. \tag{43}$$

Take the derivative of F gives

$$f_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}. \tag{44}$$

Note that $|\hat{\theta}_S| \le \epsilon \Leftrightarrow |-x_1 - x_2| \le 2\epsilon$. Therefore, we integrate $f_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i)$ over the region $|\epsilon_1 - \epsilon_2| \le 2\epsilon$ to obtain $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right]$. However, note that $0 \le \epsilon_1, \epsilon_2 \le 1$, thus for $\epsilon > \frac{1}{2}$, the region $|\epsilon_1 - \epsilon_2| \le 2\epsilon$ becomes the whole $[0,1] \times [0,1]$ and the integration is 1. Then (42) becomes $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[E_i\right] = \mathbf{P}\left[E\right] = 1 - (\frac{1}{2})^{n-1}$. For $\epsilon \le \frac{1}{2}$, by (42) we have

$$\begin{aligned} \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= \sum_{i=1}^{n-1}\mathbf{P}\left[E_i\right]\int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1 \\ &= \sum_{i=1}^{n-1} C_i^n(\frac{1}{2})^n\int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1 \\ &= (\frac{1}{2})^n\int_{|\epsilon_1-\epsilon_2|\le 2\epsilon}\sum_{i=1}^{n-1} C_i^n i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1. \end{aligned} \tag{45}$$

Note that $C_i^n i(n-i) = n(n-1)C_{i-1}^{n-2}$, (45) becomes

$$\begin{aligned}
\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} \sum_{i=1}^{n-1} C_{i-1}^{n-2}(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} \sum_{i=0}^{n-2} C_{i}^{n-2}(1-\epsilon_1)^{n-2-i}(1-\epsilon_2)^{i} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n [\int_{[0,1]\times[0,1]} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 - \int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1].
\end{aligned} \tag{46}$$

Now we compute the two integration in (46)

$$\begin{aligned}
&\int_{[0,1]\times[0,1]} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^1 [-\frac{1}{n-1}(2-\epsilon_1-\epsilon_2)^{n-1}|_0^1] d\epsilon_1 \\
&= \int_0^1 \frac{1}{n-1}[(2-\epsilon_1)^{n-1} - (1-\epsilon_1)^{n-1}] d\epsilon_1 = [-\frac{1}{n(n-1)}(2-\epsilon)^n + \frac{1}{n(n-1)}(1-\epsilon_1)^n] \;|_0^1 = \frac{2^n-2}{n(n-1)}.
\end{aligned} \tag{47}$$

For the second integration, note that it can decomposed as

$$\int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^{1-2\epsilon}\int_{\epsilon_1+2\epsilon}^{1} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 + \int_{2\epsilon}^{1}\int_0^{\epsilon_1-2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1. \tag{48}$$

Since the two sub-integration's are identical because the two sub regions are symmetric. We only show the computation for the first.

$$\begin{aligned}
&\int_0^{1-2\epsilon}\int_{\epsilon_1+2\epsilon}^{1} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^{1-2\epsilon} [-\frac{1}{n-1}(2-\epsilon_1-\epsilon_2)^{n-1}|_{\epsilon_1+2\epsilon}^1] d\epsilon_1 \\
&= \int_0^{1-2\epsilon} [-\frac{1}{n-1}(1-\epsilon_1)^{n-1} + \frac{2^{n-1}}{n-1}(1-\epsilon_1-\epsilon)^{n-1}] d\epsilon_1 \\
&= [\frac{1}{n(n-1)}(1-\epsilon_1)^n - \frac{2^{n-1}}{n(n-1)}(1-\epsilon_1-\epsilon)^n] \;|_0^{1-2\epsilon} \\
&= \frac{2^{n-1}}{n(n-1)}[\epsilon^n + (1-\epsilon)^n] - \frac{1}{n(n-1)}.
\end{aligned} \tag{49}$$

Thus we have

$$\begin{aligned}
&\int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = 2\int_0^{1-2\epsilon}\int_{\epsilon_1+2\epsilon}^{1} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 \\
&= \frac{2^n}{n(n-1)}[\epsilon^n + (1-\epsilon)^n] - \frac{2}{n(n-1)}.
\end{aligned} \tag{50}$$

Combine (47) and (50), we can compute (46) as follows.

$$\begin{aligned}
\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= = n(n-1)(\frac{1}{2})^n [\frac{2^n-2}{n(n-1)} - \frac{2^n}{n(n-1)}(\epsilon^n + (1-\epsilon)^n) + \frac{2}{n(n-1)}] \\
&= \frac{2^n-2}{2^n} - \epsilon^n - (1-\epsilon)^n + (\frac{1}{2})^{n-1} \\
&= 1 - \epsilon^n - (1-\epsilon)^n
\end{aligned} \tag{51}$$

Therefore we have

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \begin{cases} 1 - \epsilon^n - (1-\epsilon)^n & \text{if } \epsilon \le \frac{1}{2} \\ 1 - (\dfrac{1}{2})^{n-1} & \text{if } \frac{1}{2} < \epsilon \le 1. \end{cases} \tag{52}$$

Now combine (41) and (52) we have

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right] = \begin{cases} 1 - \epsilon^n - (1-\epsilon)^n & \text{if } \epsilon \le \frac{1}{2} \\ 1 - (\dfrac{1}{2})^{n-1} & \text{if } \frac{1}{2} < \epsilon < 1 \\ 1 & \text{if } \epsilon = 1. \end{cases} \tag{53}$$

which is equivalent to (29). $\square$

Lemma 9. *Let $n = 4m$, where m is an integer. Let S be an n-item iid sample drawn from $p_{\mathbb{Z}}$. $\forall \epsilon > 0, \forall \delta \in (0,1)$, $\exists \mathbb{M}(\epsilon, \delta) = \max\{\frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta}, (\frac{1}{\epsilon} \ln \frac{3}{\delta})^{\frac{1}{2}}\}$ such that $\forall m \ge \mathbb{M}(\epsilon, \delta)$, $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \le \epsilon\right] > 1 - \delta$.*

Proof. Let $S_1 = \{x \mid (x, 1) \in S\}$ and $S_2 = \{x \mid (x, -1) \in S\}$ respectively. Then we have $|S_1| + |S_2| = 4m$. Define event $E_1 : \{|S_1| \ge m \wedge |S_2| \ge m\}$. Then we have

$$\mathbf{P}\left[E_1\right] = 1 - 2\sum_{i=0}^{m-1} C_i^{4m} (\frac{1}{2})^{4m}. \tag{54}$$

where we rule out all possible sequences of $4m$ points which lead to $|S_1| < m$ or $|S_2| < m$. By standard result [37] (Lemma A.5) $\sum_{k=0}^{d} C_k^m \le (\frac{em}{d})^d$, we have

$$\mathbf{P}\left[E_1\right] \ge 1 - 2(\frac{4em}{m-1})^{m-1}(\frac{1}{2})^{4m} = 1 - \frac{1}{2}\frac{e^{m-1}}{4^m}(\frac{m}{m-1})^{m-1} \ge 1 - \frac{1}{2}(\frac{e}{4})^m \ge 1 - (\frac{e}{4})^m \tag{55}$$

where the 2nd-to-last inequality follows from the fact that $e \ge (1 + \frac{1}{m-1})^{m-1}$. Note that by definition $m \ge \frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta} > \frac{1}{\ln 4 - 1} \ln \frac{3}{\delta}$, thus $(\frac{e}{4})^m < \frac{\delta}{3}$ and $\mathbf{P}\left[E_1\right] > 1 - \frac{\delta}{3}$. Since $|S_1| + |S_2| = 4m$, then either $|S_1| \ge 2m$ or $|S_2| \ge 2m$. Without loss of generality we assume $|S_1| \ge 2m$. We then divide the interval [0, 1] equally into $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor$ segments. The length of each segment is $\frac{1}{N} = O(\frac{1}{m^2})$ as Figure 4 shows. Note that $m \ge \frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta} > 3e \ln \frac{3}{\delta}$, thus $N \ge \lfloor 3em \rfloor > 2em > m$.

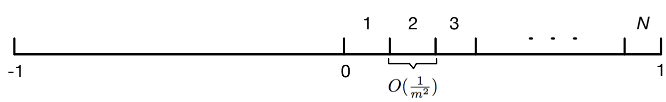


Figure 4: segments

Let N_o be the number of segments that are occupied by the points in S_1. Note that N_o is a random variable. Let E_2 be the event that $N_o \ge m$. Now we lower bound $\mathbf{P}\left[E_2\right]$. This is a variant of the coupon collector's problem: there are N distinct coupons, and in $|S_1|$ trials we want to collect at least m distinct coupons. Note that $\mathbf{P}\left[E_2\right] = 1 - \mathbf{P}\left[E_2^c\right] = 1 - \sum_{i=1}^{m-1} \mathbf{P}\left[N_o = i\right]$. Let T_i be the number of all possible coupon sequences of S_1 such that S_1 occupies exactly i segments (i.e. distinct coupons). We have C_i^n ways of choosing i segments among a total of N. Also, for each choice of i segments, the number of all possible coupon sequences of S_1 such that S_1 fully occupies those i segments without empty is upper bounded by $i^{|S_1|}$. Thus $T_i \le C_i^n i^{|S_1|}$ and we have

$$\mathbf{P}\left[N_o = i\right] = \frac{T_i}{N^{|S_1|}} \le C_i^n (\frac{i}{N})^{|S_1|}. \tag{56}$$

Since $m \ge \frac{3e}{\ln 4 - 1} \ln \frac{3}{\delta} > \log_2 \frac{3}{\delta}$, $|S_1| \ge 2m$, and $N > 2em$, thus

$$\begin{aligned} \mathbf{P}\left[E_2^c\right] &= \sum_{i=1}^{m-1} \mathbf{P}\left[N_o = i\right] \le \sum_{i=1}^{m-1} C_i^n (\frac{i}{N})^{|S_1|} \le \sum_{i=1}^{m-1} C_i^n (\frac{m}{N})^{2m} \\ &< \sum_{i=0}^{m} C_i^n (\frac{m}{N})^{2m} \le (\frac{eN}{m})^m (\frac{m}{N})^{2m} = (\frac{em}{N})^m < (\frac{em}{2em})^m = (\frac{1}{2})^m < \frac{\delta}{3}. \end{aligned} \tag{57}$$

Thus $\mathbf{P}\left[E_2\right] \geq 1 - \frac{\delta}{3}$. Applying union bound, $\mathbf{P}\left[E_1, E_2\right] \geq 1 - \frac{2\delta}{3}$.

Let E_3 be the following event: there exist a point x_2 in S_2 such that $-x_2$, the flipped point, lies in the same segment as some point x_1 in S_1. If E_3 happens, then $|x_1 + x_2| = |x_1 - (-x_2)| \leq \frac{1}{N}$. Note that $\mathbf{P}\left[E_3\right] \geq \mathbf{P}\left[E_1, E_2, E_3\right] = \mathbf{P}\left[E_3 \mid E_1, E_2\right] \mathbf{P}\left[E_1, E_2\right]$. Now we lower bound $\mathbf{P}\left[E_3 \mid E_1, E_2\right]$. Given E_1 and E_2 happen, we have $|S_2| \geq m$ and $N_o \geq m$. Since $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor \leq m^2 (\ln \frac{3}{\delta})^{-1}$, we have

$$\mathbf{P}\left[E_3^c \mid E_1, E_2\right] = (1 - \frac{N_o}{N})^{|S_2|} \leq (1 - \frac{m}{N})^{|S_2|} \leq (1 - \frac{m}{N})^m \leq e^{-\frac{m^2}{N}} \leq \frac{\delta}{3}. \tag{58}$$

Thus, $\mathbf{P}\left[E_3 \mid E_1, E_2\right] = 1 - \mathbf{P}\left[E_3^c \mid E_1, E_2\right] > 1 - \frac{\delta}{3}$. $\mathbf{P}\left[E_3\right] \geq \mathbf{P}\left[E_1, E_2, E_3\right] = \mathbf{P}\left[E_3 \mid E_1, E_2\right] \mathbf{P}\left[E_1, E_2\right] \geq (1 - \frac{\delta}{3})(1 - \frac{2\delta}{3}) > 1 - \delta$. Thus with probability at least $1 - \delta$, there exist $x_2 \in S_2$ and $x_1 \in S_1$ such that $|x_1 + x_2| \leq \frac{1}{N}$.

We now bound $\frac{1}{N}$. $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor \geq \frac{1}{2} m^2 (\ln \frac{3}{\delta})^{-1}$. Therefore $\frac{1}{N} \leq \frac{2}{m^2} \ln \frac{3}{\delta}$. Recall by definition $m \geq (\frac{1}{\epsilon} \ln \frac{3}{\delta})^{\frac{1}{2}}$, thus $\frac{1}{N} \leq 2\epsilon$.

We now have $|x_1 + x_2| \leq 2\epsilon$. Finally, since $\{s_-, s_+\}$ selected by teacher B_{ms} is the most symmetric pair, it must satisfy $|s_- + s_+| \leq |x_1 + x_2| \leq 2\epsilon$. Putting together, with probability at least $1 - \delta$, $R(\hat{\theta}_{B_{ms}(S)}) = \frac{1}{2}|s_- + s_+| \leq \epsilon$. $\square$