Learning from Synthetic Data: Limitations of ERM

Kareem Amin KAMIN@GOOGLE.COM
Google Research, New York

Alex Bie ALEXBIE@GOOGLE.COM
Google Research, New York

Weiwei Kong WEIWEIKONG@GOOGLE.COM
Google Research, New York

Umar Syed USYED@GOOGLE.COM
Google Research, New York

Sergei Vassilvitskii SERGEIV@GOOGLE.COM
Google Research, New York

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

The prevalence and low cost of LLMs have led to a rise of synthetic content. From review sites to court documents, "natural" content has been contaminated by data points that appear similar to natural data, but are in fact LLM-generated. In this work we revisit fundamental learning theory questions in this, now ubiquitous, setting. We model this scenario as a sequence of learning tasks where the input is a mix of natural and synthetic data, and the learning algorithms are oblivious to the origin of any individual example.

We study the possibilities and limitations of ERM in this setting. For the problem of estimating the mean of an arbitrary d-dimensional distribution, we find that while ERM converges to the true mean, it is outperformed by an algorithm that assigns non-uniform weights to examples from different generations of data. For the PAC learning setting, the disparity is even more stark. We find that ERM does not always converge to the true concept, echoing the model collapse literature. However, we show there are algorithms capable of learning the correct hypothesis for arbitrary VC classes and arbitrary amounts of contamination.

1. Introduction

Large language models (LLMs) represent cutting-edge advances of AI, and developers of LLMs routinely consume publicly-available datasets such as Common Crawl[1] to improve the capabilities of their models. The first generation of LLMs were largely trained on human-generated data. However, the success of LLMs and their increased adoption has had an unexpected consequence of AI-generated content appearing in places where there was previously none. Thus machine learning practitioners should be aware that there is an increased chance that their training data is contaminated by LLM-generated content.

Previous work has looked into the value of synthetic (*i.e.*, AI-generated) data, and showed that while naively adding this data to the training mix may lead to model collapse, being more diligent about which data is added, the amount of curation it undergoes, and the specifics of the training process may mitigate that risk, or reverse it, leading to improved performance. These works almost

1. http://commoncrawl.org

uniquely focus on the LLM setting, trying to improve state of the art performance on a set of benchmarks.

In contrast, in this work we take a traditional learning theory view on this problem. We begin by formalizing the setting and developing a framework that captures the invariants of having natural training data contaminated by synthetic additions. Specifically, we see three salient points:

- **Groundtruth.** There exists a (potentially small) set of natural data, coming from the true data generation distribution.
- **Contamination.** Natural data is repeatedly supplemented with synthetic data, which attempts to mimic or imitate the groundtruth. Moreover, one cannot be certain about the origin of any individual example, making synthetic data difficult to filter out.
- **Repetition.** The process is continuous, the training set grows over time, adding data contaminated with the latest model in each iteration.

Faced with this contaminated data, our goal is to find a learning algorithm which outputs models that continuously improve generalization error in every iteration.

1.1. Our Contributions

Parameterized Contamination. Learning in the presence of synthetic data has been studied theoretically by multiple authors (see Section 1.2). Many of these works consider an extreme, *purely recursive*, setting. Aside from an initial seed set of uncontaminated data, the learner is exposed to synthetic data in each subsequent iteration of learning that has been generated by the previous iteration's model. We extend and generalize these works by introducing a parameter α characterizing the degree of model contamination in each iteration, where $\alpha = 1$ corresponds to the pure recursive setting. The challenge in our setting will be that natural data and synthetic data are indistinguishable.

Mean estimation. Within this framework, we begin by studying the most fundamental statistical problem of estimating the mean of an unknown distribution. The sample average is well-known to be the empirical risk minimizer for l_2-error. In the synthetic data and model collapse literature this approach, gathering a dataset across multiple generations of models and weighing its examples uniformly, is well-studied and is sometimes known as an *augmentation workflow* (Dey and Donoho, 2024) or *uniform weighting*.

Our first result (Theorem 2) is an *exact* characterization of the variance of uniform weighting. The finding generalizes existing results along two key dimensions: the exact characterization holds for all contamination parameters α, and makes almost no distributional assumptions. Our second key result (Theorem 8) is that, for all distributions in our setting, there exists a non-trivial $(\alpha < 1)$ contamination parameter, where uniform weighting is *not* the minimum-variance unbiased estimate (MVUE) of the mean.

PAC learning. The mean estimation problem provides a glimpse into the kinds of problems synthetic data can introduce even in simple settings. To better understand the implications of synthetic data, we study this question in the PAC learning setting (Valiant, 1984). We give a lower bound that demonstrates that the natural algorithm, repeatedly computing the ERM hypothesis on the previous round's data, does not continually improve the generalization error of the classifier (Theorem 9). It is notable that the construction holds for a very simple problem: learning 1-d thresholds in the realizable setting, but the lower bound extends to VC classes of arbitrary dimensions.

Complementing the lower bound we construct two universal algorithms, for problems with arbitrary VC dimension, that, ignoring computational constraints, produce classifiers with vanishing generalization error (Theorems 12, 13) regardless of the amount of data contamination.

1.2. Related work

Beginning with the paper by Shumailov et al. (2024) that introduced the notion of "model collapse," researchers have shown how to avoid it by being careful about how and where to use synthetic data, and have provided some theoretical justifications for its success. Much of this work concerns studying the effect of synthetic data in the context of generative models and LLMs. See, for instance, the work by Alemohammad et al. (2024); Hataya et al. (2023); Gerstgrasser et al. (2024) on model collapse, and Amin et al. (2025); Bertrand et al. (2024); Seddik et al. (2024); Feng et al. (2024); Ferbach et al. (2024); Firdoussi et al. (2025) on ways to combine synthetic and natural data. Our work departs from these by studying the effect of contaminated synthetic data on two fundamental learning settings: *mean estimation*, and *PAC learning*.

Mean estimation. Inspired by the model collapse literature, several authors (Suresh et al., 2024; Dey and Donoho, 2024; Barzilai and Shamir, 2025; Kanabar and Gastpar, 2025), have considered theoretical learning problems relevant to mean estimation in settings with model-contaminated data. Dey and Donoho (2024) provide a useful taxonomy, distinguishing between the *augmentation workflow* where examples are combined uniformly across learning iterations, and the *discard workflow* where only the last iteration's examples are used for learning. To the best of our knowledge, our work is the first to exactly characterize the variance of mean estimation for a d-dimensional distribution in the augmentation workflow, with minimal distributional assumptions and arbitrary model contamination.

Most of the relveant work (Suresh et al., 2024; Dey and Donoho, 2024; Barzilai and Shamir, 2025) focuses on the purely recursive setting where the data produced on each iteration comes from the previous generation's model ($\alpha = 1$, in the language of this work). The work of Kanabar and Gastpar (2025) stands apart in considering general forms of data contamination ($\alpha \neq 1$). However, they consider only discrete distributions.

Within the purely-recursive literature, we can compare our results. Suresh et al. (2024) study the discard workflow, but for specific distributions: discrete distributions, 1-d Gaussians, and 1-d mixtures of Gaussians. Kazdan et al. (2024) characterize the variance of 1-d Gaussian mean estimation in the augmentation workflow. Barzilai and Shamir (2025) study the augmentation workflow in a setting that subsumes Gaussian mean estimation. However, their main result requires that sample sizes grow with the number of iterations of learning. In the special case of $\alpha = 1$, we recover the $\pi^2/6$ variance noted by Dey and Donoho (2024), who analyze Gaussian mean estimation. While Dey and Donoho (2024) note the benefits of the augmentation workflow over the discard workflow, our work goes a step further and establishes the sub-optimality of even the augmentation workflow.

PAC learning. To the best of our knowledge, we are the first to consider PAC learning in the context of learning from synthetic data. Valiant (1985) and Kearns and Li (1988) modified the PAC learning setting to include an adversary who arbitrarily perturbs a fraction of the training data, while Angluin and Laird (1988) introduced a random probability that the true label for each training example is flipped. Unlike those noise models — which can be described, respectively, as malicious

and oblivious — in our setting the source of the noise is the learner itself, via the models it outputs in previous iterations. Finally, our main results in the PAC setting make use of results of Angluin and Laird (1988); Liu et al. (2002) and Mansouri and Ben-David (2025) via reductions.

2. Mean Estimation

We begin with one of the simplest learning question: given a set of samples from a distribution, estimate the mean of the distribution. In the traditional setting, it is well known that among all unbiased estimates, the empirical mean of the samples minimizes the squared loss – the ℓ_2^2 norm between the estimate and the truth. For many distributions (e.g. Gaussian) it is also the minimum variance unbiased estimator (MVUE) of the mean. Our first results exactly characterize the variance of the empirical mean when data is contaminated with synthetic data, and demonstrate that there is no distribution for which it is the MVUE in general.

2.1. Setup

Suppose that the natural data is drawn from a d-dimensional distribution with mean μ. At each time $t = 1, 2, \ldots$, we will produce an estimate Y_t for μ. We are interested in studying a setting where this estimate contaminates the data produced in the subsequent round.

A simple version of this is as follows. Suppose that on round t we are given access to a pool of examples, where a $(1-\alpha)$ fraction comes from a distribution D_0 with mean μ and an α fraction comes from a distribution D_1 with mean Y_{t-1}. Let X_t denote the average of these examples, which has the same distribution as $\alpha Y_{t-1} + (1-\alpha)\mu + U$, where $U \sim D(\Sigma)$ and $D(\Sigma)$ has mean 0 and covariance Σ (that depends on α, D_0, D_1 and the size of the pool).

Traditionally, one would estimate μ by taking Y_t to be the average of all examples observed up to time t. Letting $U_t \sim D(\Sigma)$, this results in the following stochastic process:

$$\begin{aligned} X_1 = \mu + U_t, \quad X_t = \alpha Y_{t-1} + (1-\alpha)\mu + U_t \qquad (t \geq 2), \\ Y_t = \frac{1}{t}\sum_{s=1}^{t} X_s \qquad (t \geq 1). \end{aligned} \tag{1}$$

More generally, a learner may combine observations $\{X_t\}$ non-uniformly across rounds to form estimates $\{Y_t\}$. That is, for

$$w = (w^1, w^2, \ldots) \text{ for } w^s \in \mathbb{R}^s,$$

and X_t is as in (1), the learner could form the sequence

$$Y_t(w) = \sum_{s=1}^{t} w_s^t X_s(w) \qquad (t \geq 1). \tag{2}$$

It can be shown, inductively, that for any weighting strategy in the probability simplex, the set $\{Y_t(w)\}$ consists of unbiased estimators.

Remark 1 *If for every t, the vector w^t is in the probability simplex, then $\mathbb{E}[X_t(w)] = \mathbb{E}[Y_t(w^t)] = \mu$ for any $\alpha > 0$ and $t \geq 1$.*

Moreover, we are interested in minimizing $\mathrm{Var}(Y_t(w))$, since:

$$\begin{aligned}\mathbb{E}\left[\|Y_t(w)-\mu\|_2^2\right] &= \mathbb{E}\left[\mathrm{tr}\left(\left[Y_t(w)-\mu\right]\left[Y_t(w)-\mu\right]^*\right)\right] \\ &= \mathrm{tr}\mathbb{E}\left[\left[Y_t(w)-\mu\right]\left[Y_t(w)-\mu\right]^*\right] = \mathrm{tr}\ \mathrm{Var}\left[Y_t(w)\right].\end{aligned}$$

2.2. Uniform Weights

Our first result characterizes the variance of Y_t in equation (1) with equality. The proof is straightforward; we express the variance of Y_t in terms of the variance of Y_{t-1}, establishing the recursion $\mathrm{Var}(Y_t) = \left(\frac{t-1+\alpha}{t}\right)^2 \mathrm{Var}(Y_{t-1}) + \frac{1}{t^2}\Sigma$. We then solve this recursion in terms of the gamma function. The full proof is provided in Appendix A.1.

Theorem 2 *It holds that*

$$\mathrm{Var}(Y_t) = \left\{\frac{1}{t^2} + \left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2\right\}\Sigma.$$

Before analyzing this expression further, we can already specialize Theorem 2 to the edge cases $\alpha \in \{0,1\}$ using the elementary recursion $\Gamma(z+1) = z\Gamma(z)$ for $z>0$. When $\alpha = 0$, we recover that the variance of the sample mean on i.i.d. data decays like $O(1/t)$. When $\alpha = 1$ we instantiate a pure recursive learning setting, where the learner has access to natural data only on one round, and subsequently re-trains on a dataset augmented by its own predictions. This results in a far worse estimator, which has constant error regardless of how many rounds elapse, recovering the $\pi^2/6$ constant observed by Dey and Donoho (2024).

Remark 3 *We have the following edge cases. For $\alpha = 0$,*

$$\mathrm{Var}(Y_t) = \left(\frac{1}{t^2} + \left[\frac{\Gamma(t)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k)}\right]^2\right)\Sigma = \frac{1}{t}\Sigma.$$

For $\alpha = 1$,

$$\mathrm{Var}(Y_t) = \left(\frac{1}{t^2} + \left[\frac{\Gamma(t+1)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+1)}\right]^2\right)\Sigma = \left(\sum_{k=1}^{t}\frac{1}{k^2}\right)\Sigma \overset{t\to\infty}{\to} \frac{\pi^2}{6}\Sigma.$$

We can also use Gautschi's inequality to further analyze the expression for $\mathrm{Var}(Y_t)$. Remarkably, we can characterize the covariance of this vector up to constant factors. The full proof is in Appendix A.1.

Lemma 4 *(Gautschi's inequality) For $z>0$ and $\lambda \in (0,1)$, it holds that*

$$z^{1-\lambda} \le \frac{\Gamma(z+1)}{\Gamma(z+\lambda)} \le (z+1)^{1-\lambda}.$$

Theorem 5 *For $t \ge 3$, we have that*

$$\frac{1}{2}\left[\frac{1}{t} + \frac{1}{t^2} + \frac{1}{t^{2(1-\alpha)}}\right]\Sigma \preceq \mathrm{Var}(Y_t) \preceq 4\left[\frac{1}{t} + \frac{1}{t^2} + \frac{1}{t^{2(1-\alpha)}}\right]\Sigma. \tag{3}$$

where $\preceq$ indicates Loewner ordering.

Theorem 3 allows us to make multiple observations. First, when $\alpha \leq 1/2$, i.e. the data contains more natural examples than synthetic examples, the variance is $\Theta(1/t)$. At $\alpha = 1/2$ there is a phase change, and the dominant term becomes $\Theta(t^{2(1-\alpha)})$ when $\alpha > 1/2$.

We also observe that for any distribution $D(\Sigma)$ where the sample average is the MVUE for μ in the standard i.i.d. setting (such as the Gaussian distribution), there is no α and alternative weighting strategy w that can enjoy a rate that is better than $\Theta(1/t)$ in t. The proof follows by contradiction. Suppose that a $o(1/t)$ strategy existed. Then, it can be used to attain the same rate for $\alpha = 0$ (i.e. the standard i.i.d. setting). This follows by receiving i.i.d. samples $\mu + U_t$, for each round t, and simulating the recursion in (2) with the weighting strategy w. This, in turn, contradicts the fact that uniform weighting is the MVUE in the standard i.i.d. setting.

Theorem 6 *Suppose the sample average is the MVUE for estimating μ from i.i.d. data $\{X_t \sim \mu + U_t\}$, $U_t \sim D(\Sigma)$. Then there is no α, weighting scheme w, and function $f_\alpha(t)$ such that $\mathrm{Var}(Y_t(w)) \preceq f_\alpha(t)\Sigma$ and $f_\alpha(t) = o(1/t)$.*

Although uniform weighting sometimes achieves the optimal dependence on t when $\alpha \leq 1/2$, we can also show that uniform weighting is *not* the MVUE in general; in particular, it is not the MVUE for large α. We remark that when $\alpha = 1$, there is a separation between the uniform weighting strategy and the simple strategy that assigns, for every t, $w^t_{\text{simple}} = (1, 0, 0, \ldots, 0)$, which follows from Remark 3.

Remark 7 *For $\alpha = 1$, and sufficiently large t, $\mathrm{Var}(Y_t(w_{\text{simple}})) = \Sigma \prec \mathrm{Var}(Y_t(w_{\text{uniform}}))$.*

The next theorem investigates whether the sub-optimality of uniform weighting is a phenomenon that happens only in the degenerate case when $\alpha = 1$. We show that this is false. There are non-trivial $\alpha \in (0, 1)$ for which uniform weighting is also not the MVUE. For the following theorem, let $Y^\alpha_t(w)$ indicate the estimator under weighting scheme w and data-contamination parameter α.

Theorem 8 *There exists a family of weighting schemes $\{w_\alpha\}_{\alpha\in[0,1]}$ such that there exists an $\alpha^* < 1$ where $\mathrm{Var}(Y^\alpha_t(w_\alpha)) \prec \mathrm{Var}(Y^\alpha_t(w_{\text{uniform}}))$ for every $\alpha \in (\alpha^*, 1]$.*

The proof follows by observing that uniform weighting is not optimal for $\alpha = 1$. We then consider the family of strategies $\hat{w}_\alpha$ that upweight data from the first round, by taking $w^1_\alpha = 1$ and:

$$\hat{w}^{s+1}_\alpha = \left(\tfrac{1}{1+(s-1)(1-\alpha)}, \tfrac{1-\alpha}{1+(s-1)(1-\alpha)}, \ldots \tfrac{1-\alpha}{1+(s-1)(1-\alpha)}\right).$$

We can show that the variance of $Y^\alpha_t(\hat{w}_\alpha)$ is a continuous function in α from which we can conclude that there are non-trivial $\alpha \in (0, 1)$ such that $\hat{w}_\alpha$ has lower variance than uniform weighting. The full proof can be found in Appendix A.2.

3. PAC Learning

Our mean estimation results demonstrate that ERM is not an optimal solution for even a simple distributional estimation problem. We now study the same question from the perspective of classification in the PAC setting. We show that for more complex settings, synthetic data can wreak havoc on traditional learning approaches, even when it's mixed with natural data at each iteration.

One of the most powerful theorems in classical learning theory is that ERM guarantees a generalization rate for concept classes with finite VC dimension. Moreover, the excess error of the ERM classifier over the true optimal classifier in a hypothesis class is vanishing with more data. Our first main theorem in this section will demonstrate that, in general, this is not true for settings where data labeled according to previous hypotheses are mixed in with data labeled according to nature. For a very simple concept class, linear separators, repeatedly finding the empirical risk minimizer can result in generalization error that stalls. Mirroring the mean estimation setting, this occurs once the contamination rate of the data surpasses the critical point $\alpha > \frac{1}{2}$.

Subsequently, we will show that there are other algorithms (not repeated ERM) that achieve vanishing generalization error for concept classes with finite VC dimension regardless of the data-contamination rate.

3.1. Setup

Let $\mathcal{X}$ be the domain of examples, and consider the label space $\{0, 1\}$. Let $F \subseteq \{0, 1\}^{\mathcal{X}}$ be a hypothesis class. We are interested in learning algorithms that are called as part of an iterative training loop, and whose output can depend on the hypotheses that were output in previous iterations.

Accordingly, we define a learning algorithm to be a (possibly randomized) function $A : 2^{\mathcal{X}\times\{0,1\}} \times (\{0, 1\}^{\mathcal{X}})^* \to \{0, 1\}^{\mathcal{X}}$. In other words, a learning algorithm A takes as input a dataset $S \subseteq \mathcal{X} \times \{0, 1\}$ and a sequence of hypotheses $f_0, \ldots, f_{t-1} \in F$, and outputs a (possibly randomized) hypothesis $f_t = A(S, f_0, \ldots, f_{t-1})$.

Let D be a distribution on $\mathcal{X}$. In this paper, we consider the *realizable* case, where some $f^* \in F$ is the true concept that labels examples. The *loss* of a hypothesis $f \in \{0, 1\}^{\mathcal{X}}$ is defined

$$L(f) := \Pr_{x\sim D}[f(x) \neq f^*(x)].$$

For any dataset $S \subseteq \mathcal{X} \times \{0, 1\}$ the *empirical loss* of hypothesis $f \in F$ on S is defined

$$\hat{L}(f, S) = \Pr_{(x,y)\sim \mathrm{Unif}(S)}[f(x) \neq y].$$

A common learning algorithm is *empirical risk minimization*, denoted A_{ERM}, and defined

$$A_{\mathrm{ERM}}(S, f_0, \ldots, f_{t-1}) = \arg\min_{f\in F} \hat{L}(f, S).$$

Note that ERM ignores the previous hypotheses and depends only on the dataset.

Algorithm 1 formalizes our learning problem. Learning algorithm A is used to estimate an initial hypothesis f_0 from dataset $\bar{S}_0$, where each sample in $\bar{S}_0$ is drawn from D, and its label is given by the true concept f^*. In each subsequent round $t \geq 1$, a dataset S_t is added to $\bar{S}_{t-1}$ to form $\bar{S}_t$, where again each sample in S_t is drawn from D, but its label is given by f_{t-1} with probability $\alpha \in [0, 1]$ and otherwise given by f^*. Learning algorithm A is then used to estimate the next hypothesis f_t from $\bar{S}_t$ and $f_0, \ldots, f_{t-1}$.

Problem Statement. Our **goal** is to specify a learning algorithm A such that $\lim_{t\to\infty} \mathrm{E}[L(f_t)] = 0$, where the expectation is with respect to the randomness in Algorithm 1 (which includes the randomness in A).

Algorithm 1: Recursive learning

Input: Sample space $\mathcal{X}$, label space $\{0,1\}$, distribution D on $\mathcal{X}$, function class $F \subseteq \{0,1\}^{\mathcal{X}}$, true concept $f^* \in F$, sample size n, synthetic data rate $\alpha \in [0,1]$, learning algorithm A.

$\bar{S}_0 \leftarrow \{(x_i^0, y_i^0) : i \in [n], x_i^0 \sim D, y_i \sim f^*(x_i)\}$
$f_0 \leftarrow A(\bar{S}_0)$
for $t = 1, 2, 3, \ldots$ **do**
  for $i = 1, \ldots, n$ **do**
    $x_i^t \sim D$
    $b_i^t \sim \text{Bernoulli}(\alpha)$
    $y_{\text{model}} \sim f_{t-1}(x_i^t), y_{\text{true}} \sim f^*(x_i^t)$
    $y_i^t = b_i^t y_{\text{model}} + (1 - b_i^t) y_{\text{true}}$
  end
  $S_t \leftarrow \{(x_i^t, y_i^t) : i \in [n]\}$
  $\bar{S}_t \leftarrow \bar{S}_{t-1} \cup S_t$
  $f_t \leftarrow A(\bar{S}_t, f_0, \ldots, f_{t-1})$
end

Remark on set notation We use the variable S to define sets that we actually want to think of as sequences, *i.e.*, they preserve multiplicity and order. We denote them as sets because doing so affords us the convenience of using the element operator $\in$, the union operator $\cup$, and set builder notation.

3.2. Repeated ERM and Threshold Functions

Our first theorem shows that repeated ERM does not attain our goal in general. While the proof is somewhat technical, its key ideas are straight-forward. Consider a very simple concept class: learning one dimensional thresholds on the interval. Define D according to the following p.m.f:

$$p(x) = \begin{cases} \frac{1}{2} - \frac{1}{2n} & \text{if } x = +1 \\ \frac{1}{n} & \text{if } x = 0 \\ \frac{1}{2} - \frac{1}{2n} & \text{if } x = -1 \end{cases}$$

There is a constant probability $(1 - \frac{1}{n})^n$ that the learner does not see the example $x = 0$ in the 0-th round of learning. Suppose this happens. In the absence of any information on how this example should be labeled, the learner mislabels it with constant probability.

Next, as long as $\bar{S}_t$ contains more mis-labeled than correctly labeled examples for $x = 0$, ERM will continue to mislabel it. Thus, the only way for the learner to recover is to see more naturally-labeled examples than model-labeled examples for the example $x = 0$.

The final element of the proof is to consider a process $\{\Delta_k\}$ which increments whenever the learner encounters a model-provided label and decrements whenever the learner encounters a nature-provided label for $x = 0$. This process can be understood as a biased random walk on the integers. Thus, the probability that the learner never recovers can be directly related to the probability that the biased random walk is transient. Finally, the one-dimensional problem can be embedded in

a linear separator of arbitrary dimension, resulting in Theorem 9. The full proof is provided in Appendix B.1.

Let $\{W_t\}$ denote a random walk with bias parameter α. Let $W_0 = 0$, $W_{t+1} = W_t + 1$ with probability α and $W_{t+1} = W_t - 1$ with probability $(1 - \alpha)$.

Theorem 9 *Define $c^*_\alpha = \Pr[\forall t \geq 1, W_t \geq 1]$. For any sample size $n \geq 2$ and $d \geq 2$, there exists a distribution D, a hypothesis class F with VC dimension d, and $f^* \in F$, such that if $A = A_{\mathrm{ERM}}$ then*

$$\liminf_{t\to\infty} \mathrm{E}[L(f_t)] \geq \frac{c^*_\alpha}{8n}.$$

The following Lemma establishes that for $\alpha > 1/2$, $c^*_\alpha > 0$. It is a fairly standard result in probability that a biased random walk is not recurrent. However, we provide the full proof in the Appendix for completeness.

Lemma 10 *When $\alpha > 1/2$, the event that the random walk $\{W_t\}$ stays positive forever occurs with non-zero probability: $\mathbb{P}[\forall t \geq 1, W_t \geq 1] > 0$*

Corollary 11 *For any sample size $n \geq 2$ and $d \geq 2$, there exists a distribution D, a hypothesis class F with VC dimension d, and $f^* \in F$, and $\alpha > 1/2$ such that if $A = A_{\mathrm{ERM}}$ then*

$$\liminf_{t\to\infty} \mathrm{E}[L(f_t)] > 0$$

3.3. Simple Universal Algorithm

The next theorem establishes that there exists a learning algorithm that achieves our goal for any VC class, and any α. The algorithm requires access to f_{uniform}: a classifier that ignores its input and returns a random label. f_{uniform} is used to occasionally collect labels that are distributed like $f^*(x)$ with random classification noise added. On other rounds, we rely on classical learning theory results (Angluin and Laird, 1988) to learn with label noise. We show that by tuning the probability of playing the uniform classifier, vanishing generalization error can be achieved.

The algorithm has a number of drawbacks. Its use of the classifier f_{uniform} means that the algorithm occasionally deploys a classifier with maximum loss. While the probability of deploying the random classifier vanishes, this is nevertheless unrealistic and unnatural. More importantly, its generalization error vanishes at the relatively slow rate of $O(t^{-1/4})$.

Nevertheless, we include this result as it serves as a simple witness, with an easy proof, that moving away from repeated ERM allows one to learn in this setting. Our next result requires a more complicated algorithm, but achieves a more standard $O(t^{-1/2})$ rate.

Theorem 12 *Let $F \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension d. There exists a learning algorithm A outputting classifiers f_t such that for all $t \geq 0$*

$$\mathrm{E}[L(f_t)] \leq O\left(\frac{\sqrt{d\log(nt)}}{(1-\alpha)(nt)^{1/4}}\right) + \frac{1}{\sqrt{n(t+1)}} + \exp(-\Omega(\sqrt{t/n})).$$

Proof For any $t \geq 0$, given dataset $\bar{S}_t$ and hypotheses $f_0, \ldots, f_{t-1}$ as input, we define the output $f_t = A(\bar{S}_t, f_0, \ldots, f_{t-1})$ of learning algorithm A as follows. With probability $\frac{1}{\sqrt{n(t+1)}}$ let $f_t = f_{\mathrm{uniform}}$. Otherwise, let

$$\bar{S}'_t = \{(x_i, y_i) \in S_r : f_{r-1} = f_{\mathrm{uniform}} \text{ for } r \in [t]\}$$

and $f_t = A_{\mathrm{ERM}}(\bar{S}'_t, f_0, \ldots, f_{t-1})$. In other words, with probability $1 - \frac{1}{\sqrt{n(t+1)}}$, let f_t be the empirical risk minimizer for the dataset consisting of labeled samples drawn in previous rounds where the labels were generated by either f_{uniform} or f^*. Let $U_t = \sum_{r=1}^{t} \mathbf{1}\{f_{r-1} = f_{\mathrm{uniform}}\}$ be the number of these rounds, a random variable. Observe that $\bar{S}'_t$ contains nU_t samples, and each $(x, y) \in \bar{S}'_t$ is independent and distributed as follows: Draw x from D and let $y = f^*(x)$ with probability $(1-\alpha) + \frac{\alpha}{2} = \frac{1}{2} + \frac{1-\alpha}{2}$, and otherwise let $y = 1 - f^*(x)$. By Theorem 2 of Angluin and Laird (1988) we have

$$\mathrm{E}[L(f_t) \mid f_t \neq f_{\mathrm{uniform}} \wedge nU_t \geq m] \leq O\left(\frac{1}{1-\alpha}\sqrt{\frac{d \log m}{m}}\right)$$

Note that $\mathrm{E}[U_t] = \sum_{r=1}^{t} \frac{1}{\sqrt{nr}} = \Theta(\sqrt{t/n})$. Also, since each $\mathbf{1}\{f_t = f_{\mathrm{uniform}}\}$ is an independent Boolean random variable, by the multiplicative Chernoff bound we have $\Pr[U_t < O(\mathrm{E}[U_t])] \leq \exp(-\Omega(\mathrm{E}[U_t]))$. Therefore letting $m = \Theta(\sqrt{nt})$ we have

$$\begin{aligned}
\mathrm{E}[R(f_t)] &\leq \mathrm{E}[L(f_t)] \\
&\leq \mathrm{E}[L(f_t) \mid f_t \neq f_{\mathrm{uniform}} \wedge nU_t \geq m] + \Pr[f_t = f_{\mathrm{uniform}} \vee nU_t < m] \\
&\leq \mathrm{E}[L(f_t) \mid f_t \neq f_{\mathrm{uniform}} \wedge nU_t \geq \Omega(\sqrt{nt})] + \Pr[f_t = f_{\mathrm{uniform}} \vee nU_t < O(\sqrt{nt})] \\
&\leq \mathrm{E}[L(f_t) \mid f_t \neq f_{\mathrm{uniform}} \wedge nU_t \geq \Omega(\sqrt{nt})] + \Pr[f_t = f_{\mathrm{uniform}}] + \Pr[nU_t < O(\sqrt{nt})] \\
&\leq \mathrm{E}[L(f_t) \mid f_t \neq f_{\mathrm{uniform}} \wedge nU_t \geq \Omega(\sqrt{nt})] + \Pr[f_t = f_{\mathrm{uniform}}] + \Pr[U_t < O(\mathrm{E}[U_t])] \\
&\leq O\left(\frac{\sqrt{d \log(nt)}}{(1-\alpha)(nt)^{1/4}}\right) + \frac{1}{\sqrt{n(t+1)}} + \exp(-\Omega(\mathrm{E}[U_t])) \\
&\leq O\left(\frac{\sqrt{d \log(nt)}}{(1-\alpha)(nt)^{1/4}}\right) + \frac{1}{\sqrt{n(t+1)}} + \exp(-\Omega(\sqrt{t/n})).
\end{aligned}$$

■

3.4. Learning Disagreements from Positive Examples

In this section, we present an algorithm that achieves vanishing $O(t^{-1/2})$ generalization error for any contamination-rate α. Unlike the prior simple algorithm which only learns from examples in select, designated rounds, this section's algorithm learns from samples collected in all rounds. It does so by making better use of the fact that synthetic examples are labeled by a known classifier output in the previous round.

Theorem 13 *Let $F \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension d. There exists a learning algorithm A outputing classifiers f_t such that for all $t \geq 0$*

$$\mathrm{E}[R(f_t)] \leq \tilde{O}\left(\sqrt{\frac{d}{(1-\alpha)nt}}\right).$$

The analysis proceeds via a reduction to *learning with positive and unlabeled examples* (also referred to as *PU learning*). We make use of a result from the seminal work of Liu et al. (2002), as re-stated as Corollary 6 in Mansouri and Ben-David (2025).

Lemma 14 (PU learner (Liu et al., 2002); Corollary 6 from Mansouri and Ben-David (2025)) *Let $F \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension d. For a distribution D over $\mathcal{X}$ and true concept $f^* \in F$, denote by D^+ the distribution $D(x|f^*(x) = 1)$. There exists an algorithm such that when given $m_P(\varepsilon,\delta) = O(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon})$ positively labeled samples from D^+ and $m_U(\varepsilon,\delta) = O(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon})$ unlabeled samples from D, it outputs a hypothesis f such that $L(f) \leq \varepsilon$ with probability $\geq 1-\delta$.*

We discuss the connection between the settings, which conveys the main idea of the proof. In PU learning, the learner receives a random sample of labeled examples from the positive class, and a random sample of unlabeled examples from both classes. In round t of recursive learning, examples received by the learner are labeled by either the true concept f^* or the previous model f_{t-1} (which the learner has access to); for each individual example received, the learner does not know which rule it was labeled by.

Now, consider the task of learning the disagreement between the previous model and the true concept $\{x \in \mathcal{X} : f_{t-1}(x) \neq f^*(x)\}$. Since we know the f_{t-1} *exactly*, successfully learning the disagreement means we have learned the true concept f^*. Whenever we observe a disagreement between an example's label and the previous model's prediction on the point, we can be sure the label comes from the true concept – this corresponds to a positive example for the task of learning the disagreement. When the example's label and the previous model's prediction agree, we cannot distinguish between: (a) the previous model and true concept agree on the point (it is a negative example for the task of learning the disagreement); or (b) the previous model and true concept disagree on the point, but we fell into the α probability case where the example was labeled with the previous model. Hence the example is considered unlabeled for the task of learning the disagreement.

The above discussion suggests a natural algorithm for recursive learning: learn the disagreement between the previous round's model and the true concept via PU learning, and use the disagreement to form the next round's hypothesis. In the following, we formalize this idea.

Algorithm. The algorithm attaining the bound of Theorem 13 proceeds in collections of rounds we call *epochs*. Epochs are indexed by $k \geq 1$. For all the rounds t in the same epoch k, the algorithm will output the same classifier which we denote as $g_k \in F$. Note that each epoch may comprise of a different number of rounds. Formally, let $\mathsf{k}(t) : \mathbb{Z}^+ \to \mathbb{Z}^+$ denote the mapping of a round t to its corresponding epoch k. $\mathsf{k}(t)$ is a monotonically increasing step function, with $\mathsf{k}(1) = 1$. Our algorithm outputs $f_t = g_{\mathsf{k}(t)}$, with $g_1 = f_0$ as defined in Algorithm 1.

Update rule: After the designated number of rounds in epoch k using classifier g_k, we use an update rule to produce g_{k+1}, as a function of g_k and data points collected in the epoch $T_k = \bigcup_{\{t:\mathsf{k}(t)=k\}} S_t$, where S_t are the examples collected in round t.

The full update rule is given in Algorithm 2. It proceeds by mapping the learning problem in F into a learning problem in the XOR class which we denote as $g_k \oplus F$. For bits $\in \{0,1\}$ we denote by $\oplus$ the XOR operation. For two binary functions $f, g : \mathcal{X} \to \{0,1\}$, we define $(f \oplus g)(x) := f(x) \oplus g(x)$ for all $x \in \mathcal{X}$. For a class F and a binary function g, $g \oplus F := \{g \oplus f : f \in F\}$. The learning problem in the XOR class is a PU learning problem, whose solution can be converted to a solution to the original problem.

Schedule: We choose the epoch schedule $\mathsf{k}(t)$ as well as a target error schedule ε_k to achieve our desired bounds. The algorithm has the property that for all k sufficiently large, $\mathrm{E}[L(g_k)] \leq 2\varepsilon_k$. Specifically, we take $\varepsilon_k = \frac{1}{2^{k-1}}$ and $\mathsf{k}(t)$ such that each epoch is $O(\frac{d\log^2(1/\varepsilon_{k+1})}{(1-\alpha)n\varepsilon_{k+1}^2})$ rounds.

Algorithm 2: Epoch hypothesis update rule

Input: Function class $F \subseteq \mathcal{X} \to \mathcal{Y}$. Positive unlabeled learner A (see Lemma 14), target number of positive samples p_k, target number of unlabeled samples u_k. Epoch hypothesis g_k, epoch samples $T_k = \{(x_i, y_i)\}_{i=1}^{m+u_k}$.

Result: Next epoch hypothesis g_{k+1}.

$T_k^{\oplus} \leftarrow \{(x_i, y_i \oplus g_k(x_i)) : (x_i, y_i) \in T_k\}$

$P_k^{\oplus} \leftarrow \{(x_i, 1) : (x_i, y_i^{\oplus}) \in T_k^{\oplus}, y_i^{\oplus} = 1, i \in [1, m]\}$

$U_k \leftarrow \{x_i : i \in [m+1, m+u_k]\}$

if $|P_k^{\oplus}| \geq p_k$ **and** $|U_k| \geq u_k$ **then**

  $h \leftarrow A(P_k, U_k; g_k \oplus F)$

  $g_{k+1} \leftarrow h \oplus g_k$

else

  $g_{k+1} \leftarrow g_k$

end

Lemma 15 *Suppose we are in epoch $k \geq 1$. Let ε_{k+1} be the next epoch's target error, and suppose the current epoch samples T_k produced in the manner described in Algorithm 1 with g_k satisfies*

$$|T_k| \geq m_P(\varepsilon_{k+1}, \tfrac{\delta}{2}) \cdot \frac{8\log\left(\frac{2}{(\delta)}\right)}{(1-\alpha)\varepsilon_{k+1}} + m_U(\epsilon_{k+1}, \tfrac{\delta}{2}) = O\left(\frac{d\log(1/\varepsilon)\log(1/\delta)+\log^2(1/\delta)}{(1-\alpha)\varepsilon_{k+1}^2}\right)$$

where $m_P, m_U : (0,1)^2 \to \mathbb{N}$ are sample complexity functions from Lemma 14. Setting $p_k = m_P(\varepsilon_{k+1}, \frac{\delta}{2}) = O(\frac{d\log(1/\varepsilon_{k+1})+\log(1/\delta)}{\varepsilon_{k+1}})$ and $u_k = m_U(\varepsilon_{k+1}, \frac{\delta}{2}) = O(\frac{d\log(1/\varepsilon_{k+1})+\log(1/\delta)}{\varepsilon_{k+1}})$ in Algorithm 2, we have that with probability $\geq 1-\delta$, Algorithm 2 outputs g_{k+1} satisfying $L(g_{k+1}) \leq \varepsilon_{k+1}$.

Proof of Theorem 13 Given Lemma 15, we prove the result. We consider the target error schedule $\varepsilon_k = \frac{1}{2^{k-1}}$. Note that the only hypothesis we have no control over, g_1, trivially satisfies this requirement. We define the epoch schedule $\mathsf{k}(t)$ such that each round gets at least enough samples to satisfy the hypothesis of Lemma 15. Indeed, denote by r_k the required size for $|T_k|$ in Lemma 15.

$$r_k \leq C_1 \frac{d\log(1/\varepsilon_{k+1})\log(1/\delta) + \log^2(1/\delta)}{(1-\alpha)\varepsilon_{k+1}^2}$$

for some absolute constant C_1 and k sufficiently large. We set $\delta = \varepsilon_{k+1}$. Under this setting of δ, the high probability bound of Lemma 15 implies $\mathrm{E}[g_{k+1}] \leq 2\varepsilon_{k+1}$.

We choose $\mathsf{k}(t)$ such that for all $k \geq 1$, $|\{t : \mathsf{k}(t) = k\}| = \lceil r_k/n \rceil$. Since inside an epoch, we output the same hypothesis g_k, it suffices to check if the final round in an epoch satisfies the stated error rate in Theorem 13. For $k \geq 1$, the last round of a given epoch $t_k := \max\{t : \mathsf{k}(t) = k\}$ satisfies

$$t_k = \sum_{j=1}^{k} \lceil r_j/n \rceil \leq C_2 \frac{d}{(1-\alpha)n} \sum_{j=1}^{k} \log^2(\tfrac{1}{\varepsilon_{j+1}}) \cdot \frac{1}{\varepsilon_{j+1}^2} = C_2 \frac{d}{(1-\alpha)n} \sum_{j=1}^{k} (j2^j)^2 \leq C_3 \frac{d}{(1-\alpha)n} (2^k k)^2$$

for absolute constants C_2, C_3 and k sufficiently large. To conclude, we have for k sufficiently large

$$\mathrm{E}[L(f_{t_k})] = \mathrm{E}[L(g_k)] \leq 4 \cdot \tfrac{1}{2^k} \leq 4 \cdot \sqrt{\tfrac{C_3 d}{(1-\alpha)nt_k}} \cdot k = \tilde{O}\left(\sqrt{\tfrac{d}{(1-\alpha)nt_k}}\right).$$

■

Proof of Lemma 15 First, we show reducing to positive unlabeled learning is valid.

The preconditions for invoking the PU learner are satisfied. The class we are invoking the PU learner on is $g_k \oplus F$, which has the same VC dimension as F. Therefore the stated sample complexity suffices for learning $g_k \oplus F$. The samples $T_k^{\oplus}$ come from the original marginal distribution D, and are labeled $0 = (g_k \oplus f^*)(x)$ everywhere $g_k(x) = f^*(x)$. For x with $g_k(x) \neq f^*(x) \iff (g_k \oplus f^*)(x) = 1$, the label is 1 with probability $1-\alpha$ and 0 with probability α independently. Hence we can conclude that $P_k^{\oplus}$ is indeed an i.i.d. sample from $D(x|(g_k \oplus f^*)(x) = 1)$. Also, U_k is indeed an i.i.d. sample from D.

The output satisfies our desired error guarantee with respect to the original problem. With probability $\geq 1-\delta$, our PU learner for $g_k \oplus F$ outputs h with $\Pr_{x\sim D}[h(x) \neq (g_k \oplus f^*)(x)] \leq \varepsilon_{k+1}$. Since for all x, $h(x) \neq (g_k \oplus f^*)(x) \iff (h \oplus g_k)(x) \neq f^*(x)$, we have $\Pr_{x\sim D}[(h \oplus g_k)(x) \neq f^*(x)] \leq \varepsilon_{k+1}$.

Hence, whenever the learner is invoked, our update rule succeeds with probability $\geq 1 - \frac{\delta}{2}$. To finish the proof, we consider two cases:

Case 1: $L(g_k) \leq \varepsilon_{k+1}$. If the error of the current hypothesis g_k meets the target error threshold, we either: (1) gather enough samples to invoke the learner, which means our update rule succeeds with high probability; or (2) fail to gather enough and pass on g_k to the next round, which by assumption, achieves the target error.

Case 2: $L(g_k) > \varepsilon_{k+1}$. By applying a Chernoff bound and plugging in the requested size of T_k and using the assumption $L(g_k) > \varepsilon_{k+1}$, we have that with probability $\geq 1 - \frac{\delta}{2}$, $|P_k^{\oplus}| > m_P(\varepsilon_{k+1}, \frac{\delta}{2})$, that is, we obtain enough positive examples required to invoke the PU learner. This calculation is deferred to Appendix B.2. Taking a union bound over the failure probability of the PU learner concludes the proof. ■

Discussion. The algorithm requires knowledge of α, the contamination rate, since it is used to define each epoch as lasting $\tilde{O}(\frac{d}{(1-\alpha)n\varepsilon_{k+1}^2})$ rounds. We note that there are simple techniques for estimating α from data. For example, by outputting the all-zeros and all-ones classifiers across two rounds, one expects to observe αn disagreements in total. We do not fully explore whether knowledge of α can be removed, and assume that it is known exactly for simplicity.

We also observe that it is possible to get an $O(1/nt)$ rate in Theorem 13 (which is more natural to expect in the realizable setting we study) if we make an adjustment to the setup. In Theorem 13, the expected accuracy requirement is for *all* rounds. This choice in our modeling framework reflects the (realistic) objective of steady improvement from one model generation to the next. At a high level, Theorem 13's algorithm functions by collecting feedback from errors, and hence, releasing a good model every round limits the amount of feedback received and increases the overall sample complexity. If our concern was to do well only with respect to the *final* round t, we could release a model with constant error in rounds $1, 2, ..., t-1$, resulting in more feedback received, and the desired $O(1/nt)$ error rate for f_t. We leave open the question of whether it is possible to achieve the $O(1/nt)$ error rate for all rounds simultaneously.

4. Conclusion

The prevalence of synthetic data introduces a new class of learning questions, where the training set is contaminated by additional examples. We formally study this phenomenon, for two fundamental problems: mean estimation and PAC learning.

For mean estimation, we give a full characterization of the variance of the most fundamental estimator, the empirical mean. We show that this is not the MVUE in general, including for distributions where the empirical mean is the MVUE in i.i.d. settings (e.g. Gaussian distributions), and when the data is not fully contaminated ($\alpha < 1$). In the PAC setting, we show that repeated ERM can experience generalization error that stalls. We complement this with two algorithms that do achieve vanishing generalization.

Interesting open problems for mean estimation include fully characterizing the minimum variance unbiased estimator, and allowing the mean to depend on a vector of covariates instead of remaining fixed in every round. For PAC learning, open problems include expanding the results to the agnostic setting, and obtaining sample complexity lower bounds for generic algorithms.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024.

Kareem Amin, Sara Babakniya, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii. Escaping collapse: The strength of weak data for large language model training, 2025. URL https://arxiv.org/abs/2502.08924.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine learning*, 2:343–370, 1988.

Daniel Barzilai and Ohad Shamir. When models don't collapse: On the consistency of iterative mle. *arXiv preprint arXiv:2505.19046*, 2025.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024.

Apratim Dey and David Donoho. Universality of the $\pi^2/6$ pathway in avoiding model collapse. *arXiv preprint arXiv:2410.22812*, 2024.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, Julia Kempe, and FAIR Meta. Beyond model collapse: Scaling up with syn-thesized data requires verification. *arXiv preprint arXiv:2406.07515*, 2024.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences, 2024. URL https://arxiv.org/abs/2407.09499.

Aymane El Firdoussi, Mohamed El Amine Seddik, Soufiane Hayou, Reda Alami, Ahmed Alzubaidi, and Hakim Hacid. Maximizing the potential of synthetic data: Insights from random matrix theory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20555–20565, October 2023.

Millen Kanabar and Michael Gastpar. Model non-collapse: Minimax bounds for recursive discrete distribution estimation, 2025. URL https://arxiv.org/abs/2501.19273.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*, 2024.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 267–280, 1988.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 387–394, San Francisco, CA, USA, 2002.

Farnam Mansouri and Shai Ben-David. Learning from positive and unlabeled examples -finite size sample bounds. *CoRR*, abs/2507.07354, 2025.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759, 2024.

Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. Rate of model collapse in recursive training. *arXiv preprint arXiv:2412.17646*, 2024.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566, 1985.

Appendix A. Appendix

A.1. Mean Estimation Proofs

Theorem 16 *It holds that*

$$\mathrm{Var}(Y_t) = \left\{ \frac{1}{t^2} + \left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 \right\} \Sigma. \tag{4}$$

Proof Let $V_t = \mathrm{Var}(Y_t(\bar{w}))$. By definition, we have

$$\begin{aligned}
V_t &= \mathrm{Var}\left(\frac{1}{t}\sum_{s=1}^{t} X_s\right) = \mathrm{Var}\left(\frac{X_t}{t} + \frac{1}{t}\sum_{s=1}^{t-1} X_s\right) \\
&= \mathrm{Var}\left(\frac{X_t}{t}\right) + \mathrm{Var}\left(\frac{1}{t}\sum_{s=1}^{t-1} X_s\right) + 2\mathrm{Cov}\left(\frac{X_t}{t}, \frac{1}{t}\sum_{s=1}^{t-1} X_s\right) \\
&= \mathrm{Var}\left(\frac{\alpha Y_{t-1} + U_{t-1}}{t}\right) + \mathrm{Var}\left(\frac{t-1}{t} Y_{t-1}\right) + 2\mathrm{Cov}\left(\frac{X_t}{t}, \frac{t-1}{t} Y_{t-1}\right) \\
&= \frac{\alpha^2 V_{t-1} + \Sigma}{t^2} + \left(\frac{t-1}{t}\right)^2 V_{t-1} + 2\mathrm{Cov}\left(\frac{\alpha Y_{t-1} + U_{t-1}}{t}, \frac{t-1}{t} Y_{t-1}\right) \\
&= \frac{\alpha^2 V_{t-1} + \Sigma}{t^2} + \left(\frac{t-1}{t}\right)^2 V_{t-1} + \frac{2\alpha(t-1)}{t^2} V_{t-1} \\
&= \left[\frac{\alpha^2}{t^2} + \left(\frac{t-1}{t}\right)^2 + \frac{2\alpha(t-1)}{t^2}\right] V_{t-1} + \frac{1}{t^2}\Sigma \\
&= \left(\frac{t-1+\alpha}{t}\right)^2 V_{t-1} + \frac{1}{t^2}\Sigma.
\end{aligned}$$

We now proceed by induction. The case of $t = 1$ follows from the definition of Y_1, and the case of $t = 2$ follows from the above identity. Suppose our hypothesis holds at some $t \geq 3$. Then, using the identity $\Gamma(z+1) = z\Gamma(z)$, we have

$$\begin{aligned}
V_{t+1} &= \left(\frac{t+\alpha}{t+1}\right)^2 V_t + \frac{1}{(t+1)^2}\Sigma. \\
&= \left\{ \left(\frac{t+\alpha}{t+1}\right)^2 \left(\frac{1}{t^2} + \left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 \right) + \frac{1}{(t+1)^2} \right\} \Sigma \\
&= \left\{ \left(\frac{t+\alpha}{t+1}\right)^2 \frac{1}{t^2} + \left[\frac{\Gamma(t+1+\alpha)}{\Gamma(t+2)}\right]^2 \sum_{k=1}^{t-1} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 + \frac{1}{(t+1)^2} \right\} \Sigma \\
&= \left\{ \left[\frac{\Gamma(t+1+\alpha)}{\Gamma(t+2)}\right]^2 \sum_{k=1}^{t} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 + \frac{1}{(t+1)^2} \right\} \Sigma.
\end{aligned}$$

■

Theorem 17 *For $t \geq 3$, we have that*

$$\frac{1}{2}\left[\frac{1}{t}+\frac{1}{t^2}+\frac{1}{t^{2(1-\alpha)}}\right]\Sigma \preceq \mathrm{Var}(Y_t(\bar{w})) \preceq 4\left[\frac{1}{t}+\frac{1}{t^2}+\frac{1}{t^{2(1-\alpha)}}\right]\Sigma. \tag{5}$$

Proof We first prove the upper bound. Using Lemma 4, we first have that

$$\begin{aligned}
\left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right] &\leq t^{2(\alpha-1)}\sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 \\
&= t^{2(\alpha-1)}\sum_{k=1}^{t-1}\left[\frac{(k+1)^{1-\alpha}}{k}\right]^2 \\
&\leq 4t^{2(\alpha-1)}\sum_{k=1}^{t-1}\left[\frac{(k+1)^{1-\alpha}}{k+1}\right]^2 \qquad \because k \geq \frac{k+1}{2} \\
&= 4t^{2(\alpha-1)}\sum_{k=1}^{t-1}\frac{1}{k^{2\alpha}}.
\end{aligned}$$

Now, since $\alpha \in (0,1)$ and $x \mapsto x^{-2a}$ is monotonically decreasing, we have

$$\sum_{k=1}^{t-1}\frac{1}{k^{2\alpha}} \leq 1+\int_1^{t-1}\frac{1}{x^{2\alpha}}\,dx = 1+\left.\frac{1}{x^{2\alpha+1}}\right]_1^{t-1} = \left[1+(t-1)^{-(2\alpha+1)}\right] \leq \left[1+t^{-2\alpha+1}\right].$$

Combining this with the previous bounds yields

$$\left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right] \leq 4t^{2(\alpha-1)}\left[1+t^{-2\alpha+1}\right] \leq 4\left[\frac{1}{t}+\frac{1}{t^{2(1-\alpha)}}\right],$$

which gives the upper bound of (5).

The lower bound has a similar derivation. Using Lemma 4, we first have that

$$\begin{aligned}
\left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right] &\geq (t+1)^{2(\alpha-1)}\sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 \\
&\geq t^{2(\alpha-1)}\sum_{k=1}^{t-1}\left[\frac{k^{1-\alpha}}{k}\right]^2 \\
&= t^{2(\alpha-1)}\sum_{k=1}^{t-1}\frac{1}{k^{2\alpha}}.
\end{aligned}$$

Similarly, since $x \mapsto x^{-2a}$ is monotonically decreasing, we have

$$\begin{aligned}
\sum_{k=1}^{t-1}\frac{1}{k^{2\alpha}} &= 1+\sum_{k=2}^{t-1}\frac{1}{k^{2\alpha}} \geq 1+\int_2^{t}\frac{1}{x^{2\alpha}}\,dx = 1+\left.\frac{1}{x^{2\alpha+1}}\right]_2^{t} \\
&= 1+t^{-(2\alpha+1)}-\frac{1}{2^{2\alpha+1}} \geq \frac{1}{2}+t^{-(2\alpha+1)}
\end{aligned}$$

Combining this with the previous bounds yields

$$\left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1}\left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right] \geq t^{2(\alpha-1)}\left[\frac{1}{2}+t^{-(2\alpha+1)}\right] = \frac{1}{t}+\frac{1}{2t^{2(1-\alpha)}} \geq \frac{1}{2}\left[\frac{1}{t}+\frac{1}{t^{2(1-\alpha)}}\right],$$

which gives the lower bound of (5). ■

A.2. Improving on Uniform Weighting

Let us now construct a better weighting scheme for $\alpha \in (0,1)$ that is continuous in α.

Proposition 18 *Suppose* $w^1 = 1$ *and*

$$\hat{w}^{s+1} = \left(\frac{1}{1+(s-1)(1-\alpha)}, \frac{1-\alpha}{1+(s-1)(1-\alpha)}, \dots \frac{1-\alpha}{1+(s-1)(1-\alpha)}\right)$$
$$\hat{w} = (\hat{w}^1, \dots, \hat{w}^t)$$

for every $s \geq 1$. *Then,* $\mathrm{Var}(Y_1(\hat{w})) = \sigma^2$ *and, for* $t \geq 2$, *we have*

$$\mathrm{Var}(Y_t(\hat{w})) = \left\{\left(\frac{1-\alpha}{\gamma_{\alpha,t}}\right)^2 + \sum_{k=2}^{t-1}\left(\frac{C_{\alpha,t-1}}{C_{\alpha,k-1}}\right)^2\left(\frac{1-\alpha}{\gamma_{\alpha,k}}\right)^2 + C_{\alpha,t-1}^2\right\}\Sigma \tag{6}$$

where

$$\gamma_{\alpha,\ell} := 1+(\ell-1)(1-\alpha), \quad C_{\alpha,\ell} := \prod_{i=1}^{\ell}\frac{\gamma_i+\alpha(1-\alpha)}{\gamma_{i+1}} \quad \forall \ell \geq 0.$$

Proof Let $t \geq 1$ and $\alpha \in [0,1]$ be fixed and denote

$$V_t = \mathrm{Var}(Y_t(\hat{w})), \quad \gamma_\ell = \gamma_{\alpha,\ell}, \quad C_\ell = C_{\alpha,\ell}.$$

By definition, we have

$$\begin{aligned}
V_t &= \mathrm{Var}\left(\frac{X_1}{\gamma_t}+\frac{1-\alpha}{\gamma_t}\sum_{s=2}^{t-1}X_s\right)\\
&= \mathrm{Var}\left(\frac{X_1}{\gamma_t}+\frac{1-\alpha}{\gamma_t}\sum_{s=2}^{t-1}X_s+\frac{1-\alpha}{\gamma_t}X_t\right) = \mathrm{Var}\left(\frac{\gamma_{t-1}}{\gamma_t}Y_{t-1}+\frac{1-\alpha}{\gamma_t}X_t\right)\\
&= \mathrm{Var}\left(\frac{\gamma_{t-1}}{\gamma_t}Y_{t-1}\right)+\mathrm{Var}\left(\frac{1-\alpha}{\gamma_t}X_t\right)+2\mathrm{Cov}\left(\frac{\gamma_{t-1}}{\gamma_t}Y_{t-1},\frac{1-\alpha}{\gamma_t}X_t\right)\\
&= \left(\frac{\gamma_{t-1}}{\gamma_t}\right)^2 V_{t-1}+\left(\frac{1-\alpha}{\gamma_t}\right)^2\mathrm{Var}\left(\alpha Y_{t-1}+U_{t-1}\right)+2\mathrm{Cov}\left(\frac{\gamma_{t-1}}{\gamma_t}Y_t,\frac{\alpha[1-\alpha]}{\gamma_t}Y_{t-1}\right)\\
&= \left[\left(\frac{\gamma_{t-1}}{\gamma_t}\right)^2+\alpha^2\left(\frac{1-\alpha}{\gamma_t}\right)^2+\frac{2\gamma_{t-1}\alpha(1-\alpha)}{\gamma_t^2}\right]V_{t-1}+\left(\frac{1-\alpha}{\gamma_t}\right)^2\Sigma\\
&= \left[\frac{\gamma_{t-1}+\alpha(1-\alpha)}{\gamma_t}\right]^2 V_{t-1}+\left(\frac{1-\alpha}{\gamma_t}\right)^2\Sigma.
\end{aligned} \tag{7}$$

We now proceed by induction. The case of $t = 1$ follows immediately from the definition of Y_1, and the case of $t = 2$ follows from the above identity. For $t = 3$, we have

$$\begin{aligned}
V_{t+1} &= \left(\frac{t+\alpha}{t+1}\right)^2 V_t + \frac{1}{(t+1)^2}\Sigma. \\
&= \left\{\left(\frac{t+\alpha}{t+1}\right)^2 \left(\frac{1}{t^2} + \left[\frac{\Gamma(t+\alpha)}{\Gamma(t+1)}\right]^2 \sum_{k=1}^{t-1} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2\right) + \frac{1}{(t+1)^2}\right\}\Sigma \\
&= \left\{\left(\frac{t+\alpha}{t+1}\right)^2 \frac{1}{t^2} + \left[\frac{\Gamma(t+1+\alpha)}{\Gamma(t+2)}\right]^2 \sum_{k=1}^{t-1} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 + \frac{1}{(t+1)^2}\right\}\Sigma \\
&= \left\{\left[\frac{\Gamma(t+1+\alpha)}{\Gamma(t+2)}\right]^2 \sum_{k=1}^{t} \left[\frac{\Gamma(k+1)}{k\cdot\Gamma(k+\alpha)}\right]^2 + \frac{1}{(t+1)^2}\right\}\Sigma.
\end{aligned}$$

Now suppose our hypothesis holds for some $t \geq 3$ and denote

$$c_j := \frac{\gamma_j + \alpha(1-\alpha)}{\gamma_{j+1}} \implies C_{\alpha,\ell} = \prod_{i=1}^{\ell} c_j^2$$

Applying (7) and using our hypothesis, we have

$$\begin{aligned}
V_{t+1} &= \left[\frac{\gamma_t + \alpha(1-\alpha)}{\gamma_{t+1}}\right]^2 V_t + \left(\frac{1-\alpha}{\gamma_{t+1}}\right)^2 \Sigma = c_{t+1}^2 V_t + \left(\frac{1-\alpha}{\gamma_{t+1}}\right)^2 \Sigma \\
&= \left(c_t^2 \left\{\left[\frac{1-\alpha}{\gamma_t}\right]^2 + \sum_{k=2}^{t-1} \left[\frac{C_{t-1}}{C_{k-1}}\right]^2 \left[\frac{1-\alpha}{\gamma_k}\right]^2 + C_{t-1}^2\right\} + \left(\frac{1-\alpha}{\gamma_{t+1}}\right)^2\right)\Sigma. \\
&= \left(c_t^2 \left[\frac{1-\alpha}{\gamma_t}\right]^2 + \sum_{k=2}^{t-1} \left[\frac{C_t}{C_{k-1}}\right]^2 \left[\frac{1-\alpha}{\gamma_k}\right]^2 + C_t^2 + \left(\frac{1-\alpha}{\gamma_{t+1}}\right)^2\right)\Sigma \\
&= \left(\left(\frac{1-\alpha}{\gamma_{t+1}}\right)^2 + \sum_{k=2}^{t} \left[\frac{C_t}{C_{k-1}}\right]^2 \left[\frac{1-\alpha}{\gamma_k}\right]^2 + C_t^2\right)\Sigma.
\end{aligned}$$

■

Proposition 19 *For every $t \geq 1$, there exists $\alpha^* \in (0,1)$ such that for any $\alpha \in (\alpha^*, 1]$, it holds that*

$$\mathrm{Var}(Y_t(\hat{w})) \prec \mathrm{Var}(Y_t(\bar{w})).$$

Proof For every $\alpha \in [0,1]$, recall that there exists continuous $\hat{f}_t, \bar{f}_t : [0,1] \mapsto (0,\infty)$ such that

$$\mathrm{Var}(Y_t(\hat{w})) = \hat{f}_t(\alpha)\Sigma, \quad \mathrm{Var}(Y_t(\bar{w})) = \bar{f}_t(\alpha)\Sigma.$$

In particular, we have that

$$\hat{f}_t(1) = C_{1,t-1}^2 = 1 < \sum_{k=1}^{t} \frac{1}{k^2} = \bar{f}_t(1)$$

and, hence, that $\mathrm{Var}(Y_t(\hat{w})) \prec \mathrm{Var}(Y_t(\bar{w}))$. Since $g_t(\alpha) = \bar{f}_t(\alpha) - \hat{f}_t(\alpha)$ is continuous on $[0,1]$ and positive at $\alpha = 1$, the result follows immediately by continuity. ■

Appendix B. PAC Learning Proofs

B.1. Hardness Result

Theorem 20 *Define $c^*_\alpha = \Pr[\forall t \geq 1, W_t \geq 1]$. For any sample size $n \geq 2$ and $d \geq 2$, there exists a distribution D, a hypothesis class F with VC dimension d, and $f^* \in F$, such that if $A = A_{\mathrm{ERM}}$ then*

$$\liminf_{t\to\infty} \mathrm{E}[L(f_t)] \geq \frac{c^*_\alpha}{8n}.$$

Proof Consider the function class containing thresholds in 1 dimension. In other words, $X = \mathbb{R}$, and for any $\theta \in \mathbb{R}$, we define $f^\theta(x) \triangleq 1 - 2 \cdot \mathbf{1}\{x \leq \theta\}$, and $F = \{f^\theta, -f^\theta \mid \theta \in \mathbb{R}\}$.

Define D according to the following p.m.f:

$$p(x) = \begin{cases} \frac{1}{2} - \frac{1}{2n} & \text{if x = +1} \\ \frac{1}{n} & \text{if x = 0} \\ \frac{1}{2} - \frac{1}{2n} & \text{if x = -1} \end{cases}$$

On round $t = 0$, $\bar{S}_0$ contains n samples drawn from D, labeled according to the true concept f^*. Let $\bar{S}_{0,\mathcal{X}}$ denote the *set* of unlabeled samples present in the multi-set $\bar{S}_0$. Thus, if $x = 0$ appears multiple times in $\bar{S}_0$, it appears once in $\bar{S}_{0,\mathcal{X}}$ and $\bar{S}_{0,\mathcal{X}} \subset \{-1, 0, +1\}$. We first characterize the probability that we see the samples $x = -1, x = +1$ but not the sample $x = 0$ in round 0.

$$\begin{aligned}
\Pr[\bar{S}_{0,\mathcal{X}} = \{-1,+1\}] &= \Pr[0 \notin \bar{S}_{0,\mathcal{X}}] \Pr[\bar{S}_{0,\mathcal{X}} = \{-1,+1\} | 0 \notin \bar{S}_{0,\mathcal{X}}] \\
&= \left(1 - \frac{1}{n}\right)^n \Pr[\bar{S}_{0,\mathcal{X}} = \{-1,+1\} | 0 \notin \bar{S}_{0,\mathcal{X}}] \\
&= \left(1 - \frac{1}{n}\right)^n \left(1 - \Pr[\bar{S}_{0,\mathcal{X}} = \{+1\} \mid 0 \notin \bar{S}_{0,\mathcal{X}}] - \Pr[\bar{S}_{0,\mathcal{X}} = \{-1\} \mid 0 \notin \bar{S}_{0,\mathcal{X}}]\right) \\
&= \left(1 - \frac{1}{n}\right)^n \left(1 - \frac{1}{2^n} - \frac{1}{2^n}\right) \geq 1/8 \qquad (8)
\end{aligned}$$

The third equality follows since, conditioned on the event that the initial sample does not contain any 0s, it either contains both $\{-1, +1\}$, only $+1$s or only -1s. The final equality follows since, conditioned on $0 \notin \bar{S}_{0,\mathcal{X}}$, samples $+1$ and -1 are equally likely. The final inequality follows because the expressions $(1 - \frac{1}{n})^n$ and $(1 - \frac{1}{2^{n-1}})$ are both increasing in n, and $n \geq 2$ by assumption.

After round 0, we will re-index the data to ignore boundaries between rounds. Let $(\tilde{x}_k, \tilde{y}_k)$ denote the kth example encountered after round 0. In other words, writing $k = tn + i$ for $i < n$, $\tilde{x}_k = x_i^{t+1}$. Similarly, let $\tilde{b}_k$ denote whether $\tilde{x}_k$ was labeled by a previous model or nature: $\tilde{b}_k = \tilde{b}_{tn+i} = b_i^{t+1}$.

We are interested in two statistics, C_k, Δ_k, which we define as:

$$C_k = \sum_{k' \leq k} \mathbf{1}\{\tilde{x}_k = 0\}$$

and

$$\Delta_k = \sum_{k' \leq k} \mathbf{1}\{\tilde{x}_k = 0\}(\mathbf{1}\{\tilde{b}_k = 1\} - \mathbf{1}\{\tilde{b}_k = 0\}).$$

C_k increments every time a sample $x = 0$ is encountered, while Δ_k measures how much more frequently samples $x = 0$ are labeled by a previous model versus the true concept f^*. $\{\Delta_k\}$ is a random walk with both a bias and a staying probability. Defining $\Delta_0 = 0$, we have $\Delta_k = \Delta_{k-1}$ with probability $1 - 1/n$, $\Delta_k = \Delta_{k-1} + 1$ with probability α/n, and $\Delta_k = \Delta_{k-1} - 1$ with probability $(1 - \alpha)/n$.

Define W_l as a biased random walk, without any staying probability, independent of C_k. $W_0 = 0$, $W_{l+1} = W_l + 1$ with probability α and $W_{l+1} = W_l - 1$ with probability $(1 - \alpha)$. It is easy to check that Δ_k is equal in distribution to W_{C_k}, since:

$$\begin{aligned}
\Pr(W_{C_{k+1}} = W_{C_k}) &= \Pr(C_{k+1} = C_k) = \Pr(\hat{x}_k \neq 0) = 1 - 1/n \\
\Pr(W_{C_{k+1}} = W_{C_k} + 1) &= \Pr(C_{k+1} = C_k + 1, W_{C_k+1} = W_{C_k} + 1) \\
&= \Pr(C_{k+1} = C_k + 1, W_1 = W_0 + 1) \\
&= \Pr(C_{k+1} = C_k + 1) \Pr(W_1 = W_0 + 1) \\
&= \Pr(\hat{x}_k = 0) \Pr(W_1 = W_0 + 1) = \alpha/n \\
\Pr(W_{C_{k+1}} = W_{C_k} - 1) &= 1 - \Pr(W_{C_{k+1}} = W_{C_k}) - \Pr(W_{C_{k+1}} = W_{C_k} + 1) = (1 - \alpha)/n
\end{aligned}$$

Now let $\mathcal{E}_{\text{contaminated}} = \{\forall k, C_k \geq 1 \implies \Delta_k \geq 1\}$, be the event that once a sample $x = 0$ is encountered (i.e. $C_k \geq 1$), there are more model-labeled examples than naturally-labeled examples for $x = 0$ (i.e. $\Delta_k \geq 1$) for all time. Via the construction $\Delta_k = W_{C_k}$ this happens if and only if $W_l \geq 1$ for all $l \geq 1$. If $W_l \geq 1$ for all $l \geq 1$, then it's clear that $\Delta_k = W_{C_k} \geq 1$ when $C_k \geq 1$. In the other direction, if $W_l < 1$ for some some $l \geq 1$, then there is eventually an index k such that $C_k = l$, and therefore $\Delta_k = W_{C_k} < 1$. Thus,

$$\Pr(\mathcal{E}_{\text{contaminated}}) = \Pr(\forall l \geq 1, W_l \geq 1) \tag{9}$$

The random variables $\{\tilde{x}_k, \tilde{b}_k\}$ define what examples are encountered subsequent to round 0 and whether they are labeled by nature or a model. By construction of our setting (i.e. Algorithm 1), these are independent of the examples $\{x_i^0\}$ encountered in round 0. $\mathcal{E}_{\text{contaminated}}$ is measurable by first set of random variables, and $\bar{S}_{0,\mathcal{X}}$ is measurable by the second set of random variables, and therefore are also independent.

Hence combining Equations (8) and (9), we can conclude:

$$\Pr[\bar{S}_{0,\mathcal{X}} = \{-1, +1\}) \wedge \mathcal{E}_{\text{contaminated}}] \geq \Pr[\forall l \geq 1, W_l \geq 1]/8 \tag{10}$$

So far, we have have not specified algorithm A_{ERM}'s behavior if there are multiple hypotheses with the same empirical risk. Given a dataset that is separable by a threshold, we will define A_{ERM} as selecting the maximum-margin hypothesis. [2] In any other case, including when $\hat{S}$ is not separable, A_{ERM} can break ties arbitrarily.

Fix $f^* = f^{-1}$, which labels samples $x = 0$ as $+1$. We argue that on the event $\mathcal{E}_{\text{fail}} := \{\bar{S}_{0,\mathcal{X}} = \{-1, +1\}\} \wedge \mathcal{E}_{\text{contaminated}}$, the classifiers returned by repeated ERM mis-classify $x = 0$, forever, on every time step. On round $t = 0$, $A_{\text{ERM}}(\bar{S}_0)$ returns a the maximum-margin threshold f^0, which

2. In general, if $\bar{S}_t$ does not contain the example $x = 0$, any tie-breaking rule (including randomized ones) will have a constant probability of mis-classifying it. Treating this with full generality complicates the notation and exposition without providing much additional insight, and so we omit it.

assigns $f^0(0) = -1$. On subsequent rounds t, it may be that $C_{nt} = 0$, in which case $\bar{S}_t$ still only contains samples $-1, +1$, and A_{ERM} again selects the maximum-margin classifier. Otherwise, $\Delta_{nt} \geq 1$. By induction, classifiers from all previous rounds mislabeled $x = 0$, and so $\Delta_{nt} \geq 1$ implies that $\bar{S}_t$ contain more examples $(x = 0, y = -1)$ than examples $(x = 0, y = +1)$, thus the empirical risk minimizer is to return a threshold that labels $x = -1$ as -1, $x = 0$ as -1, and $x = +1$ as $+1$. Therefore, on all rounds t, $L(f_t) = 1/n$, and we can conclude that for any t:

$$\mathrm{E}[L(f_t)] \geq \mathrm{E}[L(f_t) \mid \mathcal{E}_{\text{fail}}] \Pr[\mathcal{E}_{\text{fail}}] \geq \frac{c^*_\alpha}{8n}$$

Finally, we observe that for any d, we can take $\mathcal{X} = \mathbb{R}^d$ and embed the example above in the first coordinate. We take x_1 to be drawn according to the p.m.f. D, while taking all other coordinates to be 0 with probability 1. Consider F to be the family of linear separators in d dimensions: $\operatorname{sgn}(\langle x, w\rangle + b)$ for $x \in \mathbb{R}^d$, $b \in \mathbb{R}$. For any x drawn from this distribution, $\operatorname{sgn}(\langle x, w\rangle + b) = \operatorname{sgn}(x_1 w_1 + b)$, which labels x_1 as -1 iff $x_1 < -b/w_1$. Thus, for this class too, there exists an f^* such that $\mathrm{E}[R(f_t)] \geq \frac{c^*_\alpha}{8n}$. Standard learning theory tells us that $\mathrm{VCD}(F) = d+1$, concluding the proof. ■

Lemma 21 *When $\alpha > 1/2$, the event that the random walk $\{W_n\}$ stays positive forever occurs with non-zero probability: $\mathbb{P}[\forall n \geq 1, W_n \geq 1] > 0$*

Proof For $n \geq 1$ define $E_n = \{W_n = 0\}$ as the event that the random walk returns to 0 after exactly n steps. By the Borel-Cantelli Lemma, if $\sum_{n\geq 1} \mathbb{P}[E_n] < \infty$, then the probability that the events $\{E_n\}$ occur infinitely often goes to zero: $\mathbb{P}[E_n \text{ i.o}] = 0$.

Note that on any odd n, $W_n \neq 0$ with probability 1, since the number of leftward steps and rightward steps cannot have the same parity. Therefore:

$$\begin{aligned}
\sum_{n\geq 1} \mathbb{P}[W_n = 0] &= \sum_{n\geq 1} \mathbb{P}[W_{2n} = 0] \\
&= \sum_{n\geq 1} \binom{2n}{n} \alpha^n (1-\alpha)^n \\
&\leq \sum_{n\geq 1} 4^n \alpha^n (1-\alpha)^n
\end{aligned}$$

Let $\phi = 4\alpha(1-\alpha)$. As a function of $\alpha \in [0, 1]$, ϕ is maximized at $\phi = 1$ when $\alpha = 1/2$. Therefore, $\phi < 1$ under the assumption that $\alpha > 1/2$, and $\sum_{n\geq 1} \mathbb{P}[E_n] \leq \frac{\phi}{1-\phi} < \infty$, which implies $\mathbb{P}[E_n \text{ i.o}] = 0$.

Define $T_n = \inf\{n \geq 1 \mid W_n = 0\}$, as the first return-time to zero of this random walk. Assume for the sake of contradiction that $\mathbb{P}[T_n < \infty] = 1$, then by the strong Markov property, the random walk visits zero infinitely often with probability 1, contradicting that $\mathbb{P}[E_n \text{ i.o}] = 0$. Therefore,

$$\mathbb{P}[T_n = \infty] > 0.$$

Next, define $E^+ = \{W_n \mid \forall n \geq 1, W_n > 0\}$ and $E^- = \{W_n \mid \forall n \geq 1, W_n < 0\}$, the events that the random walk stays positive forever or stays negative forever, respectively. There is a bijection

between paths in E^+ and E^- attained by reflecting the path over 0. Furthermore, since paths in E^- must have more left-ward steps than right-ward steps, and $\alpha > 1/2$, $\mathbb{P}(E^+) > \mathbb{P}(E^-)$.

Finally, since the event $\{T_n = \infty\}$ is the disjoint union of E^+ and E^-, we have $\mathbb{P}[E^+] \geq \mathbb{P}[T_n = \infty]/2 > 0$, completing the proof. ■

B.2. Omitted Calculation in Proof of Lemma 15

Lemma 22 (Chernoff bounds) [3] *Let* $X_1, ...X_m$ *be i.i.d.* $\text{Bernoulli}(p)$ *random variables for* $p \in [0,1]$. *For* $\gamma \in [0,1]$

$$\Pr\left[\sum_{i=1}^{m} X_i \leq (1-\gamma)mp\right] \leq \exp(-\gamma^2 mp/2).$$

Omitted calculation. In the setting of Proof of Lemma 15, Case 2: $L(g_k) > \varepsilon_{k+1}$, our goal is to show that $P_k^{\oplus}$ is sufficiently large. We let X_i denote the indicator variable that is 1 if and only if the ith example of $T_k^{\oplus}$, $(x_i, y_i^{\oplus})$, satisfies $y_i^{\oplus} = 1$. This occurs with probability $L(g_k)(1-\alpha)$ independently, and we draw the first m samples for $P_k^{\oplus}$. Hence $|P_k^{\oplus}| = \sum_{i=1}^{m} X_i$. We have

$$\begin{aligned}
\Pr\left[|P_k^{\oplus}| \leq \tfrac{1}{2}m\varepsilon_{k+1}(1-\alpha)\right] &= \Pr\left[\sum_{i=1}^{m} X_i \leq \tfrac{1}{2}m\varepsilon_{k+1}(1-\alpha)\right] \\
&\leq \Pr\left[\sum_{i=1}^{m} X_i \leq \tfrac{1}{2}mL(g_k)(1-\alpha)\right] && \text{(since } L(g_k) > \varepsilon_{k+1}\text{)} \\
&\leq \exp(-m(1-\alpha)L(g_k)/8) && \text{(by Lemma 22 using } \gamma = \tfrac{1}{2}\text{)} \\
&\leq \exp(-m(1-\alpha)\varepsilon_{k+1}/8) && \text{(since } L(g_k) > \varepsilon_{k+1}\text{)}
\end{aligned}$$

Therefore taking

$$m = m_P(\varepsilon_{k+1}, \tfrac{\delta}{2}) \cdot \frac{8\log\left(\frac{1}{(\delta/2)}\right)}{\alpha\varepsilon_{k+1}}$$

as in Lemma 15 ensures that $|P_k^{\oplus}| > 4\log(\frac{2}{\delta}) \cdot m_P(\varepsilon_{k+1}, \frac{\delta}{2})$ with probability $\geq 1 - \exp(-m(\varepsilon_{k+1}, \frac{\delta}{2}) \cdot \log(\frac{1}{\delta/2})) \geq 1 - \frac{\delta}{2}$.

3. https://en.wikipedia.org/wiki/Chernoff_bound#Multiplicative_form_(relative_error)