DIMENSION-FREE DATA FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting high quality data is a crucial prerequisite for training and fine-tuning large language models. While empirical filtering heuristics are widely used, they often lack rigorous guarantees regarding how well the filtered data aligns with the target task. In this paper we propose a theoretical framework for data filtering that provides explicit distribution matching guarantees. Given a small set of N_{S} high quality examples and a larger set of N_{B} lower quality examples, our goal is to design a filter to identify *more* high quality examples. Our main result is a simple mechanism that generates a distribution of examples that is within $\tilde{O}\big(\frac{1}{N_{\mathsf{S}}} + \frac{1}{N_{\mathsf{B}}}\big)$ total variation distance of the high quality distribution. We also show nearly matching information-theoretic lower bounds.

1 INTRODUCTION

Selecting high-quality data is a crucial component in the training of language models (LLMs), one that directly impacts their capabilities. The quality, diversity, and representativeness of the training data influence a model's ability to generate good and effective responses. Low-quality training data can lead to less desirable outcomes such as hallucinations or biases, impairing the effectiveness of an LLM (Bender et al., 2021). Carefully curated data has been shown to mitigate some of these problems (Dodge et al., 2021) and even help in model transferability (Weidinger et al., 2021). Since LLM quality often scales with data size (Kaplan et al., 2020; Hoffmann et al., 2022), selecting high-quality, well-annotated, and domain-relevant examples improves downstream tasks while improving computational efficiency (Brown et al., 2020). Therefore, robust data selection methodologies, including data filtering and deduplication, are essential to training capable LLMs.

Data filtering has found at least three kinds of uses in the modern LLM pipeline. First, filtering provides a way to *augment* data for a fine-tuning task by selecting data from the pre-training dataset that are helpful for the fine-tuning task (Iter & Grangier, 2021; Albalak et al., 2024; Mizrahi et al., 2025). Second, filtering can be used for *removing* irrelevant data during fine-tuning steps, improving efficiency. For instance, the recent work of Xia et al. (2024) shows that strategically selecting influential data for instruction tuning can outperform training on the entire dataset, often using as little as 5% of the data (see also (Yang et al., 2024)). Third, filtering is a powerful way to remove noise or improve data quality. For example, recent open-source datasets designed to support the training and fine-tuning of LLMs such as DCLM-baseline-1.0, RefinedWeb or RedPajama (Li et al., 2024a; Penedo et al., 2023; Weber et al., 2024), use automatic filters for data quality. These filters are based on models that are trained on clean, high-quality data (such as from Wikipedia) to recognize noisy or irrelevant content; a similar filtering method was also used in GPT-3 pretraining (Brown et al., 2020). Given its clear benefits, the practical interest in data filtering has been rapidly growing (e.g., Li et al., 2024b; Tu et al., 2025; Cleanlab, 2024; Susmelj, 2024).

While data selection and filtering have proven effective in practice, much of the work in this area has been empirical or heuristic-based. Some approaches, however, take a more principled view, treating filtering as a selection problem. This perspective allows methods from active learning (e.g., Markov et al., 2023; Susmelj, 2024), information retrieval, optimization (e.g., Killamsetty et al., 2021; Engstrom et al., 2024; Liu et al., 2024), spectral analysis (e.g., Joneidi et al., 2020), and even deep learning itself (Fan et al., 2017) to be used for assigning importance weights to training examples. Importantly, a key characteristic in many of the above lines of work is that they commit to a particular downstream task and assume that labels are available for training a classifier (filter); for a comprehensive survey of data selection methods for LLMs, see (Albalak et al., 2024). Motivated by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

these, we ask: How can we design a model-agnostic data filtering mechanism that provides rigorous, theoretical guarantees on the alignment between the filtered data and the target distribution?

1.1 Our results

Model for data filtering. Suppose we have a *target* distribution S that is of interest for the downstream learning task and we have access to N_{S} independent samples $x_1^{\mathsf{S}}, \dots x_{N_{\mathsf{S}}}^{\mathsf{S}}$ from S. Our goal is to expand this sample to a much larger size in order to improve performance on the downstream task. In order to do so, we also have access to a sample $x_1^{\mathsf{B}}, \dots x_{N_{\mathsf{B}}}^{\mathsf{B}}$ from some *different* distribution B with $N_{\mathsf{B}} \gg N_{\mathsf{S}}$. The goal is to identify a subset of these N_{B} samples to add to our dataset for the downstream task. We assume that B is a mixture of S and some other distribution O. For example, S might represent "high quality" text or images obtained from textbooks or other authoritative sources, while B might represent the general bulk distribution of data scraped from the internet, of which a small fraction is also high quality. Alternatively, the downstream learning task can be one that is very specific (e.g., chest X-ray classification), while B can be a large dataset of X-ray images (of which a small fraction is chest X-rays). Our goal is to design a *filter* that, given a sample $x \sim \mathsf{B}$, will identify whether x comes from the S component or not. We measure the quality of the filter by the total variation distance between the target distribution S and the filtered distribution, which is obtained by keeping only samples from B that pass the filter.

Intuitively, any meaningful filtering mechanism requires some degree of separation between S and O. For example, if O were a re-weighted version of S, filtering would be impossible (since all points would have to pass the filter and the resulting distribution would be a convex combination of S and O). How then should this separation be formalized?

In many settings of interest, distinguishing the target distribution S from the noise distribution O is significantly simpler than the downstream learning task. For example, while building a domain-specific model is complex, a simple classifier can often easily distinguish chest X-rays from other images, or mathematics textbooks from arbitrary documents, assuming a sufficient feature map exists. To capture this notion that the classification problem would be "easy" (specifically, linearly separable) if labels were available, we formally model the overall distribution as a mixture $\mathsf{B} = p \cdot \mathsf{S} + (1-p) \cdot \mathsf{O}$ for some $p \in (0, 1)$. We assume that feature vectors have norms bounded by R (Assumption 1) and we parameterize the difficulty of separating S from O by values ε and γ: all but a ε fraction of S and O can be separated by a hyperplane with a margin γ. We call ε, γ the *overlap parameters* (Definition 1). Intuitively, we are relying on the availability of modern very high-dimensional pre-trained feature vectors/embeddings for our task of interest. With such embeddings, we can be more confident that our simpler filtering task is separable, but at the same time it is critical that our learning results do not depend on the dimensionality of the embeddings. In Section 1.2, we discuss related works that make similar assumptions. Our main result is:

Theorem 1 (Theorem 7, informal)**.** *Let B be a mixture of distributions S and O with overlap parameters ε, γ. Suppose we are given N_{S} and N_{B} samples from S and B respectively. Then one can find a filtering function f such that, if B^f denotes the distribution given by filtering points from B that satisfy $f(x) = 1$, then with probability $\geq 1 - \delta$ over the samples:*

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \leq \tilde{O}\left(\frac{\varepsilon}{p} + \left(\frac{R^2}{\gamma^2} + \log\frac{1}{\delta}\right)\left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{pN_{\mathsf{B}}}\right)\right).$$

Let us first interpret this bound, and see why augmenting the original N_{S} samples with filtered samples from B would help. Consider the simple downstream task of classification (the same intuition carries over to more complex models). If we were to use only the N_{S} samples, we would incur an error that decreases with N_{S} as $1/(N_{\mathsf{S}})^c$, where c is the exponent in the scaling law for the task. For many classification or regression tasks, we have $c = 1/2$ (Kakade et al., 2008). However, if we were to use m additional points from B^f, we can hope to reduce the error to $1/(m + N_{\mathsf{S}})^c$. Of course, filtering is not perfect, and Theorem 1 bounds the *bias* we add. Assuming $\varepsilon = 0$ for simplicity and R, γ, δ are constants, the bias bound is $\tilde{O}\left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{m}\right)$.[1] Thus for $c = 1/2$, choosing $m \approx N_{\mathsf{S}}^2$, the bias is $\tilde{O}(1/N_{\mathsf{S}})$, as is the error from the $1/(m + N_{\mathsf{S}})^{1/2}$ term. This reduces the total error to $\tilde{O}(1/N_{\mathsf{S}})$, a *quadratic* improvement over using just samples from S.

[1] This uses the fact that $pN_{\mathsf{B}} \approx m$, the number of samples that would "pass" the filter with high probability.

One intuitive way to prove the theorem (i.e., design a filter) is by letting through points that are "close" to those in S. A natural idea may be to choose the convex hull of the given N_{S} points from S and let through points that are within a $\gamma/4$ distance of the convex hull. This choice, however, turns out to run into the curse of dimensionality: in d dimensions, we can construct examples with $\gamma = \Omega(1)$ for which this idea fails unless $N_{\mathsf{S}} = \exp(\Omega(d))$. As a second approach, one can try to find/learn the hyperplane that is guaranteed to separate S and O directly via optimization. However, although we have N_{S} positive examples from S, we do not have any known negative examples at all. Moreover, producing enough ground-truth annotations to match the number of positive examples N_{S} (e.g., by asking human experts) may be prohibitively expensive. This naturally maps to the setting of PU learning (Denis, 1998). An intuitive approach to this issue is *negative sampling*: if the mixture weight of S in B is small, then perhaps we could randomly sample some elements from B, assume all of them are not from S, and then train a classifier under this assumption. Generalizing this idea, we might consider solving the following *weighted* empirical risk minimization problem:

$$w := \operatorname*{argmin}_{w} \lambda_{\mathsf{S}} \sum_{i=1}^{N_{\mathsf{S}}} \mathbf{1}[\langle w, x_i^{\mathsf{S}} \rangle \le 0] + \lambda_{\mathsf{B}} \sum_{i=1}^{N_{\mathsf{B}}} \mathbf{1}[\langle w, x_i^{\mathsf{B}} \rangle \ge 0].$$

The weights λ_{S} and λ_{B} should indicate how confident we are in the labeling: we know that each x_i^{S} came from S, but we are only guessing that most of the x_i^{B} are not from S. Moreover, we expect that $N_{\mathsf{B}} \gg N_{\mathsf{S}}$, so that this is a highly imbalanced classification task. We can account for these facts mathematically by setting $\lambda_{\mathsf{S}} \gg \lambda_{\mathsf{B}}$. Our formal approach takes this intuition to the natural extreme: we set $\lambda_{\mathsf{S}} \to \infty$, and $\lambda_{\mathsf{B}} = 1$, which means we attempt to classify all x_i^{B} as not from S, subject to the *constraint* that we classify all x_i^{S} as coming from S. (See Section 3 for details.) For simplicity, our initial result assumes that the algorithm knows the margin parameter γ; in Section 5 we show how to obtain the same result without knowing γ.

We show that the optimization problem we need to solve is unfortunately NP-hard (see Appendix G). Obtaining computationally efficient filtering functions with similar error guarantees (possibly using more complex function classes than linear classifiers) is a challenging and important open direction. Nevertheless, simple convex surrogates suffice to solve the problem in practice.

Lower bounds. Finally, we complement our upper bounds on the error, and prove a matching lower bound. Crucially the lower bound holds in an information-theoretic sense (it is not restricted to computationally efficient algorithms), it holds even when S and O are separated by a margin γ, and when we are given *infinitely* many samples from B.

Theorem 2. *For all sufficiently small $\gamma > 0$ and $N_{\mathsf{S}} \ge 1/\gamma^2$, consider any algorithm for the data filtering problem that obtains at most N_{S} samples from the distribution S (and arbitrarily many samples from B) and outputs a distribution $\widehat{\mathsf{S}}$. There exist distributions S, O, and B with $R \le 3$ and $\varepsilon_{\mathsf{S}} = \varepsilon_{\mathsf{O}} = 0$ that satisfy Definition 1 such that with probability at least $\Omega(1)$,*

$$d_{\mathrm{tv}}(\mathsf{S}, \widehat{\mathsf{S}}) \ge \Omega\left(\frac{1}{\gamma^2 N_{\mathsf{S}}}\right).$$

(This also implies the same lower bound on $\mathbb{E}[d_{\mathrm{tv}}(\mathsf{S}, \widehat{\mathsf{S}})]$.)

The lower bound proofs are based on constructing S and O such that B is independent of the parameters defining S. This property ensures that N_{B} being arbitrarily large does not help! Second, we ensure that unless one can find the parameters defining S, one cannot obtain a good approximation to S. Finally, the result is obtained by reducing the problem to multi-hypothesis testing and using an appropriate version of the classic Fano's inequality. (See Section 4 for details.)

1.2 Related work

Our model of data filtering falls under the umbrella of binary classification with only positive and unlabeled examples (also called PU learning; see (Bekker & Davis, 2020) for atsurvey). Early work on this problem is by Denis (1998), who showed that any concept class learnable by statistical queries is also learnable using polynomially many samples from S and B. Later, Liu et al. (2002) studied a setup essentially similar to ours in the bounded VC-dimension setting. They show that when it is possible to separate S from O using a classifier from a hypothesis class with VC-dimension d,

then solving the same constrained optimization problem we consider yields a classifier with both false negative and false positive rate $\tilde{O}(d/N_{\mathsf{S}})$ so long as N_{B} is sufficiently large; this yields a total variation distance bound of $\tilde{O}(d/N_{\mathsf{S}})$. Our formulation is also related to semi-supervised novelty detection, studied in Blanchard et al. (2010). They showed that detecting a novel component in a mixture reduces to Neyman–Pearson classification between the nominal sample and the mixture sample (see also Scott & Nowak (2005)). These works obtain VC-dimension-based guarantees on the error rates, while we obtain dimension-free margin-based bounds and translate them into TV error guarantees. More recently Lee et al. (2026) has studied a significant generalization of this setting that considers notions of corruption in the data and algorithmic efficiency. They achieve results qualitatively similar to (Liu et al., 2002), but in much greater generality.

Our results differ and improve in three aspects.

1. We consider the large-d setting to obtain margin-based bounds, which is more appropriate for working with modern high-dimensional embeddings.
2. We obtain a tight matching lower bound. In fact, our lower-bound construction sets $d = 1/\gamma^2$, so it *also* shows that the results of (Liu et al., 2002) are tight.
3. By moving to the margin-setting, we have introduced a new difficulty: the margin may be unknown to the algorithm. We provide a new method to address this additional challenge.

Ours vs. margin-based VC-dimension bounds. A fact about the VC-dimension for linear classifiers with positive margin is the following: The set of hypotheses that maintain a margin of γ between positive and negative labels has VC-dimension at most $O(1/\gamma^2)$ (Vapnik & Chervonenkis, 1982; Hush & Scovel, 2001). Now, given the prior bounds of $\tilde{O}(d/N_{\mathsf{S}})$ one should guess that in our margin-based setting we can obtain $\tilde{O}(1/(\gamma^2 N_{\mathsf{S}}))$ in setting where a perfect filter exists, which is indeed exactly what we show. However, it is not so obvious how to apply these bounds here. Roughly, the issue is that since we do not actually see the labels for our points, we cannot identify easily a hypothesis class with VC dimension $O(1/\gamma^2)$. See Appendix H for more detail.

Related models. The idea of filtering or data selection to "match" the distribution of a fine-tuning task has been used in prior work, notably (Liu et al., 2024; Shachaf et al., 2021). However, they set up an optimization problem that uses an EMD-based objective, and use its solution for data selection. Mizrahi et al. (2025) also develop an efficient method to identify pre-training data points that are *close* to the fine-tuning data. But unlike our results, these works do not provide d_{tv} guarantees.

2 PRELIMINARIES

Let S and O be two distributions in $\mathbb{R}^d$ and let $\mathsf{B} = p \cdot \mathsf{S} + (1-p) \cdot \mathsf{O}$ denote their mixture with *mixing* probability $p \in (0,1)$. Let $x_1^{\mathsf{S}}, \ldots x_{N_{\mathsf{S}}}^{\mathsf{S}}$ be N_{S} independent samples drawn from S and let $x_1^{\mathsf{B}}, \ldots, x_{N_{\mathsf{B}}}^{\mathsf{B}}$ be N_{B} independent samples drawn from B. For any binary function $f : \mathbb{R}^d \to \{0,1\}$ and distribution D, let D^f denote the conditional distribution obtained by selecting samples from D that satisfy $f(x) = 1$, i.e., for any $\mathcal{E} \subset \mathbb{R}^d, \Pr_{x\sim\mathsf{D}^f}[x \in \mathcal{E}] = \Pr_{x\sim\mathsf{D}}[x \in \mathcal{E} \mid f(x) = 1]$. If $\Pr_{x\sim\mathsf{D}}[f(x) = 1] = 0$, we set $\mathsf{D}^f = \mathsf{D}$. For any two distributions P and Q, let $d_{\mathrm{tv}}(\mathsf{P},\mathsf{Q}) = \sup_{\mathcal{E}\subset\mathbb{R}^d} |\Pr_{x\sim\mathsf{P}}[x \in \mathcal{E}] - \Pr_{x\sim\mathsf{Q}}[x \in \mathcal{E}]|$ denote the *total variation* distance between them and let $d_{\mathrm{KL}}(\mathsf{P} \,\|\, \mathsf{Q}) = \int \log(d\mathsf{P}/d\mathsf{Q})\, d\mathsf{P}$ denote the *KL-divergence* of P from Q, with value $+\infty$ when P is not absolutely continuous with respect to Q. We use $[x]_1$ to denote $\max\{1, x\}$ and use $\tilde{O}(\cdot)$ to suppress constants and logarithmic factors. Finally, by a mild abuse of notation, for any distribution D, we write $x \in \mathsf{D}$ to indicate that x lies in the support of D.

We make the following assumption throughout the paper: we only consider distributions whose support have bounded norm.

Assumption 1 (Bounded features)**.** $\Pr_{x\sim\mathsf{B}}[\|x\| \le R] = 1$ *for a fixed (possibly unknown)* $R > 0$.

We also parameterize the "overlap" between S and O.

Definition 1 (Overlap parameters)**.** *There is a vector* v, $\|v\| \le 1$ *and parameters* γ, ε_{S}, ε_{O} *such that:*

$$\mathbb{E}_{x\sim\mathsf{S}}[\mathbf{1}[\langle v, x\rangle \le \gamma]] \le \varepsilon_{\mathsf{S}}, \text{ and } \mathbb{E}_{x\sim\mathsf{O}}[\mathbf{1}[\langle v, x\rangle \ge -\gamma]] \le \varepsilon_{\mathsf{O}}.$$

The values ε_{S} and ε_{O} represent the false negative and false positive rates of a margin-based classifier. For any given ε values, we are interested in the largest possible γ that satisfies the requested tolerance. For simplicity, we initially assume our algorithm knows the value of γ, but later relax this in Section 5.

All missing proofs are in the Appendix.

3 Filtering Algorithm

In this section, we design a filter $f : \mathbb{R}^d \to \{0, 1\}$ such that the distribution B^f is close to S, establishing Theorem 1. To simplify presentation, we sketch the analysis for the special case that the overlap parameters ε_{S} and ε_{O} are both zero, but our analysis applies for all parameter values.

Algorithm. Suppose that we are given N_{S} samples from S and N_{B} samples from B. As discussed earlier, we try to find a linear separator f, defined by a vector w, that (i) classifies all the points in N_{S} as 1 and (ii) classifies as many points from N_{B} as possible as 0, while respecting the overlap parameters. More formally, w is given by solving the following optimization problem (parametrized by a failure probability tolerance δ, and which is feasible with probability at least $1-\delta$ by Definition 1 and Corollary 26):

$$\hat{\varepsilon}_{\mathsf{S}} := 2\varepsilon_{\mathsf{S}} + \frac{5\log(1/\delta)}{3N_{\mathsf{S}}}, \qquad w := \operatorname*{argmin}_{\substack{\|w\|\le 1 \\ \frac{1}{N_{\mathsf{S}}}\sum_{i=1}^{N_{\mathsf{S}}} \mathbf{1}[\langle w, x_i^{\mathsf{S}}\rangle \le \gamma] \le \hat{\varepsilon}_{\mathsf{S}}}} \sum_{i=1}^{N_{\mathsf{B}}} \mathbf{1}[\langle w, x_i^{\mathsf{B}}\rangle \ge -\gamma]. \tag{1}$$

To gain some intuition, notice that in the special case that $\varepsilon_{\mathsf{S}} = \varepsilon_{\mathsf{O}} = 0$, we may instead use the simpler objective:

$$w := \operatorname*{argmin}_{\substack{\|w\|\le 1 \\ \sum_{i=1}^{N_{\mathsf{S}}} \mathbf{1}[\langle w, x_i^{\mathsf{S}}\rangle \le \gamma] = 0}} \sum_{i=1}^{N_{\mathsf{B}}} \mathbf{1}[\langle w, x_i^{\mathsf{B}}\rangle \ge -\gamma]. \tag{2}$$

We will describe our analysis in the main text in this simpler setting: all the key intuition transfers to the more general algorithm with a careful accounting for the error. The formulation (2) turns out to be NP hard when $\gamma = 0$ (see Appendix G). However, it can be solved in time $n^{\mathrm{poly}(1/\gamma)}$; see Lee et al. (2026). Let $f : \mathbb{R}^d \to \{0, 1\}$ be defined as $f(x) = \mathbf{1}[\langle w, x\rangle \ge 0]$. If the optimization problem is infeasible, set $f \equiv 1$. Our algorithm then outputs the distribution B^f, i.e., the mixture distribution but conditioned on passing the filter f. We show that f thus defined satisfies Theorem 1.

To show this, we first demonstrate that it suffices to bound the false negative rate and false positive rate of such a filter to obtain a similar bound on the distance between S and the filtered distribution. We defer the proof to Appendix B.

Lemma 3. *For $\mathsf{B} = p \cdot \mathsf{S} + (1-p) \cdot \mathsf{O}$ and for any binary function f, we have*

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \le 2 \Pr_{x\sim\mathsf{S}}[f(x) = 0] + \frac{2}{p} \Pr_{x\sim\mathsf{O}}[f(x) = 1].$$

We bound the false negative and false positive rate of our classifier using PAC-Bayes bounds for margin-based classifiers. Informally, we need to show that the actual probability of $\langle w, x\rangle \le 0$ (for some fixed vector w) is both upper bounded and lower bounded by the empirical margin loss, up to small error terms. Formally, we use the following theorem (full proof in Appendix E for completeness).

Theorem 4. *Let $x \in \mathbb{R}^d$ be a random variable with distribution D. We assume that $\|x\| \le R$, i.e., the support of the distribution D has bounded norm. Let $x_1, \ldots, x_n$ be n independent samples from D. Let $y \in \{1, -1\}$. Then for any $\gamma > 0$ and $0 < \delta < 1$, with probability $\ge 1 - \delta$, we have for all $w \in \mathbb{R}^d$ with $\|w\| \le 1$:*

$$\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] \le \frac{3}{2n} \cdot \sum_{t=1}^{n} \mathbf{1}[\langle w, x_t\rangle y \le \gamma] + \frac{69R^2}{2\gamma^2 n}\left[\log\frac{n\gamma^2}{R^2}\right]_1 + \frac{8}{n}\cdot\log\frac{1}{\delta},$$

and also we have with probability at least $1 - \delta$:

$$\frac{1}{n} \cdot \sum_{t=1}^{n} \mathbf{1}[\langle w, x_t\rangle y \le -\gamma] \le 2 \cdot \Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] + \frac{51R^2}{\gamma^2 n}\left[\log\frac{\gamma^2 n}{R^2}\right]_1 + \frac{12}{n}\cdot\log\frac{1}{\delta}.$$

False negatives. Since w is chosen so that all N_{S} samples from S are correctly classified, we can use standard generalization bounds to bound the false negative rate.

Theorem 5. *For any $\delta \in (0,1)$, with probability $\geq 1-\delta$ (over the samples $\{x_i^{\mathsf{S}}\}_{i=1}^{N_{\mathsf{S}}}$ and $\{x_i^{\mathsf{B}}\}_{i=1}^{N_{\mathsf{B}}}$):*

$$\Pr_{x\sim\mathsf{S}}[f(x)=0] \leq \frac{69R^2}{2\gamma^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\gamma^2}{R^2}\right)\right]_1 + \frac{8\log(1/\delta)}{N_{\mathsf{S}}}.$$

Proof. The same bound holds for $\Pr_{x\sim\mathsf{S}}[\langle w,x\rangle \leq 0]$. The statement follows directly by applying Theorem 4 on samples $x_1^{\mathsf{S}},\ldots,x_{N_{\mathsf{S}}}^{\mathsf{S}}$ from the distribution S and observing that $\mathbf{1}[\langle w, x_t^{\mathsf{S}}\rangle \leq \gamma] = 0,\ \forall t$. □

False positives. We can use standard concentration inequalities to bound the false positive rate $\Pr_{x\sim\mathsf{O}}[f(x)=1]$ as $O(\frac{1}{\sqrt{N_{\mathsf{B}}}})$. However, we can leverage our generalization bound and the fact that false positives need to be counterbalanced by false negatives to obtain a stronger dependence on N_B.

Theorem 6. *Suppose that $p \leq 1/2$, $N_{\mathsf{B}} \geq \max\{4/(1-p), 1/p\}\log(1/\delta)$. Then, with probability $\geq 1-5\delta$:*

$$\Pr_{x\sim\mathsf{O}}[\langle w,x\rangle \geq 0] \leq \tilde{O}\left(\frac{pR^2}{\gamma^2 N_{\mathsf{S}}} + \frac{R^2}{\gamma^2 N_{\mathsf{B}}}\right).$$

Proof sketch. Since w minimizes the number of positive predictions on samples from B while correctly classifying the samples from S, w cannot label significantly more points as positive than the true separator v does. Hence, any "excess" positive labels assigned to the O (i.e., false positives) must be effectively counterbalanced by "missing" positive labels on S (i.e., false negatives). This leads to a situation where the empirical false positive rate on the O-component of B is upper bounded by the empirical false negative rate on the S-component of B; but by Theorem 5, we know that the latter is very small. As before, to obtain the final bound, we use the PAC-Bayes generalization bounds for margin classifiers (Theorem 4). □

Now, we can finally put everything together to prove our main result in Theorem 1, provided in detail with general overlap parameters below.

Theorem 7. *Suppose that $p \leq 1/2$, $N_{\mathsf{B}} \geq \max\{4/(1-p), 1/p\}\log(1/\delta)$, and let S and O have overlap parameters ε_{S}, ε_{O}, γ. Choose w according to* (1)*. Let B^f be the distribution given by filtering points from B that satisfy $f(x) = \mathbf{1}[\langle w,x\rangle \geq 0] = 1$. Then with probability at least $1-8\delta$:*

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \leq \tilde{O}\left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \left(\frac{R^2}{\gamma^2} + \log(1/\delta)\right)\left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{pN_{\mathsf{B}}}\right)\right).$$

Proof. In the case $\varepsilon_{\mathsf{S}} = \varepsilon_{\mathsf{O}} = 0$, the result follows immediately by applying Lemma 3 to Theorem 5 and Theorem 6. The general case requires essentially the same arguments but with careful and tedious bookkeeping for keep track of the ε contributions. We provide a full proof in the Appendix. □

4 Lower Bounds on Error

In this section, we show that the dependence of $\frac{1}{\gamma^2 N_{\mathsf{S}}}$ in our error guarantee is unavoidable. Our lower bound is information-theoretic, using distributions in $\mathbb{R}^d$, for $d \approx \frac{1}{\gamma^2}$.

4.1 A warm-up

We first show a lower bound of $\frac{1}{N_{\mathsf{S}}}$ using distributions on the line, which we then embed in $\mathbb{R}^2$. Specifically, we show that there exists two settings $(\mathsf{S}^0, \mathsf{B}^0)$ and $(\mathsf{S}^1, \mathsf{B}^1)$ of distribution pairs such that (i) $d_{\mathrm{tv}}(\mathsf{S}^0, \mathsf{S}^1) > \Omega(\frac{1}{N_{\mathsf{S}}})$ and (ii) any algorithm that sees at most N_{S} samples from the corresponding small distribution (and possibly infinitely many samples from the corresponding big distribution) fails to distinguish between the two settings with at least a constant probability.

The distributions are defined as follows. Let $\epsilon = \Theta(1/N_{\mathsf{S}})$ be a parameter and let $\gamma = 1/4$.

- S^0 has mass 1 at point $x = 0$. S^1 has mass $(1 - \epsilon)$ at $x = 0$ and mass ϵ at $x = 1$.
- O^0 has mass ϵ at $x = 1$ and mass $1 - \epsilon$ at $x = 2$. O^1 has mass 1 at $x = 2$.
- The two mixture distributions are $\mathsf{B}^0 = \frac{1-\epsilon}{2-\epsilon} \cdot \mathsf{S}^0 + \frac{1}{2-\epsilon} \cdot \mathsf{O}^0$ and $\mathsf{B}^1 = \frac{1}{2-\epsilon} \cdot \mathsf{S}^1 + \frac{1-\epsilon}{2-\epsilon} \cdot \mathsf{O}^1$.

The distributions B^0 and B^1 are identical. Thus, samples from the big distributions cannot help in distinguishing between the cases. After embedding x as $(x, 1)$, both cases satisfy Definition 1 with zero overlap: use the unit vectors proportional to $(-1, 1/2)$ and $(-1, 3/2)$, respectively. The mixing weight is not supplied to the learner in this warm-up; the full construction below fixes $p = 1/2$.

We will now use Fano's inequality to prove our lower bound. Following the framework of Duchi & Wainwright (2013); Scarlett & Cevher (2021), let $V \in \{0, 1\}$ be chosen uniformly at random, and suppose the algorithm is given m samples from S^V, where $m \leq \frac{1}{2\epsilon}$. Let $Y = (Y_1, \ldots, Y_m) \in \{0, 1\}^m$ denote the sample, and let $\widehat{V}$ be the algorithm's estimate/guess for V. If p_e denotes the probability that the algorithm makes an error, i.e., $\Pr[V \neq \widehat{V}]$, then Fano's inequality states that

$$H(V \mid \widehat{V}) \leq H_2(p_e) + p_e \log(|\mathrm{dom}(V)| - 1) = H_2(p_e),$$

where $H_2(\cdot)$ is the binary entropy function. Thus, it suffices to prove that $H(V \mid \widehat{V})$ is $\Omega(1)$, and this would imply that $p_e = \Omega(1)$.

By the data processing inequality, $H(V \mid \widehat{V}) \geq H(V \mid Y)$, since $\widehat{V}$ is purely a function of the samples Y. Now, we can find a lower bound on $H(V \mid Y)$ by a direct computation, as follows:

$$\begin{aligned} H(V \mid Y) &:= \sum_{y \in \{0,1\}^m} \Pr[Y = y] \cdot H(V \mid Y = y) \\ &= \Pr[Y = 0^m] \cdot H(V \mid Y = 0^m) + \sum_{\substack{y \in \{0,1\}^m \\ y \neq 0^m}} \Pr[Y = y] \cdot H(V \mid Y = y). \end{aligned}$$

Notice that the second term on the RHS must be zero (because if $Y \neq 0^m$, we can conclude that $V = 1$). Thus we can focus on the first term. The probability term is easy to evaluate: $\Pr[Y = 0^m] \geq (1 - \epsilon)^m$, and since $m \leq 1/(2\epsilon)$, we have $\Pr[Y = 0^m] \geq 0.5$ for all $m > 2$.

In this case, we can also evaluate $H(V \mid Y = 0^m)$ numerically using Bayes rule, and see that it is ≥ 0.8 (we omit the details since this is a warm-up). Thus, we have $H(V \mid Y) \geq 0.4$, and this implies that $p_e \geq \Omega(1)$. This shows that any algorithm that tries to distinguish between the cases $V = 0$ and $V = 1$ using $m \leq 1/(2\epsilon)$ samples fails with at least a constant probability. Since $d_{\mathrm{tv}}(\mathsf{S}^0, \mathsf{S}^1) \geq \epsilon$, this also implies that if an algorithm outputs a distribution $\widehat{\mathsf{S}}$, then $\mathbb{E}[d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^V)] \geq \Omega(\epsilon)$. (This will be formalized in Lemma 16 below.)

The proof of Theorem 2 can be found in Appendix C. It extends the above ideas us a more delicate multi-way hypothesis testing construction.

5 Unknown Margin

The previous results have assumed knowledge of the margin γ in order to specify the objective (1). In this section, we consider the case that γ is unknown. Instead, given p and a bound R, we assume a user-specified tolerance ϵ_{S} and ϵ_{O} and a conservative range for γ values $[\gamma_{\min}, \gamma_{\max}]$. Given these tolerances, we build an algorithm that obtains the same bound as would be available if the *best possible* γ were chosen a-priori.

Our strategy is the following: we will construct a logarithmic grid of guesses for γ: $\gamma_1 \leq \cdots \leq \gamma_K$ where $\gamma_{i+1} = 2\gamma_i$, $\gamma_1 = \gamma_{\min}$, and $K = 1 + \lfloor \log_2(\gamma_{\max}/\gamma_{\min}) \rfloor$. Now, if we could devise a procedure to test whether a given γ_i satisfies $\gamma/2 \leq \gamma_i \leq \gamma$, then we could simply test all the K guesses and output the distribution B^{f_i} corresponding to the correct γ_i. Unfortunately, it is unclear how to formally establish that any particular γ_i satisfies $\gamma/2 \leq \gamma_i \leq \gamma$. Instead, we will produce a test that does the following with high probability:

1. For $i \geq 2$, if $\gamma_i \leq \gamma$, then γ_i will pass our test.
2. If γ_i passes the test, then $d_{\mathrm{tv}}(\mathsf{B}^{f_i}, \mathsf{B}^{f_{i-1}}) \leq \tilde{O}\left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \frac{1}{\gamma_i^2 N_{\mathsf{S}}} + \frac{1}{p\gamma_i^2 N_{\mathsf{B}}} + \frac{1}{pN_{\mathcal{T}}}\right)$.

Put together, these two properties suggest that we should start with f_1 and identify the largest i such that γ_j passes the test for all $2 \le j \le i$. We stop before any infeasible problem or zero empirical denominator; if the first problem is infeasible, we use $f \equiv 1$. Then, we output B^{f_i}. The first condition above guarantees that $\gamma_i \ge \gamma/2$, because otherwise γ_{i+1} would also pass the test. However, it also may be that $\gamma_i > \gamma$. Nevertheless, the second condition guarantees that for any $j < i$ with $\gamma_j \ge \gamma/2$, we have $d_{\mathrm{tv}}(\mathsf{B}^{f_{j+1}}, \mathsf{B}^{f_j}) \le \tilde{O}\left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \frac{1}{\gamma^2 N_{\mathsf{S}}} + \frac{1}{p\gamma^2 N_{\mathsf{B}}} + \frac{1}{pN_{\mathcal{T}}}\right)$. Thus, $d_{\mathrm{tv}}(\mathsf{B}^{f_i}, \mathsf{S}) \le \tilde{O}\left(K\varepsilon_{\mathsf{S}} + \frac{K\varepsilon_{\mathsf{O}}}{p} + \frac{K}{\gamma^2 N_{\mathsf{S}}} + \frac{K}{p\gamma^2 N_{\mathsf{B}}} + \frac{K}{pN_{\mathcal{T}}}\right)$.

It remains to describe the testing procedure. To do this, we form an estimate of the total variation distance between $\mathsf{B}^{f_{i-1}}$ and B^{f_i}. Estimating the total variation distance between two arbitrary distributions is challenging, but in our case the B^{f_i}'s have a structure that enables an efficient estimate. Specifically, if A_i is the support of B^{f_i}, then the total variation distance satisfies:

$$d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) = \max\left\{\frac{\Pr_{x\sim\mathsf{B}}[x \in A_i \setminus A_{i-1}]}{\Pr_{x\sim\mathsf{B}}[x \in A_i]}, \ \frac{\Pr_{x\sim\mathsf{B}}[x \in A_{i-1} \setminus A_i]}{\Pr_{x\sim\mathsf{B}}[x \in A_{i-1}]}\right\}.$$

The two numerators and two denominators in the above expression can be easily estimated simply by counting how many points lie in the relevant sets. This enables a simple estimate of $d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i})$.

Armed with this estimate, we notice that if $\gamma_i \le \gamma$, then also $d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) \le d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{S}) + d_{\mathrm{tv}}(\mathsf{B}^{f_i}, \mathsf{S})$, which is small by Theorem 7. So, our test uses our estimate to check if the total variation distance is smaller than some threshold specified by Theorem 7. Any $\gamma_i \le \gamma$ passes this test, and for $\gamma_i \ge \gamma$ that also passes the test, the total variation distance cannot increase very much by definition.

To accomplish this, we will assume access to fresh i.i.d. samples from B: $x_1^M, \ldots, x_{N_{\mathcal{T}}}^M$. These new samples will be used to create our total variation distance estimate. (These samples could be obtained, e.g., by splitting the original N_{B}-sized set of samples in two. We will also need $N_{\mathcal{T}}$ to be large enough to make the total variation distance estimates accurate.) Formally, we show the following:

Theorem 8. *Let $x_1^M, \ldots, x_{N_{\mathcal{T}}}^M$ be fresh i.i.d. samples from B. Let $0 < \gamma_{\min} \le \gamma_{\max}$ and ε_{S} and ε_{O} be given, with $0 \le \varepsilon_{\mathsf{S}} < 1/24$ and known $0 < p \le 1/2$ and R. Define $K = 1 + \lfloor \log_2(\gamma_{\max}/\gamma_{\min}) \rfloor$. Let $\Gamma^{\varepsilon_{\mathsf{S}}, \varepsilon_{\mathsf{O}}}$ be the nonempty set of γ values in $[\gamma_{\min}, \gamma_{\max}]$ satisfying the overlap parameter definition (1). Then, for any $0 < \delta < 1$, provided the sample sizes satisfy* (15) *with $\delta_0 = \delta/(44K)$, there exists a filter $f : \mathbb{R}^d \to \{0, 1\}$ such that we have with probability $\ge 1 - \delta$:*

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \le \tilde{O}\left(\inf_{\gamma \in \Gamma^{\varepsilon_{\mathsf{S}}, \varepsilon_{\mathsf{O}}}} K\left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \left(\frac{R^2}{\gamma^2} + \log(K/\delta)\right)\left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{pN_{\mathsf{B}}}\right)\right) + \frac{K\log(K/\delta)}{pN_{\mathcal{T}}}\right).$$

6 Experiments

We conduct a modest empirical study to confirm that our approach yields meaningful filters on synthetic data and on real text data when applied to off-the-shelf pre-trained embeddings.

The optimization in (1) is computationally hard (Appendix G). In practice, we solve by training a linear SVM on synthetic data (logistic regression for real data) with positive class weight $N_{\mathsf{B}}/N_{\mathsf{S}}$.[2] The filter f accepts a point if the classifier scores it positive. All synthetic results are averaged over 10–15 trials and standard deviations are at most a few percent of the means; we omit them for space.

6.1 Synthetic experiments

Data. Points from S are sampled from a Gaussian truncated to lie in $\{x \in \mathbb{R}^{50} : \langle e_1, x\rangle > \gamma\}$ with $\|x\| \le R$, and points from O are sampled analogously with $\langle e_1, x\rangle < -\gamma$. The remaining coordinates are isotropic Gaussian noise scaled so that $\mathbb{E}[\|x\|^2]$ stays bounded. We set $\mathsf{B} = p\mathsf{S} + (1-p)\mathsf{O}$ with $p = 0.01$, $R = 3$. We set $N_{\mathsf{S}} = 200$, $\gamma = 0.1$ and sweep N_{B} from 10^4 to $2{\cdot}10^6$.

Downstream task. To simulate an end-to-end application, we define a downstream task *within* S: predict $\mathrm{sign}(\langle e_2, x\rangle)$. We shift the second coordinate of S samples so that the label is informative ($\Pr[\langle e_2, x\rangle > 0 \mid \mathsf{S}] \approx 0.84$). For O samples the second coordinate is unshifted and the label is

[2] This is consistent with how filters are trained in practice—see, e.g.,(Weber et al., 2024; Brown et al., 2020).

Bernoulli$(1/2)$; this models the situation where a labels are meaningful only on S and are uniformative on O. We then train a logistic regression in three settings: (i) S alone, (ii) S augmented with B samples that pass the filter (*filtered*), (iii) S augmented with the same number of unfiltered B samples (*random*). Test accuracy is reported on $5{,}000$ fresh S samples.

We plot (Figure 1) the downstream application performance under these settings for different values of N_{B}. We also estimate the distance ($\widehat{d_{\mathrm{tv}}}$) between the filtered distribution and S via a histogram-based estimator. We observe applying the filter significantly decreases estimated TV distance and also increases accuracy on the downstream task. The two controls show no improvement in the downstream task. Note that including random samples from O does not *hurt* downstream task performance too much because the labels on O are random.

6.2 Real-world experiment: filtering Wikipedia from a web crawl

To test the framework with more realistic data, we use Wikipedia Wikimedia Foundation (2023) as a high-quality target S and the C4 web crawl Raffel et al. (2020) as the low-quality distribution. We construct B as a mixture of 2,020 Wikipedia paragraphs and 200,000 C4 paragraphs ($p \approx 0.01$, $N_{\mathsf{B}} = 202{,}020$); each paragraph is truncated to 500 characters and embedded with two off-the-shelf sentence-transformer models: MiniLM Reimers & Gurevych (2019) (384-dim) and BGE Xiao et al. (2023) (768-dim). The filter is L2-regularised logistic regression with the same weighting as before. We do not have a synthetic downstream task here, so we report false-negative and false-positive rates against the ground-truth source labels (Table 1).

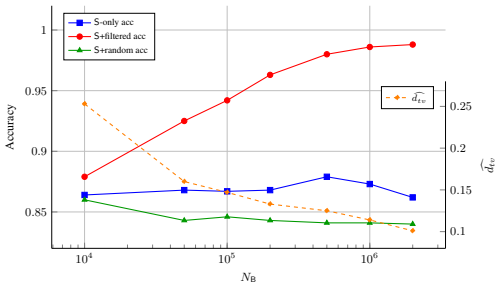

Figure 1: N_{B} vs downstream accuracy and TV-distance. $\widehat{d_{\mathrm{tv}}}$ is a TV-distance estimate generated via histogram binning.

The higher-dimensional BGE embedding yields a better filter than the MiniLM one. This supports our theory: filter quality depends directly on embedding quality rather than dimension.

7 Conclusion and Open Problems

In this work, we formalize the process of selecting "high quality" data from a large bulk dataset, a task that is becoming more and more critical in LLM applications. We provide initial results illustrating that there are significant theoretical advantages available, but this problem is still relatively immature and many open questions remain.

Table 1: Filtering Wikipedia from a Wikipedia/C4 mixture using two off-the-shelf embeddings. $N_{\mathsf{B}} = 202{,}020$, $p \approx 0.01$.

Embedding	dim	N_{S}	FN rate	FP rate
MiniLM	384	2,000	0.180	0.080
BGE	768	2,000	0.173	0.037

One direction is developing methods with comparable guarantees that are computationally efficient and robust to inexact solutions to the optimization problem used to design our filter: in practice, we identify the filter via a convex relaxation (e.g., logistic regression), and the analysis could be extended to cover this scenario. Additionally, our analysis relies on the linear structure of the filter, which aligns with common practice (Weber et al., 2024; Brown et al., 2020). Since our bounds are dimension-free, they are likely to extend to kernelized settings. Nevertheless, a truly general approach applicable to arbitrary filter classes would be even more desirable.

Finally, our model could be extended to capture a broader range of practical applications. For instance, some large datasets used for LLM pretraining provide *quality* scores for each sample, which may be imputed from a small set of data assumed to be high-quality (Weber et al., 2024). Our current framework implements only a binary filter, so developing principled methods to construct such scores is crucial for enabling applications that leverage more of the available data. We hope our work inspires further theoretical studies on this practically important problem.

8 AI Disclosure Statement

In this paper we used AI tools for researching related work, generating code to run experiments and generate the synthetic data, and for checking and correcting minor mistakes in the proofs, on the level of arithmetic errors. All algorithmic and proof ideas were developed by the authors without AI assistance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *TMLR*, 2024.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, pp. 610–623, 2021.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11, 2010.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS 2020*, 2020.

Team Cleanlab. How to detect bad data in your instruction tuning dataset, 2024. URL `https://cleanlab.ai/blog/filter-llm-tuning-data`. Accessed: 2024-02-05.

François Denis. Pac learning from positive statistical queries. In *ALT*, pp. 112–126, 1998.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *EMNLP*, pp. 1286–1305, 2021.

John Duchi. Statistics and Information Theory. `https://web.stanford.edu/class/stats311/lecture-notes.pdf`, 2024.

John C. Duchi and Martin J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv*, 1311.2669, 2013.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: model-aware dataset selection with datamodels. In *ICML*, 2024.

Yang Fan, Fei Tian, Tao Qin, Jiang Bian, and Tie-Yan Liu. Learning what data to learn. *arXiv*, 1702.08635, 2017.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39(2):742–765, 2009.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *NIPS*, pp. 30016–30030, 2022.

Don Hush and Clint Scovel. On the vc dimension of bounded margin classifiers. *Machine Learning*, 45(1):33–44, 2001.

Dan Iter and David Grangier. On the complementarity of data selection and fine tuning for domain adaptation. *arXiv*, 2109.07591, 2021.

D.S. Johnson and F.P. Preparata. The densest hemisphere problem. *TCS*, 6(1):93–107, 1978.

Mohsen Joneidi, Saeed Vahidian, Ashkan Esmaeili, Weijia Wang, Nazanin Rahnavard, Bill Lin, and Mubarak Shah. Select to better learn: Fast and accurate deep learning using data selection from nonlinear manifolds. In *CVPR*, pp. 7816–7826, 2020.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *NIPS*, 2008.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv*, 2001.08361, 2020.

KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh K. Iyer. GLISTER: generalization based data subset selection for efficient and robust learning. In *AAAI*, pp. 8110–8118, 2021.

Jane H Lee, Anay Mehrotra, and Manolis Zampetakis. Learning with positive and imperfect unlabeled data. In *STOC*, 2026.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. In *NeurIPS*, pp. 14200–14282, 2024a.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *NAACL-HLT*, pp. 7602–7635, 2024b.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised classification of text documents. In *ICML*, pp. 387–394, 2002.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: data selection for task-specific model finetuning. In *NeurIPS*, 2024.

Gábor Lugosi and Gergely Neu. Online-to-PAC conversions: Generalization bounds via regret analysis. *arXiv*, 2305.19674, 2023.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *AAAI*, pp. 15009–15018, 2023.

David Mizrahi, Anders Boesen Lindbo Larsen, Jesse Allardice, Suzie Petryk, Yuri Gorokhov, Jeffrey Li, Alex Fang, Josh Gardner, Tom Gunter, and Afshin Dehghan. Language models improve when pretraining data matches target tasks. *arXiv*, 2507.12466, 2025.

Alexander Munteanu, Simon Omlor, and Jeff M. Phillips. Hardness of high-dimensional linear classification. *arXiv*, 2603.19061, 2026.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *NeurIPS*, pp. 79155–79172, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. Dataset available at `https://huggingface.co/datasets/allenai/c4`.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, 2019. Model: `all-MiniLM-L6-v2`, `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

Jonathan Scarlett and Volkan Cevher. *An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation*, pp. 487–528. Cambridge University Press, 2021.

Clayton Scott and Robert Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11), 2005.

Gal Shachaf, Alon Brutzkus, and Amir Globerson. A theoretical analysis of fine-tuning with linear teachers. In *NeurIPS*, pp. 15382–15394, 2021.

Igor Susmelj. Improve your large language models (LLMs) with active learning, 2024. URL https://www.lightly.ai/post/improve-your-large-language-models-llms-with-active-learning. Accessed: 2024-02-05.

Zeao Tu, Xiangdi Meng, Yu He, Zihan Yao, Tianyu Qi, Jun Liu, and Ming Li. Resofilter: Fine-grained synthetic data filtering for large language models through data-parameter resonance analysis. In *NAACL (Findings)*, 2025.

Vladimir N Vapnik and A Ya Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability & Its Applications*, 26(3): 532–553, 1982.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *NeurIPS (Datasets and Benchmarks)*, 2024.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *arXiv*, 2112.04359, 2021.

Wikimedia Foundation. Wikimedia downloads. https://huggingface.co/datasets/wikimedia/wikipedia, 2023. Accessed via Hugging Face Datasets.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *ICML*, pp. 54104–54132, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general Chinese embedding, 2023. Model: BAAI/bge-base-en-v1.5, https://huggingface.co/BAAI/bge-base-en-v1.5.

Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. SmallToLarge (S2L): scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. In *NeurIPS*, pp. 83465–83496, 2024.

A Concentration Bound

We use the following simple concentration bound.

Theorem 9. *Suppose $A_1, \dots, A_T$ is a sequence of random variables supported in $[-1, 1]$, adapted to a filtration $(\mathcal{F}_t)_{t=0}^T$, with conditional mean $\mu_t = \mathbb{E}[A_t \mid \mathcal{F}_{t-1}]$. Let $\lambda \in (0, 1/2]$ and $\delta \in (0, 1)$. Then with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} (\mu_t - A_t) - \lambda \sum_{t=1}^{T} A_t^2 \le \frac{\log(1/\delta)}{\lambda},$$

and also, again with probability at least $1 - \delta$:

$$\sum_{t=1}^{T} (A_t - \mu_t) - \lambda \sum_{t=1}^{T} A_t^2 \le \frac{\log(1/\delta)}{\lambda}.$$

Proof. The first statement follows from the proof of (Lugosi & Neu, 2023, Lemma 27), conditioning on $\mathcal{F}_{t-1}$. While their lemma statement requires $A \in [0, 1]$, inspecting their proof shows that the only time this is used is to establish that $A \le 1$ and so it holds even with $A \in [-1, 1]$. Then, by applying the result to $-A$ instead of A, we obtain the second result. □

B Missing Proofs from Section 3

Lemma 3. *For $\mathsf{B} = p \cdot \mathsf{S} + (1 - p) \cdot \mathsf{O}$ and for any binary function f, we have*

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \le 2 \Pr_{x \sim \mathsf{S}}[f(x) = 0] + \frac{2}{p} \Pr_{x \sim \mathsf{O}}[f(x) = 1].$$

Proof. For convenience, we define

$$\alpha := \Pr_{x \sim \mathsf{S}}[f(x) = 0], \quad \beta := \Pr_{x \sim \mathsf{O}}[f(x) = 1].$$

If $\alpha \ge 1/2$, the RHS is ≥ 1 and the conclusion is trivial. So we can assume that $\alpha < 1/2$. Let ζ be the conditional probability that a retained sample came from the O component. By definition,

$$\zeta = \frac{(1 - p)\beta}{p(1 - \alpha) + (1 - p)\beta}.$$

In order to compare this distribution with S, note that by definition,

$$\Pr_{x \sim \mathsf{S}}[f(x) = 1] = 1 - \alpha.$$

Next, observe that:

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \le \frac{1}{2} \left(|\alpha - \zeta| + \alpha + \zeta \right).$$

Indeed, $\mathsf{B}^f = (1 - \zeta)\mathsf{S}^f + \zeta \mathsf{O}^f$ and $\mathsf{S} = (1 - \alpha)\mathsf{S}^f + \alpha \mathsf{S}^{1-f}$, so the inequality follows by the triangle inequality for signed measures. Terms with zero coefficients may be omitted. Next, we note that since $\alpha < 1/2$,

$$\zeta = \frac{(1 - p)\beta}{p(1 - \alpha) + (1 - p)\beta} \le \frac{2(1 - p)\beta}{p} \le \frac{2\beta}{p}.$$

Thus,

$$d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) \le \alpha + \zeta \le \alpha + \frac{2\beta}{p}.$$

This completes the proof. □

Theorem 6. *Suppose that $p \le 1/2$, $N_{\mathsf{B}} \ge \max\{4/(1 - p), 1/p\} \log(1/\delta)$. Then, with probability $\ge 1 - 5\delta$:*

$$\Pr_{x \sim \mathsf{O}}[\langle w, x \rangle \ge 0] \le \tilde{O}\left(\frac{pR^2}{\gamma^2 N_{\mathsf{S}}} + \frac{R^2}{\gamma^2 N_{\mathsf{B}}} \right).$$

Proof. We first set up some additional notation to ease the presentation. Let $A_{\mathsf{O}} \subset \{x_1^{\mathsf{B}}, \ldots, x_{N_{\mathsf{B}}}^{\mathsf{B}}\}$ be the set of points that came from the O component and let $A_{\mathsf{S}} = \{x_1^{\mathsf{B}}, \ldots, x_{N_{\mathsf{B}}}^{\mathsf{B}}\} \setminus A_{\mathsf{O}}$ be the points that came from the S component. Here the sample collections retain their indices and multiplicities. Conditional on the mixture-component labels, A_{O} consists of $|A_{\mathsf{O}}|$ independent samples from O and similarly for A_{S}. For convenience, we define $\ell^\gamma(z, x, y) = \mathbf{1}[\langle z, x\rangle y \le \gamma]$.

We note that the vector v (that defines the true hyperplane separating S from O) is a feasible point for the optimization problem (2) and also satisfies that $\ell^\gamma(v, x_i^{\mathsf{B}}, -1) = 1$ if and only if $x_i^{\mathsf{B}} \in A_{\mathsf{S}}$. However, since w is chosen to be the optimal point, we have:

$$\sum_{x \in A_{\mathsf{S}} \cup A_{\mathsf{O}}} \ell^\gamma(w, x, -1) \le \sum_{x \in A_{\mathsf{S}} \cup A_{\mathsf{O}}} \ell^\gamma(v, x, -1) = |A_{\mathsf{S}}|.$$

$$\begin{aligned}\sum_{x \in A_{\mathsf{O}}} \ell^\gamma(w, x, -1) &= \sum_{x \in A_{\mathsf{O}} \cup A_{\mathsf{S}}} \ell^\gamma(w, x, -1) - \sum_{x \in A_{\mathsf{S}}} \ell^\gamma(w, x, -1) \\ &\le |A_{\mathsf{S}}| - \sum_{x \in A_{\mathsf{S}}} \ell^\gamma(w, x, -1) \le \sum_{x \in A_{\mathsf{S}}} \mathbf{1}[\langle w, x\rangle \le -\gamma]. \end{aligned} \tag{3}$$

We first apply Theorem 4 to the distribution O with $y = -1$ to get the following with probability at least $1 - \delta$.

$$\begin{aligned}\Pr_{x \sim \mathsf{O}}[\langle w, x\rangle \ge 0] &= \Pr_{x \sim \mathsf{O}}[\langle w, x\rangle(-1) \le 0] \\ &\le \frac{3}{2} \cdot \frac{1}{|A_{\mathsf{O}}|} \sum_{x \in A_{\mathsf{O}}} \ell^\gamma(w, x, -1) + \frac{69R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{|A_{\mathsf{O}}|\gamma^2}{R^2}\right]_1 + \frac{8 \log(1/\delta)}{|A_{\mathsf{O}}|}. \end{aligned} \tag{4}$$

On the event in Lemma 10, $|A_{\mathsf{O}}| > 0$. If $|A_{\mathsf{S}}| = 0$, the final bound follows directly from (3) and (4), so assume $|A_{\mathsf{S}}| > 0$. Focusing on the first term, we first bound it using (3) and then apply Theorem 4 to the distribution S so that we have the following with probability at least $1 - \delta$:

$$\begin{aligned}\sum_{x \in A_{\mathsf{O}}} \ell^\gamma(w, x, -1) &\le \sum_{x \in A_{\mathsf{S}}} \mathbf{1}[\langle w, x\rangle \le -\gamma] \\ &\le |A_{\mathsf{S}}| \cdot \left(2 \Pr_{x \sim \mathsf{S}}[\langle w, x\rangle \le 0] + \frac{51R^2}{\gamma^2 |A_{\mathsf{S}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 + \frac{12 \log(1/\delta)}{|A_{\mathsf{S}}|}\right)\end{aligned}$$

using Theorem 5 to bound $\Pr_{x \sim \mathsf{S}}[\langle w, x\rangle \le 0]$ yields:

$$\begin{aligned}&\le |A_{\mathsf{S}}| \left(\frac{69R^2}{\gamma^2 N_{\mathsf{S}}} \left[\log \frac{N_{\mathsf{S}}\gamma^2}{R^2}\right]_1 + \frac{16 \log(1/\delta)}{N_{\mathsf{S}}} \right. \\ &\qquad \left. + \frac{51R^2}{\gamma^2 |A_{\mathsf{S}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 + \frac{12 \log(1/\delta)}{|A_{\mathsf{S}}|}\right).\end{aligned}$$

Substituting back into (4) and simplifying yields:

$$\begin{aligned}\Pr_{x \sim \mathsf{O}}[\langle w, x\rangle \ge 0] &\le \frac{3 \cdot 69 |A_{\mathsf{S}}| R^2}{2|A_{\mathsf{O}}|\gamma^2 N_{\mathsf{S}}} \left[\log \frac{N_{\mathsf{S}}\gamma^2}{R^2}\right]_1 + \frac{24 |A_{\mathsf{S}}| \log(1/\delta)}{N_{\mathsf{S}} |A_{\mathsf{O}}|} \\ &\quad + \frac{3 \cdot 51R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 + \frac{18 \log(1/\delta)}{|A_{\mathsf{O}}|} \\ &\quad + \frac{69R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{|A_{\mathsf{O}}|\gamma^2}{R^2}\right]_1 + \frac{8 \log(1/\delta)}{|A_{\mathsf{O}}|} \\ &\le \frac{111 \cdot R^2}{\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{(N_{\mathsf{S}} + N_{\mathsf{B}})\gamma^2}{R^2}\right]_1 \left(\frac{|A_{\mathsf{S}}|}{N_{\mathsf{S}}} + 1\right) \\ &\quad + \frac{\log(1/\delta)}{|A_{\mathsf{O}}|} \left(\frac{24 |A_{\mathsf{S}}|}{N_{\mathsf{S}}} + 26\right). \end{aligned} \tag{5}$$

We further simplify terms by bounding the sizes of A_{S} and A_{O}. Lemma 10 in Appendix B shows that with probability $1 - \delta$ (for each of the events) we have that $|A_{\mathsf{O}}| \ge \frac{2(1-p)}{3} N_{\mathsf{B}} - \frac{4 \log(1/\delta)}{3} \ge$

$\frac{(1-p)N_{\mathsf{B}}}{3}$ and $|A_{\mathsf{S}}| \le 2pN_{\mathsf{B}} + 4\log(1/\delta) \le 6pN_{\mathsf{B}}$, where the last inequalities follow since $N_{\mathsf{B}} \ge \max\{\frac{4\log(1/\delta)}{1-p}, \frac{\log(1/\delta)}{p}\}$. Substituting these bounds back into (5), we have:

$$\Pr_{x\sim\mathsf{O}}[\langle w, x\rangle \ge 0] \le \frac{333\cdot R^2}{\gamma^2(1-p)N_{\mathsf{B}}}\left[\log\frac{(N_{\mathsf{S}}+N_{\mathsf{B}})\gamma^2}{R^2}\right]_1\left(\frac{6pN_{\mathsf{B}}}{N_{\mathsf{S}}}+1\right) + \frac{3\log(1/\delta)}{(1-p)N_{\mathsf{B}}}\left(\frac{144pN_{\mathsf{B}}}{N_{\mathsf{S}}}+26\right).$$

We note that we have combined a total of five probability $1-\delta$ events, so overall the statement holds with probability at least $1-5\delta$ by a union bound. □

The following simple lemma provides bounds on the sizes of $|A_{\mathsf{S}}|$ and $|A_{\mathsf{O}}|$.

Lemma 10. *Let A_{O} and A_{S} be the subsets of $\{x_1^{\mathsf{B}}, \dots, x_{N_{\mathsf{B}}}^{\mathsf{B}}\}$ arising from O and S respectively. Then with probability at least $1-\delta$, we have:*

$$|A_{\mathsf{O}}| \ge \frac{2(1-p)}{3}N_{\mathsf{B}} - \frac{4\log(1/\delta)}{3},$$

and with probability at least $1-\delta$, we have:

$$|A_{\mathsf{S}}| \le 2pN_{\mathsf{B}} + 4\log(1/\delta).$$

Proof. Let X_i be the indicator $\mathbf{1}[x_i^{\mathsf{B}} \in A_{\mathsf{O}}]$. Then each X_i is independent and $\mathbb{E}[X_i] = 1-p$. Therefore by Theorem 9 (with $\lambda = 1/2$), we have:

$$\sum_{i=1}^{N_{\mathsf{B}}} X_i \ge \frac{2(1-p)}{3}N_{\mathsf{B}} - \frac{4\log(1/\delta)}{3}.$$

Similarly, let Y_i be the indicator $\mathbf{1}[x_i^B \in A_{\mathsf{S}}]$. Then each Y_i is independent and $\mathbb{E}[Y_i] = p$. Therefore by Theorem 9 (with $\lambda = 1/2$), we have:

$$\sum_{i=1}^{N_{\mathsf{B}}} Y_i \le 2pN_{\mathsf{B}} + 4\log(1/\delta).$$ □

C Missing Proofs from Section 4

In this section, we extend the ideas from the "warmup" in Section 4 to obtain our main lower bound result, Theorem 2.

Outline. We first construct a multi-way hypothesis testing problem, where we have distributions S^V indexed by $V \in \{0,1\}^d$. As in the warm-up, the distributions B^V will turn out to be independent of V (which would imply that irrespective of how large N_{B} is, the samples from B^V do not provide any information about V). Using an information-theoretic argument, we show that the probability of error (defined appropriately) in the hypothesis testing problem must be large with at least a constant probability. We then show a simple lemma that lets us reduce Theorem 2 to multi-way hypothesis testing, thus implying our result. In what follows, we use the following setting of parameters:

$$a := 2\gamma, \qquad \epsilon := \frac{1}{2N_{\mathsf{S}}}, \qquad d := 6\left\lfloor\frac{1}{12a^2}\right\rfloor.$$

We also write m to be the number of samples seen by the algorithm, so $m \le N_{\mathsf{S}}$. We start with a lemma that lets us reduce our lower bound to a multi-way hypothesis testing problem. Let $d_H(\cdot,\cdot)$ denote the Hamming distance.

Lower bound distributions. Our choice of d is a multiple of six and satisfies $d = \Theta(1/\gamma^2)$. The assumption on N_{S} ensures that $d\epsilon \leq 1$. All logarithms and entropies in this section are in base two. Given any $V \in \{0,1\}^d$, we construct distributions S^V and O^V over $\mathbb{R}^{d+1}$ as follows.

Let x drawn from the small distribution S^V be

$$x = \begin{cases} ae_0 + e_i & \text{w.p. } \epsilon v_i, \quad \forall i \in [d] \\ ae_0 & \text{w.p. } 1 - \|V\|_1 \epsilon. \end{cases}$$

Let x drawn from the "other" distribution O^V be

$$x = \begin{cases} ae_0 + e_i & \text{w.p. } \epsilon(1 - v_i), \quad \forall i \in [d] \\ -ae_0 + 2e_i & \text{w.p. } \frac{1}{d}\left(1 - \epsilon(d - \|V\|_1)\right), \quad \forall i \in [d]. \end{cases}$$

Next, we define $\mathsf{B} = \frac{1}{2}(\mathsf{S}^V + \mathsf{O}^V)$. Thus, a sample $x \sim \mathsf{B}^V$ is distributed as

$$x = \begin{cases} ae_0 & \text{w.p. } \frac{1 - \|V\|_1 \epsilon}{2}, \\ ae_0 + e_i & \text{w.p. } \frac{\epsilon}{2} \quad \forall i \in [d], \\ -ae_0 + 2e_i & \text{w.p. } \frac{1}{2d}\left(1 - (d - \|V\|_1)\epsilon\right) \quad \forall i \in [d]. \end{cases}$$

We start with simple observations, followed by a lemma about the margin.

Observation 11. *For all $V \in \{0,1\}^d$ with $\|V\|_1 = d/2$, the distributions B^V are identical.*

Observation 12. *Suppose $V, W \in \{0,1\}^d$, $\|V\|_1 = \|W\|_1$ and $d_H(V,W) = \theta d$, for some $\theta \in [0,1]$. Then*

$$d_{\mathrm{tv}}(\mathsf{S}^V, \mathsf{S}^W) = \frac{1}{2}\theta d\epsilon.$$

Lemma 13. *Let $V \in \{0,1\}^d$ satisfy $\|V\|_1 = d/2$. There exists $z \in \mathbb{R}^{d+1}$ with $\|z\| = 1$, such that for all x, y in the supports of S^V and O^V respectively, we have:*

$$\langle z, x \rangle \geq \frac{a}{\sqrt{2}} > \gamma \quad \text{and} \quad \langle z, y \rangle \leq -\frac{a}{\sqrt{2}} < -\gamma.$$

In the remainder of this section, we prove Theorem 2. Let

$$\mathcal{V} := \{V \in \{0,1\}^d : \|V\|_1 = d/2\}.$$

The lower bound proceeds as follows. We choose V uniformly at random from $\mathcal{V}$ and the algorithm is given m samples from S^V. The algorithm is also allowed an arbitrary number of samples from B^V, but from Observation 11 and our choice of $\mathcal{V}$, these samples do not provide any information about V. We show that for any algorithm that outputs a $\widehat{V}$,

$$\Pr[d_H(\widehat{V}, V) \geq \Omega(d)] \geq \Omega(1).$$

From Lemma 16, this proves Theorem 2. To proceed, we will use a distance-based Fano's inequality.

Lemma 14 (Distance-based Fano's inequality; Duchi & Wainwright (2013))**.** *Let $\mathcal{V}$ be any discrete set with an associated distance ρ. Let t be a distance parameter and let N_t be the number of $V' \in \mathcal{V}$ at a distance $\leq t$ from some V in $\mathcal{V}$. (We assume this number is independent of V.[3]) Then if V is drawn uniformly from $\mathcal{V}$, for any algorithm that observes samples X (depending on V) and outputs $\widehat{V}$,*

$$\Pr[\rho(\widehat{V}, V) \geq t] \geq 1 - \frac{I(V;X) + 1}{\log \frac{|\mathcal{V}|}{N_t}}.$$

To apply Lemma 14, using $\mathcal{V}$ as defined above and using the Hamming distance d_H as ρ, we have $|\mathcal{V}| = \binom{d}{d/2}$ and $N_t \leq \sum_{j=0}^{\lfloor t \rfloor} \binom{d}{j} \leq 2^{dH_2(t/d)}$ for $0 \leq t \leq d/2$. The key step will be to bound $I(V;X)$.

[3]The result of Duchi & Wainwright (2013) is more general, but we only require this *uniform* case.

Lemma 15. *Let $V \sim \mathcal{V}$ be drawn uniformly at random, and let X denote a set of $m \leq \frac{1}{2\epsilon}$ samples from S^V. Then we have $I(V;X) \leq \frac{7d}{8}$.*

The proof proceeds by showing that $H(V \mid X)$ is $\geq d/8$. The intuition for the this is that if $m \leq \frac{1}{2\epsilon}$, there is a good chance that X will "miss" many of the elements of V, and thus there is sufficient remaining entropy. Once we have the following Lemma, Theorem 2 follows by using the distance-based Fano's inequality with $t = d/100$.

Lemma 16. *Suppose $\mathcal{A}$ is an algorithm that takes N_{S} samples from S^V and with probability δ, outputs a distribution $\widehat{\mathsf{S}}$ such that $d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^V) \leq cd\epsilon$. Then there exists an algorithm $\mathcal{A}'$ that takes N_{S} samples from S^V and with probability δ, outputs $\widehat{V}$ such that $d_H(V, \widehat{V}) \leq 4cd$.*

Proof. Since the argument is information-theoretic, we do not need to worry about efficiency of $\mathcal{A}'$. Consider $\mathcal{A}'$ that simply takes the distribution $\widehat{\mathsf{S}}$ output by $\mathcal{A}$, and returns

$$\widehat{V} := \operatorname*{argmin}_{W \in \mathcal{V}} d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^W).$$

We claim that if $d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^V) \leq cd\epsilon$, then $d_H(V, \widehat{V}) \leq 4cd$. We can show this via contradiction. Indeed, by the construction of our probability distributions (Observation 12), if $d_H(V, \widehat{V}) > 4cd$, then $d_{\mathrm{tv}}(\mathsf{S}^V, \mathsf{S}^{\widehat{V}}) > 2cd\epsilon$. Now from the definition of $\widehat{V}$,

$$d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^{\widehat{V}}) \leq d_{\mathrm{tv}}(\widehat{\mathsf{S}}, \mathsf{S}^V) \leq cd\epsilon,$$

and so by the triangle inequality, $d_{\mathrm{tv}}(\mathsf{S}^V, \mathsf{S}^{\widehat{V}}) \leq 2cd\epsilon$, which is a contradiction. □

Lemma 13. *Let $V \in \{0,1\}^d$ satisfy $\|V\|_1 = d/2$. There exists $z \in \mathbb{R}^{d+1}$ with $\|z\| = 1$, such that for all x, y in the supports of S^V and O^V respectively, we have:*

$$\langle z, x \rangle \geq \frac{a}{\sqrt{2}} > \gamma \quad \textit{and} \quad \langle z, y \rangle \leq -\frac{a}{\sqrt{2}} < -\gamma.$$

Proof. Fix any V and consider the vector

$$z := e_0 - \frac{1}{\sqrt{\|V\|_1}} \sum_{i \notin V} e_i.$$

The vectors in the support of S^V are ae_0 and $ae_0 + e_i$, for $i \in V$. For both such vectors x, we have that $\langle z, x \rangle = a$.

Likewise, the vectors in the support of O^V are $ae_0 + e_i$ for $i \notin V$, and $-ae_0 + 2e_i$, for some $i \in [d]$. For y of the first kind, $\langle z, y \rangle = a - \sqrt{\frac{2}{d}}$, and for the second kind, $\langle z, y \rangle$ is either $-a$ or $-a - 2\sqrt{\frac{2}{d}}$. By our choice of d namely $d \leq \frac{1}{2a^2}$, we have $a - \sqrt{\frac{2}{d}} \leq -a$, and so in either case, we have that $\langle z, y \rangle \leq -a$.

As defined above, $\|z\|^2 = 2$, and so dividing through by $\sqrt{2}$, the lemma follows. □

Lemma 15. *Let $V \sim \mathcal{V}$ be drawn uniformly at random, and let X denote a set of $m \leq \frac{1}{2\epsilon}$ samples from S^V. Then we have $I(V;X) \leq \frac{7d}{8}$.*

Proof. We observe that $I(V;X) = H(V) - H(V \mid X)$, and since $H(V) \leq d$, it suffices to prove that $H(V \mid X) \geq d/8$. Let $x_1, \ldots, x_m$ denote the samples X. (Each x_i is a $(d+1)$-dimensional vector.) As before, we have

$$H(V \mid X) = \sum_x \Pr[X = x] \cdot H(V \mid X = x). \tag{6}$$

The main observation is that if $m \leq \frac{1}{2\epsilon}$, there is a good chance that the x will "miss" many of the elements of V. We formalize this below. For any $x = (x_1, \ldots, x_m)$, we say that x "witnesses" index

i if $x_j = ae_0 + e_i$, for some $j \in [m]$. Define $r(x)$ to be the number of distinct indices witnessed by x.

We claim that $\sum_{x\,:\,r(x)\le \frac{d}{3}} \Pr[X = x] \ge \frac{1}{4}$. To show this, we first prove that for any V,

$$\sum_{x\,:\,r(x)>\frac{d}{3}} \Pr[X = x \mid V] \le \frac{3}{4}. \tag{7}$$

Indeed, note that for any V, and any index i such that $v_i = 1$, the probability that it is *not* witnessed by x is $(1-\epsilon)^m \ge 1/2$, by our choice of m. This implies that the expected number of $i \in [d]$ witnessed by x is $\le d/4$. By Markov's inequality, (7) follows, and from the law of conditional probability, the claim follows.

Next, consider any set of samples x such that $r(x) \le d/3$. We claim that $H(V \mid X = x)$ must be large. To see this, observe that given x, there are $\binom{d-r(x)}{d/2}$ potential candidates for V that all have precisely the same likelihood of producing the samples x. This implies that

$$H(V \mid X = x) = \log \binom{d - r(x)}{d/2} \ge \log \binom{2d/3}{d/2}.$$

We use the well known inequality

$$\frac{1}{n+1} \cdot 2^{nH_2(k/n)} \le \binom{n}{k} \le 2^{nH_2(k/n)}. \tag{8}$$

Plugging in our parameters, we obtain $H(V \mid X = x) \ge \frac{2d}{3} H_2(3/4) - \log(2d/3+1) > \frac{d}{2}$, assuming d is large enough. Thus from (6), we have

$$H(V \mid X) \ge \sum_{x\,:\,r(x)\le d/3} \Pr[X = x] \cdot \frac{d}{2} \ge \frac{1}{4} \cdot \frac{d}{2} = \frac{d}{8}.$$

From the definition of $I(V; X)$, this completes the proof. □

We can now complete the proof of our main lower bound.

Theorem 2. *For all sufficiently small $\gamma > 0$ and $N_{\mathsf{S}} \ge 1/\gamma^2$, consider any algorithm for the data filtering problem that obtains at most N_{S} samples from the distribution S (and arbitrarily many samples from B) and outputs a distribution $\widehat{\mathsf{S}}$. There exist distributions S, O, and B with $R \le 3$ and $\varepsilon_{\mathsf{S}} = \varepsilon_{\mathsf{O}} = 0$ that satisfy Definition 1 such that with probability at least $\Omega(1)$,*

$$d_{\mathrm{tv}}(\mathsf{S}, \widehat{\mathsf{S}}) \ge \Omega\left(\frac{1}{\gamma^2 N_{\mathsf{S}}}\right).$$

(This also implies the same lower bound on $\mathbb{E}[d_{\mathrm{tv}}(\mathsf{S}, \widehat{\mathsf{S}})]$.)

Proof. We apply the distance-based Fano's inequality from Lemma 14 with $t = d/100$. With this choice of t, we have

$$\log \frac{|\mathcal{V}|}{N_t} \ge \log \binom{d}{d/2} - dH_2(0.01).$$

Using (8), this term is $\ge d - d \cdot H_2(0.01) - \log(d+1) > 11d/12$, for large enough d.

Thus, applying Fano's inequality, we have that for any algorithm that outputs $\widehat{V}$, if the number of samples is $\le \frac{1}{2\epsilon}$,

$$\Pr\left[d_H(\widehat{V}, V) \ge \frac{d}{100}\right] \ge 1 - \frac{7d/8 + 1}{11d/12} > \frac{1}{25},$$

for large enough d. Using Lemma 16 and $d\epsilon = \Theta(1/(\gamma^2 N_{\mathsf{S}}))$, this completes the proof of the theorem. □

D MISSING PROOFS FOR SECTION 3

False Negatives.

Theorem 17. *For any $\delta \in (0,1)$, with probability $\geq 1-\delta$ (over the samples $\{x_i^{\mathsf{S}}\}_{i=1}^{N_{\mathsf{S}}}$ and $\{x_i^{\mathsf{B}}\}_{i=1}^{N_{\mathsf{B}}}$):*

$$\Pr_{x\sim\mathsf{S}}[f(x)=0] = \Pr_{x\sim\mathsf{S}}[\langle w,x\rangle < 0] \leq \frac{3\hat{\varepsilon}_{\mathsf{S}}}{2} + \frac{69R^2}{2\gamma^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\gamma^2}{R^2}\right)\right]_1 + \frac{8\log(1/\delta)}{N_{\mathsf{S}}}.$$

Proof. The same bound holds for $\Pr_{x\sim\mathsf{S}}[\langle w,x\rangle \leq 0]$ whenever w satisfies the empirical constraint. The statement follows directly by applying Theorem 4 on samples $x_1^{\mathsf{S}},\ldots,x_{N_{\mathsf{S}}}^{\mathsf{S}}$ from distribution S and observing that $\frac{1}{N_{\mathsf{S}}}\sum_{t=1}^{N_{\mathsf{S}}}\mathbf{1}[\langle w,x_t^{\mathsf{S}}\rangle \leq \gamma] \leq \hat{\varepsilon}_{\mathsf{S}}$. □

False Positives.

Theorem 18. *Suppose that $N_{\mathsf{B}} \geq \max\{4/(1-p),1/p\}\log(1/\delta)$. Define $\hat{\varepsilon}_{\mathsf{O}} = 2\varepsilon_{\mathsf{O}} + \frac{5\log(1/\delta)}{3|A_{\mathsf{O}}|}$. Then, with probability $\geq 1-7\delta$:*

$$\begin{aligned}\Pr_{x\sim\mathsf{O}}[f(x)=1] &= \Pr_{x\sim\mathsf{O}}[\langle w,x\rangle \geq 0]\\ &\leq \frac{3}{2}\hat{\varepsilon}_{\mathsf{O}} + \frac{81\cdot p}{1-p}\hat{\varepsilon}_{\mathsf{S}}\\ &\quad + \frac{333\cdot R^2}{\gamma^2(1-p)N_{\mathsf{B}}}\left[\log\frac{(N_{\mathsf{S}}+N_{\mathsf{B}})\gamma^2}{R^2}\right]_1\left(\frac{6pN_{\mathsf{B}}}{N_{\mathsf{S}}}+1\right)\\ &\quad + \frac{3\log(1/\delta)}{(1-p)N_{\mathsf{B}}}\left(\frac{144pN_{\mathsf{B}}}{N_{\mathsf{S}}}+26\right),\end{aligned}$$

and also:

$$\hat{\varepsilon}_{\mathsf{O}} \leq 2\varepsilon_{\mathsf{O}} + \frac{15\log(1/\delta)}{2(1-p)N_{\mathsf{B}}}.$$

Assuming $p \leq 1/2$ and using $\tilde{O}$ to suppress constants and logarithmic factors in $1/\delta$, this is:

$$\Pr_{x\sim\mathsf{O}}[\langle w,x\rangle \geq 0] \leq \tilde{O}\left(\varepsilon_{\mathsf{O}} + \varepsilon_{\mathsf{S}} + \frac{pR^2}{\gamma^2 N_{\mathsf{S}}} + \frac{R^2}{\gamma^2 N_{\mathsf{B}}}\right).$$

Proof. We first set up some additional notation to help simplify the presentation. Let $A_{\mathsf{O}} \subset \{x_1^{\mathsf{B}},\ldots,x_{N_{\mathsf{B}}}^{\mathsf{B}}\}$ be the set of points that came from the O component and let $A_{\mathsf{S}} = \{x_1^{\mathsf{B}},\ldots,x_{N_{\mathsf{B}}}^{\mathsf{B}}\} \setminus A_{\mathsf{O}}$ be the points that came from the S component. Here the sample collections retain their indices and multiplicities. Conditional on the mixture-component labels, A_{O} consists of $|A_{\mathsf{O}}|$ independent samples from O and similarly for A_{S}. For convenience, we define $\ell^\gamma(z,x,y) = \mathbf{1}[\langle z,x\rangle y \leq \gamma]$. Finally, define $\hat{\varepsilon}_{\mathsf{O}} = 2\varepsilon_{\mathsf{O}} + \frac{5\log(1/\delta)}{3|A_{\mathsf{O}}|}$.

Now, by Corollary 26, we have the following, each with probability at least $1-\delta$:

$$\frac{1}{N_{\mathsf{S}}}\sum_{i=1}^{N_{\mathsf{S}}}\mathbf{1}[\langle v,x_i^{\mathsf{S}}\rangle \leq \gamma] \leq 2\varepsilon_{\mathsf{S}} + \frac{5\log(1/\delta)}{3N_{\mathsf{S}}} = \hat{\varepsilon}_{\mathsf{S}},$$

$$\frac{1}{|A_{\mathsf{O}}|}\sum_{x\in A_{\mathsf{O}}}\mathbf{1}[\langle v,x\rangle \geq -\gamma] \leq 2\varepsilon_{\mathsf{O}} + \frac{5\log(1/\delta)}{3|A_{\mathsf{O}}|} = \hat{\varepsilon}_{\mathsf{O}}.$$

This implies that, with high probability, the vector v (that defines a hyperplane approximately separating S from O) is a feasible point for the optimization problem (1). Thus, we have:

$$\sum_{x\in A_{\mathsf{S}}\cup A_{\mathsf{O}}}\ell^\gamma(w,x,-1) \leq \sum_{x\in A_{\mathsf{S}}\cup A_{\mathsf{O}}}\ell^\gamma(v,x,-1) \leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + |A_{\mathsf{S}}|.$$

$$\sum_{x\in A_{\mathsf{O}}}\ell^\gamma(w,x,-1) = \sum_{x\in A_{\mathsf{O}}\cup A_{\mathsf{S}}}\ell^\gamma(w,x,-1) - \sum_{x\in A_{\mathsf{S}}}\ell^\gamma(w,x,-1)$$

$$\leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + |A_{\mathsf{S}}| - \sum_{x \in A_{\mathsf{S}}} \ell^{\gamma}(w, x, -1) \leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + \sum_{x \in A_{\mathsf{S}}} \mathbf{1}[\langle w, x \rangle \leq -\gamma]. \tag{9}$$

We first apply Theorem 4 to the distribution O with $y = -1$ to get the following with probability at least $1 - \delta$.

$$\begin{aligned}\Pr_{x \sim \mathsf{O}}[\langle w, x \rangle \geq 0] &= \Pr_{x \sim \mathsf{O}}[\langle w, x \rangle (-1) \leq 0] \\ &\leq \frac{3}{2} \cdot \frac{1}{|A_{\mathsf{O}}|} \sum_{x \in A_{\mathsf{O}}} \ell^{\gamma}(w, x, -1) + \frac{69R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{|A_{\mathsf{O}}|\gamma^2}{R^2}\right]_1 + \frac{8 \log(1/\delta)}{|A_{\mathsf{O}}|}. \end{aligned} \tag{10}$$

On the event in Lemma 10, $|A_{\mathsf{O}}| > 0$. If $|A_{\mathsf{S}}| = 0$, the final bound follows directly from (9) and (10), so assume $|A_{\mathsf{S}}| > 0$. Focusing on the first term, we first bound it using (9) and then apply Theorem 4 to the distribution S so that we have the following with probability at least $1 - \delta$:

$$\begin{aligned}\sum_{x \in A_{\mathsf{O}}} \ell^{\gamma}(w, x, -1) &\leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + \sum_{x \in A_{\mathsf{S}}} \mathbf{1}[\langle w, x \rangle \leq -\gamma] \\ &\leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + |A_{\mathsf{S}}| \cdot \left(2 \Pr_{x \sim \mathsf{S}}[\langle w, x \rangle \leq 0] + \frac{51R^2}{\gamma^2 |A_{\mathsf{S}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 \right. \\ &\qquad \left. + \frac{12 \log(1/\delta)}{|A_{\mathsf{S}}|}\right)\end{aligned}$$

using Theorem 17 to bound $\Pr_{x \sim \mathsf{S}}[\langle w, x \rangle \leq 0]$ yields:

$$\begin{aligned}&\leq |A_{\mathsf{O}}|\hat{\varepsilon}_{\mathsf{O}} + |A_{\mathsf{S}}| \left(3\hat{\varepsilon}_{\mathsf{S}} + \frac{69R^2}{\gamma^2 N_{\mathsf{S}}} \left[\log \frac{N_{\mathsf{S}}\gamma^2}{R^2}\right]_1 + \frac{16 \log(1/\delta)}{N_{\mathsf{S}}} \right. \\ &\qquad \left. + \frac{51R^2}{\gamma^2 |A_{\mathsf{S}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 + \frac{12 \log(1/\delta)}{|A_{\mathsf{S}}|}\right).\end{aligned}$$

Substituting back into (10) and simplifying yields:

$$\begin{aligned}\Pr_{x \sim \mathsf{O}}[\langle w, x \rangle \geq 0] &\leq \frac{3}{2}\hat{\varepsilon}_{\mathsf{O}} + \frac{9|A_{\mathsf{S}}|}{2|A_{\mathsf{O}}|}\hat{\varepsilon}_{\mathsf{S}} + \frac{3 \cdot 69 |A_{\mathsf{S}}| R^2}{2|A_{\mathsf{O}}|\gamma^2 N_{\mathsf{S}}} \left[\log \frac{N_{\mathsf{S}}\gamma^2}{R^2}\right]_1 \\ &\quad + \frac{24|A_{\mathsf{S}}| \log(1/\delta)}{N_{\mathsf{S}}|A_{\mathsf{O}}|} \\ &\quad + \frac{3 \cdot 51R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{\gamma^2 |A_{\mathsf{S}}|}{R^2}\right]_1 + \frac{18 \log(1/\delta)}{|A_{\mathsf{O}}|} \\ &\quad + \frac{69R^2}{2\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{|A_{\mathsf{O}}|\gamma^2}{R^2}\right]_1 + \frac{8 \log(1/\delta)}{|A_{\mathsf{O}}|} \\ &\leq \frac{3}{2}\hat{\varepsilon}_{\mathsf{O}} + \frac{9|A_{\mathsf{S}}|}{2|A_{\mathsf{O}}|}\hat{\varepsilon}_{\mathsf{S}} \\ &\quad + \frac{111 \cdot R^2}{\gamma^2 |A_{\mathsf{O}}|} \left[\log \frac{(N_{\mathsf{S}} + N_{\mathsf{B}})\gamma^2}{R^2}\right]_1 \left(\frac{|A_{\mathsf{S}}|}{N_{\mathsf{S}}} + 1\right) \\ &\quad + \frac{\log(1/\delta)}{|A_{\mathsf{O}}|} \left(\frac{24|A_{\mathsf{S}}|}{N_{\mathsf{S}}} + 26\right).\end{aligned} \tag{11}$$

We further simplify terms by bounding the sizes of A_{S} and A_{O}. Lemma 10 in Appendix B shows that with probability $1 - \delta$ (for each of the events) we have that $|A_{\mathsf{O}}| \geq \frac{2(1-p)}{3} N_{\mathsf{B}} - \frac{4 \log(1/\delta)}{3} \geq \frac{(1-p)N_{\mathsf{B}}}{3}$ and $|A_{\mathsf{S}}| \leq 2pN_{\mathsf{B}} + 4 \log(1/\delta) \leq 6pN_{\mathsf{B}}$, where the last inequalities follow since $N_{\mathsf{B}} \geq \max\{\frac{4 \log(1/\delta)}{1-p}, \frac{\log(1/\delta)}{p}\}$. Substituting these bounds back into (11), we have:

$$\Pr_{x \sim \mathsf{O}}[\langle w, x \rangle \geq 0] \leq \frac{3}{2}\hat{\varepsilon}_{\mathsf{O}} + \frac{81 \cdot p}{1-p}\hat{\varepsilon}_{\mathsf{S}} + \frac{333 \cdot R^2}{\gamma^2 (1-p) N_{\mathsf{B}}} \left[\log \frac{(N_{\mathsf{S}} + N_{\mathsf{B}})\gamma^2}{R^2}\right]_1 \left(\frac{6pN_{\mathsf{B}}}{N_{\mathsf{S}}} + 1\right)$$

$$+ \frac{3\log(1/\delta)}{(1-p)N_{\mathsf{B}}}\left(\frac{144pN_{\mathsf{B}}}{N_{\mathsf{S}}} + 26\right).$$

We note that we have combined a total of seven probability $1-\delta$ events, so overall the statement holds with probability at least $1-7\delta$ by a union bound. □

Now, we can finally put everything together to prove our main result in Theorem 1, provided in more detail below:

Theorem 7. *Choose w according to* (1)*. Let B^f be the distribution given by filtering points from B that satisfy $f(x) = \mathbf{1}[\langle w, x\rangle \geq 0] = 1$. Then with probability at least $1-8\delta$:*

$$\begin{aligned} d_{\mathrm{tv}}(\mathsf{B}^f, \mathsf{S}) &\leq \left(6 + \frac{4\cdot 81}{1-p}\right)\varepsilon_{\mathsf{S}} + \frac{6}{p}\varepsilon_{\mathsf{O}} + \frac{69R^2}{\gamma^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\gamma^2}{R^2}\right)\right]_1 + \frac{(21 + 270/(1-p))\log(1/\delta)}{N_{\mathsf{S}}} \\ &\quad + \frac{666\cdot R^2}{\gamma^2(1-p)pN_{\mathsf{B}}}\left[\log\left(\frac{(N_{\mathsf{S}}+N_{\mathsf{B}})\gamma^2}{R^2}\right)\right]_1\left(\frac{6pN_{\mathsf{B}}}{N_{\mathsf{S}}} + 1\right) \\ &\quad + \frac{6\log(1/\delta)}{(1-p)pN_{\mathsf{B}}}\left(\frac{144pN_{\mathsf{B}}}{N_{\mathsf{S}}} + \frac{119}{4}\right) \\ &\leq \tilde{O}\left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \left(\frac{R^2}{\gamma^2} + \log(1/\delta)\right)\left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{pN_{\mathsf{B}}}\right)\right). \end{aligned}$$

Proof. The result follows immediately by applying Lemma 3 to Theorem 17 and Theorem 18. □

E PAC-Bayes and Margin Bounds

To bound the false negative rate and the false positive rate of our filter, we appeal to PAC-Bayes generalization bounds for margin-based classifiers. However, our approach requires us to show that the true population probability of some event is both upper and lower bounded by the empirical probability of the event (computed on n independent samples), up to small error terms. Unfortunately, such generalization bounds do not seem to be well-known and so we include a proof for completeness. The following discussion is inspired by the arguments from Lugosi & Neu (2023).

First, we need the following variation on the regret of an optimistic multiplicative weights algorithm.

Theorem 20. *Let $m > 0$ and $\eta \in (0, 1/(4m)]$, let X be any domain and $c_1, \ldots, c_N$ be an arbitrary sequence of cost functions from X to $[-m, m]$. Let $g_1, \ldots, g_N$ be another sequence of functions from X to $[-m, m]$. Let D_0 be any distribution over X. Then there exists an algorithm that outputs distributions D_t over X such that D_t is a function of only $c_1, \ldots, c_{t-1}$ and $g_1, \ldots, g_t$ and, for every distribution $\mathsf{D}_\star$ over X:*

$$\sum_{t=1}^{N}\left(\mathbb{E}_{x\sim\mathsf{D}_t}[c_t(x)] - \mathbb{E}_{x\sim\mathsf{D}_\star}[c_t(x)]\right) \leq \frac{d_{\mathrm{KL}}(\mathsf{D}_\star \,\|\, \mathsf{D}_0)}{\eta} + \eta\sum_{t=1}^{N}\mathbb{E}_{x\sim\mathsf{D}_\star}[(c_t(x) - g_t(x))^2].$$

Proof. By (Lugosi & Neu, 2023, Theorem 23), the above result holds for $c_i : X \to [0,1]$ and $g_i : X \to [0,1]$ with any $\eta \in (0, 1/2]$. To use this fact, define $\hat{c}_i = 1/2 + c_i/2m$ and $\hat{g}_i = 1/2 + g_i/2m$ so that $\hat{c}_i$ and $\hat{g}_t$ both have range $[0,1]$. Then for any $\hat{\eta} = 2m\eta \in (0, 1/2]$, there is a procedure that guarantees:

$$\sum_{t=1}^{N}\left(\mathbb{E}_{x\sim\mathsf{D}_t}[\hat{c}_t(x)] - \mathbb{E}_{x\sim\mathsf{D}_\star}[\hat{c}_t(x)]\right) \leq \frac{d_{\mathrm{KL}}(\mathsf{D}_\star \,\|\, \mathsf{D}_0)}{\hat{\eta}} + \hat{\eta}\sum_{t=1}^{N}\mathbb{E}_{x\sim\mathsf{D}_\star}[(\hat{c}_t(x) - \hat{g}_t(x))^2].$$

Observe that:

$$\sum_{t=1}^{N}\left(\mathbb{E}_{x\sim\mathsf{D}_t}[\hat{c}_t(x)] - \mathbb{E}_{x\sim\mathsf{D}_\star}[\hat{c}_t(x)]\right) = \frac{1}{2m}\sum_{t=1}^{N}\left(\mathbb{E}_{x\sim\mathsf{D}_t}[c_t(x)] - \mathbb{E}_{x\sim\mathsf{D}_\star}[c_t(x)]\right)$$

and

$$(\hat{c}_t(x) - \hat{g}_t(x))^2 = \frac{1}{4m^2}(c_t(x) - g_t(x))^2$$

so that we have:

$$\sum_{t=1}^{N}\left(\mathbb{E}_{x\sim \mathsf{D}_t}[c_t(x)]-\mathbb{E}_{x\sim \mathsf{D}_\star}[c_t(x)]\right)\le \frac{2md_{\mathrm{KL}}\big(\mathsf{D}_\star\,\|\,\mathsf{D}_0\big)}{\hat\eta}+\frac{\hat\eta}{2m}\sum_{t=1}^{N}\mathbb{E}_{x\sim \mathsf{D}_\star}[(c_t(x)-g_t(x))^2]$$
$$=\frac{d_{\mathrm{KL}}\big(\mathsf{D}_\star\,\|\,\mathsf{D}_0\big)}{\eta}+\eta\sum_{t=1}^{N}\mathbb{E}_{x\sim \mathsf{D}_\star}[(c_t(x)-g_t(x))^2]. \qquad \square$$

We can now establish PAC-Bayes style upper and lower bounds on the expected loss.

Theorem 21. *Suppose $\ell_1,\dots,\ell_N$ are i.i.d. functions that map X to $[0,1]$. Let $\mathcal{L}=\mathbb{E}[\ell]$. Let $\mathsf{D}_\star$ be any distribution over X (possibly depending on $\ell_1,\dots,\ell_N$), and let P be any distribution on X (independent of $\ell_1,\dots,\ell_N$). Then for any $\delta\in(0,1)$, we have with probability at least $1-\delta$:*

$$\mathbb{E}_{z\sim \mathsf{D}_\star}\left[\sum_{t=1}^{N}\ell_t(z)\right]\le 2N\cdot\mathbb{E}_{z\sim \mathsf{D}_\star}[\mathcal{L}(z)]+12d_{\mathrm{KL}}\big(\mathsf{D}_\star\,\|\,\mathsf{P}\big)+12\log(1/\delta).$$

Also, with probability at least $1-\delta$, we have:

$$N\cdot\mathbb{E}_{z\sim \mathsf{D}_\star}[\mathcal{L}(z)]\le\frac{3}{2}\mathbb{E}_{z\sim \mathsf{D}_\star}\left[\sum_{t=1}^{N}\ell_t(z)\right]+8d_{\mathrm{KL}}\big(\mathsf{D}_\star\,\|\,\mathsf{P}\big)+8\log(1/\delta).$$

Proof. Let $\eta\le 1/4$ be some constant to be fixed later. Define $c_t(z)=\mathcal{L}(z)-\ell_t(z)+\eta\ell_t(z)^2$ and $g_t(z)=\mathcal{L}(z)$. Notice that since $\ell_t(z)\in[0,1]$, $c_t(z)\ge -1$ and $c_t(z)\le 1+\sup_{a\in[0,1]}(-a+\eta a^2)=1$ so that $c_t(z)\in[-1,1]$. Let $\mathsf{D}_0=\mathsf{P}$ be the prior and $\mathsf{D}_1,\dots,\mathsf{D}_N$ be the outputs of the algorithm whose existence is claimed in Theorem 20.

We first observe that since $\eta\le 1/2$ and $\ell_t(z)\in[0,1]$, $(c_t(z)-g_t(z))^2=(\ell_t(z)-\eta\ell_t(z)^2)^2\le \ell_t(z)^2+\eta^2\ell_t(z)^4\le(1+\eta^2)\ell_t(z)^2\le 2\ell_t(z)^2$. Now, we bound $\mathbb{E}_{z\sim \mathsf{D}_\star}[\sum_t \ell_t(z)-\mathcal{L}(z)]$ as follows:

$$\mathbb{E}_{z\sim \mathsf{D}_\star}\left[\sum_{t=1}^{N}(\ell_t(z)-\mathcal{L}(z))\right]=-\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_\star}[c_t(z)]+\eta\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_\star}[\ell_t(z)^2]$$
$$=\sum_{t=1}^{N}\left(\mathbb{E}_{z\sim \mathsf{D}_t}[c_t(z)]-\mathbb{E}_{z\sim \mathsf{D}_\star}[c_t(z)]\right)+\eta\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_\star}[\ell_t(z)^2]$$
$$-\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_t}[c_t(z)]$$

applying Theorem 20 to bound the first term, and using the observed bound on $(c_t(z)-g_t(z))^2$:

$$\le\frac{d_{\mathrm{KL}}\big(\mathsf{D}_\star\,\|\,\mathsf{P}\big)}{\eta}+3\eta\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_\star}[\ell_t(z)^2]-\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_t}[c_t(z)]. \qquad (12)$$

Now to bound the third term above, let $\mathcal{F}_t=\sigma(\mathsf{P},\ell_1,\dots,\ell_t)$. Since D_t is $\mathcal{F}_{t-1}$-measurable and ℓ_t is independent of $\mathcal{F}_{t-1}$, we have that $\mathbb{E}[\mathbb{E}_{z\sim \mathsf{D}_t}[\ell_t(z)]\mid\mathcal{F}_{t-1}]=\mathbb{E}_{z\sim \mathsf{D}_t}[\mathcal{L}(z)]$. So we can apply Theorem 9 with $A_t=\mathbb{E}_{z\sim \mathsf{D}_t}[\ell_t(z)]$ and $\mu_t=\mathbb{E}_{z\sim \mathsf{D}_t}[\mathcal{L}(z)]$ to get:

$$-\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_t}[c_t(z)]=-\sum_{t=1}^{N}\mathbb{E}_{z\sim \mathsf{D}_t}[\mathcal{L}(z)-\ell_t(z)+\eta\ell_t(z)^2]=\sum_{t=1}^{N}\left(A_t-\mu_t-\eta\mathbb{E}_{z\sim \mathsf{D}_t}[\ell_t(z)^2]\right)$$

applying Jensen inequality and then Theorem 9, we have:

$$\le\sum_{t=1}^{N}\left(A_t-\mu_t-\eta A_t^2\right)\le\frac{\log(1/\delta)}{\eta}.$$

Substituting back into (12),

$$\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N(\ell_t(z)-\mathcal{L}(z))\right]\le\frac{d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})}{\eta}+3\eta\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[\ell_t(z)^2]+\frac{\log(1/\delta)}{\eta},$$

$$\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N\ell_t(z)\right]\le N\mathbb{E}_{z\sim\mathsf{D}_\star}[\mathcal{L}(z)]+\frac{d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})}{\eta}+3\eta\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[\ell_t(z)]+\frac{\log(1/\delta)}{\eta}.$$

Setting $\eta=1/6$ we obtain:

$$\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N\ell_t(z)\right]\le 2N\mathbb{E}_{z\sim\mathsf{D}_\star}[\mathcal{L}(z)]+12d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})+12\log(1/\delta).$$

Similarly, to obtain the other direction, we can define c_t instead by $c_t(z)=\ell_t(z)-\mathcal{L}(z)+\eta\ell_t(z)^2$ for some $\eta=1/8$ and define $g_t(z)=-\mathcal{L}(z)$. Notice that $c_t(z)$ and $g_t(z)$ are both in $[-2,2]$, so that $\eta=1/8$ fulfills the conditions of Theorem 20. As earlier, let $\mathsf{D}_0=\mathsf{P}$ be the prior and let $\mathsf{D}_1,\dots,\mathsf{D}_N$ be the distributions output by the algorithm of Theorem 20. Then we have:

$$\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N(\mathcal{L}(z)-\ell_t(z))\right]=-\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[c_t(z)]+\eta\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[\ell_t(z)^2]$$

$$=\sum_{t=1}^N(\mathbb{E}_{z\sim\mathsf{D}_t}[c_t(z)]-E_{z\sim\mathsf{D}_\star}[c_t(z)])+\eta\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[\ell_t(z)^2]-\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_t}[c_t(z)]$$

applying Theorem 20 and again bounding $(c_t(z)-g_t(z))^2\le 3\ell_t(z)^2$:

$$\le\frac{d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})}{\eta}+4\eta\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_\star}[\ell_t(z)^2]-\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_t}[c_t(z)].$$

Setting $\eta=1/8$ and simplifying yields:

$$N\cdot\mathbb{E}_{z\sim\mathsf{D}_\star}[\mathcal{L}(z)]\le\frac{3}{2}\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N\ell_t(z)\right]+8d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})-\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_t}[c_t(z)].$$

To bound the final term, set $A_t=\mathbb{E}_{z\sim\mathsf{D}_t}[\ell_t(z)]$ and $\mu_t=\mathbb{E}[A_t\mid\mathcal{F}_{t-1}]=\mathbb{E}_{z\sim\mathsf{D}_t}[\mathcal{L}(z)]$ as before. We now have by Theorem 9 and Jensen inequality:

$$-\sum_{t=1}^N\mathbb{E}_{z\sim\mathsf{D}_t}[c_t(z)]=\sum_{t=1}^N(\mu_t-A_t-\eta\mathbb{E}_{x\sim\mathsf{D}_t}[\ell_t(x)^2])\le\sum_{t=1}^N(\mu_t-A_t-\eta A_t^2)\le\frac{\log(1/\delta)}{\eta},$$

so that again:

$$N\cdot\mathbb{E}_{z\sim\mathsf{D}_\star}[\mathcal{L}(z)]\le\frac{3}{2}\mathbb{E}_{z\sim\mathsf{D}_\star}\left[\sum_{t=1}^N\ell_t(z)\right]+8d_{\mathrm{KL}}(\mathsf{D}_\star\,\|\,\mathsf{P})+8\log(1/\delta).\qquad\square$$

Now, we need to use our generalization bound (Theorem 21) to achieve a margin bound. To do so, we use the following lemma from (Duchi, 2024, proof of Theorem 6.2.7).

Lemma 22. *Let $w,x\in\mathbb{R}^d$ and $y\in\{\pm1\}$. Let $\mathsf{D}=\mathsf{N}(w,\sigma^2I)$ with $\sigma>0$. Let $\gamma_1\le\gamma_2\le\gamma_3$ be arbitrary real numbers (including negative numbers). For $x=0$, interpret the exponential terms below as zero. Then:*

$$\mathbf{1}[\langle w,x\rangle y\le\gamma_1]\le\mathbb{E}_{v\sim\mathsf{D}}[\mathbf{1}[\langle v,x\rangle y\le\gamma_2]]+\exp\left(-\frac{(\gamma_2-\gamma_1)^2}{2\sigma^2\|x\|^2}\right),$$

$$\mathbb{E}_{v\sim\mathsf{D}}[\mathbf{1}[\langle v,x\rangle y\le\gamma_2]]\le\mathbf{1}[\langle w,x\rangle y\le\gamma_3]+\exp\left(-\frac{(\gamma_3-\gamma_2)^2}{2\sigma^2\|x\|^2}\right).$$

Proof. The case $x = 0$ is immediate, so assume $x \neq 0$. Let $Z \sim \mathsf{N}(0, \sigma^2 I)$. Notice that $w + Z$ has distribution D. Observe that:

$$\mathbf{1}[\langle w, x\rangle y \leq \gamma_1] - \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_2 - \gamma_1]] \leq \mathbb{E}[\mathbf{1}[\langle w + Z, x\rangle y \leq \gamma_2]].$$

To see this, observe that if $\mathbf{1}[\langle w, x\rangle y \leq \gamma_1] = 0$, then the statement is true since the LHS is non-positive while the RHS is non-negative. If $\mathbf{1}[\langle w, x\rangle y \leq \gamma_1] = 1$, then $\mathbb{E}[\mathbf{1}[\langle w + Z, x\rangle y \leq \gamma_2]] = \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \leq \gamma_2 - \langle w, x\rangle y]] \geq \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \leq \gamma_2 - \gamma_1]] = 1 - \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_2 - \gamma_1]] = \mathbf{1}[\langle w, x\rangle y \leq \gamma_1] - \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_2 - \gamma_1]]$.

Finally, notice that $\langle Z, x\rangle y$ is $\sigma^2\|x\|^2$-subgaussian so that $\mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_2 - \gamma_1]] \leq \exp(-(\gamma_2 - \gamma_1)^2/2\sigma^2\|x\|^2)$ to conclude the desired bound.

Similarly, observe that:

$$\mathbb{E}[\mathbf{1}[\langle w + Z, x\rangle y \leq \gamma_2]] \leq \mathbf{1}[\langle w, x\rangle y \leq \gamma_3] + \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_3 - \gamma_2]].$$

For this statement, notice that if $\mathbf{1}[\langle w, x\rangle y \leq \gamma_3] = 1$, then the statement is true since the LHS is at most 1. So, if $\mathbf{1}[\langle w, x\rangle y \leq \gamma_3] = 0$, then we have $\mathbb{E}[\mathbf{1}[\langle w + Z, x\rangle y \leq \gamma_2]] = \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \leq \gamma_2 - \langle w, x\rangle y]] \leq \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \leq \gamma_2 - \gamma_3]] = \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_3 - \gamma_2]] = \mathbf{1}[\langle w, x\rangle y \leq \gamma_3] + \mathbb{E}[\mathbf{1}[\langle Z, x\rangle y \geq \gamma_3 - \gamma_2]]$ and again the result now follows from sub-Gaussianity of $\langle Z, x\rangle y$. □

Theorem 23. *Let $x \in \mathbb{R}^d$ be a random variable with distribution D. We assume that $\|x\| \leq R$, i.e., the support of the distribution D has bounded norm. Let $x_1, \ldots, x_n$ be n independent samples from D. Let $y \in \{1, -1\}$. Then for any $\gamma > 0, \sigma^2 > 0$ and $0 < \delta < 1$, with probability at least $(1 - \delta)$, we have for all $w \in \mathbb{R}^d$ with $\|w\| \leq 1$:*

$$\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \leq 0] \leq \frac{3}{2n} \cdot \sum_{t=1}^{n} \mathbf{1}[\langle w, x_t\rangle y \leq \gamma] + \frac{5}{2} \cdot \exp\left(-\frac{\gamma^2}{8\sigma^2 R^2}\right) + \frac{4}{n\sigma^2} + \frac{8}{n} \cdot \log\frac{1}{\delta}.$$

Proof.

$$\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \leq 0] = \mathbb{E}_{x\sim\mathsf{D}}[\mathbf{1}[\langle w, x\rangle y \leq 0]].$$

Let $\mathsf{D}_\star = \mathsf{N}(w, \sigma^2 I)$ be a spherical Gaussian with mean w and variance σ^2. Then, applying Lemma 22 with $\gamma_1 = 0$ and $\gamma_2 = \gamma/2$, and using $\|x\| \leq R$, we get

$$\leq \mathbb{E}_{x\sim\mathsf{D}}\mathbb{E}_{v\sim\mathsf{D}_\star}[\mathbf{1}[\langle v, x\rangle y \leq \gamma/2]] + \exp\left(-\frac{\gamma^2}{8\sigma^2 R^2}\right).$$

We first setup some notation to simplify the presentation. For any $v \in \mathbb{R}^d$ and $\gamma > 0$, let $\ell_t^\gamma(v) = \mathbf{1}[\langle v, x_t\rangle y \leq \gamma]$. Let $\mathcal{L}^\gamma(v) = \mathbb{E}_{x\sim\mathsf{D}}[\mathbf{1}[\langle v, x\rangle y \leq \gamma]]$. Then interchanging the expectations above, we have

$$= \mathbb{E}_{v\sim\mathsf{D}_\star}[\mathcal{L}^{\gamma/2}(v)] + \exp\left(-\frac{\gamma^2}{8\sigma^2 R^2}\right). \tag{13}$$

Let $\mathsf{P} = \mathsf{N}(0, \sigma^2 I)$. Then, applying Theorem 21, with probability at least $1 - \delta$, we have

$$n \cdot \mathbb{E}_{v\sim\mathsf{D}_\star}[\mathcal{L}^{\gamma/2}(v)] \leq \frac{3}{2} \cdot \mathbb{E}_{v\sim\mathsf{D}_\star}\left[\sum_{t=1}^{n} \ell_t^{\gamma/2}(v)\right] + 8 d_{\mathrm{KL}}\big(\mathsf{D}_\star \,\|\, \mathsf{P}\big) + 8\log\frac{1}{\delta}.$$

But since $\mathsf{D}_\star$ and P are both Gaussian distributions with the same variance, we have $d_{\mathrm{KL}}\big(\mathsf{D}_\star \,\|\, \mathsf{P}\big) = \frac{\|w\|^2}{2\sigma^2} \leq \frac{1}{2\sigma^2}$. Substituting back into (13) gives:

$$\begin{aligned}\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \leq 0] &\leq \frac{3}{2n} \cdot \mathbb{E}_{v\sim\mathsf{D}_\star}\left[\sum_{t=1}^{n} \ell_t^{\gamma/2}(v)\right] + \frac{4}{n\sigma^2} + \frac{8}{n} \cdot \log\frac{1}{\delta} + \exp\left(-\frac{\gamma^2}{8\sigma^2 R^2}\right) \\ &= \frac{3}{2n} \cdot \sum_{t=1}^{n} \mathbb{E}_{v\sim\mathsf{D}_\star}\left[\ell_t^{\gamma/2}(v)\right] + \frac{4}{n\sigma^2} + \frac{8}{n} \cdot \log\frac{1}{\delta} + \exp\left(-\frac{\gamma^2}{8\sigma^2 R^2}\right). \end{aligned} \tag{14}$$

Applying the second direction of Lemma 22 with $\gamma_2 = \gamma/2$ and $\gamma_3 = \gamma$, we get $\mathbb{E}_{v\sim\mathsf{D}_\star}[\ell_t^{\gamma/2}(v)] = \mathbb{E}_{v\sim\mathsf{D}_\star}[\mathbf{1}[\langle v, x_t\rangle y \le \gamma/2]] \le \mathbf{1}[\langle w, x_t\rangle y \le \gamma] + \exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right)$. Substituting back into (14) then gives:

$$\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] \le \frac{3}{2n}\cdot\sum_{t=1}^{n}\mathbf{1}[\langle w, x_t\rangle y \le \gamma] + \frac{5}{2}\cdot\exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right) + \frac{4}{n\sigma^2} + \frac{8}{n}\cdot\log\frac{1}{\delta}. \square$$

Theorem 24. *Let $x \in \mathbb{R}^d$ be a random variable with distribution D. We assume that $\|x\| \le R$, i.e., the support of the distribution D has bounded norm. Let $x_1, \dots, x_n$ be n independent samples from D. Let $y \in \{1, -1\}$. Then for any $\gamma > 0, \sigma^2 > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$, we have for all $w \in \mathbb{R}^d$ with $\|w\| \le 1$:*

$$\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbf{1}[\langle w, x_t\rangle y \le -\gamma] \le 2\cdot\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] + 3\cdot\exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right) + \frac{6}{n\sigma^2} + \frac{12}{n}\cdot\log\frac{1}{\delta}.$$

Proof. Let $\mathsf{D}_\star = \mathsf{N}(w, \sigma^2 I)$ be a spherical Gaussian with variance σ^2. Then, applying Lemma 22 with $\gamma_1 = -\gamma$ and $\gamma_2 = -\gamma/2$ and using $\|x\| \le R$, we get

$$\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbf{1}[\langle w, x_t\rangle y \le -\gamma] \le \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{v\sim\mathsf{D}_\star}[\mathbf{1}[\langle v, x_t\rangle y \le -\gamma/2]] + \exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right).$$

Let $\ell_t(v) = \mathbf{1}[\langle v, x_t\rangle y \le -\gamma/2]$ and $\mathcal{L}(v) = \mathbb{E}_{x\sim\mathsf{D}}\mathbf{1}[\langle v, x\rangle y \le -\gamma/2]$. Then, we have

$$\le \mathbb{E}_{v\sim\mathsf{D}_\star}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t(v)\right] + \exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right).$$

Let $\mathsf{P} = \mathsf{N}(0, \sigma^2 I)$. We can now apply Theorem 21 to obtain the following with probability at least $1 - \delta$:

$$\le 2\mathbb{E}_{v\sim\mathsf{D}_\star}[\mathcal{L}(v)] + \frac{12 d_{\mathrm{KL}}\big(\mathsf{D}_\star \,\|\, \mathsf{P}\big)}{n} + \frac{12\log(1/\delta)}{n} + \exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right).$$

Substituting $d_{\mathrm{KL}}\big(\mathsf{D}_\star \,\|\, \mathsf{P}\big) = \frac{\|w\|^2}{2\sigma^2} \le \frac{1}{2\sigma^2}$ and exchanging the order of expectations in the first term, we get:

$$\le 2\mathbb{E}_{x\sim\mathsf{D}}\mathbb{E}_{v\sim\mathsf{D}_\star}[\mathbf{1}[\langle v, x\rangle y \le -\gamma/2]] + \frac{6}{n\sigma^2} + \frac{12\log(1/\delta)}{n} + \exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right).$$

Applying the second direction of Lemma 22 with $\gamma_2 = -\gamma/2$ and $\gamma_3 = 0$ then gives:

$$\le 2\mathbb{E}_{x\sim\mathsf{D}}[\mathbf{1}[\langle w, x\rangle y \le 0]] + 3\exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right) + \frac{6}{n\sigma^2} + \frac{12\log(1/\delta)}{n}$$

$$= 2\cdot\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] + 3\cdot\exp\left(-\frac{\gamma^2}{8\sigma^2R^2}\right) + \frac{6}{n\sigma^2} + \frac{12}{n}\cdot\log\frac{1}{\delta}. \quad \square$$

We are now ready to complete the proof of Theorem 4 that we restate for convenience.

Theorem 4. *Let $x \in \mathbb{R}^d$ be a random variable with distribution D. We assume that $\|x\| \le R$, i.e., the support of the distribution D has bounded norm. Let $x_1, \dots, x_n$ be n independent samples from D. Let $y \in \{1, -1\}$. Then for any $\gamma > 0$ and $0 < \delta < 1$, with probability $\ge 1 - \delta$, we have for all $w \in \mathbb{R}^d$ with $\|w\| \le 1$:*

$$\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] \le \frac{3}{2n}\cdot\sum_{t=1}^{n}\mathbf{1}[\langle w, x_t\rangle y \le \gamma] + \frac{69R^2}{2\gamma^2 n}\left[\log\frac{n\gamma^2}{R^2}\right]_1 + \frac{8}{n}\cdot\log\frac{1}{\delta},$$

and also we have with probability at least $1 - \delta$:

$$\frac{1}{n}\cdot\sum_{t=1}^{n}\mathbf{1}[\langle w, x_t\rangle y \le -\gamma] \le 2\cdot\Pr_{x\sim\mathsf{D}}[\langle w, x\rangle y \le 0] + \frac{51R^2}{\gamma^2 n}\left[\log\frac{\gamma^2 n}{R^2}\right]_1 + \frac{12}{n}\cdot\log\frac{1}{\delta}.$$

Proof. The first statement follows from Theorem 23 and the second statement follows from Theorem 24 by setting $\sigma^2 = \frac{\gamma^2}{8R^2[\log(n\gamma^2/R^2)]_1}$. □

E.1 Simple Bernstein inequality

The following is a simple variation on the classical Bernstein inequality:

Theorem 25. *Suppose $X_1, \ldots, X_N \in [-1, 1]$ are independent mean-zero random variables with $\mathbb{E}[X_i^2] = \sigma_i^2$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\sum_{i=1}^N X_i \le \sqrt{2\log(1/\delta)\sum_{i=1}^N \sigma_i^2 + \frac{\log^2(1/\delta)}{9}} + \frac{\log(1/\delta)}{3} \le \sqrt{2\log(1/\delta)\sum_{i=1}^N \sigma_i^2} + \frac{2\log(1/\delta)}{3}.$$

Proof. Let us set $V = \sum_{i=1}^N \sigma_i^2$, $\alpha = \log(1/\delta)$ and t to be the solution to the quadratic equation:

$$t^2 - \frac{2}{3}\alpha t - 2\alpha V = 0$$

$$t = \frac{\alpha}{3} + \sqrt{\frac{\alpha^2}{9} + 2\alpha V} \le \frac{2\alpha}{3} + \sqrt{2\alpha V}.$$

Now, from Bernstein inequality, we have that for this choice of t:

$$\Pr\left[\sum_{i=1}^N X_i \ge t\right] \le \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^N \mathbb{E}[X_i^2] + \frac{1}{3}t}\right) = \delta,$$

from which the conclusion immediately follows. □

For Bernoulli X_i, we have the following immediate corollary:

Corollary 26. *Suppose $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables with $\mathbb{E}[X_i] = p$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\sum_{i=1}^N X_i \le Np + \sqrt{2\log(1/\delta)Np(1-p)} + \frac{2\log(1/\delta)}{3} \le 2Np + \frac{5\log(1/\delta)}{3}.$$

Proof. The first inequality follows from Theorem 25. The second follows since $\sqrt{2ab} \le a + b$. □

F Proofs for Unknown Margin Case

The following theorem restates Theorem 8 and expands upon the filter construction.

Theorem 27. *Suppose we have access to another dataset $x_1^M, \ldots, x_{N_\mathcal{T}}^M$ that is an i.i.d. sample from B, independent of the training data (perhaps by splitting the original sample in two). Let $0 < \gamma_{\min} \le \gamma_{\max}$ and ε_S and ε_O be given, with $0 \le \varepsilon_\mathsf{S} < 1/24$. Suppose $0 < p \le 1/2$ and that p and R are known. Define $K = 1 + \lfloor \log_2(\gamma_{\max}/\gamma_{\min}) \rfloor$ and $\delta_0 = \delta/(44K)$ for $\delta \in (0, 1)$. Let $\Gamma^{\varepsilon_\mathsf{S}, \varepsilon_\mathsf{O}}$ be the nonempty set of γ values in $[\gamma_{\min}, \gamma_{\max}]$ satisfying the overlap separation (Definition 1). Set $\gamma_1 = \gamma_{\min}$ and $\gamma_{i+1} = 2\gamma_i$ for $i < K$. Let C be the absolute constant in Lemma 31, and suppose*

$$N_\mathsf{S} \ge \max\left\{e, \frac{138CR^2}{\gamma_{\min}^2}, 64\log(1/\delta_0)\right\},$$

$$N_\mathsf{B} \ge \max\left\{\frac{4}{1-p}, \frac{1}{p}\right\}\log(1/\delta_0), \qquad N_\mathcal{T} \ge \frac{16\log(1/\delta_0)}{p}. \tag{15}$$

Set $\hat{\varepsilon}_\mathsf{S} = 2\varepsilon_\mathsf{S} + 5\log(1/\delta_0)/(3N_\mathsf{S})$. For each i for which the problem is feasible, define:

$$w_i := \arg \min_{\substack{\|w\| \le 1 \\ \frac{1}{N_\mathsf{S}}\sum_{t=1}^{N_\mathsf{S}} \mathbf{1}[\langle w, x_t^\mathsf{S}\rangle \le \gamma_i] \le \hat{\varepsilon}_\mathsf{S}}} \sum_{t=1}^{N_\mathsf{B}} \mathbf{1}[\langle w, x_t^\mathsf{B}\rangle \ge -\gamma_i].$$

Choose these solutions using only the training data. Set $f_j(x) = \mathbf{1}[\langle w_j, x\rangle \geq 0]$. *Define* $A_j = \{x | f_j(x) = 1\}$ *and, when both denominators are positive, define:*

$$\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_{j-1}}, \mathsf{B}^{f_j}) = \frac{\sum_{i=1}^{N_\mathcal{T}} \mathbf{1}[x_i^M \in A_{j-1} - A_j]}{\sum_{i=1}^{N_\mathcal{T}} \mathbf{1}[x_i^M \in A_{j-1}]} + \frac{\sum_{i=1}^{N_\mathcal{T}} \mathbf{1}[x_i^M \in A_j - A_{j-1}]}{\sum_{i=1}^{N_\mathcal{T}} \mathbf{1}[x_i^M \in A_j]} \tag{16}$$

For $\rho \in (0,1)$*, also define:*

$$\begin{aligned}
\textsc{threshold}_{\gamma_j}(\rho) = 48 \cdot \Bigg[& \left(6 + \frac{4 \cdot 81}{1-p}\right) \varepsilon_\mathsf{S} + \frac{6}{p}\varepsilon_\mathsf{O} + \frac{69R^2}{\gamma_j^2 N_\mathsf{S}} \left[\log\left(\frac{N_\mathsf{S}\gamma_j^2}{R^2}\right)\right]_1 \\
& + \frac{(21 + 270/(1-p))\log(1/\rho)}{N_\mathsf{S}} \\
& + \frac{666 \cdot R^2}{\gamma_j^2 (1-p) p N_\mathsf{B}} \left[\log\left(\frac{(N_\mathsf{S}+N_\mathsf{B})\gamma_j^2}{R^2}\right)\right]_1 \left(\frac{6pN_\mathsf{B}}{N_\mathsf{S}} + 1\right) \\
& + \frac{6\log(1/\rho)}{(1-p)pN_\mathsf{B}} \left(\frac{144pN_\mathsf{B}}{N_\mathsf{S}} + \frac{119}{4}\right) \Bigg] \\
& + \frac{96\log(1/\rho)}{pN_\mathcal{T}}
\end{aligned} \tag{17}$$

If the first problem is infeasible, output $f \equiv 1$*. Otherwise, let* $\hat{i}$ *be the largest index such that all problems up to* $\hat{i}$ *are feasible and, for all* $2 \leq j \leq \hat{i}$*, both denominators in* (16) *are positive and:*

$$\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_{j-1}}, \mathsf{B}^{f_j}) \leq \textsc{threshold}_{\gamma_{j-1}}(\delta_0).$$

Output $f_{\hat{i}}$*. Then, with probability at least* $1 - \delta$*, the first problem is feasible and, simultaneously for every* $\gamma \in \Gamma^{\varepsilon_\mathsf{S}, \varepsilon_\mathsf{O}}$*:*

$$d_{\mathrm{tv}}(\mathsf{B}^{f_i}, \mathsf{S}) \leq \tilde{O}\left(K\left(\varepsilon_\mathsf{S} + \frac{\varepsilon_\mathsf{O}}{p} + \left(\frac{R^2}{\gamma^2} + \log(1/\delta_0)\right)\left(\frac{1}{N_\mathsf{S}} + \frac{1}{pN_\mathsf{B}}\right)\right) + \frac{K\log(1/\delta_0)}{pN_\mathcal{T}}\right)$$

Proof. The proof mainly follows from technical results we will provide later in this section. Fix $\gamma \in \Gamma^{\varepsilon_\mathsf{S}, \varepsilon_\mathsf{O}}$ and let $i_\star$ be the largest index such that $\gamma_{i_\star} \leq \gamma$. Then, we have $\gamma_{i_\star} \geq \gamma/2$. Applying Theorem 7 to all admissible grid points and taking a union bound, with probability at least $1 - 8K\delta_0$, simultaneously for every such $i_\star$, the feasibility event in its proof holds and:

$$d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_{i_\star}}) \leq \tilde{O}\left(\varepsilon_\mathsf{S} + \frac{\varepsilon_\mathsf{O}}{p} + \left(\frac{R^2}{\gamma^2} + \log(1/\delta_0)\right)\left(\frac{1}{N_\mathsf{S}} + \frac{1}{pN_\mathsf{B}}\right)\right).$$

On this event, all problems with index at most $i_\star$ are feasible. Next, applying Theorem 34 with failure parameter δ_0, with probability at least $1 - 36K\delta_0$, we have that for all $2 \leq i \leq K$ for which both problems are feasible, the empirical denominators are positive and:

$$d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) \leq \frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) + \frac{96\log(1/\delta_0)}{pN_\mathcal{T}},$$

and also if $i \leq i_\star$, then

$$\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) \leq \textsc{threshold}_{\gamma_{i-1}}(\delta_0).$$

Thus, $\hat{i} \geq i_\star$. Then, we bound the total variation distance as follows:

$$\begin{aligned}
d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_i}) &\leq d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_{i_\star}}) + \sum_{i=i_\star+1}^{\hat{i}} d_{\mathrm{tv}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) \\
&\leq d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_{i_\star}}) + \sum_{i=i_\star+1}^{\hat{i}} \left(\frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_{i-1}}, \mathsf{B}^{f_i}) + \frac{96\log(1/\delta_0)}{pN_\mathcal{T}}\right)
\end{aligned}$$

$$\leq d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_{i_\star}}) + \sum_{i=i_\star+1}^{\hat{i}} \left(\frac{15}{4} \textsc{threshold}_{\gamma_{i-1}}(\delta_0) + \frac{96 \log(1/\delta_0)}{pN_{\mathcal{T}}} \right)$$

$$\leq d_{\mathrm{tv}}(\mathsf{S}, \mathsf{B}^{f_{i_\star}}) + \frac{15K}{4} \textsc{threshold}_{\gamma_{i_\star}}(\delta_0) + \frac{96K \log(1/\delta_0)}{pN_{\mathcal{T}}}$$

$$\leq \tilde{O}\left(K \left(\varepsilon_{\mathsf{S}} + \frac{\varepsilon_{\mathsf{O}}}{p} + \left(\frac{R^2}{\gamma^2} + \log(1/\delta_0) \right) \left(\frac{1}{N_{\mathsf{S}}} + \frac{1}{pN_{\mathsf{B}}} \right) \right) + \frac{K \log(1/\delta_0)}{pN_{\mathcal{T}}} \right),$$

where the last line uses $\gamma_{i_\star} \geq \gamma/2$, the monotonicity in γ of $\gamma^{-2}[\log(c\gamma^2)]_1$ for $c > 0$, and the form of $\textsc{threshold}_{\gamma_{i_\star}}(\delta_0)$. The two events above hold simultaneously with probability at least $1 - 44K\delta_0 = 1 - \delta$, which completes the proof. □

The proof builds on several technical results, which we present in order, culminating in the final result of this section, Theorem 34, which establishes the key properties of our procedure for checking different γ values. In these results, $\delta \in (0, 1)$, and the filtering solutions use only the training data.

We first show that we can estimate a fraction X/Y by individually estimating X and Y, so long as the estimate for Y is not too small. This is relevant because of estimate total variation distance has exactly this form.

Lemma 28. *Suppose $A \in [2X/3 - \epsilon, 2X + \epsilon]$ and $B \in [2Y/3 - \epsilon, 2Y + \epsilon]$ and $Y \geq 2\epsilon$ for nonnegative X, A and positive Y, B, ϵ. Then:*

$$\frac{4X}{15Y} - \frac{6\epsilon}{Y} \leq \frac{A}{B} \leq \frac{12X}{Y} + \frac{6\epsilon}{Y}.$$

Proof. For the upper bound:

$$\frac{A}{B} \leq \frac{2X + \epsilon}{2Y/3 - \epsilon} \leq \frac{2X + \epsilon}{Y/6} = \frac{12X}{Y} + \frac{6\epsilon}{Y},$$

For the lower bound, if $2X/3 - \epsilon \leq 0$, then $4X/(15Y) - 2\epsilon/(5Y) \leq 0 \leq A/B$. Otherwise:

$$\frac{A}{B} \geq \frac{2X/3 - \epsilon}{2Y + \epsilon} \geq \frac{2X/3 - \epsilon}{5Y/2} = \frac{4X}{15Y} - \frac{2\epsilon}{5Y} \geq \frac{4X}{15Y} - \frac{6\epsilon}{Y}.$$ □

We next show that we can estimate the probability of an event by counting how often that event occurs. We are interested here in *multiplicative* error rather than additive error, and this extra slack allows us obtain reasonable error with a much smaller sample size.

Lemma 29. *Let D be a distribution and E be an measurable set and $x_1, \ldots, x_N$ be an i.i.d. sample from D. Then with probability at least $1 - 2\delta$, both of the following hold:*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] \leq 2 \Pr_{x \sim \mathsf{D}}[x \in E] + \frac{4 \log(1/\delta)}{N},$$

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] \geq \frac{2}{3} \Pr_{x \sim \mathsf{D}}[x \in E] - \frac{4 \log(1/\delta)}{N}.$$

Proof. By Theorem 9, we have that for any $\lambda \in (0, 1/2]$, with probability at least $1 - 2\delta$:

$$\left| \Pr_{x \sim \mathsf{D}}[x \in E] - \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] \right| \leq \frac{\lambda}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] + \frac{\log(1/\delta)}{\lambda N}.$$

Set $\lambda = 1/2$ to conclude:

$$\left| \Pr_{x \sim \mathsf{D}}[x \in E] - \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] \right| \leq \frac{1}{2N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] + \frac{2 \log(1/\delta)}{N}$$

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] - \Pr_{x \sim \mathsf{D}}[x \in E] \leq \frac{1}{2N} \sum_{i=1}^{N} \mathbf{1}[x_i \in E] + \frac{2 \log(1/\delta)}{N}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[x_i \in E] \le 2\Pr_{x\sim\mathsf{D}}[x \in E] + \frac{4\log(1/\delta)}{N}$$

$$\Pr_{x\sim\mathsf{D}}[x \in E] - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[x_i \in E] \le \frac{1}{2N}\sum_{i=1}^{N}\mathbf{1}[x_i \in E] + \frac{2\log(1/\delta)}{N}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[x_i \in E] \ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in E] - \frac{4\log(1/\delta)}{3N}$$

$$\ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in E] - \frac{4\log(1/\delta)}{N}. \qquad \square$$

Now, we show that that we can estimate a ratio of two probabilities by the ratio of empirical estimates of the numerator and denominator, so long as neither denominator is too small.

Lemma 30. *Let* D *be a distribution, let* $x_1, \ldots, x_N$ *be an i.i.d. sample from* D*, and let* A, B *be two (measurable) subsets of its support. Suppose* $\min\{\Pr_{x\sim\mathsf{D}}[x \in A], \Pr_{x\sim\mathsf{D}}[x \in B]\} \ge 8\log(1/\delta)/N$*. Define* $\epsilon = \frac{24\log(1/\delta)}{N\min\{\Pr_{x\sim\mathsf{D}}[x\in A],\Pr_{x\sim\mathsf{D}}[x\in B]\}}$*. Then with probability at least* $1-8\delta$*:*

$$\frac{4\Pr_{x\sim\mathsf{D}}[x \in A\setminus B]}{15\Pr_{x\sim\mathsf{D}}[x \in A]} - \epsilon \le \frac{\sum_{i=1}^{N}\mathbf{1}[x_i \in A\setminus B]}{\sum_{i=1}^{N}\mathbf{1}[x_i \in A]} \le \frac{12\Pr_{x\sim\mathsf{D}}[x \in A\setminus B]}{\Pr_{x\sim\mathsf{D}}[x \in A]} + \epsilon,$$

$$\frac{4\Pr_{x\sim\mathsf{D}}[x \in B\setminus A]}{15\Pr_{x\sim\mathsf{D}}[x \in B]} - \epsilon \le \frac{\sum_{i=1}^{N}\mathbf{1}[x_i \in B\setminus A]}{\sum_{i=1}^{N}\mathbf{1}[x_i \in B]} \le \frac{12\Pr_{x\sim\mathsf{D}}[x \in B\setminus A]}{\Pr_{x\sim\mathsf{D}}[x \in B]} + \epsilon.$$

Proof. First, to make notation simpler, for any set S, let $N_S = \sum_{i=1}^{N}\mathbf{1}[x_i \in S]$. Define $r = \frac{4\log(1/\delta)}{N}$. Then by Lemma 29, the following hold simultaneously with probability at least $1-8\delta$:

$$\frac{N_A}{N} \le 2\Pr_{x\sim\mathsf{D}}[x \in A] + r$$
$$\frac{N_A}{N} \ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in A] - r$$
$$\frac{N_B}{N} \le 2\Pr_{x\sim\mathsf{D}}[B] + r$$
$$\frac{N_B}{N} \ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in B] - r$$
$$\frac{N_{A\setminus B}}{N} \le 2\Pr_{x\sim\mathsf{D}}[x \in A\setminus B] + r$$
$$\frac{N_{A\setminus B}}{N} \ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in A\setminus B] - r$$
$$\frac{N_{B\setminus A}}{N} \le 2\Pr_{x\sim\mathsf{D}}[x \in B\setminus A] + r$$
$$\frac{N_{B\setminus A}}{N} \ge \frac{2}{3}\Pr_{x\sim\mathsf{D}}[x \in B\setminus A] - r$$

Next, because $\Pr_{x\sim\mathsf{D}}[A] \ge 2r$ and $\Pr_{x\sim\mathsf{D}}[B] \ge 2r$, by Lemma 28 we have:

$$\frac{4\Pr_{x\sim\mathsf{D}}[x \in A\setminus B]}{15\Pr_{x\sim\mathsf{D}}[x \in A]} - \frac{6r}{\Pr_{x\sim\mathsf{D}}[x \in A]} \le \frac{N_{A\setminus B}}{N_A} \le \frac{12\Pr_{x\sim\mathsf{D}}[x \in A\setminus B]}{\Pr_{x\sim\mathsf{D}}[x \in A]} + \frac{6r}{\Pr_{x\sim\mathsf{D}}[x \in A]},$$

$$\frac{4\Pr_{x\sim\mathsf{D}}[x \in B\setminus A]}{15\Pr_{x\sim\mathsf{D}}[x \in B]} - \frac{6r}{\Pr_{x\sim\mathsf{D}}[x \in B]} \le \frac{N_{B\setminus A}}{N_B} \le \frac{12\Pr_{x\sim\mathsf{D}}[x \in B\setminus A]}{\Pr_{x\sim\mathsf{D}}[x \in B]} + \frac{6r}{\Pr_{x\sim\mathsf{D}}[x \in B]}.$$

Substitute in the value of r to conclude the desired result. $\square$

We next show that even if we use a poor estimate for γ, the number of points that pass the filter is still reasonably large.

Lemma 31. *For a given $\hat{\gamma}$ and ε_{S} and δ, let w be a solution, when feasible, to:*

$$\hat{\varepsilon}_{\mathsf{S}} := 2\varepsilon_{\mathsf{S}} + \frac{5\log(1/\delta)}{3N_{\mathsf{S}}}$$

$$w := \underset{\substack{\|w\|\leq 1 \\ \frac{1}{N_{\mathsf{S}}}\sum_{i=1}^{N_{\mathsf{S}}} \mathbf{1}[\langle w, x_i^{\mathsf{S}}\rangle \leq \hat{\gamma}] \leq \hat{\varepsilon}_{\mathsf{S}}}}{\textit{argmin}} \sum_{i=1}^{N_{\mathsf{B}}} \mathbf{1}[\langle w, x_i^{\mathsf{B}}\rangle \geq -\hat{\gamma}],$$

Let C be the positive number satisfying $C = 2[\log(C \cdot 2 \cdot 69)]_1$. Suppose $N_{\mathsf{S}} \geq \max[\exp(1), C \cdot \frac{2\cdot 69\cdot R^2}{\hat{\gamma}^2}, 64\log(1/\delta)]$. Suppose also $0 < \hat{\gamma}$ and $0 \leq \varepsilon_{\mathsf{S}} \leq \frac{1}{24}$. Then, with probability at least $1-\delta$, every feasible solution w satisfies:

$$\Pr_{x\sim\mathsf{B}}[\langle w, x\rangle \geq 0] \geq p/2.$$

Proof. Inspection of the proof of Theorem 17 reveals that it does not actually require the margin used to define w to be in any way related to the "true" margin γ. So, we may apply that result here to obtain that with probability at least $1-\delta$:

$$\Pr_{x\sim\mathsf{S}}[\langle w, x\rangle < 0] \leq \frac{3\hat{\varepsilon}_{\mathsf{S}}}{2} + \frac{69R^2}{2\hat{\gamma}^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\hat{\gamma}^2}{R^2}\right)\right]_1 + \frac{8\log(1/\delta)}{N_{\mathsf{S}}}.$$

Now, since $\varepsilon_{\mathsf{S}} \leq \frac{1}{24}$ and $N_{\mathsf{S}} \geq 20\log(1/\delta)$, we have:

$$\begin{aligned}\frac{3\hat{\varepsilon}_{\mathsf{S}}}{2} &= 3\cdot\varepsilon_{\mathsf{S}} + \frac{5\log(1/\delta)}{2N_{\mathsf{S}}} \\ &\leq \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4}\end{aligned}$$

Further, we also have:

$$\frac{8\log(1/\delta)}{N_{\mathsf{S}}} \leq \frac{1}{8}$$

Finally, let us bound $\frac{69R^2}{2\hat{\gamma}^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\hat{\gamma}^2}{R^2}\right)\right]_1$. Since $\hat{\varepsilon}_{\mathsf{S}} < 1$, feasibility implies $\hat{\gamma} < R$, so $\exp(1) \leq C\cdot 2\cdot 69R^2/\hat{\gamma}^2$. We have that $\frac{69R^2}{2\hat{\gamma}^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\hat{\gamma}^2}{R^2}\right)\right]_1$ is decreasing in N_{S}. Therefore, if $\exp(1) \leq C\cdot\frac{2\cdot 69\cdot R^2}{\hat{\gamma}^2} \leq N_{\mathsf{S}}$, we must have:

$$\frac{69R^2}{2\hat{\gamma}^2 N_{\mathsf{S}}}\left[\log\left(\frac{N_{\mathsf{S}}\hat{\gamma}^2}{R^2}\right)\right]_1 \leq \frac{[\log(C\cdot 2\cdot 69)]_1}{4C} = \frac{1}{8}$$

So, overall, we have that:

$$\Pr_{x\sim\mathsf{S}}[\langle w, x\rangle < 0] \leq 1/2$$

Thus, we have:

$$\Pr_{x\sim\mathsf{B}}[\langle w, x\rangle \geq 0] \geq p \Pr_{x\sim\mathsf{S}}[\langle w, x\rangle \geq 0] = p - p\Pr_{x\sim\mathsf{S}}[\langle w, x\rangle < 0] \geq p/2. \qquad \square$$

Next, we establish that the $\widehat{d_{\mathrm{tv}}}$ value computed in the our unknown margin procedure is a good estimate of the total variation distance.

Theorem 32. *Let D be a distribution and let A, B be two (measurable) subsets of the support of D. Let D_A be the distribution restricted to A and D_B be the distribution restricted to B, so*

that $\Pr_{x\sim \mathsf{D}_A}[x \in E] = \Pr_{x\sim\mathsf{D}}[x \in E|\ x \in A]$ *and similarly for* D_B. *Suppose* $\min\{\Pr_{x\sim\mathsf{D}}[x \in A], \Pr_{x\sim\mathsf{D}}[x \in B]\} \geq 8\log(1/\delta)/N$. *Let* $x_1, \ldots, x_N$ *be an i.i.d. sample from* D. *Define:*

$$\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) = \frac{\sum_{i=1}^N \mathbf{1}[x_i \in A \setminus B]}{\sum_{i=1}^N \mathbf{1}[x_i \in A]} + \frac{\sum_{i=1}^N \mathbf{1}[x_i \in B \setminus A]}{\sum_{i=1}^N \mathbf{1}[x_i \in B]}.$$

If either denominator is zero, set $\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) = 0$. *Then with probability at least* $1 - 8\delta$, *both denominators are positive and we have:*

$$d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) \leq \frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) + \frac{48\log(1/\delta)}{N \cdot \min\{\Pr_{x\sim\mathsf{D}}[x \in B], \Pr_{x\sim\mathsf{D}}[x \in A]\}},$$

and also:

$$\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) \leq 24 d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) + \frac{48\log(1/\delta)}{N \cdot \min\{\Pr_{x\sim\mathsf{D}}[x \in B], \Pr_{x\sim\mathsf{D}}[x \in A]\}}.$$

Proof. For simplicity, for any set S, let $N_S = \sum_{i=1}^N \mathbf{1}[x_i \in S]$. In this notation, we have:

$$\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) = \frac{N_{A\setminus B}}{N_A} + \frac{N_{B\setminus A}}{N_B}. \tag{18}$$

Moreover, since for all $E \subset A \cap B$, we have $\mathsf{D}_A[E] = \lambda \mathsf{D}_B[E]$ for some fixed λ, we also have:

$$\begin{aligned}
d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) &= \frac{1}{2}\left(\Pr_{x\sim\mathsf{D}_A}[x \in A \setminus B] + \Pr_{x\sim\mathsf{D}_B}[x \in B \setminus A]\right. \\
&\qquad \left. + \left|\Pr_{x\sim\mathsf{D}_A}[x \in A \cap B] - \Pr_{x\sim\mathsf{D}_B}[x \in B \cap A]\right|\right) \\
&= \frac{1}{2}\left(\Pr_{x\sim\mathsf{D}_A}[x \in A \setminus B] + \Pr_{x\sim\mathsf{D}_B}[x \in B \setminus A]\right. \\
&\qquad \left. + \left|1 - \Pr_{x\sim\mathsf{D}_A}[x \in A \setminus B] - (1 - \Pr_{x\sim\mathsf{D}_B}[x \in B - A])\right|\right) \\
&= \max\left\{\Pr_{x\sim\mathsf{D}_A}[x \in A \setminus B],\ \Pr_{x\sim\mathsf{D}_B}[x \in B \setminus A]\right\} \\
&\leq \Pr_{x\sim\mathsf{D}_A}[x \in A \setminus B] + \Pr_{x\sim\mathsf{D}_B}[x \in B \setminus A] \\
&= \frac{\Pr_{x\sim\mathsf{D}}[x \in A \setminus B]}{\Pr_{x\sim\mathsf{D}}[x \in A]} + \frac{\Pr_{x\sim\mathsf{D}}[x \in B \setminus A]}{\Pr_{x\sim\mathsf{D}}[x \in B]}.
\end{aligned} \tag{19}$$

Our goal is to show that the above two terms in (19) are roughly equal to the two terms in (18) up to a constant factor. Define $\epsilon = \frac{24\log(1/\delta)}{N\min\{\Pr_{x\sim\mathsf{D}}[x\in A], \Pr_{x\sim\mathsf{D}}[x\in B]\}}$. By the proof of Lemma 30, using the intermediate lower bound $4X/(15Y) - 2r/(5Y)$ from Lemma 28 with $r = 4\log(1/\delta)/N$, with probability at least $1 - 8\delta$, we have:

$$\begin{aligned}
\frac{4\Pr_{x\sim\mathsf{D}}[x \in A \setminus B]}{15\Pr_{x\sim\mathsf{D}}[x \in A]} - \epsilon/15 \leq \frac{\sum_{i=1}^N \mathbf{1}[x_i \in A \setminus B]}{\sum_{i=1}^N \mathbf{1}[x_i \in A]} \leq \frac{12\Pr_{x\sim\mathsf{D}}[x \in A \setminus B]}{\Pr_{x\sim\mathsf{D}}[x \in A]} + \epsilon, \\
\frac{4\Pr_{x\sim\mathsf{D}}[x \in B \setminus A]}{15\Pr_{x\sim\mathsf{D}}[x \in B]} - \epsilon/15 \leq \frac{\sum_{i=1}^N \mathbf{1}[x_i \in B \setminus A]}{\sum_{i=1}^N \mathbf{1}[x_i \in B]} \leq \frac{12\Pr_{x\sim\mathsf{D}}[x \in B \setminus A]}{\Pr_{x\sim\mathsf{D}}[x \in B]} + \epsilon.
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\frac{N_{A\setminus B}}{N_A} + \frac{N_{B\setminus A}}{N_B} &\geq \frac{4}{15}\left(\frac{\Pr_{x\sim\mathsf{D}}[x \in A \setminus B]}{\Pr_{x\sim\mathsf{D}}[x \in A]} + \frac{\Pr_{x\sim\mathsf{D}}[x \in B \setminus A]}{\Pr_{x\sim\mathsf{D}}[x \in B]}\right) - 2\epsilon/15 \\
&\geq \frac{4}{15} d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) - 2\epsilon/15.
\end{aligned}$$

Rearranging gives $d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) \leq (15/4)\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B) + \epsilon/2$, which implies the stated bound since $\epsilon/2 \leq 2\epsilon$.

For the upper bound on $\widehat{d_{\mathrm{tv}}}(\mathsf{D}_A, \mathsf{D}_B)$, observe that:

$$\begin{aligned} d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) &\geq \frac{1}{2} \Pr_{x \sim \mathsf{D}_A}[A \setminus B] + \frac{1}{2} \Pr_{x \sim \mathsf{D}_B}[B \setminus A] \\ &= \frac{1}{2} \left(\frac{\Pr_{x \sim \mathsf{D}}[x \in A \setminus B]}{\Pr_{x \sim \mathsf{D}}[x \in A]} + \frac{\Pr_{x \sim \mathsf{D}}[x \in B \setminus A]}{\Pr_{x \sim \mathsf{D}}[x \in B]} \right). \end{aligned}$$

Then, we have:

$$\begin{aligned} \frac{N_{A \setminus B}}{N_A} + \frac{N_{B \setminus A}}{N_B} &\leq \frac{12 \Pr_{x \sim \mathsf{D}}[x \in A \setminus B]}{\Pr_{x \sim \mathsf{D}}[x \in A]} + \frac{12 \Pr_{x \sim \mathsf{D}}[x \in B \setminus A]}{\Pr_{x \sim \mathsf{D}}[x \in B]} + 2\epsilon \\ &\leq 24 d_{\mathrm{tv}}(\mathsf{D}_A, \mathsf{D}_B) + 2\epsilon, \end{aligned}$$

and again, substitute the value of ϵ to conclude the desired result □

We next establish that any two filtered distributions with margin proposals that are smaller than the true margin will be such that the estimated total variation distance between them is bounded.

Theorem 33. *Suppose w_1 and w_2 are two solutions from running our filtering procedure (1) with two different margin values $0 < \gamma_1 \leq \gamma_2$ rather than using the true margin γ. Suppose that $\gamma_2 \leq \gamma$. Define $f_1(x) = \mathbf{1}[\langle w_1, x \rangle \geq 0]$ and $f_2(x) = \mathbf{1}[\langle w_2, x \rangle \geq 0]$. Let $A_1 = \{x : \langle w_1, x \rangle \geq 0\}$ and $A_2 = \{x : \langle w_2, x \rangle \geq 0\}$. Define $\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2})$ as in (16) and $\textsc{threshold}_{\gamma_1}(\delta)$ as in (17). Suppose $0 < p \leq 1/2$, $\varepsilon_{\mathsf{S}} < 1/24$, and the sample sizes satisfy (15) with $\gamma_{\min}$ replaced by γ_1 and δ_0 by δ. Then with probability at least $1 - 26\delta$, both problems are feasible and:*

$$\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) \leq \textsc{threshold}_{\gamma_1}(\delta).$$

Proof. By Theorem 7 and its feasibility event, with probability at least $1 - 16\delta$, both problems are feasible and, since the bound decreases with the proposed margin, both $d_{\mathrm{tv}}(\mathsf{B}^{f_2}, \mathsf{S})$ and $d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{S})$ are bounded as follows:

$$\begin{aligned} d_{\mathrm{tv}}(\mathsf{B}^{f_{\{1,2\}}}, \mathsf{S}) \leq{}& \left(6 + \frac{4 \cdot 81}{1-p}\right) \varepsilon_{\mathsf{S}} + \frac{6}{p} \varepsilon_{\mathsf{O}} + \frac{69 R^2}{\gamma_1^2 N_{\mathsf{S}}} \left[\log\left(\frac{N_{\mathsf{S}} \gamma_1^2}{R^2} \right) \right]_1 \\ &+ \frac{(21 + 270/(1-p)) \log(1/\delta)}{N_{\mathsf{S}}} \\ &+ \frac{666 \cdot R^2}{\gamma_1^2 (1-p) p N_{\mathsf{B}}} \left[\log\left(\frac{(N_{\mathsf{S}} + N_{\mathsf{B}}) \gamma_1^2}{R^2} \right) \right]_1 \left(\frac{6 p N_{\mathsf{B}}}{N_{\mathsf{S}}} + 1 \right) \\ &+ \frac{6 \log(1/\delta)}{(1-p) p N_{\mathsf{B}}} \left(\frac{144 p N_{\mathsf{B}}}{N_{\mathsf{S}}} + \frac{119}{4} \right) \end{aligned} \tag{20}$$

Further, by Lemma 31, there is some constant C such that for $N_{\mathsf{S}} \geq \max[\exp(1), C \cdot \frac{2 \cdot 69 \cdot R^2}{\gamma_1^2}, 64 \log(1/\delta)]$ and $N_{\mathcal{T}} \geq 16 \log(1/\delta)/p$, we have with probability at least $1 - 2\delta$: $\min\{\Pr_{x \sim \mathsf{B}}[x \in A_1], \Pr_{x \sim \mathsf{B}}[x \in A_2]\} \geq p/2$ and also $p/2 \geq 8 \log(1/\delta)/N_{\mathcal{T}}$. In this case, we can apply Theorem 32, which tells us that with probability at least $1 - 8\delta$:

$$\begin{aligned} \widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) &\leq 24 d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) + \frac{48 \log(1/\delta)}{N_{\mathcal{T}} \cdot \min\{\Pr_{x \sim \mathsf{B}}[x \in A_2], \Pr_{x \sim \mathsf{B}}[x \in A_1]\}} \\ &\leq 24 (d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{S}) + d_{\mathrm{tv}}(\mathsf{B}^{f_2}, \mathsf{S})) + \frac{48 \log(1/\delta)}{N_{\mathcal{T}} \cdot \min\{\Pr_{x \sim \mathsf{B}}[x \in A_2], \Pr_{x \sim \mathsf{B}}[x \in A_1]\}} \\ &\leq 24 (d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{S}) + d_{\mathrm{tv}}(\mathsf{B}^{f_2}, \mathsf{S})) + \frac{96 \log(1/\delta)}{p N_{\mathcal{T}}}. \end{aligned}$$

Finally, observe that $\textsc{threshold}_{\gamma_1}(\delta) = 24 \cdot 2 \cdot X + \frac{96 \log(1/\delta)}{p N_{\mathcal{T}}}$ where X is the bound in (20).

We have combined events of probability $1 - 16\delta$, $1 - 8\delta$ and $1 - 2\delta$, so overall the conclusion holds with probability at least $1 - 26\delta$. □

Finally, we present the main result powering our unknown margin procedure. Roughly speaking, it shows that for any two putative margin values $\gamma_1 \leq \gamma_2$, we can construct an estimate of the total variation distance between the filtered distributions corresponding to γ_1 and γ_2 such that the true total variation distance is bounded by the estimate, and also if $\gamma_2 \leq \gamma$, then the estimate is in turn bounded by a threshold that is on the order of the desired final total variation distance bound.

Theorem 34. *Suppose w_1 and w_2 are two solutions from running our filtering procedure (1) with two different margin values $0 < \gamma_1 \leq \gamma_2$ rather than using the true margin γ. Define $f_1(x) = \mathbf{1}[\langle w_1, x\rangle \geq 0]$ and $f_2(x) = \mathbf{1}[\langle w_2, x\rangle \geq 0]$. Let $A_1 = \{x : \langle w_1, x\rangle \geq 0\}$ and $A_2 = \{x : \langle w_2, x\rangle \geq 0\}$. Define $\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2})$ as in (16) and $\textsc{threshold}_{\gamma_1}(\delta)$ as in (17).*

Suppose $0 < p \leq 1/2$, $\varepsilon_{\mathsf{S}} < 1/24$, and the sample sizes satisfy (15) *with $\gamma_{\min}$ replaced by γ_1 and δ_0 by δ. Then with probability at least $1 - 36\delta$, whenever both problems are feasible, the empirical denominators are positive and:*

$$d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) \leq \frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) + \frac{96\log(1/\delta)}{pN_{\mathcal{T}}}.$$

Furthermore, if it holds that $\gamma_2 \leq \gamma$, then both problems are feasible and also:

$$\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) \leq \textsc{threshold}_{\gamma_1}(\delta).$$

Proof. From Lemma 31, there is a constant C such that if $N_{\mathsf{S}} \geq \max[\exp(1), C \cdot \frac{2\cdot 69\cdot R^2}{\gamma_1^2}, 64\log(1/\delta)]$ and $N_{\mathcal{T}} \geq 16\log(1/\delta)/p$, then with probability at least $1-2\delta$, $\min\{\Pr_{x\sim\mathsf{B}}[x \in A_1], \Pr_{x\sim\mathsf{B}}[x \in A_2]\} \geq p/2$ and also $p/2 \geq 8\log(1/\delta)/N_{\mathcal{T}}$. Conditional on the training data, the validation sample is independent of A_1, A_2. Therefore, we may apply Theorem 32 to conclude that we have with probability at least $1 - 8\delta$:

$$\begin{aligned} d_{\mathrm{tv}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) &\leq \frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) + \frac{48\log(1/\delta)}{N_{\mathcal{T}}\min\{\Pr_{x\sim\mathsf{B}}[x \in A_1], \Pr_{x\sim\mathsf{B}}[x \in A_2]\}} \\ &\leq \frac{15}{4}\widehat{d_{\mathrm{tv}}}(\mathsf{B}^{f_1}, \mathsf{B}^{f_2}) + \frac{96\log(1/\delta)}{pN_{\mathcal{T}}}, \end{aligned}$$

which establishes the first part of the Theorem with probability at least $1 - 10\delta$.

The second part follows directly from Theorem 33 with probability at least $1 - 26\delta$. So overall, both parts hold with probability at least $1 - 36\delta$. □

G Hardness of Constrained Classification

Our algorithm from Section 3 is based on solving the following problem:

Definition 2 (ConstrainedClassification)**.** *Given two sets of points $A, B \subset \mathbb{R}^d$, find $w \in \mathbb{R}^d, \theta \in \mathbb{R}$ so as to*

$$\textit{maximize } \sum_{x\in B} \mathbf{1}[\langle w, x\rangle > \theta] \textit{, subject to } \langle w, x\rangle \leq \theta \quad \forall x \in A.$$

In the corresponding decision version, given parameter k, we say that (A, B) is a YES instance iff the objective value above is $\geq |B| - k$.

We show that even with $|A| = 1$ (i.e., one constraint), this problem is at least as hard as MinLin-Classification, which we will now define. In what follows, an affine hyperplane h in $\mathbb{R}^d$ will be defined by (w, θ), where $w \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$. It is also associated with the halfspace:

$$h = \{x : \langle w, x\rangle \geq \theta\}.$$

The complement of h will be denoted $\overline{h} := \{x : \langle w, x\rangle < \theta\}$.

Definition 3 (MinLinClassification)**.** *Given two sets of points $P, N \subset \mathbb{R}^d$, find an affine hyperplane h so as to minimize*

$$|P \cap \overline{h}| + |N \cap h|.$$

In the corresponding decision version of the problem, we are given a parameter K, and (P, N) is a YES instance iff $\exists h$ such that the objective value above is $\leq K$.

In other words, the MINLINCLASSIFICATION problem seeks to find a linear function for which the number of "positive" points (P) assigned a negative value plus the number of "negative" points (N) assigned a non-negative value, is minimized. The problem is known to be NP-hard in high dimensions (Johnson & Preparata, 1978; Guruswami & Raghavendra, 2009). Very recently, Munteanu et al. (2026) showed that, assuming the k-SUM conjecture, the problem cannot be solved in time $O(n^{\frac{d}{2}-\gamma})$ for any $\gamma > 0$, where $n := |P| + |N|$. We show that MINLINCLASSIFICATION reduces to CONSTRAINEDCLASSIFICATION.[4]

Theorem 35. *Even for $|A| = 1$, the decision version of* CONSTRAINEDCLASSIFICATION *is NP-hard. Moreover, assuming the k-SUM conjecture, for any $\gamma > 0$, there is no algorithm for* CONSTRAINED-CLASSIFICATION *with running time $O(n^{\frac{d-1}{2}-\gamma})$.*

Proof. We will give a reduction of linear size that incurs an increase in dimension of 1, and thus both the parts of the Theorem follow from the known hardness results for MINLINCLASSIFICATION.

Consider any instance (N, P, k) of MINLINCLASSIFICATION. Now construct the following instance of CONSTRAINEDCLASSIFICATION. The set A is simply $\mathbf{0}_{d+1}$. The set B is

$$B := \{(x, 1) : x \in P\} \cup \{(-x, -1) : x \in N\}. \tag{21}$$

Completeness. First, suppose (N, P, k) is a YES instance of MINLINCLASSIFICATION. Then by definition, there exists an affine hyperplane $h = (w, \theta)$ (as above) such that

$$|P \cap \overline{h}| + |N \cap h| \leq k.$$

This can also be written as

$$|\{x \in P : \langle w, x\rangle < \theta\}| + |\{x \in N : \langle w, x\rangle \geq \theta\}| \leq k.$$

First, by changing θ to $\theta' = \theta - \eta$ for a sufficiently small $\eta > 0$, we can ensure that for any $x \in N \cup P$ such that $\langle w, x\rangle < \theta$ (i.e., strictly smaller than θ), we also have $\langle w, x\rangle < \theta'$. For such a θ', we will have

$$|\{x \in P : \langle w, x\rangle \leq \theta'\}| = |\{x \in P : \langle w, x\rangle < \theta\}|.$$

This is because any $x \in P$ with $\langle w, x\rangle < \theta$ (strictly) will appear both on the LHS and RHS, whereas x with $\langle w, x\rangle \geq \theta$ will not appear in either. Similarly, we can argue that

$$|\{x \in N : \langle w, x\rangle \geq \theta'\}| = |\{x \in N : \langle w, x\rangle \geq \theta\}|.$$

This is because by the choice of θ', there is no $x \in N$ such that $\langle w, x\rangle \in [\theta', \theta)$. This implies:

$$|\{x \in P : \langle w, x\rangle \leq \theta'\}| + |\{x \in N : \langle w, x\rangle \geq \theta'\}| \leq k. \tag{22}$$

Now consider the $(d + 1)$-dimensional weight vector w' that is the concatenation of w with $(-\theta')$, i.e., $w' = (w, -\theta')$. Clearly, $\langle w', \mathbf{0}_{d+1}\rangle = 0$. Next, consider any point in the set B defined in (21). For any $x \in P$, $\langle w', (x, 1)\rangle = \langle w, x\rangle - \theta'$. Also for any $x \in N$, $\langle w', (-x, -1)\rangle = -(\langle w, x\rangle - \theta')$. Thus, from (22), at most k of the points $y \in B$ have $\langle w', y\rangle \leq 0$. Thus the objective value for this instance of CONSTRAINEDCLASSIFICATION is at least $|B| - k$.

Soundness. Next, suppose the instance of CONSTRAINEDCLASSIFICATION given by $A = \{\mathbf{0}_{d+1}\}$ and B as defined in (21) is a YES instance, i.e., it has an objective value $\geq |B| - k$. Thus, there exists $w' \in \mathbb{R}^{d+1}$ and a $\lambda \in \mathbb{R}$ such that $\langle w', \mathbf{0}_{d+1}\rangle \leq \lambda$ and $\langle w', y\rangle > \lambda$ for at least $|B| - k$ of the points in B. Notice that the first constraint implies that $\lambda \geq 0$.

Let us write $w' = (w, -\theta)$ for $w \in \mathbb{R}^d$. Then for any $y \in B$ of the form $y = (x, 1)$ with $x \in P$,

$$\langle w', y\rangle > \lambda \implies \langle w, x\rangle - \theta > \lambda \geq 0 \implies \langle w, x\rangle > \theta.$$

Likewise, for any $y \in B$ of the form $y = (-x, -1)$ with $x \in N$,

$$\langle w', y\rangle > \lambda \implies -\langle w, x\rangle + \theta > \lambda \geq 0 \implies \langle w, x\rangle < \theta.$$

Since the condition $\langle w', y\rangle > \lambda$ holds for at least $|B| - k$ of the vectors $y \in B$, the classifier $h = (w, \theta)$ correctly classifies at least $|P| + |N| - k$ of the vectors in $P \cup N$. This implies that (P, N) is a YES instance of MINLINCLASSIFICATION.

This completes the proof. □

[4]Gemini and ChatGPT were used to help generate ideas for this proof.

To obtain the homogeneous formulation in (2) with $\gamma = 0$, embed each point x as $(x, 1)$. Given a feasible (w, θ) for CONSTRAINEDCLASSIFICATION, use the vector $(-w, \theta + \eta)$ for sufficiently small $\eta > 0$. All constraint points then have strictly positive score, while every point counted by the objective of CONSTRAINEDCLASSIFICATION has negative score. Conversely, a feasible vector (u, t) for the homogeneous problem gives the feasible CONSTRAINEDCLASSIFICATION solution $(-u, t)$ with the same objective complement. Normalizing the vector does not change its signs, so the constraint $\|w\| \leq 1$ does not affect this reduction.

H VC DIMENSION OF PARTIAL SEPARATORS

One might be tempted to apply standard VC dimension bounds to our problem by considering the hypothesis class H_ε of hyperplanes h for which all but a ε of points in $x_1^{\mathsf{B}}, \ldots, x_{\mathsf{B}}^N$ are at a distance γ away from h, and then applying the VC-dimension based results of Liu et al. (2002). Note that we cannot select from a class that actually separates the points in our training set from B from those in S up to a slack of ε because we do not know apriori which points in the training set come from which distribution. This class H_ε which does not depend on the unknown training labels is a reasonable alternative. The true separating vector v lies in $H_{\varepsilon_{\mathsf{S}}+\varepsilon_{\mathsf{B}}}$.

If $\varepsilon_{\mathsf{S}} = \varepsilon_{\mathsf{B}} = 0$, then the VC dimension of this class is $O(R/\gamma^2)$. If d is the VC dimension of H_ε, then the results of Liu et al. (2002) would recover bounds of the form $O(\varepsilon_{\mathsf{S}} + \varepsilon_{\mathsf{B}} + d/N_{\varepsilon_{\mathsf{S}}})$.

However, when $\varepsilon > 0$ the VC dimension of $\mathcal{H}_\varepsilon$ need not be $O(R^2/\gamma^2)$. Let $k = \varepsilon N$ and $d = k + 1$, and consider the sample $x_i = e_i$ for $i \leq k$ and $x_i = \pm e_d$ for $i > k$, so that all the points are in a ball of radius $R = 1$. For any signs $y \in \{\pm 1\}^k$, let v be the unit vector with $v_i = y_i/\sqrt{2k}$ for $i \leq k$ and $v_d = 1/\sqrt{2}$. Every point $x_i = \pm e_d$ satisfies $|\langle v, x_i \rangle| = 1/\sqrt{2}$, so $v \in \mathcal{H}_\varepsilon$ with $\gamma = 1/\sqrt{2}$. But $\operatorname{sign}\langle v, e_i \rangle = y_i$ for $i \leq k$, so $\mathcal{H}_\varepsilon$ shatters $e_1, \ldots, e_k$. Hence the VC dimension is at least $k = \varepsilon N$, even though $R^2/\gamma^2 = 2$.