Self-training Converts Weak Learners to Strong Learners in Mixture Models

Spencer Frei* UC Berkeley
Difan Zou* UCLA
Zixiang Chen* UCLA
Quanquan Gu UCLA

Abstract

We consider a binary classification problem when the data comes from a mixture of two rotationally symmetric distributions satisfying concentration and anti-concentration properties enjoyed by log-concave distributions among others. We show that there exists a universal constant $C_{\text{err}} > 0$ such that if a pseudolabeler $\boldsymbol{\beta}_{\text{pl}}$ can achieve classification error at most C_{err}, then for any $\varepsilon > 0$, an iterative self-training algorithm initialized at $\boldsymbol{\beta}_0 := \boldsymbol{\beta}_{\text{pl}}$ using pseudolabels $\widehat{y} = \text{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x} \rangle)$ and using at most $\tilde{O}(d/\varepsilon^2)$ unlabeled examples suffices to learn the Bayes-optimal classifier up to ε error, where d is the ambient dimension. That is, self-training converts weak learners to strong learners using only unlabeled examples. We additionally show that by running gradient descent on the logistic loss one can obtain a pseudolabeler $\boldsymbol{\beta}_{\text{pl}}$ with classification error C_{err} using only $O(d)$ labeled examples (i.e., independent of ε). Together our results imply that mixture models can be learned to within ε of the Bayes-optimal accuracy using at most $O(d)$ labeled examples and $\tilde{O}(d/\varepsilon^2)$ unlabeled examples by way of a semi-supervised self-training algorithm.

1 Introduction

Current state-of-the-art methods for computer vision and natural language understanding have relied upon *self-training* methods. These methods are generally unsupervised or semi-supervised learning approaches that take advantage of massive unlabeled datasets to improve performance on benchmark machine learning tasks (Devlin et al., 2019; Chen et al., 2020a). As human-annotated labeled data is expensive to collect, any approach which can reduce the number of labeled examples necessary for good performance is very desirable.

One common approach in semi-supervised and self-supervised learning is the usage of a *pseudolabeler*, which generates labels for unlabeled data $\mathbf{x}$ by using the outputs of a classifier $\mathbf{x} \mapsto \widehat{y} := \text{sgn}(f(\mathbf{x}; \boldsymbol{\beta}))$ where the pseudolabeler $f(\mathbf{x}; \boldsymbol{\beta})$ has weights $\boldsymbol{\beta}$ that have been pre-trained on labeled data (or a combination of labeled and unlabeled data). This approach has been remarkably successful in improving performance on image recognition tasks (Pham et al., 2021; Rizve et al., 2021), although there is very little theoretical understanding for why this method can improve performance or reduce the labeled sample complexity of the learning problem.

In this work, we provide algorithmic guarantees for the error of linear classifiers trained using only unlabeled samples using a standard self-training framework. We assume the learner has access to an initial classifier $\boldsymbol{\beta}_{\text{pl}}$ which could be generated in an arbitrary manner. Given unlabeled examples $\{\mathbf{x}_i\}_{i=1}^n$, initial classifier $\boldsymbol{\beta}_0 := \boldsymbol{\beta}_{\text{pl}}$, at each time t we generate pseudolabels $\widehat{y}_i = \text{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i \rangle)$ that are then used in a standard gradient-based optimization of a weight-normalized loss of the form $\ell(\widehat{y}_i \langle \boldsymbol{\beta}_t, \mathbf{x}_i \rangle / \|\boldsymbol{\beta}_t\|)$. We assume the data is generated by a mixture model with two modes in the sense that, for labels $y \in \{\pm 1\}$ and mean parameter $\boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbf{x}|y$ is a random variable with mean $y\boldsymbol{\mu}$ and the distribution of $\mathbf{z} := \mathbf{x} - y\boldsymbol{\mu}$ is unimodal, spherically symmetric, and satisfies some mild concentration and anti-concentration properties.

Our main contributions are as follows.

(1) Provided the classification error of the initial pseudolabeler is smaller than some absolute constant C_{err}, self-training with $\tilde{O}(d/\varepsilon^2)$ unlabeled examples produces a classifier that has classification error at most ε larger than the Bayes-optimal error.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

(2) If the mixture model is sufficiently separated (i.e., $\|\boldsymbol{\mu}\| \geq C_{\boldsymbol{\mu}}$ for some absolute constant $C_{\boldsymbol{\mu}}$), then in the *supervised* learning setting, gradient descent on the logistic loss finds a classifier with classification error at most C_{err} using only $O(d)$ labeled examples—i.e., independent of ε.

(3) Putting (1) and (2) together implies that in the semi-supervised setting, mixture models can be learned to within ε of the Bayes-optimal accuracy using $O(d)$ labeled examples and $\tilde{O}(d/\varepsilon^2)$ unlabeled examples by using self-training with weight normalization.

Organization of the paper. We first discuss related work in Section 2. We provide our main results on self-training with unlabeled examples in Section 3. In Section 4, we describe our results in the supervised setting and combine this with our results from Section 3 to get guarantees in the semi-supervised setting. In Section 5, we provide a proof sketch for our results on self-training. We conclude in Section 6, and leave detailed proofs for the appendices.

Notation. We note here the notational conventions adopted in the paper. We use bold letters to denote vectors. We use $\|\mathbf{x}\|$ to denote the ℓ^2 Euclidean norm of a vector $\mathbf{x}$. We say that $f(x) = O(g(x))$ if there exist universal constants C, C' such that $f(x) \leq Cg(x)$ for $x \geq C'$; $f(x) = \Omega(g(x))$ if there exist C, C' such that $f(x) \geq Cg(x)$ for $x \geq C'$; and $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$. We use $\tilde{O}, \tilde{\Omega}$, and $\tilde{\Theta}$ to additionally ignore logarithmic factors. For a vector $\mathbf{v}$, we denote by $\text{err}(\mathbf{v}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y \neq \text{sgn}(\langle \mathbf{v}, \mathbf{x}\rangle)$, where the distribution $\mathcal{D}$ will be understood from the context in which this term appears. We use $\mathbb{1}(A)$ to denote the indicator function of an event A, i.e. equal to one when the event A occurs and zero otherwise. The function $\text{sgn}(z) = \mathbb{1}(z > 0) - \mathbb{1}(z < 0)$ is the sign function, equal to the sign of a real number with $\text{sgn}(0) = 0$. We use the notation $a \wedge b$ to denote the minimum of a and b, and the notation $a \vee b$ to denote the maximum of a and b. For a linear classifier $\mathbf{x} \mapsto \text{sgn}(\langle \mathbf{x}, \boldsymbol{\beta}\rangle)$ with parameter $\boldsymbol{\beta}$, we will interchangeably refer to $\boldsymbol{\beta}$ as the classifier or as the parameter. We will likewise interchangeably refer to the parameters $\boldsymbol{\beta}_{\text{pl}}$ defining a pseudolabeler, a pseudolabeler $\mathbf{x} \mapsto \langle \mathbf{x}, \boldsymbol{\beta}_{\text{pl}}\rangle$, and the classifier induced by a pseudolabeler $\mathbf{x} \mapsto \text{sgn}(\langle \mathbf{x}, \boldsymbol{\beta}_{\text{pl}}\rangle)$ itself, with the particular sense being clear in context.

2 Related Work

Although the usage of the term 'pseudolabel' dates back to as recently as 2013 (Lee, 2013), the usage of self-supervised (unsupervised) methods to improve performance in supervised learning tasks has a long history in machine learning (Scudder, 1965; Yarowsky, 1995). It is only over the past few years that self-supervised 'pre-training' methods have become a standard approach for improving performance in supervised learning tasks like image recognition and natural language understanding (Devlin et al., 2019; Chen et al., 2020a; Pham et al., 2021). Such methods are particularly appealing in the age of big data where in an increasing number of domains it is possible to collect massive unlabeled datasets.

From a theoretical perspective, much less is known about self-supervised and semi-supervised learning than in the supervised setting. Early works by Castelli and Cover (1995, 1996) looked at the relative value of labeled examples over unlabeled examples when the underlying marginal distribution of the features satisfies a parametric identifiability assumption. A series of works have sought to clarify under what conditions semi-supervised learning can have provably better sample complexity or generalization performance in comparison with using solely supervised learning techniques (Ben-David et al., 2008; Singh et al., 2009; Balcan and Blum, 2010; Darnstädt et al., 2013; Göpfert et al., 2019). For surveys on early work in semi-supervised learning, we refer the reader to Zhu and Goldberg (2009) and Chapelle et al. (2010).

More related to this work, a number of theorists have sought to better understand the mechanisms underlying the types of self-training algorithms used for deep neural networks. This includes an analysis of contrastive learning (Tosh et al., 2021), consistency regularization (Wei et al., 2021; Cai et al., 2021), robust self-training (Raghunathan et al., 2020), knowledge distillation (Hsu et al., 2021) and masked feature prediction (Lee et al., 2020), to mention a few. A number of works on the theory of self-training methods have focused on their applications in transfer learning and domain adaptation (Kumar et al., 2020; Chen et al., 2020b; Xie et al., 2021).

A closely related paper is by Oymak and Gulcu (2021). They considered the Gaussian mixture model setting and considered a self-training algorithm based on updating the estimate $\frac{1}{n}\sum_{i=1}^{n}[y_i\mathbf{x}_i] \approx \boldsymbol{\mu}$ for the mean of the mixture. By replacing the labels y_i with pseudolabels produced by some initial pseudolabeler, they are able to show that in the high dimensional limit, the predictors found by self-training are correlated with the Bayes-optimal predictor $\boldsymbol{\mu}$. In contrast to our results, they did not provide a guarantee that their self-training algorithm converged to the Bayes-optimal predictor. Additionally, the averaging operator they consider does not have analogues used in deep learning, which stands in contrast to the gradient-based training of the logistic loss we consider in this paper.

Kumar et al. (2020), in a broad work on the usage of self-training methods for domain adaptation, worked on a similar problem to the one we consider in this paper. They showed that in a Gaussian mixture model setting, assuming (1) iterative self-training solves an appropriate constrained nonconvex optimization problem and (2) access to infinite unlabeled data, then iterative self-training can yield the Bayes-optimal classifier provided it is initialized with a pseudolabeler with sufficiently small error. By contrast, we directly show that the nonconvex optimization algorithm consisting of self-training with a finite set of unlabeled samples via weight-normalized gradient descent yields Bayes-optimal classifiers for a more general class of distributions.

Chen et al. (2020b) showed that for a mixture model where some coordinates are 'spurious' and are distributed according to a (possibly anisotropic) Gaussian while the remaining coordinates satisfy mild distributional assumptions and fully determine the 'signal' of the label y, self-training via *projected* gradient descent learns to avoid the spurious features, provided the initial pseudolabeler does not depend much on the spurious features. Under the additional assumption that all of the coordinates are Gaussian with only one coordinate determining the label, they are able to show self-training converges to an optimal classifier. In comparison to our work, we have a more complete characterization of the sample complexity of the semi-supervised learning problem in that we show that a constant number of labeled examples is sufficient for learning pseudolabelers for which self-training learns optimal classifiers; we show convergence to the optimal classifier for more general distributions; and we consider the dynamics of self-training with *weight-normalized* gradient descent, which are different from that of projected gradient descent.

3 Self-training Converts Weak Learners to Strong Learners

In this section we show one of our key results, namely for data coming from an isotropic mixture model, there exists a universal constant $C_{\text{err}} > 0$ such that if an initial pseudolabeler $\boldsymbol{\beta}_{\text{pl}}$ has classification error at most C_{err}, then self-training using only unlabeled examples yields a classifier with classification error arbitrarily close to the Bayes-optimal error. Before we begin, let us introduce some definitions which we will need to define the mixture model we consider. Our first set of definitions are that of sub-exponential distributions and that of anti-concentration.

Definition 3.1 (Sub-exponential distributions)**.** We say $\mathcal{D}_{\mathbf{x}}$ is K-*sub-exponential* if every $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$ is a sub-exponential random vector with sub-exponential norm at most K. In particular, for any $\bar{\mathbf{v}}$ with $\|\bar{\mathbf{v}}\| = 1$, $\mathbb{P}_{\mathcal{D}_x}(|\langle \bar{\mathbf{v}}, \mathbf{x}\rangle| \geq t) \leq \exp(-t/K)$.

Definition 3.2. For $\bar{\mathbf{v}}, \bar{\mathbf{v}}' \in \mathbb{R}^d$, denote by $p_{\bar{\mathbf{v}}}(\cdot)$ the density function of the projection of $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$ on the one dimensional subspace spanned by $\bar{\mathbf{v}}$, and by $p_{\bar{\mathbf{v}},\bar{\mathbf{v}}'}(\cdot)$ the density function of the projection of $\mathbf{x}$ on the subspace spanned by $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}'$. We say the distribution satisfies U-*anti-concentration* if there exists $U > 0$ such that for any unit norm $\bar{\mathbf{v}} \in \mathbb{R}^d$, $p_{\bar{\mathbf{v}}}(t) \leq U$ for all $t \in \mathbb{R}$. We say (U', R)-*anti-anti-concentration* holds if there exists $U', R > 0$ such that for any unit norm $\bar{\mathbf{v}}, \bar{\mathbf{v}}' \in \mathbb{R}^d$, it holds that $p_{\bar{\mathbf{v}},\bar{\mathbf{v}}'}(\mathbf{a}) \geq 1/U'$ for all $\mathbf{a} \in \mathbb{R}^2$ satisfying $\|\mathbf{a}\|_2 \leq R$.

The sub-exponential definition is standard and satisfied by log-concave isotropic distributions among others. Anti-concentration and anti-anti-concentration are fairly benign distributional assumptions, the former stating that the distribution cannot assign unbounded probability mass along one dimensional projections and the latter stating that the projection of the features onto low dimensional subspaces have probability density functions which assign at least a constant amount of mass near the origin. A number of recent works have developed guarantees for learning halfspaces with noise under these distributional assumptions to avoid computational complexity lower bounds that exist without such assumptions (Diakonikolas et al., 2019, 2020b; Frei et al., 2021a,b; Zou et al., 2021).

We can now define the mixture distribution we consider in this work.

Definition 3.3. A joint distribution $\mathcal{D}$ over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ is defined as follows. Let $\boldsymbol{\mu} \in \mathbb{R}^d$, and $y = 1$ with probability $1/2$ and $y = -1$ with probability $1/2$. Then we generate $\mathbf{x}|y \sim \mathbf{z} + y\boldsymbol{\mu}$ where $\mathbf{z}$ is an isotropic, rotationally symmetric[1] and K-sub-exponential distribution satisfying U-anti-concentration and (U', R)-anti-anti-concentration. Further assume $\mathbf{z}$ is unimodal in the sense that its probability density function $p_{\mathbf{z}}(z)$ is decreasing function of $\|z\|_2$. We call $(\mathbf{x}, y) \sim \mathcal{D}$ a *mixture distribution with mean* $\boldsymbol{\mu}$ *and parameters* K, U, U', R.

We note that log-concave isotropic distributions like the standard Gaussian are K-sub-exponential and satisfy U-anti-concentration and (U', R)-anti-anti-concentration with $K, U, U', R = \Theta(1)$ (see Diakonikolas et al. (2020a, Fact 19)). Thus, our generative model is a natural generalization of the Gaussian mixture model that can accommodate a broader class of distributions. We further note that the Bayes-optimal

[1] By isotropic we mean $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = I$, and by rotationally symmetric we mean $\mathbf{x}$ has the same distribution as $Q\mathbf{x}$ for any orthogonal matrix Q.

classifier for the mixture models we consider in this work is given by the linear classifier $\mathbf{x} \mapsto \operatorname{sgn}(\langle \boldsymbol{\mu}, \mathbf{x} \rangle)$.

Fact 3.4. For mixture models satisfying Definition 3.3, the Bayes-optimal classifier is given by $\mathbf{x} \mapsto \operatorname{sgn}(\langle \boldsymbol{\mu}, \mathbf{x} \rangle)$.

A proof for Fact 3.4 is given in Appendix C. With the above in place, we can begin to describe the self-training algorithm we will use to amplify weak learners to strong learners using only unlabeled data. We assume we have access to a pseudolabeler $\boldsymbol{\beta}_{\mathrm{pl}}$ which is able to achieve a sufficiently small, but constant, population-level classification error. We then use a weight-normalized logistic regression method to train starting from the initial predictor $\boldsymbol{\beta}_{\mathrm{pl}}$ using only unlabeled examples. Our results will rely upon loss functions that are well-behaved in the following sense.

Definition 3.5. We say a loss function ℓ is *well-behaved* for some $C_\ell \geq 1$ if the loss $\ell(z)$ is 1-Lipschitz and decreasing on the interval $[0, \infty)$, and additionally $-\ell'(z) \geq \frac{1}{C_\ell} \exp(-z)$ for $z > 0$.

The exponential loss $\ell(z) = \exp(-z)$ and the logistic loss $\ell(z) = \log(1 + \exp(-z))$ are well-behaved with $C_\ell = 1$ and $C_\ell = 2$ respectively. Note that our analysis will *not* require that the loss used is convex, merely that it is decreasing, Lipschitz, and that $-\ell'$ is bounded from below by a constant times the exponential loss. Additionally note that we only specify the behavior of the loss on the interval $[0, \infty)$. As we will see, this is because in the self-training algorithm we consider, the input to the loss function is always non-negative.

We can now formally describe the self-training algorithm. Let $\sigma > 0$ be a parameter which we shall call the *temperature*. We assume we have access to $n = TB$ samples $\{\mathbf{x}_i^{(t)}\}_{i=1,\dots,B,\ t=0,\dots,T-1}$, which we partition into T batches of size B. With a well-behaved loss ℓ, we define the (unsupervised) empirical risk

$$\widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}) := \frac{1}{B} \sum_{i=1}^{B} \ell\left(\frac{1}{\sigma} \cdot \operatorname{sgn}\left(\left\langle \mathbf{x}_i^{(t)}, \boldsymbol{\beta} \right\rangle \right) \cdot \left\langle \mathbf{x}_i^{(t)}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle \right)$$
$$= \frac{1}{B} \sum_{i=1}^{B} \ell\left(\frac{1}{\sigma} \left| \left\langle \mathbf{x}_i^{(t)}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle \right| \right). \tag{3.1}$$

That is, we use a typical weight-normalized logistic regression-type loss with pseudolabels given by $\widehat{y} = \operatorname{sgn}(\langle \boldsymbol{\beta}, \mathbf{x} \rangle)$, with an additional factor given by the temperature σ. We start with the predictor $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{\mathrm{pl}} / \left\| \boldsymbol{\beta}_{\mathrm{pl}} \right\|$ and then use updates

$$\tilde{\boldsymbol{\beta}}_{t+1} = \boldsymbol{\beta}_t - \eta \nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t),$$
$$\boldsymbol{\beta}_{t+1} = \tilde{\boldsymbol{\beta}}_{t+1} / \left\| \tilde{\boldsymbol{\beta}}_{t+1} \right\|.$$

Notice the usage of weight normalization in the definition of the unsupervised loss. This can be viewed as a form of regularization for the learning algorithm, since if we do not normalize the weights it is possible that $\widehat{L^{\mathrm{u}}}(\boldsymbol{\beta})$ could be minimized by simply taking $\|\boldsymbol{\beta}\| \to \infty$. The usage of a temperature term is common in self-training algorithms (Hinton et al., 2015; Zou et al., 2019), and has also previously been used for learning halfspaces with noise (Diakonikolas et al., 2020b; Zou et al., 2021). We summarize the above into Algorithm 1.

Algorithm 1 Self-training with pseudolabels and weight normalization

input: Training dataset $S = \{\mathbf{x}_i^{(t)}\}_{i=1,\dots,B,\ t=0,\dots,T-1}$,
step size η, temperature $\sigma > 0$, pseudolabeler $\boldsymbol{\beta}_{\mathrm{pl}}$
$\boldsymbol{\beta}_0 := \boldsymbol{\beta}_{\mathrm{pl}} / \left\| \boldsymbol{\beta}_{\mathrm{pl}} \right\|$
for $t = 0, \dots, T-1$ **do**
  Generate pseudolabels $\widehat{y}_i^{(t)} = \operatorname{sgn}(\langle \mathbf{x}_i^{(t)}, \boldsymbol{\beta}_t \rangle)$
  $\tilde{\boldsymbol{\beta}}_{t+1} = \boldsymbol{\beta}_t - \frac{\eta}{B} \sum_{i=1}^{B} \nabla \ell(\frac{1}{\sigma} \cdot \widehat{y}_i^{(t)} \cdot \langle \mathbf{x}_i^{(t)}, \boldsymbol{\beta}_t / \|\boldsymbol{\beta}_t\| \rangle)$
  $\boldsymbol{\beta}_{t+1} = \tilde{\boldsymbol{\beta}}_{t+1} / \left\| \tilde{\boldsymbol{\beta}}_{t+1} \right\|$
end for
output: $\boldsymbol{\beta}_{T-1}$

Our main result is the following theorem. We will present its proof in Section 5.

Theorem 3.6. Suppose that $(\mathbf{x}, y) \sim \mathcal{D}$ follows a mixture distribution with mean $\boldsymbol{\mu}$ satisfying $\|\boldsymbol{\mu}\| = \Theta(1)$ and parameters $K, U, U', R = \Theta(1)$. Let ℓ be well-behaved for some $C_\ell \geq 1$, and assume the temperature satisfies $\sigma \geq R \vee \|\boldsymbol{\mu}\|$. Assume access to a pseudolabeler $\boldsymbol{\beta}_{\mathrm{pl}}$ which satisfies $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \operatorname{sgn}(\langle \boldsymbol{\beta}_{\mathrm{pl}}, \mathbf{x} \rangle)\big) \leq C_{\mathrm{err}}$, where $C_{\mathrm{err}} = R^2/(72 C_\ell U')$. Let $\varepsilon, \delta \in (0, 1)$, and assume that

$$B = \tilde{\Omega}\left(\varepsilon^{-1}\right), \quad T = \tilde{\Omega}\left(d\varepsilon^{-1}\right), \quad \eta = \tilde{\Theta}\left(d^{-1}\varepsilon\right).$$

Then with probability at least $1 - \delta$, by running Algorithm 1 with step size η and batch size B, the last iterate satisfies $\operatorname{err}(\boldsymbol{\beta}_{T-1}) \leq \operatorname{err}(\boldsymbol{\mu}) + \varepsilon$. In particular, $T = \tilde{O}(d/\varepsilon)$ iterations using at most $TB = \tilde{O}(d/\varepsilon^2)$ unlabeled samples suffices to be within ε error of the Bayes-optimal classifier.

Theorem 3.6 shows that provided we have a pseudolabeler which achieves a constant level of classification error, then by using only unlabeled examples, self-training with pseudolabels and weight normalization will amplify the pseudolabeler from a weak learner (achieving a constant level of accuracy) to a strong learner (achieving accuracy arbitrarily close to that of the best possible). Note that for the mixture model, $\operatorname{sgn}(\langle \boldsymbol{\mu}, \cdot \rangle)$ is the Bayes-optimal classifier over the distribution (see Fact 3.4), and if $\|\boldsymbol{\mu}\|$ is small then the best error achievable might be quite large, as the region

Algorithm 2 Logistic regression with online stochastic gradient descent

input: Failure probability $\delta \in (0,1)$, Training dataset $S = \{(\mathbf{x}_t^{(i)}, y_t^{(i)})\}$ for $t = 0,\dots,T-1$, $i = 1,\dots,4\lceil\log(1/\delta)\rceil$, step size η.
$\boldsymbol{\beta}_0^{(i)} := 0$.
for $i = 1,\dots,4\lceil\log(1/\delta)\rceil$ **do**
  for $t = 0,\dots,T-1$ **do**
    $\boldsymbol{\beta}_{t+1}^{(i)} = \boldsymbol{\beta}_t^{(i)} - \eta\nabla\log(1+\exp(-y_t^{(i)}\cdot\langle\mathbf{x}_t^{(i)},\boldsymbol{\beta}_t^{(i)}\rangle))$
  end for
end for
output: $\{\boldsymbol{\beta}_t^{(i)}\}_{t\in[T],\ i\in[\log(1/\delta)]}$

near the origin could have a large mass of samples that are just as likely to be from the $y=+1$ cluster and the $y=-1$ cluster (consider a mixture of two isotropic 2D Gaussians with means $(+1,0)$ and $(-1,0)$). Thus in some settings it may not be possible for a pseudolabeler to have error smaller than C_{err}. However, we will see in the next section that provided $\|\boldsymbol{\mu}\|$ is bounded below by a universal constant, we can ensure that a classifier trained by gradient descent using only $O(d)$ labeled examples has classification error at most C_{err}.

4 Semi-supervised Learning with $O(d)$ Labeled Examples via Self-training

Theorem 3.6 tells us that provided the self-training procedure (Algorithm 1) starts with a pseudolabeler that has classification error smaller than some absolute constant C_{err}, self-training will boost this weak learner to a strong learner quickly. In this section, we show that a standard logistic regression procedure produces a pseudolabeler that can achieve the desired constant accuracy by using only $O(d)$ samples—that is, a constant number of samples with respect to ε. The particular algorithm we consider is online SGD used to minimize the logistic loss $\ell(z) = \log(1+\exp(-z))$ defined over a linear classifier, and is given in Algorithm 2. We use $O(\log(1/\delta))$ independent runs of online SGD to amplify a constant probability guarantee to a high probability guarantee.

Theorem 4.1. Suppose that $(\mathbf{x},y) \sim \mathcal{D}$ follows a mixture distribution with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R > 0$. Let C_{err} be the constant from Theorem 3.6 and assume $\|\boldsymbol{\mu}\| \geq 3K\max(\log(8/C_{\text{err}}), 22K)$. By running Algorithm 2 with $\eta = (\|\boldsymbol{\mu}\|^2 + d)^{-1}C_{\text{err}}/8$ and $T = 8\eta^{-1}C_{\text{err}}^{-1}\|\boldsymbol{\mu}\|^2$ iterations, there exists $i \leq 4\log(1/\delta)$ and $t < T$ such that with probability at least $1-\delta$,

$$\mathbb{P}(y \neq \operatorname{sgn}(\langle\boldsymbol{\beta}_t^{(i)}, \mathbf{x}\rangle)) \leq C_{\text{err}}.$$

The proof of Theorem 4.1 follows standard stochastic convex optimization arguments and can be found in Appendix B.

Theorem 4.1 implies that if we have access to $O\big((\|\boldsymbol{\mu}\|^2 + d)\|\boldsymbol{\mu}\|^2\big)$ labeled examples, where $O(\cdot)$ hides universal constants depending on K, U, U', and R, we can learn a pseudolabeler $\boldsymbol{\beta}_{\text{pl}}$ with classification error at most C_{err}. In particular, for $\|\boldsymbol{\mu}\| = \Theta(1)$, using only $O(d)$ labeled examples suffices to learn a pseudolabler with error at most C_{err}. We can then use this pseudolabeler in Theorem 3.6 with $O(d/\varepsilon^2)$ *unlabeled* examples to perform self-training and yield a classifier which achieves classification error at most ε larger than the best-possible error. We collect these results into the following corollary.

Corollary 4.2. Let $(\mathbf{x},y) \sim \mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R = \Theta(1)$. Assume $\|\boldsymbol{\mu}\| = \Theta(1)$ satisfies

$$\|\boldsymbol{\mu}\| \geq 3K\max\left(\log(144U'/R^2), 22K\right),$$

Then for any $\varepsilon, \delta \in (0,1)$, with probability at least $1-\delta$, using $O(d)$ labeled examples in Algorithm 2 and $\tilde{O}(d/\varepsilon^2)$ unlabeled examples in Algorithm 1 suffices to learn a predictor $\boldsymbol{\beta}$ to within ε error of the Bayes-optimal classification error, where $O(\cdot)$ hides constants depending on K, U, U', R, and $\log(1/\delta)$ only, and $\tilde{O}$ additionally suppresses logarithmic dependence on ε^{-1} and d.

To the best of our knowledge, Corollary 4.2 is the first result to show that a semi-supervised self-training algorithm can learn an optimal classifier using only a constant number of labeled examples.

On a related note, we want to acknowledge that for Gaussian mixture models, there exist purely unsupervised techniques (based on clustering methods) for which $\tilde{O}(d/\varepsilon)$ unlabeled examples suffices to learn within ε of the *clustering error* $\min\big(\mathbb{P}(y \neq \operatorname{sgn}(\langle\boldsymbol{\mu},\mathbf{x}\rangle)), 1-\mathbb{P}(y \neq \operatorname{sgn}(\langle\boldsymbol{\mu},\mathbf{x}\rangle))\big)$ (Li et al., 2017). Thus, under more restrictive distributional assumptions and using algorithms designed for mixture models, it is possible to optimally learn a mixture model using only unlabeled examples. We note this to emphasize that in this work we do not make the claim that self-training with pseudolabels is the optimal algorithm for learning mixture models. Rather, our aim is to develop a better understanding of how self-training with pseudolabels can achieve good performance using few labeled examples.

5 Proof of Main Results

In this section we provide a proof for Theorem 3.6. The key to our proof comes from deriving a lower bound

that takes the form

$$\langle \bar{\boldsymbol{\mu}}, -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle \geq C_0 \sin^2(\theta_t), \tag{5.1}$$

where $\theta_t \in [0, \pi/2]$ is the angle between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$ and C_0 is some absolute constant. To see the importance of such an inequality, let us look at the increments between the weights found using Algorithm 1 and those of the (normalized) ideal predictor $\bar{\boldsymbol{\mu}} := \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Denote $\Delta_t^2 = \|\boldsymbol{\beta}_t - \bar{\boldsymbol{\mu}}\|^2$. Let $\tilde{\Delta}_t^2 = \|\tilde{\boldsymbol{\beta}}_t - \bar{\boldsymbol{\mu}}\|_2^2$. Then,

$$\begin{aligned}
\Delta_t^2 - \Delta_{t+1}^2 &\overset{(i)}{\geq} \Delta_t^2 - \tilde{\Delta}_{t+1}^2 \\
&= 2\eta\langle \nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t), \boldsymbol{\beta}_t - \bar{\boldsymbol{\mu}}\rangle - \eta^2 \left\|\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\right\|^2 \\
&\overset{(ii)}{=} 2\eta\langle -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t), \bar{\boldsymbol{\mu}}\rangle - \eta^2 \left\|\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\right\|^2 \\
&\overset{(iii)}{\geq} 2\eta\langle -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t), \bar{\boldsymbol{\mu}}\rangle - \varepsilon.
\end{aligned} \tag{5.2}$$

Inequalities (i) and (ii) follow from the fact that $\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)$ is orthogonal to $\boldsymbol{\beta}_t$, as can be seen by the identity

$$\begin{aligned}
\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t) = \frac{1}{\sigma B \|\boldsymbol{\beta}_t\|_2} \sum_{i=1}^{B} \ell'\left(\frac{1}{\sigma}\frac{|\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle|}{\|\boldsymbol{\beta}_t\|_2}\right) \\
\cdot \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle) \cdot \left(I - \frac{\boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top}{\|\boldsymbol{\beta}_t\|_2^2}\right)\mathbf{x}_i^{(t)}.
\end{aligned} \tag{5.3}$$

In particular, (i) follows from the identity $\|\tilde{\boldsymbol{\beta}}_{t+1}\|^2 = \|\boldsymbol{\beta}_t\|^2 + \eta^2 \|\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\|^2 > 1$. Inequality (iii) comes from taking η sufficiently small. Thus, if we have a lower bound like (5.1), then (5.2) shows that whenever the angle θ_t between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$ is large, the distance between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$ will decrease. Perhaps surprisingly, Lemma 5.1 below shows that one can guarantee this condition holds *provided the predictor $\boldsymbol{\beta}_t$ has classification error smaller than some absolute constant C_{err}.*

Lemma 5.1. Let $\mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R > 0$. Let ℓ be well-behaved for some $C_\ell \geq 1$, and assume the temperature satisfies $\sigma \geq R \vee \|\boldsymbol{\mu}\|$. Suppose that $\|\boldsymbol{\beta}_t\| = 1$ is an initial estimate. Denote θ_t as the angle between $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}$, and assume that $\theta_t \in [0, \pi/2]$. Assume the classification error of $\boldsymbol{\beta}_t$ satisfies

$$\mathrm{err}_t := \mathbb{P}\big(y \neq \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle)\big) \leq \frac{R^2}{72 C_\ell U'} =: C_{\mathrm{err}}.$$

Then we have

$$\langle \boldsymbol{\mu}, -\mathbb{E}[\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle \geq \frac{R^2 \|\boldsymbol{\mu}\|^2}{36 \sigma C_\ell U'} \cdot \sin^2(\theta_t).$$

Moreover, there exists a universal constant $C_B > 0$ such that for any $\varepsilon, \delta \in (0,1)$,

$$B \geq C_B \left(\frac{K C_\ell U'}{R^2}\right)^2 \varepsilon^{-1} \log(2/\delta),$$

then with probability at least $1 - \delta$,

$$\langle \boldsymbol{\mu}, -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle \geq \frac{R^2 \|\boldsymbol{\mu}\|^2}{72 \sigma C_\ell U'} \sin^2 \theta_t - \varepsilon/2.$$

The proof of Lemma 5.1 is a somewhat involved calculation, and the complete details are left for Appendix A. Below, we sketch some of the high-level ideas.

Lemma 5.1 Proof Sketch. Since $\|\boldsymbol{\beta}_t\| = 1$, using the gradient formula (5.3),

$$\begin{aligned}
&-\mathbb{E}[\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|\big) \cdot \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \big(I - \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top\big)\mathbf{x}\big].
\end{aligned}$$

Denote

$$\bar{\boldsymbol{\mu}} := \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \quad \tilde{\boldsymbol{\mu}}_t := (I - \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top)\bar{\boldsymbol{\mu}}.$$

We can then write

$$\begin{aligned}
&\langle \boldsymbol{\mu}, -\mathbb{E}\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle \\
&= \|\boldsymbol{\mu}\|\, \mathbb{E}\big[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|\big) \cdot \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}\big].
\end{aligned}$$

Define the event S_t where a sample $(\mathbf{x}, y)$ is correctly classified by $\boldsymbol{\beta}_t$,

$$S_t := \{y = \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle)\}.$$

Then let S_t^c be the complement of the event S_t, we can calculate

$$\begin{aligned}
&\mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|/\sigma\big) \cdot \operatorname{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}] \\
&= \mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t)] \\
&\quad + \mathbb{E}[\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t^c)] \\
&= \mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x})] \\
&\quad + 2\mathbb{E}[\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t^c)].
\end{aligned} \tag{5.4}$$

The proof relies upon deriving a lower bound on the first quantity and an upper bound on the absolute value of the second quantity. We proceed with the lower bound for the first term as follows. Since the quantity depends only on the projection of $\mathbf{z}$ onto the space spanned by $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$, we work in this two dimensional space. Since $\mathbf{z}$ is rotation invariant, we can rotate the coordinate system so that $\boldsymbol{\beta}_t = \mathbf{e}_2$, $\bar{\boldsymbol{\mu}} = (\sin\theta_t, \cos\theta_t)$, and $\tilde{\boldsymbol{\mu}} = (\sin\theta_t, 0)$, where θ_t is the angle between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$. Denote $p_{\boldsymbol{\beta}_t, \boldsymbol{\mu}}(\cdot, \cdot) : \mathbb{R}^2 \to [0, \infty)$ as the probability density function of the projection of $\mathbf{z}$ onto the 2D subspace spanned by $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}$. Then, using that $y\mathbf{x}$

has the same distribution as $\mathbf{z}+\boldsymbol{\mu}$,

$$\begin{aligned}
&\mathbb{E}[-\ell'\big(|\langle\boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} -\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\theta_t\cdot\ldots \\
&\quad\cdot(u_1+\|\boldsymbol{\mu}\|\sin\theta_t)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{-\infty}^{\infty} -\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\theta_t\cdot\ldots \\
&\quad\cdot\left[\int_{-\infty}^{\infty}(u_1+\|\boldsymbol{\mu}\|\sin\theta_t)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\right]\mathrm{d}u_2 \\
&\overset{(i)}{=} \|\boldsymbol{\mu}\|\sin^2\theta_t\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} -\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\ldots \\
&\quad\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}$$

In (i) we use that $\mathbf{z}$ is isotropic and so projections of it onto one dimensional subspaces are mean zero. From here, we use the well-behaved property of ℓ to bound $-\ell'$ from below by an exponential-type loss, and anti-anti-concentration to bound $p_{\beta_t,\bar{\mu}}(u_1,u_2)$ from below. After some lines of calculus, we get (see Appendix A for details)

$$\begin{aligned}
&\mathbb{E}[-\ell'\big(|\langle\boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})] \\
&\geq \frac{\|\mu\|\,R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'}.
\end{aligned} \tag{5.5}$$

Using a similar line of argument, we can show an upper bound for the second term of (5.4),

$$\begin{aligned}
&\left|\mathbb{E}\left[\ell'(|\langle\boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma)\tilde{\boldsymbol{\mu}}^\top y\mathbf{x}\cdot\mathbb{1}(S_t^c)\right]\right| \\
&\leq \|\boldsymbol{\mu}\|\sin^2\theta_t\cdot\mathbb{P}(S_t^c) = \|\boldsymbol{\mu}\|\sin^2\theta_t\cdot\mathrm{err}_t.
\end{aligned} \tag{5.6}$$

Substituting (5.5) and (5.6) into (5.4), we get

$$\begin{aligned}
&\langle\boldsymbol{\mu}, -\mathbb{E}[\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle \\
&\geq \frac{\|\mu\|^2\sin^2\theta_t}{\sigma}\left[\frac{R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'} - 2\mathrm{err}_t\right] \\
&\geq \frac{\|\mu\|^2\sin^2\theta_t}{\sigma}\left[\frac{R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma})}{6C_\ell U'} - 2\mathrm{err}_t\right].
\end{aligned}$$

Thus, we see that by choosing $\sigma\geq\|\boldsymbol{\mu}\|$, provided err_t is smaller than an absolute constant, $\langle\boldsymbol{\mu}, -\mathbb{E}[\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle$ is bounded from below by a constant multiple of $\sin^2\theta_t$, as claimed. To translate the result from the population-level estimate to that for batches of samples, we use concentration. For details, see Appendix A. □

As we described at the beginning of this section, Lemma 5.1 is the key to showing that self-training is able to learn the Bayes-optimal classifier over the distribution using only unlabeled examples. In order to apply it for the proof of Theorem 3.6, we need to confirm that the classification error of the learned classifier satisfies $\mathrm{err}_t\leq C_{\mathrm{err}}$ almost surely for each iterate of the algorithm. We also need to ensure that we have generalization guarantees for the last iterate of the algorithm, as the standard cross-validation trick used in supervised learning is not desirable in the semi-supervised setting due to the large labeled sample complexity of such an approach. These constitute the main technical hurdles in the following proof.

Proof of Theorem 3.6. For notational simplicity, in the remainder of the proof let us denote

$$C_g := \frac{72\sigma C_\ell U'}{R^2\|\mu\|},\ C_d := 2\|\boldsymbol{\mu}\|^2 + 2dK^2\log^2\left(\frac{dBT}{\delta}\right).$$

Note that C_g (the g denoting gradient; see Lemma 5.1) is a universal constant independent of the dimension while C_d depends on the dimension. In the remainder of the proof we will use $\eta=\varepsilon/(16C_dC_g)$.

Let $\bar{\boldsymbol{\mu}} := \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Denote $\Delta_t^2 = \|\boldsymbol{\beta}_t-\bar{\boldsymbol{\mu}}\|^2$. Let $\tilde{\Delta}_t^2 = \|\tilde{\boldsymbol{\beta}}_t-\bar{\boldsymbol{\mu}}\|_2^2$. Using the same argument from (5.2), we have

$$\Delta_t^2-\Delta_{t+1}^2 \geq 2\eta\langle-\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t),\bar{\boldsymbol{\mu}}\rangle - \eta^2\left\|\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\right\|^2. \tag{5.7}$$

To control the gradient norm term, we use concentration. Standard concentration of sub-exponential random variables gives (see Lemma D.1 for the full details) with probability at least $1-\delta$, for all $i\in[B]$ and $t\in[T]$,

$$\|\mathbf{x}_i^{(t)}\|^2 \leq 2\|\boldsymbol{\mu}\|^2 + 2dK^2\log^2(dBT/\delta) =: C_d. \tag{5.8}$$

By Jensen's inequality,

$$\begin{aligned}
\|\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\|^2 &\leq \frac{1}{\sigma^2B}\sum_{i=1}^{B}|\ell'(|\langle\boldsymbol{\beta}_t,\mathbf{x}_i^{(t)}\rangle|/\sigma)|^2\|\mathbf{x}_i^{(t)}\|^2 \\
&\leq \frac{C_d}{\sigma^2}.
\end{aligned} \tag{5.9}$$

Substituting (5.9) into (5.7), we get

$$\Delta_t^2-\Delta_{t+1}^2 \geq 2\eta\left[\langle-\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t),\bar{\boldsymbol{\mu}}\rangle - \eta C_d/\sigma^2\right]. \tag{5.10}$$

For the first part of our proof, we claim that for all $t=0,1,\ldots,$ it holds that $\Delta_t\leq\Delta_0$, $\theta_t\in[0,\pi/2]$, and $\mathrm{err}(\boldsymbol{\beta}_t)\leq C_{\mathrm{err}}$. We show this result by induction. For the base case, we have for any $\boldsymbol{\beta}$ of unit norm,

$$\begin{aligned}
\mathbb{P}(y\neq\mathrm{sgn}(\langle\boldsymbol{\beta},\mathbf{x}\rangle)) &= \mathbb{P}(\langle\boldsymbol{\beta},y\mathbf{x}\rangle<0) \\
&= \mathbb{P}(\langle\boldsymbol{\beta},y\mathbf{x}-\boldsymbol{\mu}\rangle<-\langle\boldsymbol{\beta},\boldsymbol{\mu}\rangle) \\
&= \mathbb{P}(\langle\boldsymbol{\beta},\mathbf{z}\rangle<-\|\boldsymbol{\mu}\|\cos\theta),
\end{aligned} \tag{5.11}$$

where θ denotes the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. Since $\mathrm{err}(\boldsymbol{\beta}_{\mathrm{pl}})\leq C_{\mathrm{err}}<1/2$ and $\mathbf{z}$ is mean zero, we must

have $\theta_0 \in [0, \pi/2]$. Thus the base case $t = 0$ holds. Now assume the result holds for $t \in \mathbb{N}$ and consider the case $t+1$. Since $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$ are each of unit norm, we have the identity

$$\Delta_t^2 = \|\boldsymbol{\beta}_t - \bar{\boldsymbol{\mu}}\|^2 = 2(1 - \cos\theta_t) = 4\sin^2(\theta_t/2)$$
$$\implies \Delta_t = 2\sin(\theta_t/2). \tag{5.12}$$

By the induction hypothesis, $\mathrm{err}(\boldsymbol{\beta}_t) \le C_{\mathrm{err}}$ and $\theta_t \in [0, /\pi/2]$. We can thus use Lemma 5.1 (with ε from the lemma statement replaced with $\varepsilon/8C_g$) and (5.10) to get

$$\begin{aligned}\Delta_t^2 - \Delta_{t+1}^2 &\ge 2\eta\left[\frac{1}{C_g}\sin^2(\theta_t) - \frac{\varepsilon}{16C_g} - \frac{\eta C_d}{\sigma^2}\right] \\ &\ge 2\eta\left[\frac{1}{4C_g}\Delta_t^2 - \frac{\varepsilon}{16C_g} - \frac{\eta C_d}{\sigma^2}\right]. \end{aligned} \tag{5.13}$$

In the last line we have used (5.12) and that $\theta_t \in [0, \pi/2]$. We therefore have

$$\begin{aligned}\Delta_{t+1}^2 &\le \left(1 - \frac{\eta}{2C_g}\right)\Delta_t^2 + \eta\varepsilon/(8C_g) + 2\eta^2 C_d/\sigma^2 \\ &\overset{(i)}{\le} \left(1 - \frac{\eta}{2C_g}\right)\Delta_0^2 + \eta\varepsilon/(8C_g) + 2\eta^2 C_d/\sigma^2 \\ &= \Delta_0^2 - \eta\left(\frac{\Delta_0^2}{2C_g} - \frac{\varepsilon}{8C_g} - \frac{2\eta C_d}{\sigma^2}\right).\end{aligned}$$

In (i) we have used that $\eta = \varepsilon/(16C_dC_g\sigma^2)$ implies $1 - \eta/2C_g > 0$ and the inductive hypothesis that $\Delta_t^2 \le \Delta_0^2$. Thus, we see that the choice of η implies (where we assume $\Delta_0^2 > \varepsilon$ without loss of generality),

$$\frac{\Delta_0^2}{2C_g} - \frac{\varepsilon}{8C_g} - \frac{2\eta C_d}{\sigma^2} = \frac{\Delta_0^2}{2C_g} - \frac{\varepsilon}{4C_g} > 0.$$

Hence $\Delta_{t+1}^2 \le \Delta_0^2$. Using (5.11) and (5.12) and the induction hypothesis, this implies $\mathrm{err}(\boldsymbol{\beta}_{t+1}) \le C_{\mathrm{err}}$ and $\theta_{t+1} \in [0, \pi/2]$. This completes the induction and hence we have that for all t, $\mathrm{err}(\boldsymbol{\beta}_t) \le C_{\mathrm{err}}$ and $\theta_t \in [0, \pi/2]$ holds so that we may apply Lemma 5.1 for every t. In particular, for every t, (5.13) holds. We re-arrange (5.13) to get for any $T \in \mathbb{N}$,

$$\Delta_T^2 \le (1 - \eta/2C_g)\Delta_{T-1}^2 + \eta\varepsilon/(8C_g) + 2C_d\eta^2/\sigma^2.$$

One can verify (see Lemma D.2 for the detailed calculation) that for $\eta = \varepsilon/(16C_dC_g\sigma^2)$ and provided the number of iterations satisfies $T \ge 32C_dC_g^2\sigma^2\varepsilon^{-1}\log(32C_dC_g^2\sigma^2\varepsilon^{-1})$, this implies

$$4\sin^2(\theta_T/2) = \Delta_T^2 \le \varepsilon.$$

To convert the guarantee for the angle between $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}$ into one on the gap of the classification error between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$, we use (5.11) to write

$$\begin{aligned}\mathrm{err}(\boldsymbol{\beta}_t) - \mathrm{err}(\bar{\boldsymbol{\mu}}) &= \mathbb{P}(\langle\boldsymbol{\beta}_t, \mathbf{z}\rangle < -\|\boldsymbol{\mu}\|\cos\theta_t) \\ &\quad - \mathbb{P}(\langle\bar{\boldsymbol{\mu}}, \mathbf{z}\rangle < -\|\boldsymbol{\mu}\|) \qquad (5.14) \\ &\overset{(i)}{=} \mathbb{P}(\langle\mathbf{v}, \mathbf{z}\rangle \in [-\|\boldsymbol{\mu}\|, -\|\boldsymbol{\mu}\|\cos\theta_t]) \\ &\overset{(ii)}{\le} U\|\boldsymbol{\mu}\|[1 - \cos\theta_t] \\ &\overset{(iii)}{\le} U\|\boldsymbol{\mu}\|\sin^2\theta_t.\end{aligned}$$

In (i) we use that $\mathbf{z}$ is rotationally invariant and that $\|\boldsymbol{\beta}_t\| = \|\bar{\boldsymbol{\mu}}\| = 1$ so that $\langle\boldsymbol{\beta}_t, \mathbf{z}\rangle$ and $\langle\bar{\boldsymbol{\mu}}, \mathbf{z}\rangle$ have the same distribution as $\langle\mathbf{v}, \mathbf{z}\rangle$ for a vector $\mathbf{v}$ satisfying $\|\mathbf{v}\| = 1$. In (ii) we have used the definition of U-anti-concentration. In (iii) we have used the inequality $1 - \cos\theta_t = 1 - \sqrt{1 - \sin^2\theta_t} \le \sin^2\theta_t$. Since $\sin^2\theta_T \le 4\sin^2(\theta_T/2) \le \epsilon$, by rescaling ε to $\varepsilon/(U\|\boldsymbol{\mu}\|)$, we get the desired result.

□

6 Conclusion

In this work we theoretically analyzed an increasingly popular semi-supervised learning method: self-training with pseudolabels via gradient based optimization of the cross-entropy loss following supervised learning with a limited number of samples. We considered the setting of general mixture models satisfying benign concentration and anti-concentration properties. We showed that provided the initial pseudolabeler has classification error smaller than some absolute constant C_{err}, using $\tilde{O}(d/\varepsilon^2)$ unlabeled samples suffices for the self-training procedure to get within ε classification error of the Bayes-optimal classifier for the distribution. By showing that the standard gradient descent algorithm can learn a pseudolabeler with classification error at most C_{err} using only $O(d)$ labeled examples, our results provide the first proof that a constant (with respect to ε) number of labeled examples suffices for optimal performance in a semi-supervised self-training algorithm.

As we mentioned above, we show that a self-training algorithm achieves a sample complexity of $\tilde{O}(d/\varepsilon^2)$ samples while Li et al. (2017) are able to achieve a sample complexity of $O(d/\varepsilon)$ in the Gaussian setting by using a clustering-based algorithm. We are unsure if the algorithm we consider can achieve $O(d/\varepsilon)$ sample complexity, but we believe there are significant technical hurdles to doing so even if we were to assume a Gaussian mixture model. Typically, to achieve fast rates for stochastic optimization algorithms one needs a type of 'smoothness' result where the gradient of the loss can be bounded from above by the loss itself. With smoothness, if one can show something akin

to a proxy PL inequality (Frei and Gu, 2021) (since $\|\nabla\widehat{L}(\beta)\| \geq \langle -\nabla\widehat{L}(\beta), \bar{\mu}\rangle$, our Lemma 5.1 establishes such an inequality), then one can get linear convergence; with our $O(1/\varepsilon)$ batch size, this would give a total of $\tilde{O}(1/\varepsilon)$ sample complexity. The challenge in our setting is that the optimization objective (the unsupervised loss) is different from the objective we truly wish to minimize (the supervised classification error), and so the 'smoothness' and 'PL inequality' must be shown in terms of the objective we truly wish to minimize (in our work, we use $\sin^2\theta_t$ as a proxy for this objective). Lemma 5.1 provides the proxy PL inequality, and we believe that a 'fast rate' result might be possible if one could show an analogous 'smoothness' result, although this is a highly challenging prospect as the unsupervised loss we consider involves an absolute value term and hence has a discontinuous derivative.

For future research, we are interested in understanding if some of the methods we have developed can translate to settings where the optimal classifier is nonlinear. We expect this analysis to require novel non-convex optimization analyses. Our hope is that such settings will allow for better insight into the usage of self-training in neural networks.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. SF acknowledges the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639; DZ is supported by the Bloomberg Data Science Ph.D. Fellowship; ZC and QG are partially supported by the National Science Foundation CAREER Award 1906169, IIS-1855099 and IIS-2008981. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Balcan, M.-F. and Blum, A. (2010). A discriminative model for semi-supervised learning. *Journal of the ACM* .

Ben-David, S., Lu, T. and Pál, D. (2008). D.: Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Conference on Learning Theory (COLT)*.

Cai, T., Gao, R., Lee, J. D. and Lei, Q. (2021). A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning (ICML)*.

Castelli, V. and Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters* **16** 105–111.

Castelli, V. and Cover, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition in the regular parametric case," in preparation. *IEEE Transactions on Information Theory* **42** 2102–2117.

Chapelle, O., Schlkopf, B. and Zien, A. (2010). *Semi-Supervised Learning*. 1st ed. The MIT Press.

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Chen, Y., Wei, C., Kumar, A. and Ma, T. (2020b). Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Darnstädt, M., Simon, H. U. and Szörényi, B. (2013). Unlabeled Data Does Provably Help. In *Symposium on Theoretical Aspects of Computer Science (STACS)*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Diakonikolas, I., Gouleakis, T. and Tzamos, C. (2019). Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Diakonikolas, I., Kontonis, V., Tzamos, C. and Zarifis, N. (2020a). Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory (COLT)*.

Diakonikolas, I., Kontonis, V., Tzamos, C. and Zarifis, N. (2020b). Non-convex sgd learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Frei, S., Cao, Y. and Gu, Q. (2021a). Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning (ICML)*.

Frei, S., Cao, Y. and Gu, Q. (2021b). Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning (ICML)*.

Frei, S. and Gu, Q. (2021). Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Göpfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I. and Urner, R. (2019). When can unlabeled data improve the learning rate? In *Conference on Learning Theory (COLT)*.

Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*.

Hsu, D., Ji, Z., Telgarsky, M. and Wang, L. (2021). Generalization bounds via distillation. In *International Conference on Learning Representations (ICLR)*.

Kumar, A., Ma, T. and Liang, P. (2020). Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning (ICML)*.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Challenges in Representation Learning Workshop*.

Lee, J. D., Lei, Q., Saunshi, N. and Zhuo, J. (2020). Predicting what you already know helps: Provable self-supervised learning. *Preprint, arXiv:2008.01064* .

Li, T., Yi, X., Carmanis, C. and Ravikumar, P. (2017). Minimax Gaussian Classification & Clustering. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Oymak, S. and Gulcu, T. C. (2021). Statistical and algorithmic insights for semi-supervised learning with self-training. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Pham, H., Dai, Z., Xie, Q., Luong, M.-T. and Le, Q. V. (2021). Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. and Liang, P. (2020). Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*.

Rizve, M. N., Duarte, K., Rawat, Y. S. and Shah, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Scudder, H. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* **11** 363–371.

Singh, A., Nowak, R. and Zhu, J. (2009). Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tosh, C., Krishnamurthy, A. and Hsu, D. (2021). Contrastive learning, multi-view redundancy, and linear models. In *International Conference on Algorithmic Learning Theory (ALT)*.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint* **arXiv:1011.3027**.

Wei, C., Shen, K., Chen, Y. and Ma, T. (2021). Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations (ICLR)*.

Xie, S. M., Kumar, A., Jones, R., Khani, F., Ma, T. and Liang, P. (2021). In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Association for Computational Linguistics (ACL)*.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **3** 1–130.

Zou, D., Frei, S. and Gu, Q. (2021). Provable robustness of adversarial training for learning halfspaces with noise. In *International Conference on Machine Learning (ICML)*.

Zou, Y., Yu, Z., Liu, X., Kumar, B. V. and Wang, J. (2019). Confidence regularized self-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Supplementary Material: Self-training Converts Weak Learners to Strong Learners in Mixture Models

A Proofs from Section 3

A.1 Proof of Lemma 5.1: expected value

In this section we prove the first part of Lemma 5.1, involving the lower bound given for $\langle \boldsymbol{\mu}, -\mathbb{E}\nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)\rangle$. Our proof relies upon similar ideas used by Diakonikolas et al. (2020b) and Zou et al. (2021) for learning halfspaces with agnostic noise. At a high level, the noise in the halfspace setting considered by (Diakonikolas et al., 2020b; Zou et al., 2021) corresponds to the error made by the pseudolabeler in our setting.

Lemma A.1 (Lemma 5.1, expected value)**.** Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R > 0$. Let ℓ be well-behaved for some $C_\ell \geq 1$, and assume the temperature satisfies $\sigma \geq R \vee \|\boldsymbol{\mu}\|$. Suppose that $\|\boldsymbol{\beta}_t\| = 1$ is an initial estimate. Denote θ_t as the angle between $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}$, and assume that $\theta_t \in [0, \pi/2]$. Assume the classification error of $\boldsymbol{\beta}_t$ satisfies

$$\mathrm{err}_t := \mathbb{P}(y \neq \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle)) \leq \frac{R^2}{72 C_\ell U'} =: C_{\mathrm{err}}.$$

Then

$$\langle \bar{\boldsymbol{\mu}}, -\mathbb{E}\nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)\rangle \geq \frac{R^2 \|\boldsymbol{\mu}\|^2}{36 \sigma C_\ell U'} \cdot \sin^2(\theta_t).$$

Proof of Lemma 5.1. Since $\|\boldsymbol{\beta}_t\| = 1$, using the gradient formula (5.3),

$$-\mathbb{E}[\nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|\big) \cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \big(I - \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top\big)\mathbf{x}\big].$$

Denote

$$\bar{\boldsymbol{\mu}} := \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \quad \tilde{\boldsymbol{\mu}}_t := (I - \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top)\bar{\boldsymbol{\mu}}.$$

We can then write

$$\langle \boldsymbol{\mu}, -\mathbb{E}\nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)\rangle = \|\boldsymbol{\mu}\| \, \mathbb{E}\big[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|\big) \cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \tilde{\boldsymbol{\mu}}^\top \mathbf{x}\big].$$

Define the event

$$S_t := \{y = \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle)\}.$$

Then we can calculate

$$\begin{aligned}
&\mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle|/\sigma\big) \cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}\rangle) \cdot \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}] \\
&= \mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t)] + \mathbb{E}[\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t^c)] \\
&= \mathbb{E}[-\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x})] + 2\mathbb{E}[\ell'\big(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}_t^\top (y\mathbf{x}) \mathbb{1}(S_t^c)]. \qquad \text{(A.1)}
\end{aligned}$$

In the remainder of the proof, we will derive a lower bound on the first quantity and an upper bound on the absolute value of the second quantity.

We proceed with the lower bound for the first term as follows. Since the quantity depends only on the projection of $\mathbf{z}$ onto the space spanned by $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$, we work in this two dimensional space. Since $\mathbf{z}$ is rotationally invariant, we can rotate the coordinate system so that $\boldsymbol{\beta}_t = \mathbf{e}_2$, $\bar{\boldsymbol{\mu}} = (\sin\theta_t, \cos\theta_t)$, and $\tilde{\boldsymbol{\mu}} = (\sin\theta_t, 0)$, where θ_t is the angle

between $\boldsymbol{\beta}_t$ and $\bar{\boldsymbol{\mu}}$. Denote $p_{\boldsymbol{\beta}_t,\boldsymbol{\mu}}(\cdot,\cdot):\mathbb{R}^2\to[0,\infty)$ as the probability density function of the projection of $\mathbf{z}$ onto the 2D subspace spanned by $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}$. Then, using that $y\mathbf{x}$ has the same distribution as $\mathbf{z}+\boldsymbol{\mu}$,

$$\begin{aligned}
&\mathbb{E}[-\ell'\big(|\langle\boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})]\\
&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}-\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\theta_t\cdot(u_1+\|\boldsymbol{\mu}\|\sin\theta_t)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2\\
&=\int_{-\infty}^{\infty}-\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\theta_t\cdot\left[\int_{-\infty}^{\infty}(u_1+\|\boldsymbol{\mu}\|\sin\theta_t)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\right]\mathrm{d}u_2\\
&\overset{(i)}{=}\|\boldsymbol{\mu}\|\sin^2\theta_t\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}-\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}\tag{A.2}$$

In (i) we use the fact that $\mathbf{z}$ isotropic implies that the projection of $\mathbf{z}$ onto one dimensional subspaces are mean zero, and thus for all u_2 we have $\int_{-\infty}^{\infty}u_1\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1=0$. We now calculate a lower bound on the remaining quantity. Since $\mathbf{z}$ is rotationally invariant, we know that $p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)$ depends only on the distance from the origin $\sqrt{u_1^2+u_2^2}$. Thus, we can convert to polar coordinates and using an abuse of notation write $p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)$ to emphasize that the density only depends on the distance r from the origin in polar coordinates. Continuing, this means we can write

$$\begin{aligned}
&\int_{-\infty}^{\infty}-\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2.\\
&=\int_{r=0}^{\infty}rp_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)\int_{\phi=-\pi}^{\pi}-\ell'(|r\cos\phi+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\mathrm{d}r\mathrm{d}\phi\\
&\overset{(i)}{\geq}\int_{r=0}^{\infty}rp_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)\int_{\phi=0}^{\pi/2}-\ell'(|r\cos\phi+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\phi\cdot\mathrm{d}r\mathrm{d}\phi\\
&\overset{(ii)}{\geq}\int_{r=0}^{\infty}rp_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)\int_{\phi=0}^{\pi/2}\frac{1}{C_\ell}\exp(-|r\cos\phi+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\sin\phi\cdot\mathrm{d}r\mathrm{d}\phi\\
&\overset{(iii)}{=}\frac{\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{C_\ell}\int_{r=0}^{\infty}rp_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)\int_{\phi=0}^{\pi/2}\exp(-r\cos(\phi)/\sigma)\cdot\sin\phi\cdot\mathrm{d}r\mathrm{d}\phi\\
&\overset{(iv)}{=}\frac{\sigma\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{C_\ell}\int_{r=0}^{\infty}p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(r)(1-\exp(-r/\sigma))\cdot\mathrm{d}r\\
&\overset{(v)}{\geq}\frac{\sigma\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t))}{C_\ell U'}\int_0^R[1-\exp(-r/\sigma)]\mathrm{d}r\\
&\overset{(vi)}{\geq}\frac{\sigma\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{2C_\ell U'}\int_0^R\frac{r}{\sigma}\mathrm{d}r\\
&=\frac{R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'}.
\end{aligned}\tag{A.3}$$

In (i) we use that ℓ is decreasing and hence $-\ell'\geq 0$, as well as $\sin\phi\in[0,1]$ for $\phi\in[0,\pi/2]$. In (ii) we use Definition 3.5. In (iii) we have used the assumption that $\theta_t\in[0,\pi/2]$ and that $\cos\theta_t\geq 0$ for $\theta_t\in[0,\pi/2]$. In (iv) we use that $\int_0^{\pi/2}\exp(-a\cos x)\sin x\mathrm{d}x=(1-\exp(-a))/a$. In (v) we have used the definition of anti-anti-concentration. In (vi) we use that $\sigma\geq R$ and that $1-\exp(-x)\geq x/2$ on $[0,1]$. Putting (A.2) together with (A.3), we get

$$\mathbb{E}[-\ell'(|\langle\boldsymbol{\beta}_t,y\mathbf{x}\rangle|)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})]\geq\frac{R^2\|\boldsymbol{\mu}\|\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'}\cdot\sin^2(\theta_t).\tag{A.4}$$

We now want an upper bound on the second term in (A.1). Using the same coordinate system defined in terms of $\boldsymbol{\beta}_t=\mathbf{e}_2$ and $\bar{\boldsymbol{\mu}}=(\sin\theta_t,\cos\theta_t)$, we have that

$$S_t^c=\{\langle\boldsymbol{\beta}_t,y\mathbf{x}\rangle<0\}=\{\langle\boldsymbol{\beta}_t,\mathbf{z}+\boldsymbol{\mu}\rangle<0\}=\{u_2+\|\boldsymbol{\mu}\|\cdot\cos\theta_t<0\}.$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\ell'(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma)\tilde{\boldsymbol{\mu}}^\top y\mathbf{x}\cdot \mathbb{1}(S_t^c)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ell'(|u_2 + \|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot(u_1\sin\theta_t + \|\boldsymbol{\mu}\|\sin^2\theta_t)\mathbb{1}(u_2+\|\boldsymbol{\mu}\|\cos\theta_t\le 0)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{-\infty}^{\infty} \ell'(|u_2 + \|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot \mathbb{1}(u_2+\|\boldsymbol{\mu}\|\cos\theta_t\le 0)\left[\int_{-\infty}^{\infty}(u_1\sin\theta_t + \|\boldsymbol{\mu}\|\sin^2\theta_t)\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\right]\mathrm{d}u_2 \\
&\overset{(i)}{=} \|\boldsymbol{\mu}\|\sin^2\theta_t\int_{-\infty}^{\infty}\ell'(|u_2+\|\boldsymbol{\mu}\|\cos\theta_t|/\sigma)\cdot\mathbb{1}(u_2+\|\boldsymbol{\mu}\|\cos\theta_t\le 0)\int_{u_1=-\infty}^{\infty} p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}$$

In (i) we use the fact that $\mathbf{z}$ isotropic implies that for all u_2 we have $\int_{-\infty}^{\infty} u_1\cdot p_{\boldsymbol{\beta}_t,\bar{\boldsymbol{\mu}}}(u_1,u_2)\mathrm{d}u_1 = 0$. Using $|\ell'(z)|\le 1$ on $[0,\infty)$, we can therefore bound

$$\left|\mathbb{E}\left[\ell'(|\langle \boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma)\tilde{\boldsymbol{\mu}}^\top y\mathbf{x}\cdot\mathbb{1}(S_t^c)\right]\right| \le \|\boldsymbol{\mu}\|\sin^2\theta_t\cdot\mathbb{P}(S_t^c) = \|\boldsymbol{\mu}\|\sin^2\theta_t\cdot\mathrm{err}_t. \tag{A.5}$$

We then return to (A.1). By combining (A.5) with (A.4), we see that

$$\begin{aligned}
\langle \boldsymbol{\mu}, -\mathbb{E}[\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle &= \frac{\|\boldsymbol{\mu}\|}{\sigma}\mathbb{E}[-\ell'\big(|\langle\boldsymbol{\beta}_t, y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})] + \frac{2\|\boldsymbol{\mu}\|}{\sigma}\mathbb{E}\left[\ell'\big(|\langle\boldsymbol{\beta}_t,y\mathbf{x}\rangle|/\sigma\big)\tilde{\boldsymbol{\mu}}^\top(y\mathbf{x})\mathbb{1}(S_t^c)\right] \\
&\ge \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\frac{R^2\|\boldsymbol{\mu}\|\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'}\cdot\sin^2(\theta_t)\right] - 2\frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\|\boldsymbol{\mu}\|\sin^2\theta_t\cdot\mathrm{err}_t\right] \\
&= \frac{\|\boldsymbol{\mu}\|^2\cdot\sin^2\theta_t}{\sigma}\left[\frac{R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\cos\theta_t)}{6C_\ell U'} - 2\mathrm{err}_t\right] \\
&\ge \frac{\|\boldsymbol{\mu}\|^2\cdot\sin^2\theta_t}{\sigma}\left[\frac{R^2\exp(-\frac{\|\boldsymbol{\mu}\|}{\sigma})}{6C_\ell U'} - 2\mathrm{err}_t\right].
\end{aligned}$$

Thus, by choosing $\sigma\ge\|\boldsymbol{\mu}\|$, we have

$$\begin{aligned}
\langle \boldsymbol{\mu}, -\mathbb{E}[\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle &\ge \frac{\|\boldsymbol{\mu}\|^2\cdot\sin^2\theta_t}{\sigma}\left[\frac{R^2\exp(-1)}{6C_\ell U'} - 2\mathrm{err}_t\right] \\
&\ge \frac{\|\boldsymbol{\mu}\|^2\cdot\sin^2\theta_t}{\sigma}\left[\frac{R^2}{18C_\ell U'} - 2\mathrm{err}_t\right].
\end{aligned}$$

In particular, if

$$\mathrm{err}_t \le \frac{R^2}{72C_\ell U'} =: C_{\mathrm{err}},$$

then we have

$$\langle \boldsymbol{\mu}, -\mathbb{E}[\nabla\widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)]\rangle \ge \frac{\|\boldsymbol{\mu}\|^2 R^2}{36\sigma C_\ell U'}\sin^2\theta_t.$$

□

A.2 Proof of Lemma 5.1, batch of samples

We now prove the second part of Lemma 5.1, where we show that provided the batch size is large enough, then the same lower bound that holds from the expected value holds when using finite samples.

Lemma A.2 (Lemma 5.1, batch of samples)**.** Let $(\mathbf{x},y)\sim\mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}\in\mathbb{R}^d$ and parameters $K,U,U',R>0$. Let ℓ be well-behaved for some $C_\ell\ge 1$ and assume the temperature satisfies $\sigma\ge R\vee\|\boldsymbol{\mu}\|$. Suppose that $\theta_t\in[0,\pi/2]$ is the angle between $\boldsymbol{\mu}$ and $\boldsymbol{\beta}_t$ where $\|\boldsymbol{\beta}_t\|=1$. Then there exists a universal constant $C_B>0$ such that for any $\varepsilon,\delta\in(0,1)$,

$$B\ge C_B\left(\frac{KC_\ell U'}{R^2}\right)^2\varepsilon^{-1}\log(2/\delta),$$

then with probability at least $1-\delta$,

$$\langle \boldsymbol{\mu}, -\nabla \widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)\rangle \geq \frac{R^2\left\|\boldsymbol{\mu}\right\|^2}{72\sigma C_\ell U'}\sin^2\theta_t - \varepsilon/2.$$

Proof of Lemma 5.1, batch of samples. Denote by $\mathbf{z}_i = y_i\mathbf{x}_i - \boldsymbol{\mu}$ and $S_t = \{y_i = \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i\rangle)\}$. Let $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|$ and $\tilde{\boldsymbol{\mu}} = (I - \boldsymbol{\beta}_t\boldsymbol{\beta}_t^\top)\bar{\boldsymbol{\mu}}$. Let us define

$$A_t = \frac{1}{B}\sum_{i=1}^{B} -\ell'\left(|\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle|/\sigma\right)\cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle)\cdot \tilde{\boldsymbol{\mu}}^\top\mathbf{x}_i^{(t)},$$

so that $\langle \boldsymbol{\mu}, -\mathbb{E}\nabla\widehat{L_t^{\mathrm{u}}}(\boldsymbol{\beta}_t)\rangle = \frac{\|\boldsymbol{\mu}\|}{\sigma}\mathbb{E}A_t$. We can use $\langle \tilde{\boldsymbol{\mu}}, \mathbf{x}_i^{(t)}\rangle = \langle \tilde{\boldsymbol{\mu}}, \mathbf{z}_i + y\boldsymbol{\mu}\rangle$ to write the above as the sum of two terms:

$$A_t^{(1)} = \frac{1}{B}\sum_{i=1}^{B} u_i^{(1)} := \frac{1}{B}\sum_{i=1}^{B} -\ell'\left(|\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle|\right)\cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle)\cdot \langle \tilde{\boldsymbol{\mu}}, \mathbf{z}_i\rangle,$$
$$A_t^{(2)} = \frac{1}{B}\sum_{i=1}^{B} u_i^{(2)} := \frac{1}{B}\sum_{i=1}^{B} -\ell'\left(|\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle|\right)\cdot \mathrm{sgn}(\langle \boldsymbol{\beta}_t, \mathbf{x}_i^{(t)}\rangle)\cdot \|\boldsymbol{\mu}\|\cdot y\langle \tilde{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}\rangle.$$

First note

$$\begin{aligned}\|\tilde{\boldsymbol{\mu}}\|^2 &= \|\bar{\boldsymbol{\mu}} - \boldsymbol{\beta}_t\langle \boldsymbol{\beta}_t, \bar{\boldsymbol{\mu}}\rangle\|^2\\ &= \|\bar{\boldsymbol{\mu}} - \boldsymbol{\beta}_t\cos\theta_t\|^2\\ &= 1 + \cos^2\theta_t - 2\cos^2\theta_t\\ &= \sin^2\theta_t.\end{aligned}$$

Since $\mathbf{z}_i$ are i.i.d. isotropic, each $u_i^{(1)}$ are i.i.d. mean zero random variables with sub-exponential norm at most $\|\tilde{\boldsymbol{\mu}}\|\cdot\left\|\langle \|\tilde{\boldsymbol{\mu}}\|/\|\tilde{\boldsymbol{\mu}}\|, \mathbf{z}\rangle\right\|_{\psi_2} \leq \|\tilde{\boldsymbol{\mu}}\|K = K\sin\theta_t$. Thus, using sub-exponential concentration (Vershynin, 2010, Proposition 5.16), there exists a universal constant $C>0$ such that for any $\xi>0$,

$$\mathbb{P}(|A_t^{(1)} - \mathbb{E}A_t^{(1)}| \geq \xi) \leq 2\exp\left(-C\min\left(\frac{\xi^2 B}{K^2\sin^2\theta_t}, \frac{\xi B}{K\sin\theta_t}\right)\right).$$

By taking $\xi = CK\sin\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}$ for a sufficiently large constant $C>0$, this implies that with probability at least $1-\delta/2$,

$$|A_t^{(1)} - \mathbb{E}A_t^{(1)}| \leq CK\sin\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}. \tag{A.6}$$

On the other hand, each of $u_i^{(2)}$ are i.i.d. sub-exponential random variables with mean $\mathbb{E}A_t$ and sub-exponential norm at most $\|\boldsymbol{\mu}\|\left\|\langle \tilde{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}\rangle\right\|_{\psi_2} = \|\boldsymbol{\mu}\|(1-\cos\theta_t) \leq \|\boldsymbol{\mu}\|\sin^2\theta_t$ (using $\theta_t \in [0,\pi/2]$) and thus for any $\xi>0$,

$$\mathbb{P}(|A_t^{(2)} - \mathbb{E}A_t^{(2)}| \geq \xi) \leq 2\exp\left(-C\min\left(\frac{\xi^2 B}{\|\boldsymbol{\mu}\|^2\sin^4\theta_t}, \frac{\xi B}{\|\boldsymbol{\mu}\|\sin^2\theta_t}\right)\right).$$

By taking $\xi = C\|\boldsymbol{\mu}\|\sin^2\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}$ for C sufficiently large universal constant, we get that with probability at least $1-\delta/2$,

$$|A_t^{(2)} - \mathbb{E}A_t^{(2)}| \leq C\|\boldsymbol{\mu}\|\sin^2\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}. \tag{A.7}$$

Putting (A.6) and (A.7) together and applying union bound, we have that with probability at least $1-\delta$,

$$\begin{aligned}\langle \boldsymbol{\mu}, -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle &= \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[A_t^{(1)} + A_t^{(2)}\right] \\ &\overset{(i)}{\geq} \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\mathbb{E}A_t - CK\sin\theta_t\sqrt{\frac{1}{B}\log(2/\delta)} - C\|\boldsymbol{\mu}\|\sin^2\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}\right] \\ &\overset{(ii)}{\geq} \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\frac{\|\boldsymbol{\mu}\|R^2}{36C_\ell U'}\sin^2\theta_t - CK\sin\theta_t\sqrt{\frac{1}{B}\log(2/\delta)} - C\|\boldsymbol{\mu}\|\sin^2\theta_t\sqrt{\frac{1}{B}\log(2/\delta)}\right]. \end{aligned} \tag{A.8}$$

In (i) we use (A.6) and (A.7). In (ii) we use Lemma A.1. Now, to complete the proof, we consider separately the case that $\sin^2\theta_t > \varepsilon$ and the case that $\sin^2\theta_t \leq \varepsilon$. In the first instance, the batch size satisfies of

$$B \geq \left(\frac{144KC^2C_\ell U'}{R^2}\right)^2 \varepsilon^{-1}\log(2/\delta) \geq \left(\frac{144KC^2C_\ell U'}{R^2\sin\theta_t}\right)^2 \log(2/\delta).$$

Thus using $K \geq 1$ and $\|\boldsymbol{\mu}\| \geq 1$, we have

$$CK\sin\theta_t\sqrt{\frac{1}{B}\log(2/\delta)} \leq \frac{R^2}{144C_\ell U'}\sin^2\theta_t \leq \frac{\|\boldsymbol{\mu}\|R^2}{144C_\ell U'}\sin^2\theta_t,$$

and

$$C\|\boldsymbol{\mu}\|\sin^2\theta_t\sqrt{\frac{1}{B}\log(2/\delta)} \leq \frac{\|\boldsymbol{\mu}\|R^2}{144KC_\ell U'}\sin^3\theta_t \leq \frac{\|\boldsymbol{\mu}\|R^2}{144C_\ell U'}\sin^2\theta_t,$$

Substituting this into (A.8), we get

$$\langle \boldsymbol{\mu}, -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle \geq \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\frac{\|\boldsymbol{\mu}\|R^2}{72C_\ell U'}\sin^2\theta_t\right] \geq \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\frac{\|\boldsymbol{\mu}\|R^2}{72C_\ell U'}\sin^2\theta_t - \varepsilon/2.\right] \tag{A.9}$$

On the other hand, if $\sin^2\theta_t \leq \varepsilon$, then notice that (A.8) becomes

$$\langle \boldsymbol{\mu}, -\nabla \widehat{L}_t^{\mathrm{u}}(\boldsymbol{\beta}_t)\rangle \geq \frac{\|\boldsymbol{\mu}\|}{\sigma}\left[\frac{\|\boldsymbol{\mu}\|R^2}{36C_\ell U'}\sin^2\theta_t - CK\varepsilon^{1/2}\sqrt{\frac{1}{B}\log(2/\delta)} - C\|\boldsymbol{\mu}\|\varepsilon\sqrt{\frac{1}{B}\log(2/\delta)}\right]. \tag{A.10}$$

Then $B = \Omega(\varepsilon^{-1})$ implies that (A.9) holds in this case as well. This completes the proof. □

B Proofs from Section 4

In this section, we prove the following theorem.

Theorem B.1 (Theorem 4.1, restated)**.** Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a mixture distribution with mean μ and parameters $K, U, U', R > 0$. Let $C_{\mathrm{err}} > 0$ be arbitrary, and assume $\|\boldsymbol{\mu}\| \geq 3K\max(\log(8/C_{\mathrm{err}}), 22K)$. By running Algorithm 2 with $\eta = (\|\boldsymbol{\mu}\|^2 + d)^{-1}C_{\mathrm{err}}/8$ and $T = 8\eta^{-1}C_{\mathrm{err}}^{-1}\|\boldsymbol{\mu}\|^2$ iterations, there exists $i \leq 4\log(1/\delta)$ and $t < T$ such that with probability at least $1-\delta$,

$$\mathbb{P}(y \neq \mathrm{sgn}(\langle \boldsymbol{\beta}_t^{(i)}, \mathbf{x}\rangle)) \leq C_{\mathrm{err}}.$$

To show this theorem, we will first need an upper bound for the classification error that is achieved by the classifier $\mathbf{x} \mapsto \mathrm{sgn}(\langle \bar{\boldsymbol{\mu}}, \mathbf{x}\rangle)$. For the standard isotropic Gaussian mixture model $N(y\boldsymbol{\mu}, I_d)$, it is easy to show that $\mathbb{P}(y \neq \mathrm{sgn}(\langle \bar{\boldsymbol{\mu}}, \mathbf{x}\rangle)) = \Phi(-\|\boldsymbol{\mu}\|)$, where Φ is the standard normal CDF. For sub-exponential mixture models, we have a similar bound.

Lemma B.2. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R > 0$. Then we have,

$$\mathbb{P}\left(y \neq \mathrm{sgn}(\langle \mathbf{x}, \boldsymbol{\mu}\rangle)\right) \leq K\exp(-\|\boldsymbol{\mu}\|/K).$$

Proof. For simplicity denote $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|$. We have

$$\begin{aligned}
\mathbb{P}\left(y \neq \operatorname{sgn}(\langle \mathbf{x}, \boldsymbol{\mu}\rangle)\right) &= \mathbb{P}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0) \\
&= \mathbb{P}(\langle y\mathbf{x} - \boldsymbol{\mu}, \bar{\boldsymbol{\mu}}\rangle < -\left\|\boldsymbol{\mu}\right\|) \\
&= \mathbb{P}(\langle \mathbf{z}, \bar{\boldsymbol{\mu}}\rangle < -\left\|\boldsymbol{\mu}\right\|) \\
&= \int_{-\infty}^{-\|\boldsymbol{\mu}\|} \mathbb{P}(\langle \mathbf{z}, \bar{\boldsymbol{\mu}}\rangle < -t)\mathrm{d}t \\
&\leq \int_{-\infty}^{-\|\boldsymbol{\mu}\|} \exp(-|t|/K)\mathrm{d}t \\
&= K\exp(-\left\|\boldsymbol{\mu}\right\|/K).
\end{aligned}$$

The inequality uses the definition of sub-exponential. □

The next intermediate result we need will be a characterization of the population loss under a surrogate for the 0-1 loss.

Lemma B.3. Let ℓ be 1-Lipschitz, decreasing, with $\ell(z) \leq \exp(-z)$ for $z > 0$. Let $(\mathbf{x}, y) \sim \mathcal{D}$ be a mixture model with mean $\boldsymbol{\mu}$ and parameters $K, U, U', R > 0$. Then

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(y\langle\boldsymbol{\mu}, x\rangle) \leq (1 + \left\|\boldsymbol{\mu}\right\| + 2\left\|\boldsymbol{\mu}\right\|^2)K\exp(-\left\|\boldsymbol{\mu}\right\|/K) + \exp(-\left\|\boldsymbol{\mu}\right\|/2K) + \exp(-\left\|\boldsymbol{\mu}\right\|/2).$$

In particular, provided $\left\|\boldsymbol{\mu}\right\| \geq 64K^2$, we have

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell(y\langle\boldsymbol{\mu}, \mathbf{x}\rangle) \leq \exp(-\left\|\boldsymbol{\mu}\right\|/3K).$$

Proof. Denote $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|$ for simplicity, and let $\gamma = 1/2$. Our proof uses an argument similar to Frei et al. (2021a, Lemma 5.9) and Zou et al. (2021, Lemma 2.7), where we decompose

$$\begin{aligned}
\mathbb{E}\ell(\langle y\mathbf{x}, \boldsymbol{\mu}\rangle) &= \mathbb{E}[\ell(\langle y\mathbf{x}, \boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0)] \\
&\quad + \mathbb{E}[\ell(\langle y\mathbf{x}, \boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle \in [0, \gamma))] \\
&\quad + \mathbb{E}[\ell(\langle y\mathbf{x}, \boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle > \gamma)].
\end{aligned} \tag{B.1}$$

We first bound the first term. Denote by OPT the classification error using $\boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|$,

$$\mathsf{OPT} = \mathbb{P}(y \neq \operatorname{sgn}(\langle \mathbf{x}, \boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|\rangle)) \leq K\exp(-\left\|\boldsymbol{\mu}\right\|/K),$$

where the inequality follows by Lemma B.2. Let $\xi = 2\left\|\boldsymbol{\mu}\right\|$. We have

$$\begin{aligned}
\mathbb{E}[\ell(\langle y\mathbf{x}, \boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0)] &\overset{(i)}{\leq} \mathbb{E}[(1 + |\langle y\mathbf{x}, \boldsymbol{\mu}\rangle|)\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0)] \\
&= \mathsf{OPT} + \left\|\boldsymbol{\mu}\right\|\mathbb{E}[|\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle|\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0,\ |\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle| \leq \xi) \\
&\quad + \left\|\boldsymbol{\mu}\right\|\mathbb{E}[|\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle|\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle < 0,\ |\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle| > \xi) \\
&\leq (1 + \left\|\boldsymbol{\mu}\right\|\xi)\mathsf{OPT} + \left\|\boldsymbol{\mu}\right\|\mathbb{E}[|\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle|\mathbb{1}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle > \xi)] \\
&= (1 + \left\|\boldsymbol{\mu}\right\|\xi)\mathsf{OPT} + \left\|\boldsymbol{\mu}\right\|\int_{\xi}^{\infty}\mathbb{P}(\langle y\mathbf{x}, \bar{\boldsymbol{\mu}}\rangle > t)\mathrm{d}t \\
&= (1 + \left\|\boldsymbol{\mu}\right\|\xi)\mathsf{OPT} + \left\|\boldsymbol{\mu}\right\|\int_{\xi}^{\infty}\mathbb{P}(\langle y\mathbf{x} - \boldsymbol{\mu}, \bar{\boldsymbol{\mu}}\rangle > t - \boldsymbol{\mu})\mathrm{d}t \\
&\overset{(ii)}{\leq} (1 + \left\|\boldsymbol{\mu}\right\|\xi)\mathsf{OPT} + \left\|\boldsymbol{\mu}\right\|\int_{\xi}^{\infty}\exp(-(t - \left\|\boldsymbol{\mu}\right\|)/K)\mathrm{d}t \\
&= (1 + \left\|\boldsymbol{\mu}\right\|\xi)\mathsf{OPT} + K\left\|\boldsymbol{\mu}\right\|\exp((\left\|\boldsymbol{\mu}\right\| - \xi)/K) \\
&= (1 + 2\left\|\boldsymbol{\mu}\right\|^2)\mathsf{OPT} + K\left\|\boldsymbol{\mu}\right\|\exp(-\left\|\boldsymbol{\mu}\right\|/K).
\end{aligned} \tag{B.2}$$

In (i) we use Cauchy–Schwarz and that ℓ is 1-Lipschitz and decreasing. In (ii) we use that $t \geq \xi \geq \left\|\boldsymbol{\mu}\right\|$ and the definition of sub-exponential.

For the second term of (B.1), we have

$$\begin{aligned}\mathbb{E}[\ell(\langle y\mathbf{x},\boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x},\bar{\boldsymbol{\mu}}\rangle\in[0,\gamma))] &\le \ell(0)\mathbb{P}(\langle y\mathbf{x},\bar{\boldsymbol{\mu}}\rangle\in[0,\gamma))\\ &\le \mathbb{P}(\langle y\mathbf{x}-\boldsymbol{\mu},\bar{\boldsymbol{\mu}}\rangle\in[-\|\boldsymbol{\mu}\|,-\|\boldsymbol{\mu}\|+\gamma))\\ &= \mathbb{P}(\langle y\mathbf{x}-\boldsymbol{\mu},\bar{\boldsymbol{\mu}}\rangle\le-\|\boldsymbol{\mu}\|+\gamma)-\mathbb{P}(\langle y\mathbf{x}-\boldsymbol{\mu},\bar{\boldsymbol{\mu}}\rangle\le-\|\boldsymbol{\mu}\|)\\ &\overset{(i)}{\le} \mathbb{P}((\langle y\mathbf{x}-\boldsymbol{\mu},\bar{\boldsymbol{\mu}}\rangle\le-\frac{1}{2}\|\boldsymbol{\mu}\|)\\ &\overset{(ii)}{\le} \exp(-\|\boldsymbol{\mu}\|/2K). \end{aligned}\tag{B.3}$$

where in (i) we use $\gamma<1\le\frac{1}{2}\|\boldsymbol{\mu}\|$ and in (ii) we have used the definition of sub-exponential.

Finally, for the third term of (B.1), we use that ℓ is decreasing and has exponential tail so that

$$\mathbb{E}[\ell(\langle y\mathbf{x},\boldsymbol{\mu}\rangle)\mathbb{1}(\langle y\mathbf{x},\bar{\boldsymbol{\mu}}\rangle>\gamma)]\le\ell(\|\boldsymbol{\mu}\|)\le\exp(-\gamma\|\boldsymbol{\mu}\|)=\exp(-\|\boldsymbol{\mu}\|/2).\tag{B.4}$$

Putting (B.2), (B.3), and (B.4) all together, we get

$$\begin{aligned}\mathbb{E}[\ell(y\langle\boldsymbol{\mu},x\rangle)] &\le (1+K\|\boldsymbol{\mu}\|^3)\mathsf{OPT}+K\|\boldsymbol{\mu}\|\exp(-\|\boldsymbol{\mu}\|^2/2)+\exp(-\|\boldsymbol{\mu}\|/(2K))+\exp(-\|\boldsymbol{\mu}\|/2)\\ &\le (1+2\|\boldsymbol{\mu}\|^2)K\exp(-\|\boldsymbol{\mu}\|/K)+K\|\boldsymbol{\mu}\|\exp(-\|\boldsymbol{\mu}\|^2/2)+2\exp(-\|\boldsymbol{\mu}\|/(2K))\\ &\overset{(i)}{\le} (3K+2)\exp(-\|\boldsymbol{\mu}\|/2K)\\ &\overset{(ii)}{\le} \exp(-\|\boldsymbol{\mu}\|/3K).\end{aligned}$$

In (i) we use that $x/\log x\ge\sqrt{x}$ and thus $2\|\boldsymbol{\mu}\|^2\exp(-\|\boldsymbol{\mu}\|/K)=\exp(-\frac{1}{k}\|\boldsymbol{\mu}\|+4\log\|\boldsymbol{\mu}\|)\le\exp(-\|\boldsymbol{\mu}\|/2K)$ for $\|\boldsymbol{\mu}\|\ge 64K^2$, and in (ii) we again use that $\|\boldsymbol{\mu}\|\ge 64K^2$. □

With the above in hand we can complete the proof of Theorem 4.1. We will show that provided the means of the mixture model are sufficiently well-separated (by an absolute constant), then the population risk under the convex surrogate ℓ can be as small as $\Theta(C_{\mathrm{err}})$. This leads to an upper bound for the classification error using supervised learning that is at most C_{err}.

Proof of Theorem 4.1. Fix $i\in\{1,\dots,\log(1/\delta)\}$ as given in Algorithm 2. As the cross-entropy loss is convex and 1-Lipschitz, and as $\mathbb{E}[\|\mathbf{x}\|^2]\le 2\|\boldsymbol{\mu}\|^2+2\mathbb{E}[\|\mathbf{z}\|^2]=2(\|\boldsymbol{\mu}\|^2+d)$, by Frei et al. (2021a, Lemma C.1), for $\eta\le(\|\boldsymbol{\mu}\|^2+d)^{-1}\varepsilon/4$, we know there exists $t_i<T=4\eta^{-1}\varepsilon^{-1}\|\boldsymbol{\mu}\|^2$ such that $\mathbb{E}[\ell(y\langle\boldsymbol{\beta}t_i,\mathbf{x}\rangle]\le\mathbb{E}[\ell(y\langle\boldsymbol{\mu},\mathbf{x}\rangle)]+\varepsilon/2$. By Markov's inequality, for each i, with probability at least $1-\frac{1}{1+\delta_0}$ over $\{(\mathbf{x}_t^{(i)},y_t^{(i)})\}_{t=0,\dots,T}$, $\mathbb{E}[\ell(y\langle\boldsymbol{\beta}_{t_i},\mathbf{x}\rangle)]\le(1+\delta_0)[\mathbb{E}\ell(y\langle\boldsymbol{\mu},\mathbf{x}\rangle)+\varepsilon]$. As the $\{(\mathbf{x}_t^{(i)},y_t^{(i)})\}$ are independent for different i, the probability of failure for I independent such i is $[1/(1+\delta_0)]^I$. As $1/x\le 1/\log(1+x)\le 2/x$ on $[0,1]$, this implies that as long as $I\ge 2\delta_0^{-1}\log(1/\delta)$, then with probability at least $1-\delta$, there exists $i\in I$ such that $\mathbb{E}[\ell(y\langle\boldsymbol{\beta}_{t_i},\mathbf{x}\rangle)]\le(1+\delta_0)[\mathbb{E}\ell(y\langle\boldsymbol{\mu},\mathbf{x}\rangle)+\varepsilon]$. In particular, for $I=4\lceil\log(1/\delta)\rceil$, we have with probability at least $1-\delta$, for some i and $t_i<T$,

$$\mathbb{E}\ell(y\langle\boldsymbol{\beta}_{t_i},\mathbf{x}\rangle)\le 2\mathbb{E}\ell(y\langle\boldsymbol{\mu},\mathbf{x}\rangle)+\varepsilon.$$

By Lemma B.3, we know that for $\|\boldsymbol{\mu}\|\ge 64K^2$, we have

$$\mathbb{E}\ell(y\langle\boldsymbol{\mu},\mathbf{x}\rangle)\le\exp(-\|\boldsymbol{\mu}\|/3K).$$

To guarantee $2\exp(-\|\boldsymbol{\mu}\|/3K)\le C_{\mathrm{err}}\log(2)/2$ it suffices to take $\|\boldsymbol{\mu}\|\ge 3K\log(8/C_{\mathrm{err}})$. Thus, provided $\|\boldsymbol{\mu}\|\ge 3K\max(\log(8/C_{\mathrm{err}}),22K)$, we have that with probability at least $1-\delta$,

$$\mathbb{P}(y\ne\mathrm{sgn}(\langle\boldsymbol{\beta}_{t_i},\mathbf{x}\rangle))\le\frac{1}{\ell(0)}\mathbb{E}\ell(y\langle\boldsymbol{\beta}_{t_i},\mathbf{x}\rangle)\le\frac{1}{2}C_{\mathrm{err}}+\varepsilon.$$

Taking $\varepsilon=C_{\mathrm{err}}/2$ completes the proof. □

C Bayes-optimal Classifier for Mixture Distributions

Here we prove a more general version of Fact 3.4 that relies only upon rotational symmetry and unimodality.

Fact C.1. Let $\boldsymbol{\mu} \in \mathbb{R}^d$. Suppose $\mathbf{z}$ is continuous, rotationally symmetric and unimodal in the sense that its density function $p_{\mathbf{z}}(\mathbf{z})$ is a decreasing function of $\|\mathbf{z}\|_2$. Assume $Y \sim \mathrm{Unif}(\{1, -1\})$ and $\mathbf{x}|Y = y \sim \mathbf{z} + y\boldsymbol{\mu}$. Then the Bayes-optimal classifier is given by $\mathbf{x} \mapsto \mathrm{sgn}(\langle \boldsymbol{\mu}, \mathbf{x} \rangle)$.

Proof. Let us introduce some notation. We denote by $\boldsymbol{X}, \boldsymbol{Z}, Y$ as random variables and by $\mathbf{z}, \mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, 1\}$ as possible realizations of those random variables. The Bayes-optimal classifier chooses a label for a feature $\mathbf{x} \in \mathbb{R}^d$ by taking the maximum value of $\mathbb{P}(Y = y|\mathbf{x})$ over $y \in \{\pm 1\}$. Thus, we can write the Bayes-optimal classifier $h_{\mathrm{Bayes}}(\mathbf{x})$ as

$$h_{\mathrm{Bayes}}(\mathbf{x}) = \mathrm{argmax}_{y \in \{\pm 1\}} \mathbb{P}(Y = y|\mathbf{x}).$$

Note that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$. Thus, by Bayes' theorem,

$$\begin{aligned}\mathbb{P}(Y = y|\mathbf{x}) &= \frac{\mathbb{P}(\mathbf{x}|Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(\mathbf{x}|Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(\mathbf{x}|Y = -1)\mathbb{P}(Y = -1)} \\ &= \frac{\mathbb{P}(\mathbf{x}|Y = y)}{\mathbb{P}(\mathbf{x}|Y = 1) + \mathbb{P}(\mathbf{x}|Y = -1)}.\end{aligned}$$

Thus, we see that the Bayes-optimal classifier chooses the label for a feature $\mathbf{x}$ which maximizes the likelihood of observing $\mathbf{x}$:

$$h_{\mathrm{Bayes}}(\mathbf{x}) = \mathrm{argmax}_{y \in \{\pm 1\}} \mathbb{P}(\mathbf{x}|Y = y).$$

If we denote by $p_{\mathbf{z}}(\cdot)$ as the probability density function of $\mathbf{z}$, since $\mathbf{x}|Y = y \sim \mathbf{z} + y\boldsymbol{\mu}$, using the properties of the probability density function under linear transformations, we have

$$\mathbb{P}(\mathbf{x}|Y = y) = p_{\mathbf{z}}(\mathbf{x} - y\boldsymbol{\mu}).$$

Thus,

$$h_{\mathrm{Bayes}}(\mathbf{x}) = \begin{cases} 1 & \text{if } p_{\mathbf{z}}(\mathbf{x} - \boldsymbol{\mu}) > p_{\mathbf{z}}(\mathbf{x} + \boldsymbol{\mu}), \\ -1 & \text{if } p_{\mathbf{z}}(\mathbf{x} - \boldsymbol{\mu}) \leq p_{\mathbf{z}}(\mathbf{x} + \boldsymbol{\mu}). \end{cases} \tag{C.1}$$

By assumption, there exists a decreasing function $g : [0, \infty) \to [0, \infty)$ such that the density function of $\mathbf{z}$ satisfies

$$p_{\mathbf{z}}(\mathbf{z}) = g(\|\mathbf{z}\|_2^2).$$

By (C.1), the Bayes-optimal decision boundary is determined by comparing $p_{zb}(\mathbf{x} - \boldsymbol{\mu})$ and $p_{\mathbf{z}}(\mathbf{x} + \boldsymbol{\mu})$. We have,

$$\begin{aligned} p_{\mathbf{z}}(\mathbf{x} - \boldsymbol{\mu}) > p_{\mathbf{z}}(\mathbf{x} + \boldsymbol{\mu}) &\overset{(i)}{\iff} g(\|\mathbf{x} - \boldsymbol{\mu}\|_2^2) > g(\|\mathbf{x} + \boldsymbol{\mu}\|_2^2) \\ &\overset{(ii)}{\iff} \|\mathbf{x} - \boldsymbol{\mu}\|_2^2 \leq \|\mathbf{x} + \boldsymbol{\mu}\|^2 \\ &\iff \|\mathbf{x}\|_2^2 + \|\boldsymbol{\mu}\|_2^2 - 2\langle \mathbf{x}, \boldsymbol{\mu} \rangle \leq \|\mathbf{x}\|_2^2 + \|\boldsymbol{\mu}\|_2^2 + 2\langle \mathbf{x}, \boldsymbol{\mu} \rangle \\ &\iff \langle \mathbf{x}, \boldsymbol{\mu} \rangle \geq 0. \end{aligned}$$

Above, (i) follows by using rotational symmetry of $\mathbf{z}$, and (ii) follows by using that $\mathbf{z}$ is unimodal so g is decreasing.

□

D Remaining Proofs

Lemma D.1. If $(\mathbf{x}, y) \sim \mathcal{D}$ is from a K-sub-exponential mixture model with mean $\boldsymbol{\mu}$, then for any $\delta > 0$, with probability at least $1 - \delta$, for any $i \in [B]$ and $t \in [T]$,

$$\|\mathbf{x}_i^{(t)}\|^2 \leq 2\|\boldsymbol{\mu}\|^2 + 2dK^2 \log^2(dBT/\delta).$$

Proof. Since the $\mathbf{z}_i^{(t)}$ are K-subexponential, we have that for each component $j \in [d]$, for any $\xi > 0$,

$$\mathbb{P}([\mathbf{z}_i^{(t)}]_j^2 \geq \xi) \leq \exp(-\sqrt{\xi}/K).$$

Since we have the inclusion for $\rho > 0$,

$$\{\|\mathbf{z}_i^{(t)}\|^2 \geq \rho\} \subset \cup_{j=1}^d \{[\mathbf{z}_i]_j^2 > \rho/d\},$$

we have that for any i,

$$\mathbb{P}(\|\mathbf{z}_i^{(t)}\|^2 \geq \rho) \leq d\mathbb{P}([\mathbf{z}_i^{(t)}]_j^2 \geq \rho/d) \leq d\exp\left(-\frac{\sqrt{\rho}}{K\sqrt{d}}\right),$$

where we have used the fact that $\mathbf{z}$ is K-sub-exponential. By taking $\rho = dK^2\log^2(dB/\delta)$, we get that with probability at least $1-\delta$, for any $i \in [B]$ and fixed t,

$$\|\mathbf{z}_i^{(t)}\|^2 \leq dK^2\log^2(dB/\delta).$$

Using Young's inequality, this implies

$$\|\mathbf{x}_i^{(t)}\|^2 \leq 2\|\boldsymbol{\mu}\|^2 + 2dK^2\log^2(dB/\delta).$$

Scaling $\delta \mapsto \delta/T$ and using a union bound completes the proof. □

Lemma D.2. Suppose that we have the recursion

$$\Delta_t^2 \leq (1-\eta/2C_g)\Delta_{t-1}^2 + \frac{\eta\varepsilon}{8C_g} + \frac{2C_d\eta^2}{\sigma^2}, \text{for } t = 1,\ldots,T,$$

where $C_dC_g^2\sigma^2 \geq 1$ and $\Delta_0 \leq 2$.[2] Then, for $\eta = \varepsilon/(16C_dC_g\sigma^2)$ and $T \geq 32C_dC_g^2\sigma^2\varepsilon^{-1}\log(32C_dC_g^2\sigma^2\varepsilon^{-1})$, we have $\Delta_T^2 \leq \varepsilon$.

Proof. We unroll the recursion and use the geometric series formula to get

$$\begin{aligned}
\Delta_T^2 &\leq \left(1-\frac{\eta}{2C_g}\right)^T\Delta_0^2 + \left(\frac{\eta\varepsilon}{8C_g} + \frac{2C_d\eta^2}{\sigma^2}\right)\sum_{i=0}^{T-1}\left(1-\frac{\eta}{2C_g}\right)^i \\
&= \left(1-\frac{\eta}{2C_g}\right)^T\Delta_0^2 + \left(\frac{\eta\varepsilon}{8C_g} + \frac{2C_d\eta^2}{\sigma^2}\right)\frac{1-\left(1-\frac{\eta}{2C_g}\right)^T}{\eta/2C_g} \\
&\leq \left(1-\frac{\eta}{2C_g}\right)^T\Delta_0^2 + \varepsilon/4 + 4C_dC_g\eta/\sigma^2.
\end{aligned}$$

Substituting the value for $\eta = \varepsilon/(16C_dC_g\sigma^2)$, we get

$$\Delta_T^2 \leq \left(1-\frac{\varepsilon}{32C_dC_g^2\sigma^2}\right)^T\Delta_0^2 + \varepsilon/2.$$

Thus, for $T \geq 32C_dC_g^2\sigma^2\varepsilon^{-1}\log(32C_dC_g^2\sigma^2\varepsilon^{-1})$ and using the identity $(1-x)^{x^{-1}\log(1/x)} \leq x$ for $x \in (0,1)$, we get that (using $\Delta_0 \leq 2$)

$$\Delta_T^2 \leq \frac{\Delta_0^2\varepsilon}{32C_dC_g^2\sigma^2} + \varepsilon/2 \leq \varepsilon.$$

□

[2] Note that $C_dC_g^2\sigma^2 \geq 1$ for C_d, C_g as in the proof of Theorem 3.6, and that $\Delta_0 = \|\beta_0 - \bar{\mu}\| \leq \|\beta_0\| + \|\bar{\mu}\| = 2$.