Efficient Active Learning Halfspaces with Tsybakov Noise: A Non-convex Optimization Approach

Yinan Li **Chicheng Zhang**
University of Arizona

Abstract

We study the problem of computationally and label efficient PAC active learning d-dimensional halfspaces with Tsybakov Noise (Tsybakov, 2004) under structured unlabeled data distributions. Inspired by Diakonikolas et al. (2020c), we prove that any approximate first-order stationary point of a smooth nonconvex loss function yields a halfspace with a low excess error guarantee.
In light of the above structural result, we design a nonconvex optimization-based algorithm with a label complexity of $\tilde{O}(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}})$[1], under the assumption that the Tsybakov noise parameter $\alpha \in (\frac{1}{3}, 1]$, which narrows down the gap between the label complexities of the previously known efficient passive or active algorithms (Diakonikolas et al., 2020b; Zhang and Li, 2021) and the information-theoretic lower bound in this setting.

1 INTRODUCTION

Active learning (Settles, 2009) is a practical machine learning paradigm motivated by the expensiveness of label annotation costs and the wide availability of unlabeled data. Consider the binary classification setting, where given an instance space $\mathcal{X}$ and a binary label space $\mathcal{Y} = \{-1, +1\}$ and a data distribution D over $\mathcal{X} \times \mathcal{Y}$, we would like to learn a classifier that accurately predicts the labels of examples drawn from D. As the performance measure of a classifier h, we define its error rate to be $\mathrm{err}(h) := \mathbb{P}_{(x,y)\sim D}(h(x) \neq y)$. Given access to unlabeled examples and the ability to interactively query a labeling oracle (oftentimes a human annotator), an active learning algorithm aims to output a model $\hat{h}$ from a hypothesis class $\mathcal{H}$ that has a low error rate with a small number of label queries. It has been shown both theoretically (e.g. Settles, 2009; Dasgupta, 2005; Balcan et al., 2007; Hanneke, 2011, 2014; Balcan and Long, 2013; Hanneke and Yang, 2015; Zhang and Chaudhuri, 2014) and empirically (e.g. Siddhant and Lipton, 2018; Dor et al., 2020) that, under many learning settings, by utilizing interaction, active learning algorithms can enjoy much better label efficiency compared with conventional supervised learning.

In this work, we study the problem of computationally and label efficient PAC active learning halfspaces (Valiant, 1985) with noise under structured unlabeled data distributions, where the hypothesis class $\mathcal{H} := \left\{h_w(x) := \mathrm{sign}(\langle w, x\rangle) : w \in \mathbb{R}^d\right\}$ is the set of linear classifiers, and D_X, the marginal distribution of D over $\mathcal{X}$, satisfies certain structural assumptions (Diakonikolas et al., 2020c) (see Definition 2 in Section 3). The goal of the learner is to (ϵ, δ)-PAC learn $\mathcal{H}$ and D, i.e. to output a classifier $\hat{h}$ such that with probability at least $1-\delta$, its excess error, $\mathrm{err}(\hat{h}) - \min_{h' \in \mathcal{H}} \mathrm{err}(h')$ is at most ϵ; the total number of label queries the learners makes as a function of ϵ, δ is referred to as its *label complexity*.

In this work, the specific label noise condition we are interested in is the Tsybakov noise condition (TNC) (Mammen and Tsybakov, 1999; Tsybakov, 2004), stated below:

Definition 1 (Tsybakov noise condition)**.** *Given $A > 0$ and $\alpha \in (0, 1]$, a distribution D over $\mathbb{R}^d \times \{-1, +1\}$ is said to satisfy the* (A, α)-Tsybakov noise condition *with respect to halfspace $w^\star \in \mathbb{R}^d$, if for all $t \in [0, \frac{1}{2}]$, $\mathbb{P}_D\left(\frac{1}{2} - \eta(x) \leq t\right) \leq A t^{\frac{\alpha}{1-\alpha}}$, where $\eta(x) := \mathbb{P}_D(y \neq \mathrm{sign}(\langle w^\star, x\rangle) \mid x)$ is the label flipping probability on example x.*

Definition 1 has two important implications on the data distribution D. First, setting $t = 0$, we get that

[1] In the main body of this work, we use $\tilde{O}(\cdot), \tilde{\Theta}(\cdot)$ to hide factors of the form $\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

$\eta(x) \leq \frac{1}{2}$ almost surely, which implies that the halfspace w^* is Bayes optimal with respect to D. Second, the fraction of examples x that has a large conditional label flipping probability ($\frac{1}{2} - \eta(x) \leq t$) is small (at most $At^{\frac{\alpha}{1-\alpha}}$). As A decreases and α increases, the noise assumption on D becomes more benign, and the learning problem becomes easier. Since the initial definition of TNC, the learning theory community has witnessed extensive effort in understanding the necessary and sufficient amount of labels for learning under it, from both statistical and computational perspectives (Hanneke, 2014; Hanneke and Yang, 2015; Balcan et al., 2007; Balcan and Long, 2013; Zhang and Chaudhuri, 2014; Wang and Singh, 2016; Diakonikolas et al., 2020d,b; Zhang and Li, 2021). Specialized to the setting of active learning halfspaces with TNC under structured unlabeled data distributions:

- From a statistical perspective, a line of works (Balcan et al., 2007; Balcan and Long, 2013; Zhang and Chaudhuri, 2014; Wang and Singh, 2016; Huang et al., 2015) propose algorithms that have a label complexity of $\tilde{O}((\frac{1}{\epsilon})^{2-2\alpha})$, which matches information-theoretic lower bounds (Wang and Singh, 2016) in terms of target excess error rate ϵ. However, these algorithms rely on explicit enumeration of classifiers from $\mathcal{H}$ or performing empirical 0-1 loss minimization, which is known to be NP-Hard in general.

- To design a computationally efficient algorithm for active learning halfspaces under Tsybakov noise, a first natural idea is to combine the well-known "margin-based active learning" framework (e.g. Balcan and Long, 2013) with convex surrogate loss minimization. Specifically, we can have an algorithm that iteratively, for each phase k: (1) learns a halfspace w_k based on labeled examples S_k using convex surrogate loss minimization; (2) actively collects a new set of labeled examples S_{k+1} in a region close to the decision boundary of w_k. Although this algorithm design and analysis framework has made some progress in learning halfspaces under Massart noise (Awasthi et al., 2015, 2016), extending it to learning under Tsybakov noise is challenging, in that the Bayes classifier $h_{w^\star}$ can behave arbitrarily poorly (just better than a random guess) in a region with a small probability.

- A recent line of pioneering works aim at designing efficient algorithms for passive learning halfspace with Tsybakov noise (Diakonikolas et al., 2020d,b). Their key insight is that, learning halfspaces can be reduced to the problem of certifying the non-optimality of a candidate halfspace. Using this, Diakonikolas et al. (2020d) developed a quasi-polynomial time learning algorithm with label complexity $d^{O(\frac{1}{\alpha^2}\log^2(\frac{1}{\epsilon}))}$; and subsequent work Diakonikolas et al. (2020b) designed a polynomial time algorithm with label complexity $(\frac{d}{\epsilon})^{O(\frac{1}{\alpha})}$ under well-behaved distributions, and $\mathrm{poly}(d)\cdot(\frac{1}{\epsilon})^{O(\frac{1}{\alpha^2})}$ under log-concave distributions

- The first active halfspace learning algorithm for Tsybakov noise that exhibits nontrivial improvements over passive learning is due to Zhang and Li (2021). Their algorithm, based on a nonstandard application of online learning regret inequalities, iteratively optimizes a proximity measure between the iterates and w^*. When the Tsybakov noise parameter $\alpha \in (\frac{1}{2}, 1]$, their algorithm has a label complexity of $\tilde{O}(d(\frac{1}{\epsilon})^{\frac{2-2\alpha}{2\alpha-1}})$.

In summary, for active learning halfspaces with TNC under structured unlabeled data distributions, there still remains a large gap between the label complexity upper bounds achieved by computationally efficient algorithms and the information-theoretic lower bound $\tilde{\Omega}((\frac{1}{\epsilon})^{2-2\alpha})$.

Our contributions. In this work, we narrow the above gap by providing an efficient active learning algorithm with a label complexity of $\tilde{O}(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}})$, under the assumption that the noise parameter $\alpha \in (\frac{1}{3}, 1]$. In the sample complexity $(\frac{d}{\epsilon})^{O(\frac{1}{\alpha})}$ of the passive algorithm from Diakonikolas et al. (2020b), the constant hidden in the Big-Oh notation in the exponent is not clear, and this drawback is more significant in terms of the dependence on d. On the other hand, our label complexity has a linear dependence on the dimensionality d. Compared to the first and only efficient active algorithm existing in this setting (Zhang and Li, 2021), our algorithm expands the feasibility of the noise parameter α from $(\frac{1}{2}, 1]$ to $(\frac{1}{3}, 1]$; when $\alpha \in [\frac{1}{2}, 0.566), (\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}} < (\frac{1}{\epsilon})^{\frac{2-2\alpha}{2\alpha-1}}$. So our algorithm outperforms Zhang and Li (2021) when $\alpha \in [\frac{1}{3}, 0.566)$. We present the label complexity and computational efficiency of all algorithms in this setting in Table 1.

Our algorithm relies on a few key technical ideas, which we elaborate on below.

Key idea 1: Computationally efficient non-convex optimization for noise tolerance. The work of Diakonikolas et al. (2020c) shows that, under the Massart noise condition, optimizing a carefully-chosen non-convex loss $L_\sigma(w) = \mathbb{E}[\ell_\sigma(w, (x, y))]$ over the noisy labeled data distribution D yield a classifier with low excess error. Importantly, Diakonikolas et al. (2020c) shows that one does not have to

Table 1: A Comparison Of The State-Of-The-Art Label Complexity And Efficiency On Learning Halfspaces With TNC Under Structured Unlabeled Data Distributions

Work	Label complexity upper bound	Passive/Active	Efficient?
Balcan and Long (2013)	$\tilde{O}(d(\frac{1}{\epsilon})^{2-2\alpha})$	Passive	No
Diakonikolas et al. (2020b)	$(\frac{d}{\epsilon})^{O(\frac{1}{\alpha})}$	Passive	Yes
Zhang and Li (2021)	$\tilde{O}(d(\frac{1}{\epsilon})^{\frac{2-2\alpha}{2\alpha-1}})$ for $\alpha \in (\frac{1}{2}, 1]$	Active	Yes
this work	$\tilde{O}(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}})$ for $\alpha \in (\frac{1}{3}, 1]$	Active	Yes

find the global minimum to achieve the above guarantee; instead, finding a first-order stationary point suffices, which admits computationally efficient procedures (e.g. Ghadimi and Lan, 2013). Inspired by this, we show that under Tsybakov noise with $\alpha > \frac{1}{3}$, for the same nonconvex loss function, a qualitatively-similar structural result holds (Lemma 4). This, when combined with standard results on efficient stochastic optimization methods for finding first-order stationary points Ghadimi and Lan (2013), yields a passive learning procedure with sample complexity of $T = O((\frac{1}{\epsilon})^{\frac{8-4\alpha}{3\alpha-1}})$ that can output a classifier that is close to one of $\{w^*, -w^*\}$ with constant probability.

Key idea 2: Label efficient first-order oracle for the non-convex objective. Our second insight is that, the optimization-based learning algorithm outlined above can be made more label-efficient in our active learning setting. At each iteration of the above algorithm, we call the stochastic gradient oracle of the population loss once. A naive implementation of this oracle requires one labeled example per call: drawing one example x from D_X, query the labeling oracle for its label y, and return $\nabla \ell_\sigma(w, (x, y))$, the gradient of the loss of the model on example (x, y). Inspired by Guillory et al. (2009), we design a much more label-efficient implementation of the stochastic gradient oracle; specifically, each call to the oracle queries $O(\sigma) = O(\epsilon^{\frac{2\alpha}{3\alpha-1}}) \ll 1$ labels in expectation. Moreover, the new implementation of the stochastic gradient oracle preserves the bound on the expected squared norm of the stochastic gradient, resulting in the same iteration complexity T. This yields a learning procedure with label complexity of $O(T\sigma) = O((\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}})$ that can output a classifier that is close to one of $\{w^*, -w^*\}$ with constant probability.

Key idea 3: Label-efficient classifier selection. The above active learning procedure is yet to achieve the (ϵ, δ)-PAC learning guarantee, in that: (1) its success probability is constant; (2) if it succeeds, it is possible that its output classifier is close to $-w^*$ as opposed to w^*. To address issue (1) and boost the success probability to $1 - \delta$, we use a repeat-and-validate procedure similar to Ghadimi and Lan (2013) to obtain multiple independent outputs $\{w_s : s \in [S]\}$ one of which is close to $\{w^*, -w^*\}$, call the stochastic gradient oracle to estimate $\|\nabla L_\sigma(w_s)\|$ for $s \in [S]$, and choose $\tilde{w}$ to be the w_s with the smallest gradient estimate. Thanks again to the label efficient first-order oracle, this step has a label complexity of $O(d(\frac{1}{\epsilon})^{\frac{4-2\alpha}{3\alpha-1}})$. To address the issue (2), we observe that under Tsybakov noise, $\tilde{w}$ and $-\tilde{w}$'s error rates differ by a constant; therefore, using a simple 0-1 loss based validation procedure suffices to find a classifier $O(\epsilon)$-close to w^*, which has an excess error of ϵ.

2 RELATED WORK

Statistical complexity for active learning halfspace under Tsybakov noise condition. The statistical complexity for active learning halfspaces under Tsybakov noise condition has been largely characterized over the past two decades (Hanneke, 2011, 2014; Hanneke and Yang, 2015). For general minimax lower bound of active learning under Tsybakov noise not specific to the hypothesis class of halfspaces, (Hanneke, 2014) provides a minimax label complexity lower bound of $\Omega\left(d(\frac{1}{\epsilon})^{2-2\alpha}\right)$ Hanneke and Yang (2015) establishes minimax label complexity upper and lower bounds for general hypothesis class, in terms of the star number and VC dimension. Specific to the class of homogeneous halfspace, when $\alpha \in (0, \frac{1}{2}]$, the minimax label complexity has a lower bound of $\Omega\left(d(\frac{1}{\epsilon})^{2-2\alpha}\right)$; when $\alpha \in [\frac{1}{2}, 1)$, the minimax label complexity has a lower bound of $\Omega\left((\frac{1}{\epsilon})^{2-2\alpha}(d + (\frac{1}{\epsilon})^{2\alpha-1})\right)$. For a more specific setting for active learning halfspaces under well-behaved distributions, (Wang and Singh, 2016) shows a minimax label complexity lower bound of $\Omega\left((\frac{1}{\epsilon})^{2-2\alpha}\right)$.

On the label complexity upper bound side, assuming the unlabeled distribution is isotropic log-concave, Balcan and Long (2013) and Wang and Singh (2016)'s active learning algorithms achieve label complexity of order $\tilde{O}\left(d(\frac{1}{\epsilon})^{2-2\alpha}\right)$. These works use a margin-based active learning framework, which is a celebrated algorithmic idea of inductively learning halfspaces under benign unlabeled distribution as-

sumptions. However, these algorithms suffer from computational intractability, since they perform empirical 0-1 risk minimization, which is known to be computationally hard (Arora et al., 1997).

Efficient passive learning halfspaces under Tsybakov noise condition. In spite of the extensive effort for efficiently learning in the presence of Tsybakov noise condition (Mammen and Tsybakov, 1999; Tsybakov, 2004), it has been an outstanding open problem to obtain an efficient learning algorithm for any natural hypothesis class (e.g., parities) until very recent years , where the first breakthrough is witnessed in passively learning halfspaces under Tsybakov noise condition under well-behaved distributions (Diakonikolas et al., 2020b,d). Those works adopt the principle of "reduction from learning to certifying the non-optimality of a candidate halfspace", and developed quasi-polynomial time certificate algorithm or polynomial time certificate algorithm, resulting in quasi-polynomial time halfspace learning algorithm (with sample complexity $d^{O(\frac{1}{\alpha^2}\log^2(\frac{1}{\epsilon}))}$) or polynomial time algorithm (with sample complexity $(\frac{d}{\epsilon})^{O(\frac{1}{\alpha})}$), respectively.

Efficient active learning halfspaces under Tsybakov noise condition. Efficient active learning under Tsybakov noise condition is conceptually more difficult than efficient passive learning. The difficulty largely lies in the "conflict" between the nature of Tsybakov noise condition - it allows even the Bayes classifier $w^\star$ to have an error rate arbitrarily close to $1/2$ in a region with a small enough probability - and the common analysis technique adopted in active learning. In more detail, the combination of localized sampling and iterative convex surrogate loss minimization technique used in many efficient active learning algorithms is hard to analyze in this setting, as they oftentimes require learning model with a small constant error rate in localized regions, which is hard to establish under Tsybakov noise. Many efficient active learning algorithms (Awasthi et al., 2014, 2015, 2016) adopt the idea of localization to some extent. To overcome this barrier and obtain an efficient active algorithm to learn halfspace, substantially novel algorithmic ideas seem to be necessary.

In this regard, there are even fewer works along the direction of active learning under Tsybakov noise. One notable work is Zhang and Li (2021), where an active learning algorithm is developed for learning halfspaces under $(A,\alpha)-$Tsybakov noise condition for $\alpha \in (\frac{1}{2}, 1]$ and well-behaved distribution, and achieves a label complexity of $\tilde{O}(d(\frac{1}{\epsilon})^{\frac{2-2\alpha}{2\alpha-1}})$. Although their label complexity results are incomparable to ours, algorithmically, Zhang and Li (2021) uses an iterative approach to find near-optimal halfspaces with increasing precision, each iteration using a different objective function; in contrast, our algorithm optimizes a fixed nonconvex objective function using stochastic gradient descent, and is thus conceptually simpler.

Due to space constraints, we defer the discussions of additional related works in Appendix A.

3 PRELIMINARIES

A (homogenous) halfspace, or a linear classifier, is a function $h_w : \mathbb{R}^d \mapsto \{\pm 1\}$ that is defined as $h_w(x) = \text{sign}(\langle w, x\rangle)$, where $w \in \mathbb{R}^d$. In this paper, we consider the standard binary classification setting, where the hypothesis class $\mathcal{H}$ is the set of halfspaces $\{h_w : w \in \mathbb{R}^d\}$. In the sequel, to ease the notation, we frequently use w to represent the halfspace h_w defined by the vector $w \in \mathbb{R}^d$. We denote by D the joint distribution of labeled examples (x, y) supported on $\mathbb{R}^d \times \{\pm 1\}$ and denote by D_X the marginal distribution of D on x. We define the empirical error rate of h on S, $\text{err}_S(h) := \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{1}(h(x) \neq y)$. For $N \in \mathbb{N}_+$, let $[N] := \{1, 2, \ldots, N\}$. Throughout this paper, for $a, b \in \mathbb{R}^d$, we use $\|a\|$ to denote $\|a\|_2$, the ℓ_2 norm of a, and use $\langle a, b\rangle$ to denote the inner product of a and b, and denote by $\theta(a, b) = \arccos\left(\frac{\langle a,b\rangle}{\|a\|\|b\|}\right) \in [0, \pi]$ the angle between them.

Following the distributional assumptions in (Diakonikolas et al., 2020b,d; Zhang and Li, 2021), this work proceeds in developing an efficient algorithm for active learning halfspace under Tsybakov noise condition. We assume D_X, the marginal distribution over the instance space, lies in the family of well-behaved distributions, which generalizes the isotropic log-concave distribution (Balcan and Long, 2013; Awasthi et al., 2014, 2016) and the uniform distribution on the d-dimensional unit sphere (Awasthi et al., 2015; Yan and Zhang, 2017; Wang and Singh, 2016). We formally define well-behaved distributions as follows:

Definition 2 (Well-behaved distributions (Diakonikolas et al., 2020b)). *Fix $L, R, U, \beta > 0$. We say a distribution D_X on $\mathbb{R}^d$ to be $(2, L, R, U, \beta)$ well-behaved, or well-behaved for short, if the following properties are satisfied: for all x randomly drawn from D_X, let x_V be the projected coordinates of x onto any 2-dimensional linear subspace V of $\mathbb{R}^d$, and p_V be the corresponding probability density function on $\mathbb{R}^2$. p_V satisfies,*

1. *$p_V(z) \geq L$, for all z such that $\|z\|_2 \leq R$;*
2. *$p_V(z) \leq U$, for all $z \in \mathbb{R}^2$;*

moreover, for any unit vector w in $\mathbb{R}^d$ and any $t > 0$,

$\mathbb{P}_{D_X}(|\langle w, x\rangle| \geq t) \leq \exp(1 - \frac{t}{\beta})$.

An important property of the objective function in iterative optimization is the smoothness property, which we define below:

Definition 3. *A twice continuous differentiable function F is L-smooth on $\mathcal{D}$, if $\|\nabla^2 F(x)\|_{\text{op}} \leq L$, for all $x \in \mathcal{D}$.*

4 ALGORITHM

Algorithm 1 Active learning halfspaces under TNC

1: **Input:** Target excess error ϵ, failure probability δ
2: $\theta_0 \leftarrow O\left(\frac{1}{\ln^2 \frac{1}{\epsilon}} \frac{\epsilon}{2}\right)$, $\sigma \leftarrow \Theta\left((\frac{1}{A})^{\frac{1-\alpha}{3\alpha-1}} \theta_0^{\frac{2\alpha}{3\alpha-1}}\right)$, $\rho \leftarrow \Theta\left((\frac{1}{A})^{\frac{2(1-\alpha)}{3\alpha-1}} \theta_0^{\frac{2(1-\alpha)}{3\alpha-1}}\right)$, $S \leftarrow \log \frac{6}{\delta}$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad w_s \leftarrow$ ACTIVE-PSGD($N = \tilde{O}(\frac{d}{\sigma^2 \rho^4}), \beta = \tilde{\Theta}(\frac{\rho^2 \sigma^2}{d})$) (see Algorithm 2)
5: **end for**
6: **for** $s = 1, 2, \ldots, S$ **do**
7: $\quad g_{s,1}, \ldots, g_{s,M_1} \leftarrow$ ACTIVE-FO(w_s) (see Algorithm 3)
8: $\quad \bar{g}_s \leftarrow \frac{1}{M_1} \sum_{i=1}^{M_1} g_{s,i}$
9: **end for**
10: $s^* \leftarrow \text{argmin}_{s \in [S]} \|\bar{g}_s\|$
11: $\tilde{w} \leftarrow w_{s^*}$
12: Draw M_2 unlabeled examples from D_X and query their labels labeled samples $\{(x_i, y_i)\}_{i=1}^{M_2}$
13: $\hat{w} \leftarrow \text{argmin}_{w \in \{\pm \tilde{w}\}} \frac{1}{M_2} \sum_{i=1}^{M_2} \mathbb{1}(\text{sign}(\langle w, x_i\rangle) \neq y_i)$
14: **Return:** $\hat{w}$

Our main algorithm (Algorithm 1) consists of three key components, namely (1) iterative non-convex optimization with active label queries (ACTIVE-PSGD, Algorithm 2); (2) label-efficient iterate selection to boost the success probability (lines 3 to 11); (3) label-efficient final iterate selection (lines 12 to 13). We now discuss each component in more detail.

4.1 Efficient non-convex optimization with active label queries (Algorithm 2)

As we will see in the Appendix A, there are several results (Awasthi et al., 2015, 2016; Diakonikolas et al., 2019) in the literature against the one-shot application of convex surrogate loss minimization in efficient learning halfspaces with noise. One possible way to get around this is to instead adopt a non-convex surrogate loss. We denote by $\phi_\sigma(t) := \frac{1}{1+e^{\frac{t}{\sigma}}}$ the softmax loss function, first proposed by Diakonikolas et al. (2020c), which can be viewed as a smooth approximation of 0-1 loss. For a halfspace w, we let $L_\sigma(w) := \mathbb{E}_{(x,y)\sim D} \phi_\sigma\left(y \frac{\langle w, x\rangle}{\|w\|}\right)$ be its normalized expected softmax loss function. Our key observation is that, in the presence of Tsybakov noise, to find a w close to w^*, it suffices to find an approximate first-order stationary point of the softmax loss L_σ. The technique of using the softmax loss in the optimization procedure and proving that an approximate stationary point suffices for the halfspace learning goal is originally developed in (Diakonikolas et al., 2020c), where it provides an efficient passive learning algorithm for learning halfspaces under Massart noise. In this work, we extend this technique to the setting of learning halfspaces under Tsybakov noise. Formally, we prove:

Algorithm 2 ACTIVE-PSGD: Projected SGD for finding a stationary point of L_σ using active learning

Input: number of steps N, step size β
$w_0 \leftarrow e_1$
for $i = 1, 2, \ldots, N$ **do**
$\quad g_i \leftarrow$ ACTIVE-FO(w_{i-1})
$\quad v_i \leftarrow w_{i-1} - \beta g_i$
$\quad w_i \leftarrow \frac{v_i}{\|v_i\|_2}$
end for
Return: w_R, where R is a random variable uniformly distributed over $\{0, \ldots, N-1\}$

Lemma 4. *Let D_X be a well behaved distribution, and D satisfies (A, α)-TNC. Denote by $L_\sigma(w) = \mathbb{E}_D\left[\phi_\sigma\left(y \frac{\langle w,x\rangle}{\|w\|_2}\right)\right]$ where ϕ_σ is softmax loss defined above. Let w be such that $\theta(w, w^*) \in (\theta, \pi - \theta)$, where $\theta \leq \Theta(A)$. Then for $\sigma = \Theta\left(\theta^{\frac{2\alpha}{3\alpha-1}}\right)$, we have that $\|\nabla_w L_\sigma(w)\|_2 \geq \Omega\left(\theta^{\frac{2(1-\alpha)}{3\alpha-1}}\right) := 2\rho$.*

Lemma 4 establishes the connection between 0-1 loss and the ℓ_2 norm of the gradient of a carefully designed non-convex loss function - softmax loss $L_\sigma(w) = \mathbb{E}_D\left[\phi_\sigma\left(y \frac{\langle w,x\rangle}{\|w\|_2}\right)\right]$. To elaborate, we prove that for any unit vector w such that $\theta(w, w^*) \in (\theta, \pi - \theta)$, we have,

$$\|\nabla_w L_\sigma(w)\|_2 \geq \Omega\left(\sigma^{\frac{1-\alpha}{\alpha}} - \frac{\sigma^2}{\theta^2}\right) \tag{1}$$

(see the proof of Lemma 16 in Appendix C); this generalizes Diakonikolas et al. (2020c)'s result from $\alpha = 1$ (Massart noise) to $\alpha \in (\frac{1}{3}, 1]$ (Tsybakov noise). Taking the contrapositive and with a careful choice of the parameters ρ and σ, this implies that if $\|\nabla L_\sigma(w)\| \leq 2\rho$, then either w or $-w$ is at an angle at most $\theta_0 := O\left(\frac{1}{\ln^2 \frac{1}{\epsilon}} \frac{\epsilon}{2}\right)$ from w^*. By Lemma 29, either w or $-w$ has an excess error at most ϵ as desired.

We conjecture that in order to obtain an efficient active learning algorithm with lower label complexity and

works for the full range of $\alpha \in (0, 1]$, a substantially new algorithmic idea would be desired. If $\alpha < 1/3$, the right-hand side of Eq. (1) can be negative, and such lower bound becomes vacuous.

To efficiently find an approximate first-order stationary point of L_σ, we adapt the iterative procedure of randomized stochastic gradient (RSG) in Ghadimi and Lan (2013) to this setting, resulting in Alg. 2. More precisely, ACTIVE-PSGD (Algorithm 2) aims at iteratively obtaining a halfspace w such that $\|\nabla L_\sigma(w)\| \leq \rho$ with a constant probability.

In more detail, ACTIVE-PSGD takes as input the number of steps N and a constant stepsize β. w_1 is initialized randomly on the unit ℓ_2-ball in $\mathbb{R}^d$. In each iteration i, ACTIVE-PSGD calls function ACTIVE-FO (Algorithm 3), which serves as a first-order stochastic gradient oracle for L_σ to obtain g_i, an unbiased estimate of $\nabla L_\sigma(w_i)$, and updates the previous iterate w_{i-1} with the step size β. As we will see, from item 1 of Lemma 9 that the direction of the stochastic gradient estimate g_i is always perpendicular to the previous iterate w_{i-1}, hence all v_i's satisfy $\|v_i\|_2 \geq 1$. The next step (line 6) is to project v_i back to the unit ℓ_2-ball to obtain w_i. Lastly, after N iterations, Algorithm 2 output one iterate from $\{w_i : i \in \{0, \ldots, N-1\}\}$ uniformly at random. We have the following performance guarantee of ACTIVE-PSGD:

Lemma 5. *Let ρ, σ be defined as in line 2 of Algorithm 1. If Algorithm 2 receives inputs $N = \tilde{O}(\frac{d}{\sigma^2\rho^4})$, $\beta = \tilde{\Theta}(\frac{\rho^2\sigma^2}{d})$, then its output w_R of Algorithm 2 satisfies, with probability at least $\frac{1}{2}$,*

$$\|\nabla L_\sigma(w_R)\| \leq \rho$$

Furthermore, during N iterations, with probability at least $1 - \frac{\delta}{6S}$ the total number of label queries is at most $T_1 := \tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right)$

Remark 6. *Arjevani et al. (2022) shows that under the assumption of smooth objective and bounded expected squared norm of the stochastic gradient, RSG achieves the optimal first-order oracle complexity. Thus, we speculate that the iteration complexity in Lemma 5 cannot be improved significantly using some other algorithm.*

The idea of ACTIVE-PSGD bears similarity with the standard iterative optimization method, with some remarkable innovation. The key insight of stochastic gradient descent is that by obtaining an unbiased stochastic gradient, each iteration is making progress toward achieving the optimization goal in expectation. In the passive learning setup, the typical way of implementing the stochastic gradient oracle is to sample (x, y) from the labeled data distribution D.

Algorithm 3 ACTIVE-FO: stochastic gradient oracle for L_σ exploiting active learning

1: **Input:**Unit vector w
2: Sample x from D_X
3: Draw $Z \sim$ Bernoulli$(q(w, x))$, where the query probability $q(w, x) := \sigma\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x \right\rangle)\right|$
4: **if** $Z = 1$ **then**
5: $\quad y \leftarrow$ query the labeling oracle on example x
6: $\quad$ **Return:** $h(w, x, y) := -\frac{1}{\sigma} y \left(\frac{x}{\|w\|_2} - \frac{\langle w, x\rangle w}{\|w\|_2^3}\right)$
7: **else**
8: $\quad$ **Return:** 0
9: **end if**

We show that in our active learning setup, the stochastic gradient oracle can be implemented more label-efficiently by the ACTIVE-FO procedure (Algorithm 3). In ACTIVE-FO, we carefully design a function of query probability $q(w, x) := \sigma\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x \right\rangle)\right|$ - not all unlabeled example x are equally important or equally informative for our optimization purpose. Intuitively, the closer the x lies to the decision boundary of halfspace w, the more informative it is (and as a consequence, we query the label for this x with higher probability), because the current w is less confident in labeling x - this idea coincides with the renowned margin-based method (Balcan et al., 2007; Balcan and Long, 2013; Wang and Singh, 2016; Awasthi et al., 2015, 2016) in the active learning literature.

Whenever ACTIVE-FO is invoked with an input unit vector w, it firstly draws an unlabeled example x from D_X, and computes the label query probability on this x according to $q(w, x)$. Note that $q(w, x)$ is a valid probability, i.e., $0 \leq q(w, x) \leq 1$ for all $w, x \in \mathbb{R}^d$, since $|\phi'_\sigma(t)| \leq \frac{1}{\sigma}$ for all $t \in \mathbb{R}$. Then it draws a Bernoulli random variable Z with success probability $q(w, x)$. If $Z = 1$, then ACTIVE-FO outputs the vector $h(w, x, y) := -\frac{1}{\sigma} y \left(\frac{x}{\|w\|_2} - \frac{\langle w, x\rangle w}{\|w\|_2^3}\right)$, otherwise, it outputs a zero vector.

Remark 7. *We show in item 4 of Lemma 9 that* ACTIVE-FO *is label-efficient; furthermore, although* ACTIVE-FO *only queries labels for a fraction of unlabeled examples x it happens to sample,* ACTIVE-FO *preserves the bound on the expected squared norm of the stochastic gradient (see Claim 31), resulting in the same iteration complexity N as passively querying the labels for all x.*

Remark 8. ACTIVE-FO *is inspired by Guillory et al. (2009), where it provides sampling rules* ***Query***(w, x) *and update rules* ***Update***(w, x, y) *for several commonly used margin-based losses. While this work exhibits some experimental results, it does not provide a theoretical analysis of this query strategy. The la-*

bel efficient implementation of the first order oracle is the first time Guillory et al. (2009)'s algorithmic idea has been formally applied to design provably efficient active learning algorithms.

Despite being simple the oracle behavior at first sight, ACTIVE-FO enjoys several properties that turn out to be essential in guaranteeing the desirable performance in our main algorithm, Algorithm 1. We present Lemma 9 for the delicate properties of ACTIVE-FO.

Lemma 9. *Let g_w be the random output of* ACTIVE-FO(w). *We have, for any unit vector w:*

1. *g_w is perpendicular to w;*
2. *g_w is an unbiased estimator of $\nabla L_\sigma(w) : \mathbb{E}[g_w] = \nabla L_\sigma(w)$;*
3. *$\mathbb{E}\left[\|g_w\|^2\right] \le \tilde{O}(\frac{d}{\sigma})$;*
4. *The expected number of label queries per call to* ACTIVE-FO *is $\tilde{O}(\sigma)$.*

Furthermore, as we will see in the next subsection, ACTIVE-FO is not only utilized in the iterative non-convex optimization procedure ACTIVE-PSGD, but also in the iterate selection, both of which help reduce the label complexity of the overall algorithm.

4.2 Label-efficient iterate selection to boost the success probability (lines 3 to 11)

Recall that ACTIVE-PSGD only guarantees that $\|\nabla L_\sigma(w)\| \le \rho$ with a constant probability. To achieve the (ϵ, δ)-PAC learning goal, lines 3 to 11 in Algorithm 1 boost the above success probability to $1-O(\delta)$.

At a high level, our method follows the classic trick of re-running an algorithm for multiple independent trials and picking the best output. One naive idea to pick the best output, is to sample a set of validation examples from D and pick the w in $\{w_1, -w_1, \ldots, w_S, -w_S\}$ that has the lowest validation error. An application of Hoeffding's inequality shows that, setting the validation sample size to $\tilde{O}\left(\frac{1}{\epsilon^2}\right)$ suffices to find a desired halfspace whose excess error at most ϵ. Together with the labeling cost in the iterative non-convex optimization, it yields a total label complexity of $\tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}} + \frac{1}{\epsilon^2}\right)$, which is substantially suboptimal to our current label complexity $\tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right)$ in Theorem 12, when $\alpha \ge \frac{5}{6}$. Here, we design a specialized procedure that achieves better label efficiency by re-utilizing our label-efficient first-order stochastic gradient oracle ACTIVE-FO.

The idea of conducting the iterate selection by the gradient norm instead of the validation error is largely inspired by the two-phase RSG (2-RSG) in Ghadimi and Lan (2013), where the analysis on the total number of first-order oracle calls is under the assumption of sub-gaussian stochastic gradient. We show that the output of ACTIVE-FO is sub-exponential (Lemma 25) and re-analyze the oracle complexity.

We re-run Algorithm 2 independently for S times, which ensures that the probability that no w in $\{w_s : s \in [S]\}$ has $\|\nabla L_\sigma(w)\| \le \rho$ is $2^{-\Theta(S)}$. The selection step using the first-order stochastic gradient oracle ACTIVE-FO (lines 6 to 11) is done as follows. After we obtain one halfspace candidate w_s in each iteration, we call ACTIVE-FO M_1 times and take the average of all outputs, to obtain a good estimate $\bar{g}_s$ of the gradient of $L_\sigma(w_s)$; therefore, $\|\bar{g}_s\|$ closely approximates $\nabla L_\sigma(w_s)$. After we collect gradient estimates for all S candidate halfspaces, we pick the one with the smallest gradient norm estimate $\|\bar{g}_s\|$. We show that our label-efficient iterate selection procedure (lines 3 to 11) enjoys the following performance guarantee:

Lemma 10. *Let ρ, σ be defined as in line 2 of Algorithm 1. Suppose $w_1, \ldots, w_S$ are such that $\min_i \|\nabla L_\sigma(w_i)\| \le \rho$, then after executing lines 6 to 11 of Algorithm 1, with*

$$M_1 = c\frac{d}{\sigma^2\rho^2}\ln\frac{S}{\delta}$$

for some constant c, with probability at least $1-\delta/6$, $\tilde{w}$ satisfies

$$\|\nabla L_\sigma(\tilde{w})\| \le 2\rho.$$

Furthermore, after M_1 calls to ACTIVE-FO*, with probability at least $1-\frac{\delta}{6S}$, the total number of label queries is at most $T_2 := \tilde{O}(d(\frac{1}{\epsilon})^{\frac{4-2\alpha}{3\alpha-1}})$.*

Lemma 10 shows that, if there exists w in $\{w_s : s \in [S]\}$ such that $\|\nabla L_\sigma(w)\| \le \rho$ (which is true with high probability after we re-run ACTIVE-PSGD independently for S times), then after executing lines 6 to 11, we have, with high probability, $\|\nabla L_\sigma(\tilde{w})\| \le 2\rho$.

To achieve label efficiency in the selection procedure, we prove in Lemma 25 that the stochastic gradients output by ACTIVE-FO are sub-exponential, and we apply a large-deviation bound of vector-valued sub-exponential random variables (Lemma 28, which is Theorem 2.1 in Juditsky and Nemirovski (2008)) to prove the performance guarantee of this iterate selection in Lemma 10. In this way, the labeling cost in the iterate selection step is of lower order than that in the iterative non-convex optimization.

4.3 Label-efficient final iterate selection (lines 12 to 13)

Up to now, combining Lemmas 5 and 10, we have successfully shown, with high probability, either $\tilde{w}$ or $-\tilde{w}$

has an excess error at most ϵ. Our last task is to pick the right one out of the pair. To this end, we draw $M_2 = \tilde{O}(1)$ iid labeled examples from D, and pick the one from $\{\pm\tilde{w}\}$ that has a lower empirical error on this sample set. We have the following lemma on the performance guarantee on the final iterate selection phase:

Lemma 11. *Suppose $\tilde{w}$ satisfies that $\exists w \in \{\pm\tilde{w}\}$, such that $\mathrm{err}(w) - \mathrm{err}(w^*) \leq \epsilon$ with $\epsilon \leq \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$, then after executing lines 12 to 13 of Algorithm 1, where $M_2 = O\left(A^{\frac{2-2\alpha}{\alpha}}\frac{1}{\alpha^2}\ln\frac{1}{\delta}\right)$, we have that with probability at least $1-\delta/3$, $\hat{w}$ satisfies*

$$\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$$

Lemma 11 shows that, if the target error ϵ satisfies $\epsilon \leq \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$ (a constant), then a constant number M_2 of labeled examples suffice to find, with high probability, the one with desired excess error guarantee.

5 PERFORMANCE GUARANTEES

Theorem 12. *Suppose D satisfies (A,α)-Tsybakov noise condition with $\alpha \in (\frac{1}{3}, 1]$ and the marginal distribution D_X is well-behaved. For any $\epsilon \leq \min(\tilde{\Theta}(A), \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}})$, and $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 outputs a halfspace $\hat{w}$, such that $\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$. In addition, its total number of label queries is at most $\tilde{O}\left(dA^{\frac{9-9\alpha}{3\alpha-1}}(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right)$.*

We compare this label complexity to the state-of-the-art sample/label complexities of passive learning (Diakonikolas et al., 2020b) and active learning (Zhang and Li, 2021) existing in the literature. Our work achieves a linear dependency on the dimensionality d, and an explicit exponent on the target error ϵ; however, in the sample complexity $(\frac{d}{\epsilon})^{O(\frac{1}{\alpha})}$ of the passive algorithm from Diakonikolas et al. (2020b), the constant hidden in the Big-Oh notation in the exponent is not clear. Compared to the first and only efficient active algorithm existing in this setting (Zhang and Li, 2021), our algorithm expands the feasibility of the noise parameter α from $(\frac{1}{2}, 1]$ to $(\frac{1}{3}, 1]$; when $\alpha \in [\frac{1}{2}, 0.566), (\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}} < (\frac{1}{\epsilon})^{\frac{2-2\alpha}{2\alpha-1}}$. So our algorithm outperforms Zhang and Li (2021) when $\alpha \in [\frac{1}{3}, 0.566)$.

Proof. Recall that at the beginning of Algorithm 1, we set the parameters $\sigma = \Theta\left(\theta_0^{\frac{2\alpha}{3\alpha-1}}\right)$, $\rho = \Theta\left(\theta_0^{\frac{2(1-\alpha)}{3\alpha-1}}\right)$, where $\theta_0 = O\left(\frac{1}{\ln^2\frac{1}{\epsilon}}\frac{\epsilon}{2}\right)$. Given our assumption that $\epsilon \leq \tilde{\Theta}(A)$, we have $\theta_0 \leq \Theta(A)$. Also, recall from Lemma 5 and Lemma 10 that $T_1 = \tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right)$, $T_2 = \tilde{O}(d(\frac{1}{\epsilon})^{\frac{4-2\alpha}{3\alpha-1}})$. We define the following events of interest,

$$E_1 = \left\{\left(\min_{s\in[S]}\|\nabla L_\sigma(w_s)\| \leq \rho\right) \wedge \left(\text{the total number of label queries at line 4 is at most } S\cdot T_1\right)\right\}$$

$$E_2 = \left\{\left(\min_{s\in[S]}\|\nabla L_\sigma(w_s)\| \leq \rho \implies \|\nabla L_\sigma(\tilde{w})\| \leq 2\rho\right) \wedge \left(\text{the total number of label queries at line 7 is at most } S\cdot T_2\right)\right\}$$

$$E_3 = \left\{\left(\exists w \in \{\pm\tilde{w}\} \text{ s.t. } \mathrm{err}(w) - \mathrm{err}(w^*) \leq \epsilon \implies \mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon\right) \wedge \left(\text{line 12 queries } M_2 \text{ labels}\right)\right\}$$

By Lemma 5, for each $s \in [S]$, with probability at least $\frac{1}{2}$, $\|\nabla L_\sigma(w_s)\| \leq \rho$. Furthermore, during N iterations, with probability at least $1 - \frac{\delta}{6S}$ the total number of label queries is at most T_1. Since Algorithm 2 is executed for $S = \log\frac{6}{\delta}$ times, and each run is independent, we have that with probability at least $1 - \frac{\delta}{6}$, $\min_{s\in[S]}\|\nabla L_\sigma(w_s)\| \leq \rho$. Applying a union bound on the total number of label queries in S runs of Algorithm 2, we have that with probability at least $1 - \frac{\delta}{6}$, the total number of label queries at line 4 is at most $S \cdot T_1$. Applying again a union bound, we have $\mathbb{P}(E_1) \geq 1 - \delta/3$.

By Lemma 10, together with a union bound on the total number of label queries in S iterations of line 7, we have $\mathbb{P}(E_2) \geq 1 - \delta/3$. By Lemma 11, $\mathbb{P}(E_3) \geq 1 - \delta/3$. Define $E = E_1 \cap E_2 \cap E_3$. By union bound, $\mathbb{P}(E) \geq 1 - \delta$. For the rest of the proof, we condition on event E happening.

Since both E_1 and E_2 happen, $\|\nabla L_\sigma(\tilde{w})\| \leq 2\rho$. Taking the contrapositive of Lemma 4, we have that if $\|\nabla L_\sigma(\tilde{w})\| \leq 2\rho$, then $\min\{\theta(\tilde{w}, w^*), \theta(-\tilde{w}, w^*)\} \leq \theta_0$.

Applying Lemma 29 with $\gamma = \frac{\epsilon}{2}$, we have that if a halfspace w satisfies $\theta(w, w^*) \leq \theta_0$, then $\mathbb{P}_{x\sim D_X}(h_w(x) \neq h_{w^*}(x)) \leq \epsilon$, which, in turn, implies such that $\mathrm{err}(w) - \mathrm{err}(w^*) \leq \epsilon$. Hence $\exists w \in \{\pm\tilde{w}\}$ s.t. $\mathrm{err}(w) -$

$\mathrm{err}(w^*) \leq \epsilon$. By the definition of E_3, $\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$. Therefore, we conclude that with probability at least $1-\delta$, the final output $\hat{w}$ of Algorithm 1 satisfies $\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$.

The total label complexity of Algorithm 1 is at most

$$\begin{aligned}
&S \cdot T_1 + S \cdot T_2 + M_2 \\
=&S \cdot \tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right) + S \cdot \tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{4-2\alpha}{3\alpha-1}}\right) + O(\ln\frac{6}{\delta}) \\
=&\tilde{O}\left(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}}\right)
\end{aligned}$$

□

6 CONCLUSIONS AND OPEN PROBLEMS

In this work, we provide a computationally and label efficient active learning algorithm that succeeds in learning a halfspace under (A, α)-Tsybakov noise condition under well-behaved unlabeled distributions. Our algorithm achieves a label complexity of $\tilde{O}(d(\frac{1}{\epsilon})^{\frac{8-6\alpha}{3\alpha-1}})$, under the assumption that the noise parameter $\alpha \in (\frac{1}{3}, 1]$.

While our algorithm narrows down the gap between the label complexities of the previously known passive or active efficient algorithms (Diakonikolas et al., 2020b; Zhang and Li, 2021) and the information-theoretic lower bound, it remains an outstanding open problem to obtain an efficient active learning algorithm that can match the label complexity of inefficient active algorithms $\tilde{O}(d(\frac{1}{\epsilon})^{2-2\alpha})$ or information-theoretic lower bound $\tilde{\Omega}((\frac{1}{\epsilon})^{2-2\alpha})$ for all $\alpha \in (0, 1]$.

References

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

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2): 317–331, 1997.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458, 2014.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190. PMLR, 2015.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pages 152–192. PMLR, 2016.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22: 35–52, 1998.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. *arXiv preprint arXiv:2006.04787*, 2020.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117, 2016.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

Ilias Diakonikolas and Daniel Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, pages 4258–4282. PMLR, 2022.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, pages 4749–4760, 2019.

Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *Advances in Neural Information Processing Systems*, 33:13586–13596, 2020a.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. A polynomial

time algorithm for learning halfspaces with tsybakov noise. *arXiv preprint arXiv:2010.01705*, 2020b.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. *arXiv preprint arXiv:2002.05632*, 2020c.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with tsybakov noise. *arXiv preprint arXiv:2006.06467*, 2020d.

Ilias Diakonikolas, Daniel M Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. *arXiv preprint arXiv:2207.14266*, 2022.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, 2020.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pages 3417–3426. PMLR, 2021.

Claudio Gentile, Zhilei Wang, and Tong Zhang. Achieving minimax rates in pool-based batch active learning. In *International Conference on Machine Learning*, pages 7339–7367. PMLR, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23 (4):2341–2368, 2013.

Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems*, 33:2147–2158, 2020.

Siddharth Gopal. Adaptive sampling for sgd by exploiting side information. In *International Conference on Machine Learning*, pages 364–372. PMLR, 2016.

Andrew Guillory, Erick Chastain, and Jeff Bilmes. Active learning as non-convex optimization. In *Artificial Intelligence and Statistics*, pages 201–208. PMLR, 2009.

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

Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Steve Hanneke. Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3), 2014.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(12):3487–3602, 2015.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Ziwei Ji, Kwangjun Ahn, Pranjal Awasthi, Satyen Kale, and Stefani Karp. Agnostic learnability of halfspaces via logistic loss. In *International Conference on Machine Learning*, pages 10068–10103. PMLR, 2022.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Wolfgang Maass and György Turán. How fast can a threshold gate learn? In *Proceedings of a workshop on Computational learning theory and natural learning systems (vol. 1): constraints and prospects: constraints and prospects*, pages 381–414, 1994.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

Burr Settles. Active learning literature survey. 2009.

Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*, 2018.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Leslie G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 560–566, 1985.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2180–2186, 2016.

Zhilei Wang, Pranjal Awasthi, Christoph Dann, Ayush Sekhari, and Claudio Gentile. Neural active learning with performance guarantees. *Advances in Neu-*

ral Information Processing Systems, 34:7510–7521, 2021.

Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces, 2017.

Chicheng Zhang. Efficient active learning of sparse halfspaces. In *Conference on Learning Theory*, pages 1856–1880. PMLR, 2018.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Chicheng Zhang and Yinan Li. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. *Proceedings of Machine Learning Research vol*, 134:1–2, 2021.

Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *arXiv*, pages arXiv–2002, 2020.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9. PMLR, 2015.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Efficient Active Learning Halfspaces with Tsybakov Noise: A Non-convex Optimization Approach Supplementary Materials

A Additional Related Work

Efficient learning halfspaces under benign noise. Besides Tsybakov noise condition, several other benign noise models have been proposed and studied in the learning theory literature. Among these, the simplest one is Random Classification Noise (RCN) (Angluin and Laird, 1988), where at each x, the label is flipped independently with the same probability. It is known that halfspaces are efficiently learnable under RCN (Blum et al., 1998).

In addition to RCN, several more realistic noise models are developed and studied, with the most distinguished one being the Massart noise condition, where at each unlabeled datapoint x, the label flipping probability is *at most* η. Several eminent works in learning theory literature are dedicated to developing efficient learning halfspace algorithms under Massart noise condition (Awasthi et al., 2015, 2016; Yan and Zhang, 2017; Zhang, 2018; Zhang et al., 2020; Diakonikolas et al., 2019, 2020c; Zhang and Li, 2021). To name a few, Awasthi et al. (2015) is among the earliest works in this thread, where an efficient algorithm is developed under the assumption that the unlabeled distribution is uniform distribution. However, this analysis is subject to the restriction that the Massart noise parameter is such that $\eta < 3 \times 10^{-6}$. Since then, subsequent works have made major improvements in label complexity in efficient learning halfspaces under Massart noise. Zhang et al. (2020) develops an efficient active learning algorithm with a label complexity of $O\left(\frac{d}{(1-2\eta)^4}\operatorname{polylog}(\frac{1}{\epsilon})\right)$, assuming the unlabeled data distribution is a log-concave distribution. Finally, the label complexity gap compared to the information-theoretic result is closed in Zhang and Li (2021), whose algorithm achieves a label complexity of $O\left(\frac{d}{(1-2\eta)^2}\operatorname{polylog}(\frac{1}{\epsilon})\right)$ under the assumption of well-behaved unlabeled distribution.

Besides distribution-specific setting under Massart noise condition, see more in Awasthi et al. (2016); Yan and Zhang (2017); Diakonikolas et al. (2020c), there are recent breakthroughs in distribution-free setting (Diakonikolas et al., 2019; Chen et al., 2020). Specifically, Diakonikolas et al. (2019) provides an efficient algorithm that can improperly learn a halfspace with a misclassification error guarantee $\eta + \epsilon$ in $\operatorname{poly}(d, \frac{1}{\epsilon})$ time. This work addresses a long-standing open problem of whether there exists a distribution-free weak learner in the presence of Massart noise. Chen et al. (2020) strengthens this result by providing an efficient proper halfspace learning algorithm that can achieve the same misclassification error guarantee, with an improved bound on the sample complexity. It also provides a black-box "distillation" procedure that converts any classifier to a proper halfspace without losing prediction accuracy.

A line of work studies selective sampling (e.g. Dekel et al., 2012; Wang et al., 2021; Gentile et al., 2022) under parametric noise models; specifically, they assume $\mathbb{P}[y = 1 \mid x] = \sigma(\langle w^*, x\rangle)$ for some function σ. This assumption is arguably strong, in that it essentially assumes that all examples at the same distance to the Bayes optimal halfspace have the same label-flipping probability. In contrast, our work does not make such assumptions, making the learning problem much more challenging.

Importance weighted sampling for stochastic gradient methods. At a high level, this work adopts the algorithmic idea of importance-weighted sampling for stochastic gradient methods. Gopal (2016) proposes a new mechanism for sampling training instances for stochastic gradient descent methods. Specifically, the sampling weights are proportional to the L2 norm of the gradient. They claim this is the way to minimize the total variance of the descent direction. Zhao and Zhang (2015) also uses importance sampling with weight proportional to the norm of the stochastic gradient, to minimize the variance of the stochastic gradient. Needell et al. (2014) shows

that in SGD for smooth and strongly convex objectives, re-weighting the sampling distribution improves the rate of convergence, where it proposes to use the weight proportional to the smoothness parameter.

Negative results on efficient learning halfspace with noise. Toward designing efficient algorithms in learning halfspaces, there have been several notable trials in understanding the possibility via convex surrogate loss minimization. Unfortunately, it has been found and discussed that such a natural and intuitive approach is unable to achieve the PAC learning guarantee.

Specifically, Diakonikolas et al. (2019) constructs discrete distribution supported on two points and argues that, for any decreasing convex loss function, the minimizer of expected loss has a misclassification error lower bound, even with a margin assumption. The arguments in Awasthi et al. (2015) and Awasthi et al. (2016) assume uniform distribution on the unit $\mathbb{R}^2$ ball, where Awasthi et al. (2015) shows that the excess error of the hinge loss minimizer will not get arbitrarily small even with unlimited sample complexity under Massart noise condition, and further, Awasthi et al. (2016) proves that convex surrogate loss minimization does not work under Massart noise condition, for a family of surrogate losses, including most commonly used loss functions.

Although hardness results have been discovered for convex surrogate loss minimization to learn halfspaces with excess error arbitrarily close to the error of Bayes optimal halfspace opt, such approaches can achieve "approximate" learning halfspaces. Specifically, Frei et al. (2021) shows that under log-concave distribution, convex surrogate loss minimization achieves a population risk of $\tilde{O}(\mathrm{opt}^{\frac{1}{2}})$, and Ji et al. (2022) provides a matching lower bound by constructing a well-behaved distribution where the minimizer of logistic loss achieves a misclassification error of $\Omega(\mathrm{opt}^{\frac{1}{2}})$.

Besides the above algorithm-specific hardness results, algorithm-independent results for learning halfspaces with noise have also been discovered. While in the realizable setting, the hypothesis class of halfspaces is efficiently learnable (Maass and Turán, 1994), with the presence of noise, the learning problem is tremendously more challenging. In the agnostic model, where the error of the Bayes classifier is known, and the corruption can be adversarial, learning halfspaces is known to be computationally hard (Guruswami and Raghavendra, 2009; Daniely, 2016). Recent results (Diakonikolas et al., 2020a; Goel et al., 2020) establish the computational hardness in agnostically learning halfspaces by showing Statistical Query lower bounds of $d^{\mathrm{poly}(1/\epsilon)}$, even under the Gaussian distribution. With Massart noise, Diakonikolas and Kane (2022) shows a distribution free lower bound that no efficient Statistical Query algorithm can achieve an error better than $\Omega(\eta)$. This result is later strengthened in Diakonikolas et al. (2022) where it proves that assuming the subexponential time hardness of the Learning with Errors (LWE) problem, no efficient algorithm can achieve an error better than $\Omega(\eta)$ in the same setting. These hardness results motivate us to define and study benign noise and unlabeled distribution conditions for efficient learning halfspace.

B Additional Notations

Throughout the Appendix Section, we use $\tilde{O}(\cdot), \tilde{\Theta}(\cdot)$ to hide factors of the form $\mathrm{polylog}(d, \frac{1}{\epsilon})$ and $\mathrm{poly}(R, U, L)$.

C Key lemmas

Lemma 13 (Restatement of Lemma 5)**.** *Let the expected loss function $L_\sigma(w) = \mathbb{E}\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|}\right)$. If Algorithm 2 receives inputs $N = \tilde{O}(\frac{d}{\sigma^2\rho^4})$, $\beta = \tilde{\Theta}(\frac{\rho^2\sigma^2}{d})$, then its output w_R is a unit vector and satisfies that, with probability at least $\frac{1}{2}$,*

$$\|\nabla L_\sigma(w_R)\| \le \rho$$

Furthermore, during N iterations, with probability at least $1 - \frac{\delta}{6S}$, the total number of label queries is at most $\tilde{O}(\frac{d}{\sigma\rho^4} + \sqrt{\frac{d}{\sigma^2\rho^4}\ln\frac{6S}{\delta}})$.

Proof. Let $L = \tilde{O}(\frac{1}{\sigma})$ be such that L_σ is L-smooth; let $B^2 = \tilde{O}(\frac{d}{\sigma})$ be such that $\mathbb{E}_{(x,y)\sim D}\left[\|g_w\|^2\right] \le B^2$; the existence of L and B is guaranteed by Lemma 24 and item 3 of Lemma 9.

Define a filtration $\{\mathcal{F}_i\}_{i=0}^N$, where $\mathcal{F}_i$ denotes the σ-field $\sigma(g_1, g_2, \ldots, g_i)$. We use $\mathbb{E}_i[\cdot]$ to denote the conditional expectation with respect to $\mathcal{F}_i$.

Denote by $\mathcal{W} = \left\{w \in \mathbb{R}^d : \|w\| \geq 1\right\}$. Note that $v_i - w_{i-1} = -\beta g_i$. Line 6 of Algorithm 2 ensures that $\|w_i\| = 1$, for all $i = 1, \ldots, N$. Further, by item 1 of Lemma 9, g_i is perpendicular to w_{i-1}, hence $\|v_i\|^2 = \|w_{i-1}\|^2 + \|\beta g_i\|^2 \geq 1$, that is, $v_i \in \mathcal{W}$. For any $t \in [0,1]$, $\|w_{i-1} + t(v_i - w_{i-1})\|^2 = \|w_{i-1}\|^2 + \|t\beta g_i\|^2 \geq 1$. Therefore, the line segment between v_i and w_{i-1} lies in $\mathcal{W}$.

Hence we have for $i = 1, 2, \ldots, N$,

$$\begin{aligned}
L_\sigma(v_i) - L_\sigma(w_{i-1}) &= \int_0^1 \left\langle \nabla L_\sigma\left(w_{i-1} + t(v_i - w_{i-1})\right), (v_i - w_{i-1})\right\rangle \mathrm{d}t \\
&= \left\langle \nabla L_\sigma\left(w_{i-1}\right), (v_i - w_{i-1})\right\rangle + \int_0^1 \left\langle \nabla L_\sigma\left(w_{i-1} + t(v_i - w_{i-1})\right) - \nabla L_\sigma\left(w_{i-1}\right), (v_i - w_{i-1})\right\rangle \mathrm{d}t \\
&\leq -\beta \left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle + \int_0^1 Lt\|v_i - w_{i-1}\|^2 \,\mathrm{d}t \\
&= -\beta \left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle + \frac{\beta^2 L}{2}\|g_i\|^2
\end{aligned}$$

where the first equality is by Newton-Leibniz formula, the inequality is by Cauchy-Schwarz inequality and the following reasoning: by multivariable mean value theorem, $\nabla L_\sigma\left(w_{i-1} + t(v_i - w_{i-1})\right) - \nabla L_\sigma\left(w_{i-1}\right) = Mt(v_i - w_{i-1})$, where M is the Hessian matrix of L_σ evaluated at some point on the line segment between v_i and w_{i-1}. Since the line segment between v_i and w_{i-1} lies in $\mathcal{W}$, together with Lemma 24, we have $\|\nabla L_\sigma\left(w_{i-1} + t(v_i - w_{i-1})\right) - \nabla L_\sigma\left(w_{i-1}\right)\| = \|Mt(v_i - w_{i-1})\| \leq \|M\|_{\mathrm{op}}\|t(v_i - w_{i-1})\| \leq Lt\|v_i - w_{i-1}\|$, for all $i \in [N]$.

Since ϕ_σ is invariant under positive scaling: for any $w \neq 0$, $\alpha > 0$, $\phi_\sigma(\alpha w) = \phi_\sigma(w)$, we have L_σ is invariant under positive scaling as well. Hence $L_\sigma(w_i) = L_\sigma(\frac{v_i}{\|v_i\|_2}) = L_\sigma(v_i)$. Therefore, for all $i = 1, 2, \ldots, N$,

$$L_\sigma(w_i) - L_\sigma(w_{i-1}) \leq -\beta \left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle + \frac{\beta^2 L}{2}\|g_i\|^2 \tag{2}$$

Summing up the above inequalities through $i = 1, \ldots, N$, we have

$$\sum_{i=1}^N \beta \left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle \leq L_\sigma(w_0) - L_\sigma(w_N) + \frac{\beta^2 L}{2}\sum_{i=1}^N \|g_i\|^2 \leq 1 + \frac{\beta^2 L}{2}\sum_{i=1}^N \|g_i\|^2 \tag{3}$$

where the last inequality follows from $0 \leq L_\sigma(w) \leq 1, \forall w \in \mathbb{R}^d$, and we have $L_\sigma(w_0) - L_\sigma(w_n) \leq 1$.

Taking expectation on both sides, and by linearity of expectation, we have:

$$\beta \sum_{i=1}^N \mathbb{E}\left[\left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle\right] \leq 1 + \frac{\beta^2 L}{2}\sum_{i=1}^N \mathbb{E}\left[\|g_i\|^2\right].$$

For the left hand side, applying item 2 of Lemma 9 and the law of iterated expectation, we have that for all $i = 1, \ldots, N$,

$$\mathbb{E}\left[\left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle\right] = \mathbb{E}\left[\mathbb{E}_{i-1}\left\langle \nabla L_\sigma\left(w_{i-1}\right), g_i\right\rangle\right] = \mathbb{E}\left[\|\nabla L_\sigma\left(w_{i-1}\right)\|^2\right]$$

For the right hand side, applying item 3 of Lemma 9 and the law of iterated expectation, we have that for all $i = 1, \ldots, N$,

$$\mathbb{E}\left[\|g_i\|^2\right] = \mathbb{E}\left[\mathbb{E}_{i-1}\|g_i\|^2\right] \leq B^2$$

Therefore, we have

$$\beta\mathbb{E}\left[\sum_{i=1}^N \|\nabla L_\sigma\left(w_{i-1}\right)\|^2\right] \leq 1 + \frac{\beta^2 L}{2}NB^2$$

Note that we are choosing R to be uniformly distributed on $\{0, \ldots, N-1\}$, hence by the law of iterated expectation,

$$\mathbb{E}\left[\|\nabla L_\sigma(w_R)\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\nabla L_\sigma(w_R)\|^2 \mid w_1, \ldots, w_{N-1}\right]\right] = \mathbb{E}\left[\frac{\sum_{i=1}^N \|\nabla L_\sigma\left(w_{i-1}\right)\|^2}{N}\right] \leq \frac{1}{\beta N}\left[1 + \frac{\beta^2 L}{2}NB^2\right]$$

By the definitions of β, N, B and L, the above inequality gives us

$$\mathbb{E}\left[\|\nabla L_\sigma(w_R)\|^2\right] \le \frac{\rho^2}{2}$$

By Markov's inequality, we have

$$\mathbb{P}\left[\|\nabla F(w_R)\|^2 \ge \rho^2\right] \le \frac{1}{2}$$

That is, with probability at least $\frac{1}{2}$,

$$\|\nabla F(w_R)\| \le \rho$$

Finally, by Lemma 18, with probability at least $1 - \frac{\delta}{6S}$, the total number of label queries after N calls to ACTIVE-FO is at most $\tilde{O}(\sigma N + \sqrt{N \ln \frac{6S}{\delta}}) = \tilde{O}(\frac{d}{\sigma\rho^4} + \sqrt{\frac{d}{\sigma^2\rho^4} \ln \frac{6S}{\delta}})$. □

Lemma 14 (Restatement of Lemma 10)**.** *Suppose $w_1, \ldots, w_S$ are such that $\min_i \|\nabla L_\sigma(w_i)\| \le \rho$, then after executing lines 6 to 11 of Algorithm 1, with*

$$M_1 = c\frac{d}{\sigma^2\rho^2} \ln \frac{S}{\delta}$$

for some constant c, with probability at least $1 - \delta/6$, $\tilde{w}$ satisfies

$$\|\nabla L_\sigma(\tilde{w})\| \le 2\rho$$

Furthermore, after M_1 calls to ACTIVE-FO*, with probability at least $1 - \frac{\delta}{6S}$, the total number of label queries is at most $\tilde{O}(\frac{d}{\sigma\rho^2} \ln \frac{S}{\delta} + \sqrt{\frac{d}{\sigma^2\rho^2} \ln \frac{S}{\delta} \ln \frac{6S}{\delta}})$.*

The algorithmic idea underlying lines 6 to 11 of Algorithm 1 for iterate selection is largely inspired by Corollary 2.5 in Ghadimi and Lan (2013). However, here we rely on the sub-exponential-ness of the stochastic gradient outputted by ACTIVE-FO (See Lemma 25), whereas Corollary 2.5 in Ghadimi and Lan (2013) assumes sub-gaussian stochastic gradient. We include the proof for completeness.

Proof. Recall from Algorithm 1 that $\bar{g}_s = \frac{1}{M_1} \sum_{i=1}^{M_1} \|g_{s,i}\|$ and $s^* = \operatorname{argmin}_{s\in[S]} \bar{g}_s$. We have

$$\begin{aligned}
\|\bar{g}_{s^*}\| &= \min_s \|\bar{g}_s\| \\
&= \min_s \|\nabla L_\sigma(w_s) + \bar{g}_s - \nabla L_\sigma(w_s)\| \\
&\le \min_s \left(\|\nabla L_\sigma(w_s)\| + \|\bar{g}_s - \nabla L_\sigma(w_s)\|\right) \\
&\le \min_s \|\nabla L_\sigma(w_s)\| + \max_s \|\bar{g}_s - \nabla L_\sigma(w_s)\| \qquad (4)
\end{aligned}$$

where the first inequality is by the triangle inequality, the second inequality is by the fact that $\min_i(a_i + b_i) \le \min_i a_i + \max_i b_i$, Further,

$$\begin{aligned}
\|\nabla L_\sigma(\tilde{w})\| &= \|\bar{g}_{s^*} + \nabla L_\sigma(\tilde{w}) - \bar{g}_{s^*}\| \\
&\le \|\bar{g}_{s^*}\| + \|\nabla L_\sigma(\tilde{w}) - \bar{g}_{s^*}\| \qquad (5)
\end{aligned}$$

where the inequality is by the triangle inequality.

Denote $\delta_{s,i} := g_{s,i} - \nabla L_\sigma(w_s)$, for $s = 1, \ldots, S, i = 1, \ldots, M_1$. We have $\bar{g}_s - \nabla L_\sigma(w_s) = \frac{1}{M_1} \sum_{i=1}^{M_1} g_{s,i} - \nabla L_\sigma(w_s) = \frac{1}{M_1} \sum_{i=1}^{M_1} \delta_{s,i}$, for all $s = 1, \ldots, S$.

By Lemma 25, for any unit vector w, $\|g_w\|$ is sub-exponential with parameter $K = \tilde{\Theta}(\frac{\sqrt{d}}{\sigma})$. Together with Proposition 2.7.1 in Vershynin (2018) on equivalent characterizations of sub-exponential random variables, we have $\mathbb{E}\left[\exp(\frac{\|g_w\|}{K}) \mid w\right] \le \exp(1)$ for any unit vector w.

Applying Lemma 28, we have for any $s = 1, \ldots, S$ and $\lambda > 0$,

$$\mathbb{P}\left\{\|\sum_{i=1}^{M_1}\delta_{s,i}\| \geq \sqrt{2}(\sqrt{e}+\lambda)\sqrt{M_1}K\right\} \leq 2\exp\left\{-\frac{1}{64}\min\left[\lambda^2, 16\sqrt{M_1}\lambda\right]\right\}$$

Taking $\lambda_0 = \max\left\{8\sqrt{\ln\frac{12S}{\delta}}, \frac{4}{\sqrt{M_1}}\ln\frac{12S}{\delta}\right\}$, we have for any $s = 1, \ldots, S$

$$\mathbb{P}\left\{\|\bar{g}_s - \nabla L_\sigma(w_s)\| \geq \sqrt{2}(\sqrt{e}+\lambda_0)\frac{1}{\sqrt{M_1}}K\right\} = \mathbb{P}\left\{\|\sum_{i=1}^{M_1}\delta_{s,i}\| \geq \sqrt{2}(\sqrt{e}+\lambda_0)\sqrt{M_1}K\right\} \leq \frac{\delta}{6S}$$

Taking a union bound over $s = 1, \ldots, S$, we have with probability at least $1 - \delta/6$,

$$\max_{s=1,\ldots,S}\|\bar{g}_s - \nabla L_\sigma(w_s)\| \leq \sqrt{2}(\sqrt{e}+\lambda_0)\frac{1}{\sqrt{M_1}}K$$

In conjunction with Equations (4) and (5), and recall that $M_1 = c\frac{d}{\sigma^2\rho^2}\ln\frac{S}{\delta}$ for some large enough constant c, we have with probability at least $1 - \delta/6$, $\tilde{w}$ satisfies

$$\begin{aligned}
\|\nabla L_\sigma(\tilde{w})\| \leq& \|g(\tilde{w})\| + \|\nabla L_\sigma(\tilde{w}) - g(\tilde{w})\| \\
\leq& \min_s \|\nabla L_\sigma(w_s)\| + \max_s \|g(w_s) - \nabla L_\sigma(w_s)\| + \|\nabla L_\sigma(\tilde{w}) - g(\tilde{w})\| \\
\leq& \rho + 2\sqrt{2}(\sqrt{e}+\lambda_0)\frac{1}{\sqrt{M_1}}K \\
\leq& \rho + 2\sqrt{2}(\sqrt{e}+8\sqrt{\ln\frac{6S}{\delta}} + \frac{4}{\sqrt{M_1}}\ln\frac{6S}{\delta})\frac{1}{\sqrt{M_1}}K \\
\leq& 2\rho
\end{aligned}$$

Lastly, by Lemma 18, with probability at least $1 - \frac{\delta}{6S}$, the total number of label queries after M_1 calls to ACTIVE-FO is at most $\tilde{O}(\sigma M_1 + \sqrt{M_1\ln\frac{6S}{\delta}}) = \tilde{O}(\frac{d}{\sigma\rho^2}\ln\frac{S}{\delta} + \sqrt{\frac{d}{\sigma^2\rho^2}\ln\frac{S}{\delta}\ln\frac{6S}{\delta}})$. □

Lemma 15 (Restatement of Lemma 11)**.** *Suppose $\tilde{w}$ satisfies that $\exists w \in \{\pm\tilde{w}\}$, such that $\mathrm{err}(w) - \mathrm{err}(w^*) \leq \epsilon$ with $\epsilon \leq \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$, then after executing lines 12 to 13 of Algorithm 1, where $M_2 = O\left((\frac{2}{\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}})^2\ln\frac{6}{\delta}\right)$, we have with probability at least $1 - \delta/3$, $\hat{w}$ satisfies*

$$\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$$

Proof. From Lemma 27, we know that $\mathrm{err}(w^*) \leq \frac{1}{2} - \alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$, hence we have $\exists\bar{w} \in \{\pm\tilde{w}\}$, such that

$$\mathrm{err}(\bar{w}) \leq \mathrm{err}(w^*) + \epsilon \leq \frac{1}{2} - \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$$

For any (x, y) and any $w \in \mathbb{R}^d$, exactly one of $\{\pm w\}$ will label (x, y) correctly. Thus for any $w \in \mathbb{R}^d$, $\mathrm{err}(w) + \mathrm{err}(-w) = 1$, and $\mathrm{err}_S(w) + \mathrm{err}_S(-w) = 1$. So we have

$$\mathrm{err}(-\bar{w}) \geq \frac{1}{2} + \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$$

By Hoeffding's inequality, drawing $M_2 = O\left((\frac{2}{\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}})^2\ln\frac{6}{\delta}\right)$ iid labeled examples from D as the validation set S, we have that with probability at least $1 - \delta/3, |\mathrm{err}(w) - \mathrm{err}_S(w)| \leq \frac{1}{2}\alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}, \forall w \in \{\pm\tilde{w}\}$.

This means that with probability at least $1 - \delta/3$, $\mathrm{err}_S(\bar{w}) < 1/2$, which implies that $\hat{w} = \bar{w}$, hence $\mathrm{err}(\hat{w}) - \mathrm{err}(w^*) \leq \epsilon$. □

Lemma 16 (Restatement of Lemma 4)**.** *Let D_X be a well-behaved distribution, and D satisfies (A,α)-TNC. Recall that $L_\sigma(w) = \mathbb{E}_D\left[\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right)\right]$ where ϕ_σ is softmax loss. Let w be a unit vector such that $\theta(w,w^*) \in (\theta, \pi-\theta)$, where $\theta \le (\frac{1}{4})^{\frac{3\alpha-1}{2(1-\alpha)}}(\frac{128U}{cR^2L})^{\frac{1}{2}} = \Theta(A)$. Then for $\sigma = \Theta\left((\frac{1}{A})^{\frac{1-\alpha}{3\alpha-1}}\theta^{\frac{2\alpha}{3\alpha-1}}\right)$, we have that $\|\nabla_w L_\sigma(w)\|_2 \ge \Omega\left((\frac{1}{A})^{\frac{2(1-\alpha)}{3\alpha-1}}\theta^{\frac{2(1-\alpha)}{3\alpha-1}}\right)$.*

Proof. With foresight, we choose $\sigma = \left(\frac{1}{768U}\cdot R^2L(\frac{RL}{A})^{\frac{1-\alpha}{\alpha}}\right)^{\frac{\alpha}{3\alpha-1}}\theta^{\frac{2\alpha}{3\alpha-1}} = \Theta\left((\frac{1}{A})^{\frac{1-\alpha}{3\alpha-1}}\theta^{\frac{2\alpha}{3\alpha-1}}\right)$. By our assumption that $\theta \le \frac{8A}{RL}\cdot(\frac{1}{4})^{\frac{3\alpha-1}{2(1-\alpha)}}(\frac{768U}{R^2L})^{\frac{1}{2}} = \Theta(A)$,

$$\sigma \le \frac{8A}{RL}(\frac{1}{4})^{\frac{\alpha}{1-\alpha}}. \tag{6}$$

Without loss of generality, suppose $w = (0,1,0,\ldots,0)$ and $w^* = (-\sin\theta, \cos\theta, 0, \ldots, 0)$, we have

$$\begin{aligned}
\|\nabla_w L_\sigma(w)\|_2 &= \left\|\nabla \mathbb{E}_D\left[\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right)\right]\right\|_2 \\
&= \left\|\mathbb{E}_D\left[\phi'_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right)y\left(\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle\, w}{\|w\|_2^3}\right)\right]\right\|_2 \\
&= \left\|\mathbb{E}_x\left[\phi'_\sigma(x_2)(1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)\left(\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle\, w}{\|w\|_2^3}\right)\right]\right\|_2 \\
&\ge \left|\mathbb{E}_x\left[\phi'_\sigma(x_2)(1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)x_1\right]\right|
\end{aligned} \tag{7}$$

where the first equality is by taking the gradient of $L_\sigma(w) = \mathbb{E}_D\left[\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right)\right]$, the second equality is by $\nabla\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right) = \phi'_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|_2}\right)y\left(\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\right)$, the third inequality is by noting that $\frac{\langle w,x\rangle}{\|w\|_2} = x_2$ and $\phi'_\sigma(t) = \phi'_\sigma(-t)$, and $\mathbb{E}[y \mid x] = (1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)$, and the last inequality is because $\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3} = (x_1, 0, x_3, \ldots)$.

Denote by $G := \left\{x \in \mathbb{R}^d : \phi'_\sigma(x_2)(1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)x_1 \ge 0\right\} = \left\{x \in \mathbb{R}^d : \operatorname{sign}(\langle w^*,x\rangle)x_1 \le 0\right\}$, and $G^C = \mathbb{R}^d \setminus G$, then we have,

$$\begin{aligned}
\|\nabla_w L_\sigma(w)\|_2 &\ge \left|\mathbb{E}_x\left[\phi'_\sigma(x_2)(1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)x_1\right]\right| \\
&= \left|\mathbb{E}_x\left[\phi'_\sigma(x_2)(1-2\eta(x))\operatorname{sign}(\langle w^*,x\rangle)x_1\left(\mathbb{1}(x\in G) + \mathbb{1}(x\in G^C)\right)\right]\right| \\
&\ge \mathbb{E}_x\left[\left|\phi'_\sigma(x_2)\right|(1-2\eta(x))|x_1|\mathbb{1}(x\in G)\right] - \mathbb{E}_x\left[\left|\phi'_\sigma(x_2)\right|(1-2\eta(x))|x_1|\mathbb{1}(x\in G^C)\right] \\
&= \mathbb{E}_x\left[\left|\phi'_\sigma(x_2)\right|(1-2\eta(x))|x_1|\right] - 2\mathbb{E}_x\left[\left|\phi'_\sigma(x_2)\right|(1-2\eta(x))|x_1|\mathbb{1}(x\in G^C)\right]
\end{aligned} \tag{8}$$

where the first inequality is from Equation (7), the equalities are because $\mathbb{1}(x\in G) + \mathbb{1}(x\in G^C) = 1$, for all $x \in \mathbb{R}^d$, the last inequality is by the triangle inequality.

We lower bound $\mathbb{E}_x\left[\left|\phi'_\sigma(x_2)\right|(1-2\eta(x))|x_1|\right]$ as follows.

Define $R_1 = \left\{x \in \mathbb{R}^d : x_1 \in [\frac{R}{4}, \frac{R}{2}], x_2 \in [0,\sigma]\right\}$, we lower bound $\mathbb{P}_x(x \in R_1)$ as follows. We project x onto the 2-dimensional subspace V spanned by $\{e_1, e_2\}$; define $\tilde{x}$ to be the coordinate of its projection, and let $\tilde{R}_1$ be the projection of R_1 onto V. Denote by $\tilde{D}_X$ the distribution of $\tilde{x}$, and denote by its probability density function p_V. Since D_X is well-behaved, we have

$$\mathbb{P}_x(x\in R_1) = \mathbb{P}_{\tilde{x}\sim\tilde{D}_X}(\tilde{x}\in\tilde{R}_1) = \int_{\tilde{R}_1} p_V(\tilde{x})d\tilde{x} \ge \frac{R}{4}\sigma L$$

Let $t = 2(\frac{R\sigma L}{8A})^{\frac{1-\alpha}{\alpha}}$; with this choice of t, $\frac{R}{8}\sigma L \ge A(\frac{t}{2})^{\frac{\alpha}{1-\alpha}}$. Also note that by Eq. (6), $t \le \frac{1}{2}$. We obtain the

following,

$$\begin{aligned}
\mathbb{E}_x\left[|\phi'_\sigma(x_2)|\,(1-2\eta(x))|x_1|\right] \geq & \mathbb{E}_x\left[|\phi'_\sigma(x_2)|\,(1-2\eta(x))|x_1|\mathbb{1}(x\in R_1)\right] \\
\geq & \frac{1}{6\sigma}\cdot t\cdot \mathbb{E}\left[\mathbb{1}(1-2\eta(x)\geq t)|x_1|\mathbb{1}(x\in R_1)\right] \\
\geq & \frac{1}{6\sigma}\cdot t\frac{R}{4}\cdot \mathbb{E}\left[\mathbb{1}(1-2\eta(x)\geq t)\mathbb{1}(x\in R_1)\right] \\
\geq & \frac{1}{6\sigma}\cdot t\frac{R}{4}\cdot \left[\mathbb{P}(x\in R_1)-\mathbb{P}(1-2\eta(x)\leq t)\right] \\
\geq & \frac{1}{6\sigma}\cdot t\frac{R}{4}\cdot \left[\mathbb{P}(x\in R_1)-A(\frac{t}{2})^{\frac{\alpha}{1-\alpha}}\right] \\
\geq & \frac{1}{6\sigma}\cdot t\frac{R}{4}\cdot \left[\frac{R}{4}\sigma L-\frac{R}{8}\sigma L\right] \\
= & \frac{1}{192}c\cdot R^2Lt \\
= & \frac{1}{192}c\cdot R^2L\cdot 2(\frac{R\sigma L}{8A})^{\frac{1-\alpha}{\alpha}} \qquad (9)
\end{aligned}$$

where the first inequality is since $R_1\subset\mathbb{R}^d$, the second inequality is by noting that when $|x_2|\leq\sigma$, $|\phi'_\sigma(x_2)|\geq\frac{1}{\sigma}\frac{e}{(1+e)^2}\geq\frac{1}{6\sigma}$. The third is because for all $x\in R_1$, $|x_1|\geq\frac{R}{4}$, the fourth is by basic logic operation, the fifth is by the definition of TNC and $t\leq\frac{1}{2}$. The sixth is by $\mathbb{P}_x(x\in R_1)\geq\frac{R}{4}\sigma L$ and $\frac{R}{8}\sigma L\geq A(\frac{t}{2})^{\frac{\alpha}{1-\alpha}}$. The other inequalities and equalities are all by algebra.

Next, we upper bound $\mathbb{E}_x\left[|\phi'_\sigma(x_2)|\,(1-2\eta(x))|x_1|\mathbb{1}(x\in G^C)\right]$.

Let $f(r\cos\varphi, r\sin\varphi)$ denote the density function after projection on the 2-d subspace spanned by $\{e_1,e_2\}$,

$$\begin{aligned}
\mathbb{E}_x\left[|\phi'_\sigma(x_2)|\,(1-2\eta(x))|x_1|\mathbb{1}(x\in G^C)\right] \leq & \mathbb{E}_x\left[|\phi'_\sigma(x_2)|\,|x_1|\mathbb{1}(x\in G^C)\right] \\
= & \mathbb{E}_x\left[\frac{1}{\sigma}\frac{e^{\frac{|x_2|}{\sigma}}}{(1+e^{\frac{|x_2|}{\sigma}})^2}|x_1|\mathbb{1}(x\in G^C)\right] \\
= & \mathbb{E}_x\left[\frac{1}{\sigma}\frac{e^{-\frac{|x_2|}{\sigma}}}{(1+e^{-\frac{|x_2|}{\sigma}})^2}|x_1|\mathbb{1}(x\in G^C)\right] \\
\leq & \mathbb{E}_x\left[\frac{1}{\sigma}e^{-\frac{|x_2|}{\sigma}}|x_1|\mathbb{1}(x\in G^C)\right] \\
= & \frac{2}{\sigma}\int_0^\infty\int_\theta^{\frac{\pi}{2}} f(r\cos\varphi,r\sin\varphi)r^2\cos\varphi e^{-\frac{r\sin\varphi}{\sigma}}\,\mathrm{d}\,\varphi\,\mathrm{d}\,r \\
\leq & \frac{2}{\sigma}U\int_0^\infty\int_\theta^{\frac{\pi}{2}} r^2\cos\varphi e^{-\frac{r\sin\varphi}{\sigma}}\,\mathrm{d}\,\varphi\,\mathrm{d}\,r \\
= & 2U\frac{\sigma^2}{\tan^2\theta} \\
\leq & 2U\frac{\sigma^2}{\theta^2} \qquad (10)
\end{aligned}$$

where the first inequality is by $\eta(x)\geq 0$, the second equality uses $\phi'_\sigma(t)=-\frac{1}{\sigma}\frac{e^{\frac{t}{\sigma}}}{(1+e^{\frac{t}{\sigma}})^2}$, the third equality is by algebra, the fourth inequality is because $(1+e^{-\frac{|x_2|}{\sigma}})^2\geq 1$, the fifth equality is writing the expectation as the integral in the polar coordinate, the sixth inequality is by the definition of well-behaved distribution: $f(r\cos\varphi,r\sin\varphi)\leq U$, for all $r\geq 0,\varphi\in[0,2\pi]$, the next equality is by algebra, the last inequality is by the elementary fact that $\tan\theta\geq\theta$, for all $\theta\in[0,\pi/2]$.

Therefore, putting together Equations (8), (9) and (10), we obtain

$$\|\nabla_w L_\sigma(w)\|_2\geq\frac{1}{192}\cdot R^2L\cdot 2(\frac{R\sigma L}{8A})^{\frac{1-\alpha}{\alpha}}-4U\frac{\sigma^2}{\theta^2}$$

Recall the choice that $\sigma = \left(\frac{1}{768U} \cdot R^2 L (\frac{RL}{A})^{\frac{1-\alpha}{\alpha}}\right)^{\frac{\alpha}{3\alpha-1}} \theta^{\frac{2\alpha}{3\alpha-1}} = \Theta\left((\frac{1}{A})^{\frac{1-\alpha}{3\alpha-1}} \theta^{\frac{2\alpha}{3\alpha-1}}\right)$, then we obtain

$$\|\nabla_w L_\sigma(w)\|_2 \geq 4U \left(\frac{1}{768U} \cdot R^2 L (\frac{RL}{A})^{\frac{1-\alpha}{\alpha}}\right)^{\frac{2\alpha}{3\alpha-1}} \theta^{\frac{2-2\alpha}{3\alpha-1}} = \Omega\left((\frac{1}{A})^{\frac{2(1-\alpha)}{3\alpha-1}} \theta^{\frac{2(1-\alpha)}{3\alpha-1}}\right)$$

□

Lemma 17 (Restatement of Lemma 9)**.** *Let g_w be the random output of* ACTIVE-FO(w)*. We have, for any unit vector w:*

1. *g_w is perpendicular to w;*
2. *g_w is an unbiased estimator of $\nabla L_\sigma(w) : \mathbb{E}[g_w] = \nabla L_\sigma(w)$;*
3. *$\mathbb{E}\left[\|g_w\|^2\right] \leq \tilde{O}(\frac{d}{\sigma})$;*
4. *The expected number of label queries per call to* ACTIVE-FO *is $\tilde{O}(\sigma)$.*

Proof. Recall that $q(w,x) = \sigma\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|$, and $h(w,x,y) = -\frac{1}{\sigma} y \left(\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\right)$.

We prove the first term as follows. Note that g_w can take the value of 0 or $h(w,x,y)$. If $g_w = 0$, then obviously, $\langle g_w, w\rangle = 0$. If $g_w = h(w,x,y)$, we have

$$\langle h(w,x,y), w\rangle = -\frac{1}{\sigma}\left\langle \frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}, w\right\rangle = -\frac{1}{\sigma}\left(\frac{\langle x,w\rangle}{\|w\|_2} - \frac{\langle x,w\rangle \|w\|^2}{\|w\|_2^3}\right) = 0.$$

Hence we conclude that in both cases, g_w is perpendicular to w.

For the second item, let $w \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}[g_w] =& \mathbb{E}_{(x,y)\sim D, Z\sim \text{Bernoulli}(q(w,x))}[h(w,x,y)Z] \\
=& \mathbb{E}_{(x,y)\sim D}[h(w,x,y)q(w,x)] \\
=& \mathbb{E}_{(x,y)\sim D}\left[-\frac{1}{\sigma} y \left(\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\right)\sigma\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|\right] \\
=& \nabla L_\sigma(w)
\end{aligned}$$

where the first equality is by the definition of ACTIVE-FO, the second equality uses the tower rule, the third equality plugs in the value of $h(w,x,y)$ and $q(w,x)$, the last equality is by the definition of $L_\sigma(w)$.

Now we prove the third item.

1. If $\sigma < \frac{1}{e}$.

Let C below be from Lemma 19. We have for any w such that $\|w\| \geq 1$, for any $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned}
\mathbb{E}\left[\|g_w\|^2\right] =& \mathbb{E}_{(x,y)\sim D, Z\sim \text{Bernoulli}(q(w,x))}\left[\|h(w,x,y)Z\|^2\right] \\
=& \mathbb{E}_{(x,y)\sim D}\left[q(w,x)\|h(w,x,y)\|_2^2\right] \\
=& \mathbb{E}_{(x,y)\sim D}\left[\sigma\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \frac{1}{\sigma^2}\|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\|^2\right] \\
=& \frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\|^2 \\
\leq& \frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \|x\|^2 \\
\leq& \frac{1}{\sigma}\left(\mathbb{E}_{(x,y)\sim D}\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^p\right)^{\frac{1}{p}} \left(\mathbb{E}_{(x,y)\sim D}\|x\|^{2q}\right)^{\frac{1}{q}} \\
\leq& \frac{1}{\sigma}\left(C\frac{1}{\sigma^{p-1}}\ln\frac{1}{\sigma}\right)^{\frac{1}{p}} \left(\Gamma(2q+1)e\beta^{2q}d^q\right)^{\frac{1}{q}} \\
=& \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{1-\frac{1}{p}}\cdot q^2 d) \\
=& \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{q}}\cdot q^2 d)
\end{aligned}$$

where the first equality is by the definition of g_w in Algorithm 3, the second equality uses the tower rule, the third equality is by the definition of $q(w,x)$ and $h(w,x,y)$, the fourth equality is by algebra, the fifth inequality is because for all $x \in \mathbb{R}^d$,

$$\|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\|^2 \leq \|\frac{x}{\|w\|_2}\|^2 \leq \|x\|^2$$

The sixth inequality is by Holder's inequality. The seventh inequality is by Lemmas 19 and 20. The eighth equality is by Lemma 30. The ninth equality uses that $\frac{1}{p} + \frac{1}{q} = 1$.

Choosing $q = \ln\frac{1}{\sigma}$, we have $q > 1$ since $\sigma < \frac{1}{e}$.

we have

$$\mathbb{E}_{(x,y)\sim D}\left[\|g_w\|^2\right] \leq \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{\ln\frac{1}{\sigma}}}\cdot(\ln\frac{1}{\sigma})^2 d) = \frac{1}{\sigma}\tilde{O}(\exp(\ln\frac{1}{\sigma}\cdot\frac{1}{\ln\frac{1}{\sigma}})\cdot(\ln\frac{1}{\sigma})^2 d) = \tilde{O}(\frac{d}{\sigma})$$

where the last two equalities are by algebra.

2. If $\sigma \geq \frac{1}{e}$. Then we can proceed as follows,

$$\begin{aligned}
\mathbb{E}\left[\|g_w\|^2\right] =& \mathbb{E}_{(x,y)\sim D, Z\sim \text{Bernoulli}(q(w,x))}\left[\|h(w,x,y)Z\|^2\right] \\
=& \mathbb{E}_{(x,y)\sim D}\left[q(w,x)\|h(w,x,y)\|_2^2\right] \\
=& \mathbb{E}_{(x,y)\sim D}\left[\sigma\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \frac{1}{\sigma^2}\|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\|^2\right] \\
=& \frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle w}{\|w\|_2^3}\|^2 \\
\leq& \frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left|\phi_\sigma'(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \|x\|^2 \\
\leq& \frac{1}{\sigma^2}\mathbb{E}_{(x,y)\sim D}\|x\|^2 \\
\leq& \frac{1}{\sigma^2}O(d) \\
=& O(d)
\end{aligned}$$

where the sixth inequality is because $|\phi_\sigma'(t)| \leq \frac{1}{\sigma}$, for all $t \in \mathbb{R}$, the seventh inequality uses Lemma 20 with $q = 2$, the eighth inequality uses $\sigma \geq \frac{1}{e}$.

Lastly, we prove the fourth item. We have,

$$\mathbb{E}_{(x,y)\sim D, Z\sim\text{Bernoulli}(q(w,x))}[Z] = \mathbb{E}_{(x,y)\sim D}[q(w,x)] = \sigma\mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right| \le \sigma C \ln\frac{1}{\sigma} = \tilde{O}(\sigma)$$

where the first equality uses the tower rule, the second equality uses the definition of $q(w,x)$, the inequality is by applying Lemma 19 with $p=1$. □

Lemma 18. *With probability at least $1-\delta$, the total number of label queries after T calls to* ACTIVE-FO *is at most $\tilde{O}(\sigma T + \sqrt{T\ln\frac{1}{\delta}})$.*

Proof. Let Z_i be the query indicator the i-th time ACTIVE-FO is called. Define a filtration $\{\mathcal{G}_i\}_{i=0}^N$, where $\mathcal{G}_i$ denotes the σ-field $\sigma(w_1, Z_1, w_2, Z_2, \ldots, w_i, Z_i)$. In this proof, We use $\mathbb{E}_i[\cdot]$ to denote the conditional expectation with respect to $\mathcal{G}_i$.

Let $M_i = \sum_{j=1}^i Z_j - \mathbb{E}[Z_j \mid \mathcal{G}_{j-1}]$. It can be seen that $\{M_i\}_{i=1}^T$ is a martingale and $|M_i - M_{i-1}| \le 1$. Applying Azuma's inequality, we have that with probability at least $1-\delta$,

$$M_T = \sum_{j=1}^T Z_j - \sum_{j=1}^T \mathbb{E}[Z_j \mid \mathcal{G}_{j-1}] \le \sqrt{2T\ln\frac{1}{\delta}}.$$

In addition, by item 4 of Lemma 9, $\mathbb{E}[Z_j \mid \mathcal{G}_{j-1}] = \tilde{O}(\sigma)$ for all $j \in \{1,\ldots,T\}$. Combining with the above inequality, we conclude that

$$\sum_{j=1}^T Z_j \le \sum_{j=1}^T \mathbb{E}[Z_j \mid \mathcal{G}_{j-1}] + \sqrt{2T\ln\frac{1}{\delta}} \le \tilde{O}(\sigma T + \sqrt{T\ln\frac{1}{\delta}}).$$

□

D Auxiliary lemmas

Lemma 19. *Let D_X be a well-behaved distribution, then there exists a constant C, such that for any w such that $\|w\| \ge 1$, and any $p \ge 1$,*

$$\mathbb{E}_{x\sim D_X}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^p \le C\frac{1}{\sigma^{p-1}}\ln\frac{1}{\sigma}$$

Proof. For $k \in \{0\}\cup\mathbb{N}$, denote by $R_k = \left\{x \in \mathbb{R}^d : \left|\left\langle \frac{w}{\|w\|}, x\right\rangle\right| \in (k\sigma, (k+1)\sigma)\right\}$, then we have

$$\begin{aligned}
\mathbb{E}_{x\sim D_X}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^p &= \sum_{k=0}^\infty \mathbb{E}_{x\sim D_X}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^p \mathbb{1}(x\in R_k) \\
&\le \sum_{k=0}^\infty \left(\frac{1}{\sigma}\frac{e^k}{(1+e^k)^2}\right)^p \mathbb{P}(x\in R_k) \\
&\le \frac{1}{\sigma^p} 4\sigma U\beta \ln\frac{2}{\sigma U\beta} \sum_{k=0}^\infty \left(\frac{e^k}{(1+e^k)^2}\right)^p \\
&= \frac{1}{\sigma^{p-1}} 4U\beta \ln\frac{2}{\sigma U\beta} \sum_{k=0}^\infty \left(\frac{e^k}{(1+e^k)^2}\right)^p \\
&\le C\frac{1}{\sigma^{p-1}}\ln\frac{1}{\sigma}
\end{aligned}$$

where the first equality is by the partition of $\mathbb{R}^d$, namely $\mathbb{R}^d = \cup_{k=0}^{\infty} R_k$. The second inequality is because for $k \in \{0\} \cup \mathbb{N}$, if $|t| \in (k\sigma, (k+1)\sigma)$, $|\phi'_\sigma(t)| = \left|\frac{1}{\sigma}\frac{e^{\frac{t}{\sigma}}}{(1+e^{\frac{t}{\sigma}})^2}\right| \leq \frac{1}{\sigma}\frac{e^k}{(1+e^k)^2}$. The third inequality is by Lemma 22. The constant C in the fourth equality exists, because

$$\frac{\frac{e^{k+1}}{(1+e^{k+1})^2}}{\frac{e^k}{(1+e^k)^2}} = e\frac{(1+e^k)^2}{(1+e^{k+1})^2} \leq e\frac{(1+e^0)^2}{(1+e^1)^2} < 1, \forall k = 0, 1, 2, \ldots$$

this means $\left\{\frac{e^k}{(1+e^k)^2} : k = 0, 1, 2, \ldots\right\}$ is decaying faster than a convergent power series, and thus this sequence is also summable. For all $p \geq 1$, $\left(\frac{e^k}{(1+e^k)^2}\right)^p \leq \frac{e^k}{(1+e^k)^2}$, so $\sum_{k=0}^{\infty}\left(\frac{e^k}{(1+e^k)^2}\right)^p \leq \sum_{k=0}^{\infty}\frac{e^k}{(1+e^k)^2}$.

□

Lemma 20. *Let D_X be a well-behaved distribution, then for all $q \geq 2$, we have $\mathbb{E}_{x\sim D_X}\|x\|^q \leq \Gamma(q+1)e\beta^q d^{\frac{q}{2}}$.*

Proof. By Holder's inequality, with $p' = \frac{q}{2}, q' = \frac{q}{q-2}$,

$$\sum_{i=1}^{d} x_i^2 = \sum_{i=1}^{d} x_i^2 \cdot 1 \leq \left(\sum_{i=1}^{d}|x_i|^q\right)^{\frac{2}{q}}\left(\sum_{i=1}^{d}1\right)^{\frac{q-2}{q}} = \left(\sum_{i=1}^{d}|x_i|^q\right)^{\frac{2}{q}} d^{\frac{q-2}{q}}$$

Hence

$$\|x\|^q = \left(\sum_{i=1}^{d} x_i^2\right)^{\frac{q}{2}} \leq \left(\sum_{i=1}^{d}|x_i|^q\right) d^{\frac{q-2}{2}}$$

$$\mathbb{E}_{x\sim D_X}\|x\|^q \leq \mathbb{E}_{x\sim D_X}\left(\sum_{i=1}^{d}|x_i|^q\right) d^{\frac{q-2}{2}} = d^{\frac{q-2}{2}}\sum_{i=1}^{d}\mathbb{E}|x_i|^q$$

For all $i \in [d], q \geq 2$,

$$\mathbb{E}|x_i|^q = \int_0^{\infty}\mathbb{P}(|x_i|^q > t)\,\mathrm{d}\,t = \int_0^{\infty}\mathbb{P}(|x_i| > t^{\frac{1}{q}})\,\mathrm{d}\,t \leq \int_0^{\infty}\exp(1 - t^{\frac{1}{q}}/\beta)\,\mathrm{d}\,t = \Gamma(q+1)e\beta^q$$

where the first equality computes the expectation by integrating over the tail probability, the inequality is by the definition of well-behaved distribution, the last equality can be calculated by the definition of the Gamma function: let $x = t^{\frac{1}{q}}/\beta$, then $t = x^q\beta^q$ and $\mathrm{d}\,t = x^{q-1}q\beta^q\,\mathrm{d}\,x$, and

$$\int_0^{\infty}\exp(1 - t^{\frac{1}{q}}/\beta)\,\mathrm{d}\,t = e\int_0^{\infty}e^{-x}x^{q-1}q\beta^q\,\mathrm{d}\,x = eq\beta^q\Gamma(q) = e\beta^q\Gamma(q+1)$$

Hence $\mathbb{E}_{x\sim D_X}\|x\|^q \leq \Gamma(q+1)e\beta^q d^{\frac{q}{2}}$. □

Claim 21. *Let D_X be a well-behaved distribution, then for any unit vector w and for all $q \geq 1$, $\mathbb{E}_{x\sim D_X}|\langle x, v\rangle|^q \leq \Gamma(q+1)e\beta^q$.*

Proof.

$$\mathbb{E}\,|\langle x, v\rangle|^q = \int_0^{\infty}\mathbb{P}(|\langle x, v\rangle|^q > t)\,\mathrm{d}\,t = \int_0^{\infty}\mathbb{P}(|\langle x, v\rangle| > t^{\frac{1}{q}})\,\mathrm{d}\,t \leq \int_0^{\infty}\exp(1 - t^{\frac{1}{q}}/\beta)\,\mathrm{d}\,t = \Gamma(q+1)e\beta^q$$

where the first equality computes the expectation by integrating over the tail probability, the inequality is by the definition of well-behaved distribution, the other equalities and inequalities are by algebra. □

Lemma 22. *Let D_X be a well-behaved distribution, then for any unit vector w, any $b_0 \geq 0, b > 0$, we have*

$$\mathbb{P}_{x\sim D_X}(b_0 < |\,\langle w, x\rangle\,| < b_0 + b) \leq 4bU\beta\ln\frac{2}{bU\beta}$$

Proof. WLOG, assume $w = (1, 0, \ldots, 0)$, then $b_0 < |\langle w, x\rangle| < b_0 + b$ is equivalent to $b_0 < |x_1| < b_0 + b$. For any $\gamma > 0$, by the definition of well-behaved distribution,

$$\mathbb{P}(b_0 < |x_1| < b_0 + b) \le \mathbb{P}(b_0 < |x_1| < b_0 + b, |x_2| \le \beta \ln \frac{e}{\gamma}) + \mathbb{P}(|x_2| \ge \beta \ln \frac{e}{\gamma}) \le 4bU\beta \ln \frac{e}{\gamma} + \gamma$$

where the first inequality is because $\{x : b_0 < |x_1| < b_0 + b\} \subset \left\{x : b_0 < |x_1| < b_0 + b, |x_2| \le \beta \ln \frac{e}{\gamma}\right\} \cup \left\{x : |x_2| \ge \beta \ln \frac{e}{\gamma}\right\}$, the second inequality is by the definition of well-behaved distribution.

Taking $\gamma = 4bU\beta$, we have

$$\mathbb{P}_{x\sim D_X}(b_0 < |\langle w, x\rangle| < b_0 + b) \le 4bU\beta \left(\ln \frac{e}{4bU\beta} + 1\right) \le 4bU\beta \ln \frac{2}{bU\beta}$$

□

Lemma 23. *$|\phi''_\sigma(t)| \le \frac{1}{\sigma}|\phi'_\sigma(t)|$ for all $t \in \mathbb{R}$.*

Proof. Since $\phi_\sigma(t) = \frac{1}{1+e^{\frac{t}{\sigma}}}$, by direct calculation, $\phi'_\sigma(t) = -\frac{1}{\sigma}\frac{e^{\frac{t}{\sigma}}}{(1+e^{\frac{t}{\sigma}})^2}$, and $\phi''_\sigma(t) = \frac{1}{\sigma^2}\frac{e^{\frac{t}{\sigma}}(e^{\frac{t}{\sigma}}-1)}{(1+e^{\frac{t}{\sigma}})^3}$. Hence for all $t \in \mathbb{R}$, we have

$$|\phi''_\sigma(t)| = \frac{1}{\sigma^2}\frac{e^{\frac{t}{\sigma}}|e^{\frac{t}{\sigma}} - 1|}{(1+e^{\frac{t}{\sigma}})^3} \le \frac{1}{\sigma^2}\frac{e^{\frac{t}{\sigma}}}{(1+e^{\frac{t}{\sigma}})^2} = \frac{1}{\sigma}|\phi'_\sigma(t)|$$

where the inequality is by the elementary fact that $|a - 1| \le a + 1$ for all $a \ge 0$. □

Lemma 24. *For all w such that $\|w\| \ge 1$, $\|\nabla^2_w L_\sigma(w)\| = \tilde{O}(\frac{1}{\sigma})$.*

Proof. Throughout this proof, we denote by $\ell(w, x) := \left\langle \frac{w}{\|w\|}, x\right\rangle$. We continue the calculation of $\nabla^2 L_\sigma(w)$ in Lemma B.2 of Diakonikolas et al. (2020c) and refine the result therein.

$$\begin{aligned}\nabla^2\phi_\sigma(y\ell(w,x)) =& \phi''_\sigma(y\ell(w,x))\left(\frac{xx^\top}{\|w\|^2} - \frac{\langle w,x\rangle}{\|w\|^4}wx^\top - \frac{\langle w,x\rangle}{\|w\|^4}xw^\top + \frac{\langle w,x\rangle^2}{\|w\|^6}ww^\top\right)\\ &+ \phi'_\sigma(y\ell(w,x))y\nabla^2\ell(w,x) \end{aligned} \tag{11}$$

Our goal here is to upper bound $\|\mathbb{E}_{(x,y)\sim D}\nabla^2\phi_\sigma(y\ell(w,x))\|_{\mathrm{op}}$. By triangle inequality, it suffices to upper bound the operator norm of each individual term.

Let $v \in \mathbb{S}^{d-1}$, C below be from Lemma 19.

For any w such that $\|w\| \ge 1$, for any $p, q \ge 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned}\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{xx^\top}{\|w\|^2}\right]v\right\rangle\right| \le& \mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^2}\langle x, v\rangle^2\right]\\ \le& \frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left[|\phi'_\sigma(\ell(w,x))|\langle x, v\rangle^2\right]\\ \le& \frac{1}{\sigma}\left[\mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle\frac{w}{\|w\|}, x\right\rangle)\right|^p\right]^{\frac{1}{p}}\left[\mathbb{E}_{(x,y)\sim D}|\langle x, v\rangle|^{2q}\right]^{\frac{1}{q}}\\ \le& \frac{1}{\sigma}\left(C\frac{1}{\sigma^{p-1}}\ln\frac{1}{\sigma}\right)^{\frac{1}{p}}\left(\Gamma(2q+1)e\beta^{2q}\right)^{\frac{1}{q}}\\ =& \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{1-\frac{1}{p}}\cdot q^2)\\ =& \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{q}}\cdot q^2)\end{aligned}$$

where the first inequality is moving the absolute value inside the expectation, the second inequality uses Lemma 23 and ϕ'_σ is even, as well as $\|w\| \geq 1$, the third inequality is by Holder's inequality. The fourth inequality is by Lemmas 19 and Claim 21. The fifth equality is by Lemma 30. The sixth equality uses that $\frac{1}{p} + \frac{1}{q} = 1$.

Taking $q = \ln \frac{1}{\sigma}$, we have

$$\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{xx^\top}{\|w\|^2}\right] v\right\rangle\right| \leq \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{q}} \cdot q^2) = \tilde{O}(\frac{1}{\sigma})$$

Similarly,

$$\begin{aligned}
&\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{\langle w,x\rangle}{\|w\|^4}wx^\top\right] v\right\rangle\right| \\
\leq&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^4}|\langle w,x\rangle||\langle v,w\rangle||\langle x,v\rangle|\right] \\
\leq&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^3}|\langle w,x\rangle||\langle x,v\rangle|\right] \\
=&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^2}|\left\langle \frac{w}{\|w\|},x\right\rangle||\langle x,v\rangle|\right] \\
\leq&\frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left[|\phi'_\sigma(\ell(w,x))||\left\langle \frac{w}{\|w\|},x\right\rangle||\langle x,v\rangle|\right] \\
\leq&\frac{1}{\sigma}\left[\mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|},x\right\rangle)\right|^p\right]^{\frac{1}{p}}\left[\mathbb{E}_{(x,y)\sim D}\left|\left\langle \frac{w}{\|w\|},x\right\rangle\right|^{2q}\right]^{\frac{1}{2q}}\left[\mathbb{E}_{(x,y)\sim D}|\langle x,v\rangle|^{2q}\right]^{\frac{1}{2q}} \\
\leq&\frac{1}{\sigma}\left(\frac{1}{\sigma^{p-1}}\ln\frac{1}{\sigma}\right)^{\frac{1}{p}}\left(\Gamma(2q+1)e\beta^{2q}\right)^{\frac{1}{2q}}\left(\Gamma(2q+1)e\beta^{2q}\right)^{\frac{1}{2q}} \\
=&\frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{1-\frac{1}{p}} \cdot q^2)
\end{aligned}$$

Again, taking $q = \ln \frac{1}{\sigma}$, we have

$$\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{\langle w,x\rangle}{\|w\|^4}wx^\top\right] v\right\rangle\right| \leq \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{q}} \cdot q^2) = \tilde{O}(\frac{1}{\sigma})$$

The same calculation and the upper bound goes for $\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{\langle w,x\rangle}{\|w\|^4}xw^\top\right] v\right\rangle\right|$.

For the fourth term in Eqn (11),

$$\begin{aligned}
\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{\langle w,x\rangle^2}{\|w\|^6}ww^\top\right] v\right\rangle\right| \leq&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^6}\langle w,x\rangle^2\langle v,w\rangle^2\right] \\
=&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^4}\left\langle \frac{w}{\|w\|},x\right\rangle^2\langle v,w\rangle^2\right] \\
\leq&\mathbb{E}_{(x,y)\sim D}\left[\frac{|\phi''_\sigma(y\ell(w,x))|}{\|w\|^2}\left\langle \frac{w}{\|w\|},x\right\rangle^2\right] \\
\leq&\frac{1}{\sigma}\mathbb{E}_{(x,y)\sim D}\left[|\phi'_\sigma(\ell(w,x))|\left\langle \frac{w}{\|w\|},x\right\rangle^2\right]
\end{aligned}$$

where the first inequality is moving the absolute value inside the expectation, the second inequality is by algebra, the third inequality uses $\left|\left\langle \frac{w}{\|w\|},v\right\rangle\right| \leq 1$, the fourth inequality uses Lemma 23 and ϕ'_σ is even, as well as $\|w\| \geq 1$.

It follows the same upper bound

$$\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\phi''_\sigma(y\ell(w,x))\frac{\langle w,x\rangle^2}{\|w\|^6}ww^\top\right] v\right\rangle\right| \leq \tilde{O}(\frac{1}{\sigma})$$

To upper bound the operator norm of the last term in Eqn (11), note that $\phi'_\sigma(t) \le \frac{1}{\sigma}$, for all $t \in \mathbb{R}$, and

$$\nabla^2 \ell(w,x) = -\frac{xw^T}{\|w\|^3} - \frac{wx^T}{\|w\|^3} - \frac{\langle w,x\rangle}{\|w\|^3} I + 3\frac{\langle w,x\rangle}{\|w\|^5} ww^T$$

Then we can upper bound the operator norm for each individual term,

$$\begin{aligned}
\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\frac{xw^\top}{\|w\|^3}\right] v\right\rangle\right| \le& \mathbb{E}_{(x,y)\sim D}\left[\frac{1}{\|w\|^3}\left|\langle x,v\rangle\langle w,v\rangle\right|\right] \\
\le& \mathbb{E}_{(x,y)\sim D}\left[\frac{1}{\|w\|^2}\left|\langle x,v\rangle\right|\right] \\
\le& \mathbb{E}_{(x,y)\sim D}\left[\left|\langle x,v\rangle\right|\right] \\
=& O(1)
\end{aligned}$$

where the first inequality is moving the absolute value inside the expectation, the second inequality uses $\left|\left\langle \frac{w}{\|w\|}, v\right\rangle\right| \le 1$, the third inequality uses $\|w\| \ge 1$, the last equality applies Claim 21 with $q = 1$.

Similarly,

$$\begin{aligned}
\left|\left\langle v, \mathbb{E}_{(x,y)\sim D}\left[\frac{\langle w,x\rangle}{\|w\|^3} I\right] v\right\rangle\right| \le& \mathbb{E}_{(x,y)\sim D}\left[\frac{1}{\|w\|^2}\left|\left\langle \frac{w}{\|w\|}, x\right\rangle\right|\right] \\
\le& \mathbb{E}_{(x,y)\sim D}\left[\left|\left\langle \frac{w}{\|w\|}, x\right\rangle\right|\right] \\
=& O(1)
\end{aligned}$$

Putting above terms together, we have

$$\|\nabla_w^2 L_\sigma(w)\|_{\mathrm{op}} = \tilde{O}(\frac{1}{\sigma})$$

□

Lemma 25 (Stochastic gradient is sub-exponential)**.** *Let g_w be the random output of* Active-FO(w)*. Then for any unit vector w, $\|g_w\|$ is sub-exponential with parameter $K = \tilde{\Theta}(\frac{\sqrt{d}}{\sigma})$, that is, the tails of $\|g_w\|$ satisfy*

$$\mathbb{P}(\|g_w\| \ge t) \le 2\exp(-\frac{t}{K}), \forall t \ge 0$$

Proof. Assume WLOG that $w = (1, 0, \ldots, 0)$,

$$\begin{aligned}
\mathbb{P}(\|g_w\| \ge t) \le& \mathbb{P}(\|h(w,x,y)\| \ge t) \\
\le& \mathbb{P}(\frac{1}{\sigma}\|x\| \ge t) \\
=& \mathbb{P}(\|x\| \ge \sigma t) \\
\le& \mathbb{P}(\sqrt{d}\|x\|_\infty \ge \sigma t) \\
\le& d \cdot \mathbb{P}(|x_i| \ge \frac{\sigma t}{\sqrt{d}}) \\
\le& d\exp(-\frac{\sigma t}{\sqrt{d}})
\end{aligned}$$

where the first inequality uses the fact that the events $\{\|g_w\| \ge t\} \subseteq \{\|h(w,x,y)\| \ge t\}, \forall t \ge 0$, the second inequality uses that $\|h(w,x,y)\| \le \frac{1}{\sigma}\|x\|, \forall x \in \mathbb{R}^d$, so the events $\{\|h(w,x,y)\| \ge t\} \subseteq \left\{\frac{1}{\sigma}\|x\| \ge t\right\}, \forall t \ge 0$. The third equality is by algebra. The fourth inequality uses $\|x\| \le \sqrt{d}\|x\|_\infty, \forall x \in \mathbb{R}^d$. The fifth inequality uses a union bound on d coordinates. The sixth inequality is by the definition of well-behaved distribution.

Therefore, by Claim 26, $\|g_w\|$ is $\Theta(\frac{\sqrt{d}\ln d}{\sigma})$ sub exponential. □

Claim 26. *If* $\mathbb{P}(|X| \geq t) \leq 2C\exp(-\frac{t}{K})$ *for some constant* $C \geq e^2$*, then* $\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t}{2K\ln C})$.

Proof. Let $t_0 = \frac{2K\ln^2 C}{2\ln C - 1}$.

1. If $t \leq t_0$, $\mathbb{P}(|X| \geq t) \leq 1 \leq 2\exp(-\frac{t_0}{2K\ln C}) \leq 2\exp(-\frac{t}{2K\ln C})$.

   The second inequality is true, because $t_0 = \frac{2K\ln^2 C}{2\ln C - 1} \leq 2K\ln C\ln 2$, using $C \geq e^2$.

2. If $t > t_0$, then $\mathbb{P}(|X| \geq t) \leq 2C\exp(-\frac{t}{K}) = 2\exp(\ln C - \frac{t}{K}) \leq 2\exp(-\frac{t}{2K\ln C})$.

□

Lemma 27. *Suppose* D *satisfies the* (A, α)*-Tsybakov noise condition. Then the Bayes error* $\mathrm{err}(w^*)$ *satisfies* $\mathrm{err}(w^*) \leq \frac{1}{2} - \alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$.

Proof. By the definition of Tsybakov noise, Definition 1, we know $\mathbb{P}(\eta(x) \geq \frac{1}{2} - t) \leq At^{\frac{\alpha}{1-\alpha}}$, for $t \in [0, \frac{1}{2}]$.

Taking $t = \frac{1}{2}$, we can see A, α need to satisfy $1 = \mathbb{P}(\eta(x) \geq 0) \leq A(\frac{1}{2})^{\frac{\alpha}{1-\alpha}}$. We proceed to calculate $\mathrm{err}(w^*)$ as follows.

$$\mathrm{err}(w^*) = \mathbb{E}\eta(x) = \int_0^\infty \mathbb{P}(\eta(x) \geq t)\,\mathrm{d}\,t = \int_0^{\frac{1}{2}} \mathbb{P}(\eta(x) \geq t)\,\mathrm{d}\,t \leq \int_0^{\frac{1}{2}} \min(1, A(\frac{1}{2} - t)^{\frac{\alpha}{1-\alpha}})\,\mathrm{d}\,t$$

where the first equality is by the definition of the Bayes classifier, the second equality is writing the expectation as the integral of the tail probability, the third equality uses $0 \leq \eta(x) \leq \frac{1}{2}, \forall x \in \mathbb{R}^d$, the inequality uses the trivial upper bound 1 of the probability.

Let $t_0 = \frac{1}{2} - (\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$, so $1 = A(\frac{1}{2} - t_0)^{\frac{\alpha}{1-\alpha}}$, and we have,

$$\begin{aligned}
\int_0^{\frac{1}{2}} \min(1, A(\frac{1}{2} - t)^{\frac{\alpha}{1-\alpha}})\,\mathrm{d}\,t =& t_0 + \int_{t_0}^{\frac{1}{2}} A(\frac{1}{2} - t)^{\frac{\alpha}{1-\alpha}}\,\mathrm{d}\,t \\
=& \frac{1}{2} - (\frac{1}{A})^{\frac{1-\alpha}{\alpha}} + (1-\alpha)(\frac{1}{A})^{\frac{1-\alpha}{\alpha}} \\
=& \frac{1}{2} + ((1-\alpha) - 1)(\frac{1}{A})^{\frac{1-\alpha}{\alpha}} \\
=& \frac{1}{2} - \alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}
\end{aligned}$$

Therefore, $\mathrm{err}(w^*) \leq \frac{1}{2} - \alpha(\frac{1}{A})^{\frac{1-\alpha}{\alpha}}$. □

Lemma 28 (Theorem 2.1 in Juditsky and Nemirovski (2008))**.** *Suppose martingale difference* $\{\xi_i\}_{i=1}^\infty$ *satisfies*

$$\forall i \geq 1, \mathbb{E}_{i-1}\left\{\exp\left\{\frac{\|\xi_i\|}{\nu}\right\}\right\} \leq \exp(1) \text{ almost surely}$$

then, for all $N \geq 1$ *and* $\gamma \geq 0$*, one has*

$$\mathbb{P}\left\{\|\sum_{i=1}^N \xi_i\| \geq \sqrt{2}(\sqrt{e} + \gamma)\sqrt{N}\nu\right\} \leq 2\exp\left\{-\frac{1}{64}\min\left[\gamma^2, 16\sqrt{N}\gamma\right]\right\}$$

Lemma 29 (Lemma 26 in Zhang and Li (2021))**.** *If* D *is* $(2, L, R, U, \beta)$*-well behaved, then, we have for any* u, v *in* $\mathbb{R}^d$*, for all* $\gamma > 0$*,* $\mathbb{P}_{x\sim D_X}(h_u(x) \neq h_v(x)) \leq 4U\beta^2\left(\ln\frac{6}{\gamma}\right)^2\theta(u, v) + \gamma$.

Lemma 30. *There exists* $c > 0$*, s.t. for any* $q \geq 1$*,* $\Gamma(2q+1)^{\frac{1}{q}} \leq cq^2$.

Proof. We show that for any $m > 0$, $\Gamma(m+1) \le 3\left(\frac{3m}{5}\right)^m$. The proof is originally from `https://math.stackexchange.com/questions/214422/bounding-the-gamma-function`; for completeness, we reproduce the proof here.

Let $0 < \alpha < 1, f(t) = e^{-\alpha t} t^m$ where $t > 0$, it can be checked (by taking a derivative) that $f(t)$ achieves the maximum at $t = \frac{m}{\alpha}$. Hence for any $m > 0$,

$$\Gamma(m+1) = \int_0^\infty e^{-t} t^m \,\mathrm{d}\, t = \int_0^\infty e^{-\alpha t} t^m e^{-(1-\alpha)t} \,\mathrm{d}\, t \le (\frac{m}{ae})^m \int_0^\infty e^{-(1-\alpha)t} \,\mathrm{d}\, t = (\frac{m}{ae})^m \left(\frac{1}{1-\alpha}\right)$$

where the first equality is by the definition of the Gamma function, and the other equalities and inequalities are by algebra.

Taking $\alpha = \frac{5}{3e}$ and noting that $\frac{1}{1-\frac{5}{3e}} \le 3$, we obtain that for any $m > 0$, $\Gamma(m+1) \le 3\left(\frac{3m}{5}\right)^m$.

Therefore, for any $q \ge 1$, we have

$$\Gamma(2q+1)^{\frac{1}{q}} \le 3^{\frac{1}{q}} \left(\frac{6q}{5}\right)^2 \le 3\left(\frac{6q}{5}\right)^2$$

□

We show in the following claim that ACTIVE-FO preserves the bound on the expected squared norm of the stochastic gradient as passively querying the labels for all x.

Claim 31. *For any unit vector w, $\mathbb{E}_{(x,y)\sim D}\left[\left\|\nabla \phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|}\right)\right\|^2\right] \le \tilde{O}(\frac{d}{\sigma})$.*

Proof. We follow a similar proof idea of item 3 in Lemma 17.

1. If $\sigma < \frac{1}{e}$.

   Let C below be from Lemma 19. We have for any w such that $\|w\| \ge 1$, for any $p, q \ge 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim D}\left[\left\|\nabla \phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|}\right)\right\|^2\right] =& \mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^2 \|\frac{x}{\|w\|_2} - \frac{\langle w,x\rangle\, w}{\|w\|_2^3}\|^2 \\
\le& \mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^2 \|x\|^2 \\
\le& \left(\mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle \frac{w}{\|w\|}, x\right\rangle)\right|^{2p}\right)^{\frac{1}{p}} \left(\mathbb{E}_{(x,y)\sim D}\|x\|^{2q}\right)^{\frac{1}{q}} \\
\le& \left(C\frac{1}{\sigma^{2p-1}} \ln\frac{1}{\sigma}\right)^{\frac{1}{p}} \left(\Gamma(2q+1) e \beta^{2q} d^q\right)^{\frac{1}{q}} \\
=& \tilde{O}((\frac{1}{\sigma})^{2-\frac{1}{p}} \cdot q^2 d) \\
=& \tilde{O}((\frac{1}{\sigma})^{1+\frac{1}{q}} \cdot q^2 d)
\end{aligned}$$

   where the first equality and second inequality are by algebra, the third inequality is by Holder's inequality. The fourth inequality is by Lemmas 19 and 20. The fifth equality is by Lemma 30. The sixth equality uses that $\frac{1}{p} + \frac{1}{q} = 1$.

   Choosing $q = \ln\frac{1}{\sigma}$, we have $q \ge 1$ since $\sigma \le \frac{1}{e}$. we have

$$\mathbb{E}_{(x,y)\sim D}\left[\left\|\nabla \phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|}\right)\right\|^2\right] \le \frac{1}{\sigma}\tilde{O}((\frac{1}{\sigma})^{\frac{1}{\ln\frac{1}{\sigma}}} \cdot (\ln\frac{1}{\sigma})^2 d) = \frac{1}{\sigma}\tilde{O}(\exp(\ln\frac{1}{\sigma} \cdot \frac{1}{\ln\frac{1}{\sigma}}) \cdot (\ln\frac{1}{\sigma})^2 d) = \tilde{O}(\frac{d}{\sigma})$$

   where the last two equalities are by algebra.

2. If $\sigma \geq \frac{1}{e}$. We proceed as follows.

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim D}\left[\left\|\nabla\phi_\sigma\left(y\frac{\langle w,x\rangle}{\|w\|}\right)\right\|^2\right] =& \mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle\frac{w}{\|w\|},x\right\rangle)\right|^2 \|\frac{x}{\|w\|_2}-\frac{\langle w,x\rangle\, w}{\|w\|_2^3}\|^2 \\
\leq& \mathbb{E}_{(x,y)\sim D}\left|\phi'_\sigma(\left\langle\frac{w}{\|w\|},x\right\rangle)\right|^2 \|x\|^2 \\
\leq& \frac{1}{\sigma^2}\mathbb{E}_{(x,y)\sim D}\|x\|^2 \\
\leq& \frac{1}{\sigma^2}O(d) \\
=& O(d)
\end{aligned}$$

where the third inequality is because $|\phi'_\sigma(t)| \leq \frac{1}{\sigma}$, for all $t \in \mathbb{R}$, the fourth inequality uses Lemma 20 with $q = 2$, the fifth inequality uses $\sigma \geq \frac{1}{e}$.

□