Smoothed Agnostic Learning of Halfspaces over the Hypercube

Yiwen Kou
Department of Computer Science, UCLA
Los Angeles, CA, US
evankou@cs.ucla.edu

Raghu Meka
Department of Computer Science, UCLA
Los Angeles, CA, US
raghum@cs.ucla.edu

Abstract

Agnostic learning of Boolean halfspaces is a fundamental problem in computational learning theory, but it is known to be computationally hard even for weak learning. Recent work (Chandrasekaran et al., 2024) proposed smoothed analysis as a way to bypass such hardness, but existing frameworks rely on additive Gaussian perturbations, making them unsuitable for discrete domains. We introduce a new smoothed agnostic learning framework for Boolean inputs, where perturbations are modeled via random bit flips. This defines a natural discrete analogue of smoothed optimality generalizing the Gaussian case. Under strictly subexponential assumptions on the input distribution, we give an efficient algorithm for learning halfspaces in this model, with runtime and sample complexity $\widetilde{O}(n^{\mathrm{poly}(\frac{1}{\sigma\epsilon})})$. Previously, such algorithms were known only with strong structural assumptions for the discrete hypercube—for example, independent coordinates or symmetric distributions. Our result provides the first computationally efficient guarantee for smoothed agnostic learning of halfspaces over the Boolean hypercube, bridging the gap between worst-case intractability and practical learnability in discrete settings.

1 Introduction

Halfspaces, or linear threshold functions (LTFs), are one of the most fundamental concept classes in machine learning. In the realizable setting (Valiant, 1984), they are efficiently learnable by classical algorithms such as the Perceptron (Rosenblatt, 1958; Novikoff, 1963), Winnow (Littlestone, 1987), large-margin methods like Support Vector Machines (Cortes and Vapnik, 1995), or by linear programming. These methods exploit linear separability and can perform well even in the presence of irrelevant features.

In contrast, the agnostic learning framework (Haussler, 1992; Kearns et al., 1992), which allows for arbitrary label noise, poses significant algorithmic challenges. In this setting, the goal is to find a hypothesis that competes with the best in a concept class, without assuming that the data is linearly separable. However, agnostic learning of halfspaces is computationally hard in the worst case: even weak learning—achieving error marginally better than random guessing—is NP-hard under standard complexity assumptions, both in continuous domains (Feldman et al., 2009) and on the Boolean hypercube (Guruswami and Raghavendra, 2009).

To overcome these barriers, several restricted models have been studied. For example, under random classification noise (RCN), halfspaces remain learnable using modified Perceptron algorithms or linear programming (Blum et al., 1996; Cohen, 1997). Under Massart noise, where adversarial flips are bounded in probability, recent work has led to efficient learning algorithms (Awasthi et al., 2015; Diakonikolas et al., 2019, 2020, 2021). Other lines of work exploit structure in the input distribution: Kalai et al. (2008) gave improper agnostic learning algorithms under uniform, spherical,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

or log-concave distributions by approximating halfspaces with low-degree polynomials, extending earlier Fourier-based methods (Linial et al., 1993).

A more recent and promising direction is based on smoothed analysis, which was introduced to explain the practical performance of algorithms that are worst-case hard (Spielman and Teng, 2001). In learning theory, Chandrasekaran et al. (2024) proposed a smoothed agnostic framework in which the learner competes with the best classifier under slight random perturbations of the inputs. This relaxation enables efficient algorithms for learning low-dimensional concepts, even when worst-case learning is intractable. However, their approach is tailored to continuous domains and relies on additive Gaussian noise, and hence is not suitable for discrete domains such as the Boolean hypercube.

Our contribution. We develop a discrete analogue of smoothed agnostic learning for Boolean concept classes over $\{\pm 1\}^n$, where additive Gaussian noise is ill-defined. Instead of perturbing examples in Euclidean space, we introduce bit-flipping noise: each input coordinate is independently flipped with probability σ. This gives rise to a new benchmark for learning that captures robustness to small discrete perturbations, interpolating between classical agnostic learning ($\sigma = 0$) and random guessing ($\sigma = 1/2$).

To formalize this, we begin by recalling the standard agnostic learning objective.

Definition 1.1 (Agnostic Optimality) *Let $\mathcal{X}$ be a domain and $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \{\pm 1\}$. Let $\mathcal{D}$ be a distribution over labeled examples $(\mathbf{x}, y) \in \mathcal{X} \times \{\pm 1\}$. The agnostic error of $\mathcal{F}$ under $\mathcal{D}$ is defined as*

$$\mathrm{opt} = \inf_{f \in \mathcal{F}} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y].$$

We now define a smoothed variant of this benchmark, in which each input is perturbed before evaluation. This definition is general and does not assume any particular structure of the domain or distribution.

Definition 1.2 (Smoothed Optimality) *Let $\mathcal{X}$ be a domain and $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \{\pm 1\}$. Let $\mathcal{D}$ be a distribution over labeled examples $(\mathbf{x}, y) \in \mathcal{X} \times \{\pm 1\}$ and $\mathcal{P}_\sigma(\mathbf{x})$ be a perturbation distribution of $\mathbf{x}$ over $\mathcal{X}$. Define the smoothed agnostic error as:*

$$\mathrm{opt}_\sigma = \inf_{f \in \mathcal{F}} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D};\widetilde{\mathbf{x}}\sim\mathcal{P}_\sigma(\mathbf{x})}[f(\widetilde{\mathbf{x}}) \neq y].$$

In standard agnostic learning, the goal is to compete with opt under $\mathcal{D}$. Following Chandrasekaran et al. (2024), we instead compete with opt_σ in Definition 1.2.

Definition 1.3 (Smoothed Agnostic Learning) *Fix $\epsilon, \sigma > 0$ and $\delta \in (0, 1)$. An algorithm $\mathcal{A}$ learns the class $\mathcal{F}$ in the σ-smoothed agnostic setting if, given i.i.d. samples from $\mathcal{D}$, it outputs a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ such that with probability at least $1 - \delta$:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{opt}_\sigma + \epsilon.$$

As an example, Chandrasekaran et al. (2024) study the case where $\mathcal{X} = \mathbb{R}^n$ and the perturbation distribution is additive Gaussian noise: $\mathcal{P}_\sigma(\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I}_n)$. This formulation relies on the Euclidean structure of $\mathbb{R}^n$ and does not extend to discrete domains. In our setting, we consider $\mathcal{X} = \{\pm 1\}^n$ and define perturbations via random bit flips: $\mathcal{P}_\sigma(\mathbf{x}) = \mathbf{x} \odot \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}_\sigma$ is a product distribution with $z_i = -1$ with probability σ and 1 otherwise. This defines a natural smoothed learning model for the Boolean hypercube that avoids embedding the domain into $\mathbb{R}^n$.

Under this framework, we show that halfspaces over the Boolean cube are efficiently learnable in the smoothed agnostic model under mild distributional assumptions. Our approach extends the classical L_1-polynomial regression framework to this smoothed setting. The key idea is that every Boolean halfspace, when composed with small random bit-flip perturbations, admits a low-degree polynomial approximation under the input distribution. To establish this, we analyze a smoothed version of the halfspace defined via a noise operator (Definition 3.2), and construct approximators using Berry–Esseen-type arguments combined with critical index analysis to handle irregular weight vectors. We obtain the following result:

Theorem 1.4 (Subgaussian-Informal, see also Theorem 4.8) *Let $\mathcal{D}$ be a distribution on $\{\pm 1\}^n \times \{\pm 1\}$ with sub-gaussian $\mathbf{x}$-marginal of variance proxy σ_0^2. There exists an algorithm that learns*

the class of linear threshold functions in the σ-smoothed setting with $N = n^{\mathrm{poly}(\sigma_0/\sigma\epsilon)} \log(1/\delta)$ samples and $\mathrm{poly}(n, N)$ runtime.

This is the first result that establishes efficient smoothed agnostic learning of halfspaces over the Boolean hypercube. While our algorithm is improper, it achieves strong generalization guarantees under natural distributions. Previously, such results for the hypercube were only known under very restricted distributions as discussed below.

2 Related Work

Distributional Assumptions in Halfspace Learning: It is well-understood that agnostically learning halfspaces is intractable in the worst case (Feldman et al., 2009; Guruswami and Raghavendra, 2009), even under relatively benign noise models (Diakonikolas et al., 2022). This has motivated a long line of *distribution-specific* algorithms that guarantee learnability by leveraging assumptions on the data distribution. Early work focused on uniform or product distributions, where powerful Fourier-analytic techniques yield low-degree approximations (Linial et al., 1993; Klivans et al., 2004a; Blais et al., 2010). Under the uniform hypercube distribution, halfspace concepts exhibit strong Fourier concentration and low noise sensitivity (O'Donnell, 2021), enabling efficient learning via low-degree polynomial approximation (Klivans et al., 2004a). This was extended to symmetric distributions in Wimmer (2010) and to arbitrary product distributions in Blais et al. (2010). However, beyond these there are very few general classes of distributions over they hypercube where halfspaces are agnostically learnable.

For continuous distributions, halfspaces were shown to be agnostically learnable under log-concave distributions by Kalai et al. (2008) and this was later extended to intersections and other functions of halfspaces in Kane et al. (2013). Much like the discrete setting, until the recent work of Chandrasekaran et al. (2024), most positive results required strong structural assumptions on the marginal distribution of the examples. This work introduced a new smoothed agnostic model which led to several new results for learning halfspaces and functions of halfspaces for a much broader class of distributions (e.g., sub-gaussian or sub-exponential densities). Our work continues this progression to very general distributions, but focuses on the Boolean domain and shows that only mild tail bounds (strictly sub-exponential) suffice for efficient learning in the smoothed setting.

Noise Models and Smoothed Analysis: In parallel to distributional assumptions on X, a complementary line of work has tackled label noise models and smoothed analysis. The classical noise models include random classification noise (RCN), where each label is independently flipped with some probability. Blum et al. (1996) gave the first polynomial-time algorithm for learning a halfspace under random classification noise, exploiting the fact that a halfspace's margin makes it relatively robust to independent label flips. A stronger noise model is the Massart noise model, which bounds the adversary by a flipping probability $\eta < 1/2$ on each example. Diakonikolas et al. (2020) gave an efficient algorithm for learning halfspaces with Massart noise over log-concave distributions. On the other hand, with adversarial (malicious) noise, learning halfspaces requires additional assumptions. Klivans et al. (2009) designed efficient algorithms for origin-centered halfspaces under malicious noise by assuming isotropic log-concave distribution and small noise rate. In smoothed analysis of learning, one assumes that either the data (Blum and Dunagan, 2002; Kane et al., 2013) or the target concept (Chandrasekaran et al., 2024) is randomly perturbed, so that pathological arrangements are avoided. Chandrasekaran et al. (2024) introduced a smoothed agnostic PAC model in $\mathbb{R}^d$ where the learner competes against the best classifier that is robust to slight Gaussian perturbations of examples. Our work can be seen as a Boolean analogue of this idea: rather than perturbing continuous inputs, we require the optimal halfspace to be stable under small random label flips.

3 Preliminaries

We review relevant definitions from Boolean function analysis that will allow us to define a discrete smoothing operator and justify using it in place of the original linear threshold function. We use definitions from the analysis of Boolean functions over product spaces, following the framework of Mossel et al. (2005). Let $(\Omega_1, \mu_1), \ldots, (\Omega_n, \mu_n)$ be finite probability spaces and let (Ω, μ) denote their product. In our setting, we take $\Omega_i = \{\pm 1\}$ and define μ to be the product distribution $\mathcal{N}_\sigma$, where each coordinate is 1 with probability $1 - \sigma$ and -1 with probability σ, independently.

Definition 3.1 (ρ-noisy copy) *Given $\mathbf{x} \in \Omega$ and $\rho \in [0,1]$, a ρ-noisy copy of $\mathbf{x}$ is a random vector $\mathbf{y} \sim \mathcal{N}_\rho(\mathbf{x})$, where each coordinate y_i is independently set to x_i with probability ρ and to an independent draw from μ_i with probability $1-\rho$.*

Definition 3.2 (Noise operator T_ρ) *For any function $f : \Omega \to \mathbb{R}$ and $\rho \in [0,1]$, the noise operator T_ρ is defined as*

$$(T_\rho f)(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_\rho(\mathbf{x})}[f(\mathbf{y})].$$

This definition generalizes the Bonami–Beckner operator (Kahn et al., 1988) when μ is the uniform distribution on the hypercube. Intuitively, $T_\rho f$ is a smoothed version of f, computed by averaging f over a neighborhood of $\mathbf{x}$ with geometric decay controlled by ρ. In particular, $T_1 f = f$, and as ρ decreases from 1, $T_\rho f$ suppresses high-frequency components of f. This operator will be used as our main tool for constructing smoothed approximations to Boolean threshold functions.

Definition 3.3 (Noise stability and noise sensitivity) *For any $f : \Omega \to \mathbb{R}$, the noise stability at parameter ρ is defined as*

$$S_\rho(f) = \langle f, T_\rho f\rangle_\mu.$$

If $f : \Omega \to \{\pm 1\}$, the noise sensitivity at parameter $\delta \in [0,1]$ is given by

$$NS_\delta(f) = \frac{1}{2} - \frac{1}{2}S_{1-\delta}(f) = \mathbb{P}_{\mathbf{x}\sim\mu,\mathbf{y}\sim\mathcal{N}_{1-\delta}(\mathbf{x})}[f(\mathbf{x}) \neq f(\mathbf{y})].$$

Equivalently, $NS_\delta(f) = \Pr_{x,y}[f(x) \neq f(y)]$ where x and y have Hamming correlation $1-2\delta$. This quantity captures the robustness of f to small input perturbations.

It is well-known that natural Boolean functions with low total influence or low-degree Fourier concentration exhibit low noise sensitivity. In particular, linear threshold functions are noise-stable under both uniform (Peres, 2004) and general product distributions (Blais et al., 2010). The following lemma bounds the noise sensitivity of halfspaces over arbitrary product spaces.

Lemma 3.4 (Theorem 3.2 in Blais et al. (2010)) *Let $f : \Omega \to \{\pm 1\}$ is a linear threshold function, where the domain $\Omega = \Omega_1 \times \cdots \times \Omega_n$ has the product distribution $\mu = \mu_1 \times \cdots \times \mu_n$. Then $NS_\delta(f) \leq \frac{5}{4}\sqrt{\delta}$.*

This bound implies that for $\rho = 1-\delta$ close to 1, the smoothed function $T_\rho f$ closely approximates the original threshold function f. This justifies our strategy of working with $T_\rho f$ instead of f in the smoothed learning setting: any learner that performs well on $T_\rho f$ will, up to a small error, also succeed on f.

Notation. We use small boldface characters for vectors and capital bold characters for matrices. We use $[d]$ to denote the set $\{1, 2, \cdots, d\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, x_i denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^d x_i^2}$ the ℓ_2 norm of $\mathbf{x}$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we use $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^d x_i y_i$ as the inner product between them and $\mathbf{x} \odot \mathbf{y} = (x_1 y_1, \cdots, x_d y_d)$ as the Hadamard product between them. We use $\mathbf{1}\{\mathcal{E}\}$ to be the indicator function of some event $\mathcal{E}$. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_\mathbf{x}$ to be the marginal distribution of $\mathbf{x}$.

4 Technical Overview

In this section, we outline the main steps of our analysis. Our approach follows a reduction-based strategy: we reduce smoothed agnostic learning of Boolean halfspaces to the problem of approximating a smoothed halfspace by a low-degree polynomial, which can then be learned via L_1 regression (Section 4.1). We begin by replacing the original target $f_\mathbf{x}(\mathbf{z}) = f(\mathbf{x} \odot \mathbf{z})$ with a smoothed surrogate $T_{1-\rho} f_\mathbf{x}(\mathbf{z})$ (Definition 3.2), facilitating approximation by low-degree polynomials (Section 4.2).

To handle the biased distribution arising from noise perturbation, we introduce a rerandomization and conditioning trick that rewrites each bit as a mixture involving uniform random variables. This allows us to express the smoothed function as a conditional expectation over uniformly random inputs, making it amenable to quantitative central-limit theorems (Berry–Esseen estimates; Section 4.3). We then use a case analysis facilitated by a decomposition of the weight vector (Section 4.4):

1. If a small number of large coordinates (the "head") dominate, the halfspace's output is primarily determined by those coordinates, and we can approximate the function directly.
2. Otherwise, the remaining "tail" is *regular*, and we apply the Berry–Esseen theorem to approximate the Boolean sum by a Gaussian. This reduces the problem to the continuous setting, where we leverage Gaussian-based techniques (the density ratio method from Chandrasekaran et al. (2024)) to construct low-degree polynomial approximations.

Together, these ingredients yield an efficient smoothed learner for Boolean halfspaces under strictly sub-exponential input distributions.

4.1 High-Level Approach via L_1egression

Our starting point is the L_1-polynomial regression method for agnostic learning. In particular, Kalai et al. (2008) established a powerful reduction from agnostic learnability to low-degree polynomial approximation.

Algorithm 1 L_1 Polynomial Regression Algorithm

Input: Sample $S = \{(x^1, y^1), \ldots, (x^N, y^N)\}$, degree bound d

1: Find polynomial p of degree $\leq d$ to minimize

$$\frac{1}{N}\sum_{j=1}^{N} |p(x^j) - y^j|.$$

(This can be done by expanding examples to include all monomials of degree $\leq d$ and then performing L_1 linear regression.)

2: Output hypothesis $h(x) = \mathrm{sign}(p(x) - t)$, where $t \in [-1, 1]$ is chosen to minimize the classification error on S.

Theorem 4.1 (Theorem 5 in Kalai et al. (2008)) *Suppose* $\min_{\deg(p)\leq d} \mathbb{E}_{\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{x}) - c(\mathbf{x})|] \leq \epsilon$ *for some degree* d *and any* c *in the concept class* $\mathcal{C}$*. Then, for* h *output by the degree-*d L_1 *polynomial regression algorithm with* $N = \mathrm{poly}(n^d/\epsilon)$ *examples,* $\mathbb{E}_{S\sim\mathcal{D}^N}[\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]] \leq \mathrm{opt} + \epsilon$*, where* $\mathrm{opt} = \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y]$*. If we repeat the algorithm* $r = O(\log(1/\delta)/\epsilon)$ *times with fresh examples each, and let* h *be the hypothesis with lowest error on an independent test set of size* $O(\log(1/\delta)/\epsilon^2)$*, then with probability at least* $1 - \delta$*,* $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{opt} + \epsilon$*.*

Theorem 4.1 says that if the target function f can be approximated in L_1 by a low-degree polynomial p with error at most ϵ, then one can efficiently learn f to misclassification error $\mathrm{opt} + \epsilon$, where opt is the Bayes-optimal error rate under distribution $\mathcal{D}$. Once such a polynomial is shown to exist, Theorem 4.1 implies a computationally efficient learning algorithm with sample complexity $N = \mathrm{poly}(n^d/\epsilon)\log(1/\delta)$.

4.2 Smoothed Learning as Non-Worst-Case Approximation

The challenge is that an arbitrary halfspace $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle - \theta)$ might not be well-approximated by any low-degree polynomial over worst-case input distributions. Following Chandrasekaran et al. (2024), we view smoothed learning as a form of non-worst-case approximation. In this smoothed agnostic setting, the learner's "effective" target concept is the mapping $(\mathbf{x}, \mathbf{z}) \mapsto f(\mathbf{x} \odot \mathbf{z})$, where $\mathbf{z} \in \{\pm 1\}^n$ is a random noise vector independent of $\mathbf{x}$ with σ close to 0 meaning only a tiny fraction of bits are flipped on average. We extend the L_1-regression reduction to handle this scenario. In particular, we prove an analogue of Kalai et al. (2008)'s result tailored to the smoothed model:

Theorem 4.2 *Suppose* $\min_{\deg(p_{\mathbf{z}})\leq d} \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_\sigma, \mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p_{\mathbf{z}}(\mathbf{x}) - f(\mathbf{x}\odot\mathbf{z})|] \leq \epsilon$ *for some degree* d *and any halfspace* f*, where* $\mathcal{D}_{\mathbf{x}}$ *is any distribution on* $\{\pm 1\}^n$*. Then, for* h *output by the degree-*d L_1 *polynomial regression algorithm with* $N = \mathrm{poly}(n^d/\epsilon)$ *examples,* $\mathbb{E}_{S\sim\mathcal{D}^N}[\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]] \leq \mathrm{opt}_\sigma + \epsilon$*. If we repeat the algorithm* $r = O(\log(1/\delta)/\epsilon)$ *times with fresh examples each, and let* h *be the hypothesis with lowest error on an independent test set of size* $O(\log(1/\delta)/\epsilon^2)$*, then with probability at least* $1 - \delta$*,* $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{opt}_\sigma + \epsilon$*.*

After this reduction, our task reduces to a purely approximation-theoretic problem: we need to construct, for each noise vector $\mathbf{z}$, a polynomial $p_{\mathbf{z}}(\mathbf{x})$ in the variable $\mathbf{x}$ such that the expected L_1 error over the smoothing process remains small:

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma,\mathbf{x}\sim D_{\mathbf{x}}}\left[|p_{\mathbf{z}}(\mathbf{x}) - f(\mathbf{x}\odot\mathbf{z})|\right] \le \epsilon.$$

To achieve this, we treat the smoothing noise $\mathbf{z}$ and consider $\mathbf{x}$ as a fixed parameter. This reduces the problem to approximating the function $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x}\odot\mathbf{z})$. We replace $f_{\mathbf{x}}(\mathbf{z})$ with its smooth approximation by applying the generalized Bonami-Beckner operator (Definition 3.2) on $\mathbf{z}$:

$$T_{1-\rho}f_{\mathbf{x}}(\mathbf{z}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[f_{\mathbf{x}}(\mathbf{y})].$$

Applying Lemma 3.4 with $\rho = O(\epsilon^2)$, we obtain:

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|T_{1-\rho}f_{\mathbf{x}}(\mathbf{z}) - f_{\mathbf{x}}(\mathbf{z})|] \le \epsilon.$$

Therefore, if we can find a low-degree polynomial that approximates $T_{1-\rho}f_{\mathbf{x}}(\mathbf{z})$ well in L_1, that polynomial will also succeed in approximating $f_{\mathbf{x}}(\mathbf{z})$. The remainder of our technical approach will be devoted to constructing such a polynomial approximator for the smoothed halfspace $T_{1-\rho}f_{\mathbf{x}}(\mathbf{z})$.

4.3 From Biased to Uniform Distribution on the Hypercube

To construct low-degree polynomial approximations, we analyze the noise-smoothed function $T_{1-\rho}f_{\mathbf{x}}$. Recall that for a fixed input $\mathbf{x}$, we define $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x}\odot\mathbf{z})$, and suppose $f(\cdot) = \mathrm{sign}(\langle\mathbf{w},\cdot\rangle - \theta)$. Then we have:

$$T_{1-\rho}f_{\mathbf{x}}(\mathbf{z}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\mathrm{sign}(\langle\mathbf{w}\odot\mathbf{x},\mathbf{y}\rangle - \theta)] = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\mathrm{sign}(\langle\mathbf{u},\mathbf{y}\rangle - \theta)],$$

where we define $\mathbf{u} = \mathbf{w}\odot\mathbf{x}$.

Here, $\mathbf{z}\sim\mathcal{N}_\sigma$ denotes a product distribution over $\{\pm1\}^n$ where each bit z_i is 1 with probability $1-\sigma$ and -1 with probability σ. The vector $\mathbf{y}$ is a $(1-\rho)$-noisy copy of $\mathbf{z}$ (Definition 3.1) with probability $1-\rho$, $y_i = z_i$; otherwise, y_i is redrawn independently from $\mathcal{N}_\sigma$ with probability ρ. Therefore, $\mathbf{y}\sim\mathcal{N}_\sigma$, correlated with $\mathbf{z}$, follows a biased distribution on the hypercube. To facilitate polynomial approximation, we aim to reduce this to a form where the randomness comes from a uniform distribution. To achieve this, we introduce a rerandomization trick that rewrites each coordinate y_i as:

$$y_i = (1-l_i)z_i + l_i\tau_i = (1-l_i)z_i + l_i(1-m_i) + l_im_i\epsilon_i,$$

where

$$l_i = \begin{cases}1 \text{ w.p. } \rho\\ 0 \text{ w.p. } 1-\rho\end{cases}, \tau_i = \begin{cases}1 \text{ w.p. } 1-\sigma\\ -1 \text{ w.p. } \sigma\end{cases}, m_i = \begin{cases}1 \text{ w.p. } 2\sigma\\ 0 \text{ w.p. } 1-2\sigma\end{cases},$$

with ϵ_i being a Radmacher random variable (uniform over $\{\pm1\}$).

This decomposition captures the full noise process: l_i is an indicator that determines whether the coordinate is kept as z_i (with probability $1-\rho$) or resampled as $\tau_i\sim(\mathcal{N}_\sigma)_i$ (with probability ρ). The variable m_i is then used to rerandomize τ_i, since τ_i can be viewed as taking the value 1 with probability $1-2\sigma$ (when $m_i = 0$) or a uniform random bit ϵ_i with probability 2σ (when $m_i = 1$).

A key benefit is that, conditional on $\mathbf{l}$ and $\mathbf{m}$, the random component $\boldsymbol{\epsilon}$ follows the uniform distribution on $\{\pm1\}^n$. We now condition on $(\mathbf{l},\mathbf{m})$ and express the smoothed function as:

$$T_{1-\rho}f_{\mathbf{x}}(\mathbf{z}) = \mathbb{E}_{\mathbf{l},\mathbf{m}}\Big[\mathbb{E}_{\mathbf{y}}[\mathrm{sign}(\langle\mathbf{u},\mathbf{y}\rangle - \theta)|\mathbf{l},\mathbf{m}]\Big] = \mathbb{E}_{\mathbf{l},\mathbf{m}}\Big[\mathbb{E}_{\boldsymbol{\epsilon}}[\mathrm{sign}(\langle\mathbf{u},\mathbf{l}\odot\mathbf{m}\odot\boldsymbol{\epsilon}\rangle + b - \theta)|\mathbf{l},\mathbf{m}]\Big],$$

where b is a deterministic shift depending on the coordinates fixed by $\mathbf{l},\mathbf{m}$.

Given that $\boldsymbol{\epsilon}$ is uniform distribution on hypercube, the inner sum behaves like a sum of independent $\{\pm1\}$ random variables. Under mild regularity condition (Definition 4.3) on the weight vector $\mathbf{u}$, we can apply the Berry–Esseen Theorem to approximate this inner distribution by a Gaussian. Specifically, we approximate:

$$\langle\mathbf{u},\mathbf{l}\odot\mathbf{m}\odot\boldsymbol{\epsilon}\rangle \approx \mathcal{N}(0,\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_2^2). \tag{4.1}$$

Substituting into the earlier expression yields the Gaussian-smoothed approximation:

$$\widetilde{T_{1-\rho}f_{\mathbf{x}}}(\mathbf{z}) = \mathbb{E}_{\mathbf{l},\mathbf{m}}\Big[\mathbb{E}_{s\sim\mathcal{N}(0,\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_2^2)}[\mathrm{sign}(s+b-\theta)|\mathbf{l},\mathbf{m}]\Big]$$

This reduces our setting to the Gaussian noise model analyzed in Chandrasekaran et al. (2024) for which efficient low-degree polynomial approximations are known. In particular, the density ratio method developed in that work can be applied to approximate $\widetilde{T_{1-\rho}f_{\mathbf{x}}}(\mathbf{z})$ with a small L_1 error.

4.4 Handling Irregularity via Critical Index Analysis

Recall that the approximation in (4.1) relies on the Berry–Esseen Theorem, which introduces a uniform approximation error of $O\big((\frac{\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_3}{\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_2})^3\big)$ for the cumulative density function. This can be further bounded by $O\big(\frac{\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_\infty}{\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_2}\big)$. Note that each coordinate $l_i m_i$ is equal to 1 with probability $2\rho\sigma$ and 0 otherwise. By concentration, we have $\|\mathbf{u}\odot\mathbf{l}\odot\mathbf{m}\|_2 \approx (2\rho\sigma)^{1/2}\|\mathbf{u}\|_2$, so the approximation error becomes $O\big((\rho\sigma)^{-1/2}\frac{\|\mathbf{u}\|_\infty}{\|\mathbf{u}\|_2}\big)$. This motivates the following regularity condition:

Definition 4.3 (regularity) *For vector $\mathbf{w}\in\mathbb{R}^n$, $\mathbf{w}$ is α-regular if $\|\mathbf{w}\|_\infty \le \alpha\cdot\|\mathbf{w}\|_2$.*

Given this definition, we see that if $\mathbf{u}$ is α-regular, then the approximation in (4.1) holds with L_∞ error $O((\rho\sigma)^{-1/2}\alpha)$. Since $\mathbf{u}=\mathbf{w}\odot\mathbf{x}$ and $\mathbf{x}\in\{\pm1\}^n$, the regularity of $\mathbf{u}$ is equivalent to that of $\mathbf{w}$. For such "good" (i.e., α-regular with small α) weight vectors $\mathbf{w}$, we can construct low-degree polynomial approximators by reducing to the Gaussian setting analyzed in Section 4.3 and Chandrasekaran et al. (2024).

However, we must also handle the "bad" or irregular cases, where $\langle\mathbf{u},\mathbf{l}\odot\mathbf{m}\odot\boldsymbol{\epsilon}\rangle$ deviates significantly from Gaussian behavior. To deal with such irregular $\mathbf{w}$, we employ critical index analysis, a standard tool in the analysis of Boolean halfspaces (Servedio, 2006; Matulef et al., 2010; Diakonikolas et al., 2010; Meka and Zuckerman, 2010; O'Donnell and Servedio, 2011; Diakonikolas and Servedio, 2013).

Definition 4.4 (α-critical index) *For $\mathbf{u}\in\mathbb{R}^n$, assume that $|u_1|\ge\cdots\ge|u_n|$. We define the α-critical index $\ell(\alpha)$ of a halfspace $h(\mathbf{x})=\mathrm{sign}(\langle\mathbf{u},\mathbf{x}\rangle-\theta)$ as the smallest index $i\in[n]$ for which $|u_i|\le\alpha\cdot\sigma_i$, where $\sigma_i:=\sqrt{\sum_{j=i}^n u_j^2}$.*

Intuitively, the α-critical index is the first index i such that the tail weight vector $(u_i,\cdots,u_n)$ is α-regular. Our earlier argument covers the case $i=1$, where the entire vector is regular. Using this framework, we obtain the following structural result:

Lemma 4.5 (Critical Index Decomposition) *Without loss of generality, let $\mathbf{u}=\mathbf{w}\odot\mathbf{x}$ with entries sorted in non-increasing magnitude, i.e., $|u_1|\ge\cdots\ge|u_n|$. Suppose $\mathbf{x}$ follows a (α,λ)-strictly sub-exponential distribution on $\{\pm1\}^n$. For any fixed $\mathbf{z}$, there exists a threshold $K=K(\alpha,\epsilon)=O\big(\log(1+\lambda)/\alpha^2+\log(1/\epsilon)\log(1/\alpha)/\rho\sigma\alpha^2\big)$ such that one of the following two conditions holds:*

1. *For some $H<K$, the tail vector $\mathbf{u}_T=(u_{H+1},\cdots,u_n)$ is α-regular, where α is to be choosen later.*
2. *For $H=K$ and at least $1-\epsilon$ fraction of $\mathbf{x}$, it holds that*
$$\mathbb{P}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\mathrm{sign}(\langle\mathbf{u}_H,\mathbf{y}_H\rangle+\langle\mathbf{u}_T,\mathbf{y}_T\rangle-\theta)\neq\mathrm{sign}(\langle\mathbf{u}_H,\mathbf{y}_H\rangle-\theta)]\le\epsilon, \tag{4.2}$$
where $\mathbf{u}_H:=(u_1,\cdots,u_H)$.

This lemma is proved by analyzing two cases, depending on whether the critical index $\ell(\alpha)$ satisfies $1<\ell(\alpha)<K$, or $\ell(\alpha)\ge K$. In the former case, we set $H=\ell(\alpha)-1$, and $\mathbf{u}_T$ is α-regular. In the latter case, the head vector forms a sufficiently long geometrically decaying sequence $|u_1|\ge\cdots\ge|u_H|$ to ensure that the influence of the remaining tail vector $\mathbf{u}_T$ on the halfspace output is negligible. That is, with high probability, $\mathrm{sign}(\langle\mathbf{u},\mathbf{y}\rangle-\theta)\approx\mathrm{sign}(\langle\mathbf{u}_H,\mathbf{y}_H\rangle-\theta)$.

We now show how to construct low-degree polynomial approximators in both cases.

Case 1: When $\mathbf{u}_T$ is α-regular, we condition on $\mathbf{y}_H$. For each fixed $\mathbf{y}_H$, the function becomes a regular halfspace in $\mathbf{y}_T$:
$$\widetilde{f_{\mathbf{x}}}(\mathbf{y}_T)=\mathrm{sign}(\langle\mathbf{u}_T,\mathbf{y}_T\rangle-\widetilde{\theta}),\ \text{where } \widetilde{\theta}=\theta-\langle\mathbf{u}_H,\mathbf{y}_H\rangle.$$
We apply the techniques of Chandrasekaran et al. (2024) to approximate this with a low-degree polynomial. One subtlety is that directly applying their construction leads to a degree polynomial in $|\widetilde{\theta}|/\|\mathbf{u}_T\|_2$. To address this, we use an indicator trick to define:
$$\begin{aligned}p_{\mathbf{y}_H}(\mathbf{x})=\mathrm{sign}(\langle\mathbf{u}_H,\mathbf{y}_H\rangle-\theta)\cdot\mathbb{1}\left(|\langle\mathbf{u}_H,\mathbf{y}_H\rangle-\theta|>C\cdot\|\mathbf{u}_T\|_2\right)\\+\widetilde{p}_{\mathbf{y}_H}(\mathbf{x})\cdot\mathbb{1}\left(|\langle\mathbf{u}_H,\mathbf{y}_H\rangle-\theta|\le C\cdot\|\mathbf{u}_T\|_2\right),\end{aligned}$$

where $\widetilde{p}_{\mathbf{y}_H}(\mathbf{x})$ can be constructed using the idea from Chandrasekaran et al. (2024) since $|\widetilde{\theta}|/\|\mathbf{u}_T\|_2$ is controlled. The indicator functions are low-degree polynomials of degree at most H, since they only depend on H variables and any function $f : \{\pm 1\}^k \to \mathbb{R}$ can be represented by a degree at most k multilinear polynomial.

Case 2: If the second condition of the lemma holds, we approximate $T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})$ directly using $\mathrm{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta)$. Since this depends only on the first H coordinates, it can be exactly represented as a polynomial of degree at most H.

In either case, we obtain a low-degree polynomial approximator for the smoothed function $T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})$.

4.5 Results

Using this framework, we establish the following approximation bound:

Definition 4.6 (Strictly Sub-exponential Distributions) *A distribution $\mathcal{D}$ on $\mathbb{R}^d$ is (α, λ)-strictly sub-exponential if for all $\|\mathbf{v}\|_2 = 1$, $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[|\langle \mathbf{x}, \mathbf{v} \rangle| > t] \leq 2 \cdot e^{-(t/\lambda)^{1+\alpha}}$.*

Lemma 4.7 *Fix $\epsilon > 0$ and a sufficiently large universal constant $C > 0$. Let $\mathcal{D}$ be a (α, λ)-strictly sub-exponential distribution on $\{\pm 1\}^n$. Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a linear threshold function. There exists a family of polynomials $p_{\mathbf{z}}$ parameterized by $\mathbf{z}$ of degree at most $O\Big(\big(C\sigma^{-\frac{1}{2}}\lambda \log(1/\epsilon)/\epsilon\big)^{6(1+\frac{1}{\alpha})^3}\Big)$ such that $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\mathbb{E}_{\mathbf{x}\sim D}[|p_{\mathbf{z}}(\mathbf{x}) - f_{\mathbf{x}}(\mathbf{z})|]$ is at most ϵ.*

Given the polynomial approximation and the degree upper bound, one can directly run L_1 polynomial regression (Algorithm 1) as stated in Theorem 4.2. We now can get our main theorem for strictly sub-exponential distributions.

Theorem 4.8 *Let $\mathcal{D}$ be a distribution on $\{\pm 1\}^n \times \{\pm 1\}$ such that the marginal distribution is (α, λ)-strictly sub-exponential. There exists an algorithm that draws $N = n^{\mathrm{poly}((\lambda/\sigma\epsilon)^{(1+1/\alpha)^3})} \log(1/\delta)$ samples, runs in time $\mathrm{poly}(n, N)$, and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1 - \delta$, it holds that $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$.*

Our main theorem shows that any Boolean halfspace on $\{\pm 1\}^n$ can be learned agnostically in the smoothed model under strictly sub-exponential input distributions. This result holds in a general and challenging setting where prior techniques fail, and it achieves efficient runtime and sample complexity. Table 1 compares our guarantees with the most relevant prior works. Conceptually, our contributions extend the scope of agnostic halfspace learning in two fundamental directions:

Relaxing distributional assumptions via smoothed optimality: A key technical contribution is our use of a smoothed benchmark opt_σ (Definition 1.3) instead of the worst-case error opt (Definition 1.1), enabling learning under substantially weaker distributional assumptions. In particular, we show that halfspaces remain efficiently learnable under general strictly sub-exponential marginals, which is a significant relaxation compared to the strong structural assumptions required in earlier work. For example, the Fourier-based techniques of Klivans et al. (2004b); Kalai et al. (2008); Blais et al. (2010) exploit spectral concentration under uniform or product distributions to obtain low-degree polynomial approximations. To go beyond the product setting, Wimmer (2010) generalized previous techniques to symmetric group to handle permutation-invariant distributions. However, these methods break down when the input has more dependencies or heavier tails. In contrast, our approach succeeds under strictly sub-exponential marginals by combining bit-flip smoothing with a critical index decomposition and Berry–Esseen approximation, enabling polynomial approximation without requiring coordinate independence or permutation invariant structure.

Extending the smoothed learning framework to the Boolean hypercube: A second core contribution is our extension of smoothed agnostic learning to the Boolean domain $\{\pm 1\}^n$, where additive Gaussian perturbations used in prior smoothed models are not well-defined. In continuous domains, several tools, including Gaussian surface area bounds (Klivans et al., 2008), log-concave concentration inequalities (Kane et al., 2013), and Gaussian smoothing combined with density ratio techniques (Chandrasekaran et al., 2024), enable efficient agnostic learning of halfspaces. However, none of these directly apply to the hypercube. Our analysis circumvents this barrier by performing a case analysis based on the critical index of the weight vector: either a small number of head coordinates

Table 1: Comparison of agnostic learning of halfspaces. We ignore the polynomial logarithmic factors in $1/\delta$.

Work	Domain	Distribution	Bench.	Smooth	Complexity
Kalai et al. (2008)	$\{\pm 1\}^n$	Uniform	opt	None	$n^{O(\frac{1}{\epsilon^4})}$
Kalai et al. (2008)	S^{n-1}	Uniform	opt	None	$n^{O(\frac{1}{\epsilon^4})}$
Kalai et al. (2008)	$\mathbb{R}^n$	Log-concave	opt	None	$n^{O(d(\epsilon))}$
Klivans et al. (2008)	$\mathbb{R}^n$	Gaussian	opt	None	$n^{O(\frac{1}{\epsilon^4})}$
Wimmer (2010)	$[B]^n$	Perm-Inv	opt	None	$n^{O(\frac{1}{\epsilon^4})}$
Kane et al. (2013)	$\mathbb{R}^n$	Sub-exp	opt_σ	Input noise	$n^{\exp(\frac{(\log\log(\frac{1}{\sigma\epsilon}))^{\widetilde{O}(1)}}{\sigma^4\epsilon^4})}$
Chandrasekaran et al. (2024)	$\mathbb{R}^n$	Strictly Sub-exp	opt_σ	Concept noise	$n^{\mathrm{poly}((\frac{\lambda}{\sigma\epsilon})^{(1+\frac{1}{\alpha})^3})}$
Ours	$\{\pm 1\}^n$	Strictly Sub-exp	opt_σ	Concept noise	$n^{\mathrm{poly}((\frac{\lambda}{\sigma\epsilon})^{(1+\frac{1}{\alpha})^3})}$

dominate and effectively determine the output, or the remaining tail is regular, allowing us to invoke the Berry–Esseen theorem to approximate the Boolean tail sum by a Gaussian, thereby enabling the use of continuous tools developed in prior work (Chandrasekaran et al., 2024).

5 Conclusion and Open Problems

In this work, we extended the smoothed agnostic learning framework to the Boolean hypercube, and demonstrated that halfspaces are efficiently learnable with respect to a broad class of input distributions (strictly sub-exponential marginals). Our approach combines tools from smoothing analysis, conditional polynomial approximation, and critical index decomposition to construct low-degree polynomial approximators in a discrete setting where standard analytic techniques are not applicable. By competing with a smoothed benchmark opt_σ, our guarantee circumvents known hardness results for agnostic learning over the hypercube, while matching the sample and runtime complexity of prior work in continuous domains.

Our current techniques apply only to single halfspaces, and the polynomial degree and runtime degrade as the smoothing parameter σ becomes small. In addition, our analysis requires strictly sub-exponential tail assumptions, and it remains unclear whether comparable guarantees are achievable under weaker conditions. Our results also suggest a potential link to agnostic learning under smoothed input distributions, analogous to the Gaussian framework in continuous domains (Kalai and Teng, 2008; Kalai et al., 2009; Kane et al., 2013; Chandrasekaran et al., 2024). Formalizing this connection in the Boolean setting appears subtle, due to the lack of Euclidean geometry and the discrete nature of bit-flip noise, and we leave it as an intriguing direction for future work.

An important open question is whether these techniques can be extended to intersections of multiple halfspaces. While our framework theoretically supports such generalizations under smoothed optimality, a major technical challenge arises in adapting critical index analysis to this setting. For a single halfspace, sorting the coordinates of the weight vector by magnitude plays a central role in identifying regular and irregular components. However, in the case of multiple halfspaces, each weight vector may induce a different ordering over coordinates, making it difficult to define a unified notion of "head" and "tail" variables. As a result, applying a shared conditioning or decomposition strategy becomes nontrivial. Developing new structural insights or approximation techniques that can handle this multi-directional irregularity remains an open problem.

More broadly, this raises the question of how far the smoothed learning framework can be pushed. Can it yield efficient algorithms for learning other complex Boolean concept classes (e.g., DNF formulas, decision lists, or polynomial threshold functions of higher degree) under heavy-tailed distributions? Can it be made adaptive to unknown noise levels or to distributions that do not satisfy strict tail bounds? We leave these questions for future work.

Acknowledgments and Disclosure of Funding

We sincerely thank the anonymous reviewers for their helpful comments. The authors acknowledge support in part from the National Science Foundation under Award CCF-2217033 (EnCORE: Institute for Emerging CORE Methods in Data Science).

References

AWASTHI, P., BALCAN, M.-F., HAGHTALAB, N. and URNER, R. (2015). Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*. PMLR.

BLAIS, E., O'DONNELL, R. and WIMMER, K. (2010). Polynomial regression under arbitrary product distributions. *Machine learning* **80** 273–294.

BLUM, A. and DUNAGAN, J. (2002). Smoothed analysis of the perceptron algorithm for linear programming. In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '02, Society for Industrial and Applied Mathematics, USA.

BLUM, A., FRIEZE, A. M., KANNAN, R. and VEMPALA, S. S. (1996). A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica* **22** 35–52.

CHANDRASEKARAN, G., KLIVANS, A., KONTONIS, V., MEKA, R. and STAVROPOULOS, K. (2024). Smoothed analysis for learning concepts with low intrinsic dimension. In *The Thirty Seventh Annual Conference on Learning Theory*. PMLR.

COHEN, E. (1997). Learning noisy perceptrons by a perceptron in polynomial time. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Machine learning* **20** 273–297.

DIAKONIKOLAS, I., GOPALAN, P., JAISWAL, R., SERVEDIO, R. A. and VIOLA, E. (2010). Bounded independence fools halfspaces. *SIAM Journal on Computing* **39** 3441–3462.

DIAKONIKOLAS, I., GOULEAKIS, T. and TZAMOS, C. (2019). Distribution-independent pac learning of halfspaces with massart noise. *Advances in Neural Information Processing Systems* **32**.

DIAKONIKOLAS, I., IMPAGLIAZZO, R., KANE, D. M., LEI, R., SORRELL, J. and TZAMOS, C. (2021). Boosting in the presence of massart noise. In *Conference on Learning Theory*. PMLR.

DIAKONIKOLAS, I., KANE, D., MANURANGSI, P. and REN, L. (2022). Cryptographic hardness of learning halfspaces with massart noise. *Advances in Neural Information Processing Systems* **35** 3624–3636.

DIAKONIKOLAS, I., KONTONIS, V., TZAMOS, C. and ZARIFIS, N. (2020). Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*. PMLR.

DIAKONIKOLAS, I. and SERVEDIO, R. A. (2013). Improved approximation of linear threshold functions. *computational complexity* **22** 623–677.

FELDMAN, V., GOPALAN, P., KHOT, S. and PONNUSWAMI, A. K. (2009). On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing* **39** 606–645.

GURUSWAMI, V. and RAGHAVENDRA, P. (2009). Hardness of learning halfspaces with noise. *SIAM Journal on Computing* **39** 742–765.

HAUSSLER, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation* **100** 78–150.

KAHN, J., KALAI, G. and LINIAL, N. (1988). The influence of variables on boolean functions. In *[Proceedings 1988] 29th Annual Symposium on Foundations of Computer Science*.

KALAI, A. T., KLIVANS, A. R., MANSOUR, Y. and SERVEDIO, R. A. (2008). Agnostically learning halfspaces. *SIAM Journal on Computing* **37** 1777–1805.

KALAI, A. T., SAMORODNITSKY, A. and TENG, S.-H. (2009). Learning and smoothed analysis. In *Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science*. FOCS '09, IEEE Computer Society, USA.

KALAI, A. T. and TENG, S.-H. (2008). Decision trees are pac-learnable from most product distributions: a smoothed analysis. *arXiv preprint arXiv:0812.0933* .

KANE, D., KLIVANS, A. and MEKA, R. (2013). Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *Conference on Learning Theory*. PMLR.

KEARNS, M. J., SCHAPIRE, R. E. and SELLIE, L. M. (1992). Toward efficient agnostic learning. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92, Association for Computing Machinery, New York, NY, USA.

KLIVANS, A. R., LONG, P. M. and SERVEDIO, R. A. (2009). Learning halfspaces with malicious noise. *Journal of Machine Learning Research* **10** 2715–2740.

KLIVANS, A. R., O'DONNELL, R. and SERVEDIO, R. A. (2004a). Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences* **68** 808–840.

KLIVANS, A. R., O'DONNELL, R. and SERVEDIO, R. A. (2004b). Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences* **68** 808–840. Special Issue on FOCS 2002.

KLIVANS, A. R., O'DONNELL, R. and SERVEDIO, R. A. (2008). Learning geometric concepts via gaussian surface area. *2008 49th Annual IEEE Symposium on Foundations of Computer Science* 541–550.

LINIAL, N., MANSOUR, Y. and NISAN, N. (1993). Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)* **40** 607–620.

LITTLESTONE, N. (1987). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*.

MATULEF, K., O'DONNELL, R., RUBINFELD, R. and SERVEDIO, R. A. (2010). Testing halfspaces. *SIAM Journal on Computing* **39** 2004–2047.

MEKA, R. and ZUCKERMAN, D. (2010). Pseudorandom generators for polynomial threshold functions. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*.

MOSSEL, E., O'DONNELL, R. and OLESZKIEWICZ, K. (2005). Noise stability of functions with low influences: invariance and optimality. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. IEEE.

NOVIKOFF, A. B. (1963). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*. Polytechnic Institute of Brooklyn.

O'DONNELL, R. (2021). Analysis of boolean functions. *arXiv preprint arXiv:2105.10386* .

O'DONNELL, R. and SERVEDIO, R. A. (2011). The chow parameters problem. *SIAM Journal on Computing* **40** 165–199.

PERES, Y. (2004). Noise stability of weighted majority. *arXiv preprint math/0412377* .

ROSENBLATT, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65** 386.

SERVEDIO, R. (2006). Every linear threshold function has a low-weight approximator. In *21st Annual IEEE Conference on Computational Complexity (CCC'06)*.

SPIELMAN, D. and TENG, S.-H. (2001). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*.

VALIANT, L. G. (1984). A theory of the learnable. *Commun. ACM* **27** 1134–1142.

WIMMER, K. (2010). Agnostically learning under permutation invariant distributions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*.

A Bonami-Beckner Operator Approximation

We show that for any linear threshold function $f : \{\pm 1\}^n \to \{\pm 1\}$ the approximation error L_1 of the operator $T_{1-\rho} f$ to f can be upper bounded by $O(\sqrt{\rho})$.

Lemma A.1 *For any linear threshold function $f : \{\pm 1\}^n \to \{\pm 1\}$ and $\sigma, \rho \in [0, 1]$, it holds that*

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma}[|T_{1-\rho} f(\mathbf{z}) - f(\mathbf{z})|] \leq \frac{5}{2}\sqrt{\rho}.$$

Proof By triangle inequality and special case of Lemma 3.4 when $\Omega = \{\pm 1\}^n$ and $\mu = \mathcal{N}_\sigma$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma}[|T_{1-\rho} f(\mathbf{z}) - f(\mathbf{z})|] &= \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma}\big[\big|\mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})}[f(\mathbf{y})] - f(\mathbf{z})\big|\big] \\
&\leq \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma}\big[\mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})}[|f(\mathbf{y}) - f(\mathbf{z})|]\big] \\
&= 2\mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma, \mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})}\big[\mathbb{1}[f(\mathbf{y}) \neq f(\mathbf{z})]\big] \\
&\leq 2NS_\rho(f) \\
&\leq \frac{5}{2}\sqrt{\rho}.
\end{aligned}$$

■

Therefore, choosing $\rho = O(\epsilon^2)$ makes this error at most $\epsilon/2$.

B Polynomial Approximation for $T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})$

We now approximate $T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})$ using a polynomial for the more general class of strictly sub-exponential distributions.

Definition B.1 (Strictly Sub-exponential Distributions) *A distribution $\mathcal{D}$ on $\mathbb{R}^d$ is (α, λ)-strictly sub-exponential if for all $\|\mathbf{v}\|_2 = 1$, $\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[|\langle \mathbf{x}, \mathbf{v} \rangle| > t] \leq 2 \cdot e^{-(t/\lambda)^{1+\alpha}}$.*

Our main goal in this section is to prove the following polynomial approximation result in Lemma 4.7. Suppose $f(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - \theta)$. Denote $\mathbf{w} \odot \mathbf{x}$ as $\mathbf{u}$. Without loss of generality, suppose that $|u_1| \geq |u_2| \geq \cdots \geq |u_n|$. Then, we have

$$T_{1-\rho} f_{\mathbf{x}}(\mathbf{z}) = \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})}[\operatorname{sign}(\langle \mathbf{w} \odot \mathbf{x}, \mathbf{y} \rangle - \theta)] = \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})}[\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta)].$$

To obtain a polynomial approximation of $T_\rho f_{\mathbf{x}}(\mathbf{z})$, we first prove Lemma 4.5.

Proof Let

$$\begin{aligned}
L(\alpha, \epsilon) &= \lceil \log(1/\epsilon)/\rho\sigma \rceil \cdot \lceil (4/\alpha^2) \log(1/\alpha) \rceil + 2\log(C/\alpha)/\alpha^2 \\
&= O\big(\log(1/\epsilon) \log(1/\alpha)/\rho\sigma\alpha^2\big) + O\big(\log(\max\{\lambda, 1\} \log^{\max\{\frac{1}{2}, \frac{1}{1+\alpha}\}}(1/\epsilon)/\alpha)/\alpha^2\big) \\
&= O\big(\log(1/\epsilon) \log(1/\alpha)/\rho\sigma\alpha^2\big) + O\big(\log(\max\{\lambda, 1\} \log(1/\epsilon)/\alpha)/\alpha^2\big) \\
&= O\big(\log(1+\lambda)/\alpha^2 + \log(1/\epsilon) \log(1/\alpha)/\rho\sigma\alpha^2\big),
\end{aligned}$$

where $C = (1 - 2\rho\sigma) \cdot \lambda \log^{\frac{1}{1+\alpha}}(2/\epsilon) + \sqrt{2\log(2/\epsilon)}$.

If α-critical index $\ell(\alpha) < L(\alpha, \epsilon)$, then the first condition holds by taking $H = \ell(\alpha) - 1$. If the α-critical index $\ell(\epsilon) \geq L(\alpha, \epsilon)$, we will show that the second condition holds.

By Lemma 5.5 in Diakonikolas et al. (2010), there exist a set of nicely separated coordinates $G = \{i_1, i_2, \cdots, i_t\} \subseteq H$ where $i_1 < i_2 < \cdots < i_t$ and $i_{k+1} - i_k = \lceil 4\log(1/\alpha)/\alpha^2 \rceil$ such that $|u_{i_{k+1}}| \leq |u_{i_k}|/3$ for any $k \in [t-1]$. Then by Claim 5.7 in Diakonikolas et al. (2010), for any two points $\mathbf{x}_1 \neq \mathbf{x}_2 \in \{\pm 1\}^t$, we have $|\langle \mathbf{u}_G, \mathbf{x}_1 \rangle - \langle \mathbf{u}_G, \mathbf{x}_2 \rangle| \geq |u_{i_t}|$. Take $t = \lceil \log(1/\epsilon)/\rho\sigma \rceil$. For any

fixed assignment to the variables in $H \setminus G$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathbf{y}}\Big[\Big|\sum_{i\in H} u_i y_i - \theta\Big| \le \frac{|u_{i_t}|}{4}\Big] \\
&= \mathbb{P}_{\mathbf{y}}\Big[\sum_{i\in G} u_i y_i \in \Big[\theta - \sum_{i\in H\setminus G} u_i y_i - \frac{|u_{i_t}|}{4}, \theta - \sum_{i\in H\setminus G} u_i y_i + \frac{|u_{i_t}|}{4}\Big]\Big] \\
&\le \max_{\mathbf{x}_1\in\{\pm 1\}^t} \mathbb{P}_{\mathbf{y}_G}[\mathbf{y}_G = \mathbf{x}_1] \\
&\le (1-\rho\sigma)^t \le e^{-\rho\sigma t} \le \epsilon,
\end{aligned}$$

where the second inequality is because there's at most one point in an interval of length $|u_{i_t}|$ given that $\langle \mathbf{u}_G, \mathbf{x}_1\rangle$ are well-separated. The third inequality is because

$$\begin{aligned}
&\max\{\mathbb{P}[y_i = 1], \mathbb{P}[y_i = -1]\} \le \max\{1-\rho\sigma, \rho\sigma\} = 1-\rho\sigma, && \text{if } z_i = 1, \\
&\max\{\mathbb{P}[y_i = 1], \mathbb{P}[y_i = -1]\} \le \max\{\rho(1-\sigma), 1-\rho(1-\sigma)\} \le 1-\rho\sigma, && \text{if } z_i = -1.
\end{aligned}$$

By our choice of $L(\alpha,\epsilon), t, i_t$, we have

$$L(\alpha,\epsilon) - i_t \ge L(\alpha,\epsilon) - \lceil \log(1/\epsilon)/\rho\sigma\rceil \cdot \lceil (4/\alpha^2)\log(1/\alpha)\rceil \ge 2\log(C/\alpha)/\alpha^2.$$

By applying Lemma 5.5 in Diakonikolas et al. (2010), we have

$$\|\mathbf{u}_T\|_2 = \sigma_T \le (\sqrt{1-\alpha^2})^{L(\alpha,\epsilon)-i_t} \cdot |u_{i_t}|/\alpha \le C^{-1}\cdot |u_{i_t}|.$$

Therefore, by Lemma C.1, for at least $1-\epsilon$ fraction of $\mathbf{x}$, it holds with probability at least $1-\epsilon$ of $\mathbf{y}$ that

$$|\langle \mathbf{u}_H, \mathbf{y}_H\rangle - \theta| = \Big|\sum_{i\in H} u_i y_i - \theta\Big| \ge \frac{|u_{i_t}|}{4} \ge \frac{C\cdot\|\mathbf{u}_T\|_2}{4} \ge \frac{|\langle \mathbf{u}_T, \mathbf{y}_T\rangle|}{4}.$$

Then, it follows that with probability at least $1-\epsilon$ of $\mathbf{y}$ we have

$$\operatorname{sign}\big(\langle \mathbf{u}_H, \mathbf{y}_H\rangle + \langle \mathbf{u}_T, \mathbf{y}_T\rangle - \theta\big) = \operatorname{sign}\big(\langle \mathbf{u}_H, \mathbf{y}_H\rangle - \theta\big).$$

■

Now we are ready to construct our polynomial. We consider the two cases in Lemma 4.5 separately.

Case 1: If the weight vector $\mathbf{w}$ of the LTF falls into the second case of Lemma 4.5, notice that $\operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H\rangle - \theta) = \operatorname{sign}(\langle \mathbf{w}_H \odot \mathbf{x}_H, \mathbf{y}_H\rangle - \theta)$ can be represented as a polynomial $p_{\mathbf{y}}(\mathbf{x})$ of degree at most $H = K$ since only H coordinates of $\mathbf{x}$ are relevant. In this case, we take our final polynomial as

$$p_{\mathbf{z}}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[p_{\mathbf{y}}(\mathbf{x})] = \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\operatorname{sign}(\langle \mathbf{w}_H \odot \mathbf{x}_H, \mathbf{y}_H\rangle - \theta)].$$

Let $\Delta(\mathbf{x})$ be defined as the error term $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|p_{\mathbf{z}}(\mathbf{x}) - T_{1-\rho}f_{\mathbf{x}}(\mathbf{z})|]$. We have that for at least $1-\epsilon$ fraction of $\mathbf{x}$ it holds that

$$\begin{aligned}
\Delta(\mathbf{x}) &= \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\Big|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\operatorname{sign}(\langle \mathbf{w}_H \odot \mathbf{x}_H, \mathbf{y}_H\rangle - \theta)] - \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\operatorname{sign}(\langle \mathbf{w}\odot\mathbf{x}, \mathbf{y}\rangle - \theta)]\Big|\Big] \\
&\le \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\big[\big|\operatorname{sign}(\langle \mathbf{w}_H \odot \mathbf{x}_H, \mathbf{y}_H\rangle - \theta) - \operatorname{sign}(\langle \mathbf{w}\odot\mathbf{x}, \mathbf{y}\rangle - \theta)\big|\big]\Big] \\
&= 2\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{P}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\big[\operatorname{sign}(\langle \mathbf{w}_H \odot \mathbf{x}_H, \mathbf{y}_H\rangle - \theta) \ne \operatorname{sign}(\langle \mathbf{w}\odot\mathbf{x}, \mathbf{y}\rangle - \theta)\big]\Big] \\
&\le 2\epsilon.
\end{aligned}$$

It follows that the final L_1 approximation error

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\big[\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|p_{\mathbf{z}}(\mathbf{x}) - T_{1-\rho}f_{\mathbf{x}}(\mathbf{z})|]\big] \\
&= \mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\Delta(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\Delta(\mathbf{x})|\text{“good” }\mathbf{x}]\cdot\mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[\text{“good” }\mathbf{x}] + \mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\Delta(\mathbf{x})|\text{“bad” }\mathbf{x}]\cdot\mathbb{P}_{\mathbf{x}\sim D_{\mathbf{x}}}[\text{“bad” }\mathbf{x}] \\
&\le 2\epsilon\cdot 1 + 2\cdot\epsilon = 4\epsilon.
\end{aligned}$$

Here “good” $\mathbf{x}$ refers to the at least $1-\epsilon$ fraction of $\mathbf{x}$ such that the approximation in (4.2) holds.

Case 2: If the weight vector $\mathbf{w}$ of the LTF falls into the first case of Lemma 4.5, we consider the following approximation

$$p_{\mathbf{y}_H}(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta) \cdot \mathbb{1}\left(|\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta| > C \cdot \|\mathbf{u}_T\|_2\right) + \widetilde{p}_{\mathbf{y}_H}(\mathbf{x}) \cdot \mathbb{1}\left(|\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta| \le C \cdot \|\mathbf{u}_T\|_2\right),$$

where $C = (1-2\rho\sigma) \cdot \lambda \log^{\frac{1}{1+\alpha}}(2/\epsilon) + \sqrt{2\log(2/\epsilon)}$ and $\widetilde{p}_{\mathbf{y}_H}(\mathbf{x})$ will be choosen later as (B.5). In this case, we take our final polynomial as

$$p_{\mathbf{z}}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}_H \sim \mathcal{N}_{1-\rho}(\mathbf{z})|_H}[p_{\mathbf{y}_H}(\mathbf{x})].$$

Let $\Delta(\mathbf{x})$ be defined as the L_1 error term $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|p_{\mathbf{z}}(\mathbf{x}) - T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})|]$. For notation simplicity, we denote $\{\mathbf{y}_H : |\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta| > C \cdot \|\mathbf{u}_T\|_2\}$ as event $\mathcal{E}$. Then, we have

$$\begin{aligned}
&\Delta(\mathbf{x}) \\
&= \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[p_{\mathbf{y}}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[f_{\mathbf{x}}(\mathbf{y})]\right|\right] \\
&= \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[(p_{\mathbf{y}}(\mathbf{x}) - f_{\mathbf{x}}(\mathbf{y})) \cdot \mathbb{1}[\mathcal{E}]\right] + \mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[(p_{\mathbf{y}}(\mathbf{x}) - f_{\mathbf{x}}(\mathbf{y})) \cdot \mathbb{1}[\mathcal{E}^c]\right]\right|\right] \\
&\le \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[(p_{\mathbf{y}}(\mathbf{x}) - f_{\mathbf{x}}(\mathbf{y})) \cdot \mathbb{1}[\mathcal{E}]\right]\right| + \left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[(p_{\mathbf{y}}(\mathbf{x}) - f_{\mathbf{x}}(\mathbf{y})) \cdot \mathbb{1}[\mathcal{E}^c]\right]\right|\right] \\
&= \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[\left(\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) - \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta)\right) \cdot \mathbb{1}[\mathcal{E}]\right]\right|\right]}_{\Delta_1(\mathbf{x})} \\
&\quad + \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[\left(\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) - \widetilde{p}_{\mathbf{y}}(\mathbf{x})\right) \cdot \mathbb{1}[\mathcal{E}^c]\right]\right|\right]}_{\Delta_2(\mathbf{x})}
\end{aligned}$$

Notice that by Lemma C.1, for at least $1-\epsilon$ fraction of $\mathbf{x}$, $|\langle \mathbf{u}_T, \mathbf{y}_T \rangle| \le C \cdot \|\mathbf{u}_T\|_2$ holds for at least $1-\epsilon$ fraction of $\mathbf{y}$. For such $\mathbf{x}$ and $\mathbf{y}$, under event $\mathcal{E}$, we have

$$|\langle \mathbf{u}_T, \mathbf{y}_T \rangle| \le C \cdot \|\mathbf{u}_T\|_2 < |\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta|,$$

then it follows that

$$\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) \cdot \mathbb{1}[\mathcal{E}] = \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle + \langle \mathbf{u}_T, \mathbf{y}_T \rangle - \theta) \cdot \mathbb{1}[\mathcal{E}] = \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta) \cdot \mathbb{1}[\mathcal{E}].$$

Then, for at least $1-\epsilon$ fraction of $\mathbf{x}$ we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[\left(\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) - \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta)\right) \cdot \mathbb{1}[\mathcal{E}]\right]\right| \\
&\le \left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[\left(\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) - \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta)\right) \cdot \mathbb{1}[\mathcal{E}] \middle| \text{“good” } \mathbf{y}\right]\right| \\
&\qquad \cdot \mathbb{P}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\text{“good” } \mathbf{y}] \\
&\quad + \left|\mathbb{E}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}\left[\left(\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) - \operatorname{sign}(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta)\right) \cdot \mathbb{1}[\mathcal{E}] \middle| \text{“bad” } \mathbf{y}\right]\right| \\
&\qquad \cdot \mathbb{P}_{\mathbf{y}\sim\mathcal{N}_{1-\rho}(\mathbf{z})}[\text{“bad” } \mathbf{y}] \\
&\le 0 + 2 \cdot \epsilon = 2\epsilon.
\end{aligned}$$

Here “good” $\mathbf{y}$ refers to the at least $1-\epsilon$ fraction of $\mathbf{y}$ such that $|\langle \mathbf{u}_T, \mathbf{y}_T \rangle| \le C \cdot \|\mathbf{u}_T\|_2$ holds. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\Delta_1(\mathbf{x})] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\Delta_1(\mathbf{x}) | \text{“good” } \mathbf{x}] \cdot \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\text{“good” } \mathbf{x}] \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\Delta_1(\mathbf{x}) | \text{“bad” } \mathbf{x}] \cdot \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\text{“bad” } \mathbf{x}] \\
&\le \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\Delta_1(\mathbf{x}) | \text{“good” } \mathbf{x}] + 2 \cdot \epsilon \\
&\le 2\epsilon + 2\epsilon = 4\epsilon.
\end{aligned}$$

Here “good” $\mathbf{x}$ refers to the at least $1-\epsilon$ fraction of $\mathbf{x}$ such that $\mathbb{P}_{\mathbf{y}}[|\langle \mathbf{u}_T, \mathbf{y}_T \rangle| \le C \cdot \|\mathbf{u}_T\|_2] \ge 1-\epsilon$ holds.

Next we consider bounding $\Delta_2(\mathbf{x})$ by constructing proper low-degree polynomial $\widetilde{p}_{\mathbf{y}}(\mathbf{x})$ as follows. Recall that $\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})$ where $y_i = z_i$ with probability $1-\rho$ and y_i randomly drawn from $\mathcal{N}_\sigma$ with

probability ρ, we use the following rerandomization trick for random vector $\mathbf{y}_T$: for each coordinate of $\mathbf{y}_T$, let

$$\left.\begin{array}{l} y_i = (1-l_i)z_i + l_i\tau_i \\ \tau_i = (1-m_i) + m_i\epsilon_i \end{array}\right\} \Longrightarrow y_i = (1-l_i)z_i + l_i(1-m_i) + l_i m_i \epsilon_i, \tag{B.1}$$

where

$$l_i = \begin{cases} 1 \text{ w.p. } \rho \\ 0 \text{ w.p. } 1-\rho \end{cases}, \tau_i = \begin{cases} 1 \text{ w.p. } 1-\sigma \\ -1 \text{ w.p. } \sigma \end{cases}, m_i = \begin{cases} 1 \text{ w.p. } 2\sigma \\ 0 \text{ w.p. } 1-2\sigma \end{cases},$$

and ϵ_i is a Radmacher random variable. Let random variable $A = \langle \mathbf{u}, \mathbf{y} \rangle - \theta$. Then by (B.1) we have

$$\begin{aligned} A &= \langle \mathbf{u}, \mathbf{y} \rangle - \theta \\ &= \langle \mathbf{u}_H, \mathbf{y}_H \rangle + \langle \mathbf{u}_T, \mathbf{y}_T \rangle - \theta \\ &= \langle \mathbf{u}_H, \mathbf{y}_H \rangle + \langle \mathbf{u}_T, (\mathbf{1}_T - \mathbf{l}_T) \odot \mathbf{z}_T \rangle + \langle \mathbf{u}_T, \mathbf{l}_T \odot (\mathbf{1}_T - \mathbf{m}_T) \rangle - \theta + \langle \mathbf{u}_T, \mathbf{l}_T \odot \mathbf{m}_T \odot \boldsymbol{\epsilon}_T \rangle \\ &= \underbrace{\langle \mathbf{u}_H, \mathbf{y}_H \rangle + \langle \mathbf{u}_T, \mathbf{v}_T \rangle - \theta}_{:=b} + \underbrace{\langle \mathbf{u}_T, \mathbf{l}_T \odot \mathbf{m}_T \odot \boldsymbol{\epsilon}_T \rangle}_{:=B}. \end{aligned}$$

where $\mathbf{v}_T := (\mathbf{1}_T - \mathbf{l}_T) \odot \mathbf{z}_T + \mathbf{l}_T \odot (\mathbf{1}_T - \mathbf{m}_T)$. Then we have

$$\begin{aligned} \widetilde{T_{1-\rho} f_{\mathbf{x}}}(\mathbf{z}) &:= \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})} \big[\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] \\ &= \mathbb{E}_{\mathbf{y}_H, A} [\operatorname{sign}(A) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]] \\ &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \big[\mathbb{E}_A [\operatorname{sign}(A) | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] \\ &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \big[\big(1 - 2\mathbb{P}_A [A \le 0 | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] \big) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] \\ &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \big[\big(1 - 2\mathbb{P}_B [B \le -b | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] \big) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \big], \end{aligned}$$

and

$$\begin{aligned} &\Delta_2(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma} \Big[\Big| \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})} \big[\operatorname{sign}(\langle \mathbf{u}, \mathbf{y} \rangle - \theta) \cdot \mathbf{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] - \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})} \big[\widetilde{p}_{\mathbf{y}}(\mathbf{x}) \cdot \mathbf{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] \Big| \Big] \\ &= \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_\sigma} \Big[\Big| \widetilde{T_{1-\rho} f_{\mathbf{x}}}(\mathbf{z}) - \mathbb{E}_{\mathbf{y} \sim \mathcal{N}_{1-\rho}(\mathbf{z})} \big[\widetilde{p}_{\mathbf{y}}(\mathbf{x}) \cdot \mathbf{1}[\mathbf{y}_H \in \mathcal{E}^c] \big] \Big| \Big]. \end{aligned}$$

By Theorem C.3, we have

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left[\frac{B}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \le x \middle| \mathbf{l}_T, \mathbf{m}_T \right] - \Phi(x) \right| \le C' \cdot \left(\frac{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_3}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \right)^3.$$

where C' is a constant. Therefore, we have

$$\begin{aligned} &\left| \widetilde{T_{1-\rho} f_{\mathbf{x}}}(\mathbf{z}) - \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \left[\left(1 - 2\Phi \left(\frac{-b}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \right) \right) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \right] \right| \\ &\le 2\mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \left[\left| \mathbb{P}[B \le -b | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] - \Phi \left(\frac{-b}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \right) \right| \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \right] \\ &\le 2C' \cdot \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \left[\left(\frac{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_3}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \right)^3 \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \right] \\ &\le 2C' \cdot \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T} \left[\frac{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_\infty}{\|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \right]. \end{aligned}$$

Notice that in this case $\mathbf{u}_T$ is α-regular and $l_i m_i$ takes 1 with probability $2\rho\sigma$ and 0 with probability $1 - 2\rho\sigma$, then by Lemma C.2, for at least $1 - \epsilon$ fraction of $(\mathbf{l}_T, \mathbf{m}_T)$ it holds that

$$\|\mathbf{u}_T\|_\infty \le \alpha \cdot \|\mathbf{u}_T\|_2 \le (\rho\sigma)^{-\frac{1}{2}} \cdot \alpha \cdot \|\mathbf{u}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2, \tag{B.2}$$

as long as the condition $\alpha \le \rho\sigma / \sqrt{\log(1/\epsilon)/2}$ holds.

Notice that

$$1-2\Phi\bigg(\frac{-b}{\|\mathbf{u}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2}\bigg)=\mathbb{E}_{X\sim\mathcal{N}(b,\|\mathbf{u}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2^2)}[\mathrm{sign}(X)|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]$$
$$=\mathbb{E}_{X\sim\mathcal{N}(b,\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2^2)}[\mathrm{sign}(X)|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T],$$

let

$$\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}:=\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\mathbb{E}_{X\sim\mathcal{N}(b,\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2^2)}[\mathrm{sign}(X)|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]$$
$$\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Big],$$

where $\mathcal{E}_1$ is the event regarding the randomness of $(\mathbf{l}_T,\mathbf{m}_T)$ such that (B.2) holds, then

$$\begin{aligned}
&\big|\widetilde{T_\rho f_{\mathbf{x}}(\mathbf{z})}-\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}\big|\\
&\le 2\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\bigg[\Big|\mathbb{P}[B\le -b|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]-\Phi\Big(-\frac{b}{\|\mathbf{u}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2}\Big)\Big|\\
&\qquad\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\bigg]\\
&\quad+\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\big|1-2\mathbb{P}[B\le -b|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]\big|\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1^c]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Big]\\
&\le 2C'\cdot\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\bigg[\frac{\|\mathbf{u}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_\infty}{\|\mathbf{u}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2}\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\bigg]\\
&\quad+\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1^c]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Big]\\
&\le 2C'\cdot(\rho\sigma)^{-\frac{1}{2}}\alpha+\epsilon\\
&\le 2C'\cdot(2\rho\sigma/\log(1/\epsilon))^{\frac{1}{2}}+\epsilon\\
&\le 2\epsilon,
\end{aligned}$$

as long as $\rho=O(\epsilon^2\log(1/\epsilon)/\sigma)$. Then, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\big[\big|\widetilde{T_\rho f_{\mathbf{x}}(\mathbf{z})}-\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}\big|\big]\le 2\epsilon.$$

Now we only need to consider polynomial approximation for $\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}$. We can recenter the expectation around zero as follows:

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{N}(b,\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2^2)}[\mathrm{sign}(X)|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]\\
&=\mathbb{E}_{s\sim\mathcal{N}(b/\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2,1)}[\mathrm{sign}(\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{u}_T\|_2\cdot s)|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T]\\
&=\mathbb{E}_{s\sim Q}\bigg[\mathrm{sign}(\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2\cdot s)\cdot\frac{\mathcal{N}(s;b/\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2,1)}{Q(s)}\bigg|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T\bigg]\\
&=\frac{e^{-\frac{b^2}{2\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2^2}}}{\sqrt{2\pi}}\cdot\mathbb{E}_{s\sim Q}\bigg[\mathrm{sign}(\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2\cdot s)\cdot e^{-\frac{s^2}{2}-\log Q(s)}\\
&\qquad\cdot e^{\frac{b\cdot s}{\|\mathbf{w}_T\odot\mathbf{l}_T\odot\mathbf{m}_T\|_2}}\bigg|\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T\bigg],
\end{aligned}$$

where $Q(s)=e^{-|s|}/2$.

For the simplicity of the following analysis, we define random variable

$$x:=\frac{\langle\mathbf{u}_T,\mathbf{v}_T\rangle}{\|\mathbf{u}_T\odot\mathbf{v}_T\|_2}=\frac{\langle\mathbf{w}_T\odot\mathbf{x}_T,\mathbf{v}_T\rangle}{\|\mathbf{u}_T\odot\mathbf{v}_T\|_2}=\frac{\langle\mathbf{w}_T\odot\mathbf{v}_T,\mathbf{x}_T\rangle}{\|\mathbf{w}_T\odot\mathbf{v}_T\|_2}.$$

Since $\mathbf{x}$ is (α,λ)-strictly sub-exponential, then x satisfies the following concentration inequality:

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\big[|x|>t\big]\le 2\cdot e^{-(t/\lambda)^{1+\alpha}}.$$

Then, we can rewrite:

$$\frac{b}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} = \frac{(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta) + \langle \mathbf{u}_T, \mathbf{v}_T \rangle}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} = \frac{(\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta) + \|\mathbf{w}_T \odot \mathbf{v}_T\|_2 \cdot x}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2}.$$

For simplicity, denote

$$a = \frac{\|\mathbf{w}_T \odot \mathbf{v}_T\|_2}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2}, \qquad c = \frac{\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2}.$$

Notice that under condition $(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1$ and condition $\mathbf{y}_H \in \mathcal{E}^c$, we have

$$|c| = \frac{|\langle \mathbf{u}_H, \mathbf{y}_H \rangle - \theta|}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \le C \cdot \frac{\|\mathbf{u}_T\|_2}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \le C \cdot (\rho\sigma)^{-1/2} := c', \tag{B.3}$$

$$\begin{aligned} |a| &\le \frac{\|\mathbf{w}_T \odot (\mathbf{1}_T - \mathbf{l}_T) \odot \mathbf{z}_T\|_2 + \|\mathbf{w}_T \odot \mathbf{l}_T \odot (\mathbf{1}_T - \mathbf{m}_T)\|_2}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \\ &\le \frac{\|\mathbf{w}_T \odot (\mathbf{1}_T - \mathbf{l}_T)\|_2 + \|\mathbf{w}_T \odot \mathbf{l}_T\|_2}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \\ &= \frac{\|\mathbf{w}_T\|_2}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} \\ &\le (\rho\sigma)^{-1/2} := a', \end{aligned} \tag{B.4}$$

where $C = (1 - 2\rho\sigma) \cdot \lambda \log^{\frac{1}{1+\alpha}}(4/\epsilon) + \sqrt{2\log(2/\epsilon)}$. We also have

$$\frac{b}{\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2} = ax + c,$$

and hence

$$\begin{aligned} &\mathbb{E}_{X \sim \mathcal{N}(b, \|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2^2)}[\mathrm{sign}(X) | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] \\ &= \frac{e^{-\frac{(ax+c)^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s \sim Q}\Big[\mathrm{sign}(\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2 \cdot s) \cdot e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{(ax+c)s} \Big| \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T \Big] \\ &= \frac{e^{-\frac{a^2x^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s \sim Q}\Big[\mathrm{sign}(\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2 \cdot s) \cdot e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{a(s-c)x + c(s-\frac{1}{2}c)} \Big| \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T \Big]. \end{aligned}$$

Then, we have

$$\begin{aligned} &\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})} \\ &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T}\Big[\mathbb{E}_{X \sim \mathcal{N}(b, \|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2^2)}[\mathrm{sign}(X) | \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T] \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \Big] \\ &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T}\Bigg[e^{-\frac{a^2x^2}{2}} \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c] \\ &\qquad \cdot \mathbb{E}_{s \sim Q}\Big[\mathrm{sign}(\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2 \cdot s) \cdot e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{a(s-c)x + c(s-\frac{1}{2}c)} \Big| \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T \Big] \Bigg]. \end{aligned}$$

We now define a polynomial $\widetilde{p}_{\mathbf{z}}(\mathbf{x})$ approximating $\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}$. To do this, we approximate $e^{-\frac{1}{2}a^2x^2}$ and $e^{a(s-c)x}$ using polynomials in x. First, we use a polynomial $p_1(x)$ to approximate $e^{-\frac{1}{2}a^2x^2}$. This polynomial is given by the following lemma. We choose the parameters later.

Lemma B.2 *Let $t \in \mathbb{Z}_+$. Let x be a random variable satisfying the (α, λ)-strictly sub-exponential tail bound. Then there exists a polynomial q of degree*

$$O\Big((a^2\lambda^2/2)^{1+1/\alpha} \big(Cb \log(ab\lambda\sqrt{\log(1/\epsilon)}) \big)^{2/\alpha} \big(\log(1/\epsilon) \big)^{\max\{1, 1/2+1/\alpha\}} C^{1/\alpha^2} \Big)$$

where C is a sufficiently large constant such that the approximation error $\mathbb{E}_x[(q(x) - e^{-\frac{1}{2}a^2x^2})^b]$ is upper bounded by 2ϵ.

Second, to approximate $e^{a(s-c)x}$, we use the function $p_2(x,s) = p_k(a(s-c)x)\,\mathbb{1}[|s| \le T]$ where $p_k(x) = 1 + \sum_{i=1}^{k-1} \frac{x^i}{i!}$ is the degree $k-1$ Taylor approximation of e^x. We choose degree k and threshold T later. Thus our final approximation of $\overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}$ is

$$\begin{aligned}
\widetilde{p}_{\mathbf{x}}(\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) &= p_1(x) \cdot \mathbb{E}_{s\sim Q}\big[\operatorname{sign}(\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2 \cdot s) \cdot e^{-\frac{s^2}{2} - \log Q(s)} \\
&\qquad \cdot e^{c(s-\frac{1}{2}c)} \cdot p_2(x,s) \big| \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T\big], \\
\widetilde{p}_{\mathbf{y}_H}(\mathbf{x}) &= \mathbb{E}_{\mathbf{l}_T, \mathbf{m}_T}\big[\widetilde{p}_{\mathbf{x}}(\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1]\big], \\
\widetilde{p}_{\mathbf{z}}(\mathbf{x}) &= \mathbb{E}_{\mathbf{y}_H}\big[\widetilde{p}_{\mathbf{y}_H}(\mathbf{x}) \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\big].
\end{aligned} \tag{B.5}$$

We now want to bound the L_1 error term $\mathbb{E}_{\mathbf{x}\sim D_{\mathbf{x}}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|\widetilde{p}_{\mathbf{z}}(\mathbf{x}) - \overline{T_\rho f_{\mathbf{x}}(\mathbf{z})}|]$. To help us analyse the error, we define the "hybrid" function $\bar{p}_{\mathbf{z}}(\mathbf{x})$ such that

$$\begin{aligned}
\bar{p}_{\mathbf{x}}(\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) &= \frac{e^{-\frac{a^2x^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s\sim\mathcal{N}(0,1)}\big[\operatorname{sign}(\|\mathbf{w}_T \odot \mathbf{l}_T \odot \mathbf{m}_T\|_2 \cdot s) \cdot e^{-\frac{s^2}{2} - \log Q(s)} \\
&\qquad \cdot e^{c(s-\frac{1}{2}c)} \cdot p_2(x,s) \big| \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T\big], \\
\bar{p}_{\mathbf{z}}(\mathbf{x}) &= \mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T}\big[\bar{p}_{\mathbf{x}}(\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\big].
\end{aligned}$$

We have that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|\widetilde{p}_{\mathbf{z}}(\mathbf{x}) - \overline{T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})}|] \\
&\le 2 \cdot \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\Big[\underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|\bar{p}_{\mathbf{z}}(\mathbf{x}) - \overline{T_{1-\rho} f_{\mathbf{x}}(\mathbf{z})}|]}_{\Delta_3(\mathbf{x})} + \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}[|\widetilde{p}_{\mathbf{z}}(\mathbf{x}) - \bar{p}_{\mathbf{z}}(\mathbf{x})|]}_{\Delta_4(\mathbf{x})}\Big].
\end{aligned}$$

We now bound $\Delta_3(\mathbf{x})$ and $\Delta_4(\mathbf{x})$ separately. We have that

$$\begin{aligned}
\Delta_3(\mathbf{x}) \le \mathbb{E}_{\mathbf{z}}\Bigg[&\mathbb{E}_{\mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T}\Big[e^{-\frac{1}{2}a^2x^2} \cdot \mathbb{E}_{s\sim\mathcal{N}(0,1)}\big[e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{c(s-\frac{1}{2}c)} \cdot |e^{a(s-c)x} - p_2(x,s)|\big] \\
&\cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\Big]\Bigg].
\end{aligned}$$

Observe that $\Delta_3(\mathbf{x})$ can be bounded as the expected sum of the following two terms:

$$\begin{aligned}
&\Delta_{31}(\mathbf{x}, \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \\
&= \frac{e^{-\frac{a^2x^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s\sim Q}\left[e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{c(s-\frac{1}{2}c)} \cdot \frac{e^{|a(s-c)x|}}{k!} \cdot |a(s-c)x|^k \cdot \mathbb{1}[|s| \le T]\right], \\
&\Delta_{32}(\mathbf{x}, \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \\
&= \frac{e^{-\frac{a^2x^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s\sim Q}\big[e^{-\frac{s^2}{2} - \log Q(s)} \cdot e^{c(s-\frac{1}{2}c)} \cdot e^{a(s-c)x} \cdot \mathbb{1}[|s| > T]\big],
\end{aligned}$$

where the first term's bound comes from the fact that $|p_k(x) - e^x| \le \frac{e^{|x|}}{k!} \cdot |x|^k$.

We first bound Δ_{31}. We have that

$$\begin{aligned}
&\Delta_{31}(\mathbf{x}, \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \\
&\leq \frac{e^{-\frac{a^2x^2}{2}}}{\sqrt{2\pi}} \cdot \mathbb{E}_{s\sim Q}\left[e^{-\frac{s^2}{2}-\log Q(s)} \cdot e^{c(s-\frac{1}{2}c)} \cdot e^{|a(s-c)x|} \cdot \mathbb{1}[|s| \leq T]\right] \cdot \frac{(|acx|+|aTx|)^k}{k!} \\
&\leq \frac{(|acx|+|aTx|)^k}{k!} \cdot \mathbb{E}_{s\sim Q}\left[\frac{e^{-\frac{1}{2}(s-(ax+c))^2}}{\sqrt{2\pi}\cdot Q(s)} \cdot \mathbb{1}[|s| \leq T]\right] \\
&\quad + \frac{(|acx|+|aTx|)^k}{k!} \cdot \mathbb{E}_{s\sim Q}\left[\frac{e^{-\frac{1}{2}(s-(c-ax))^2}}{\sqrt{2\pi}\cdot Q(s)} \cdot \mathbb{1}[|s| \leq T]\right] \\
&\leq \frac{(|acx|+|aTx|)^k}{(k!)^2} \cdot \mathbb{E}_{s\sim Q}\left[\frac{\mathcal{N}(s; ax+c, 1)}{Q(s)}\right] \\
&\quad + \frac{(|acx|+|aTx|)^{2k}}{(k!)^2} \cdot \mathbb{E}_{s\sim Q}\left[\frac{\mathcal{N}(s; c-ax, 1)}{Q(s)}\right] \\
&= \frac{2(|acx|+|aTx|)^k}{k!}
\end{aligned}$$

where the second inequality is by $e^{|a(s-c)x|} \leq e^{a(s-c)x} + e^{-a(s-c)x}$.

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\left[\Delta_{31}(\mathbf{x}, \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\right]\right]\right] \\
&\leq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\left[\frac{(|acx|+|aTx|)^k}{k!} \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\right]\right]\right] \\
&= 2\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\mathbb{E}_{\mathbf{y}_H,\mathbf{m}_T}\left[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\frac{(|acx|+|aTx|)^k}{k!} \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\right]\right]\right].
\end{aligned}$$

Under condition $(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1$ and condition $\mathbf{y}_H \in \mathcal{E}^c$, we have (B.3) and (B.4) and hence

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\frac{(|acx|+|aTx|)^k}{k!} \cdot \mathbb{1}[(\mathbf{l}_T, \mathbf{m}_T) \in \mathcal{E}_1] \cdot \mathbb{1}[\mathbf{y}_H \in \mathcal{E}^c]\right] \\
&\leq \frac{|a'|^k(|c'|+|T|)^k}{k!} \cdot \mathbb{E}_{\mathbf{x}\sim D}[|x|^k] \\
&\leq \frac{|a'|^k(|c'|+|T|)^k}{k!} \cdot C\lambda^k \cdot \Big(\frac{k}{1+\alpha}\Big)^{\frac{k}{1+\alpha}+\frac{1}{2}} e^{-\frac{k}{1+\alpha}+\frac{1+\alpha}{12k}} \\
&\leq \epsilon,
\end{aligned}$$

when

$$k \geq \big(C'|a'|(|c'|+|T|)\lambda\log(1/\epsilon)\big)^{1+\frac{1}{\alpha}} + (1+\alpha) \tag{B.6}$$

and C' is a large enough constant. The second inequality used Lemma C.5.

We now bound Δ_{32}. We have that

$$\begin{aligned}
\Delta_{32}(\mathbf{x}, \mathbf{y}_H, \mathbf{l}_T, \mathbf{m}_T) &= \mathbb{E}_{s\sim Q}\left[\frac{e^{-\frac{1}{2}(s-(ax+c))^2}}{Q(s)} \cdot \mathbb{1}[|s| > T]\right] \\
&\leq \sqrt{\mathbb{E}_{s\sim Q}\left[\left(\frac{\mathcal{N}(s; ax+c, 1)}{Q(s)}\right)^2\right] \cdot \mathbb{P}_{s\sim Q}\big[|s| > T\big]} \\
&\leq \sqrt{Ce^{|ax+c|} \cdot e^{-T}} \\
&\leq C'e^{\frac{1}{2}(|ax|+|c|)} \cdot e^{-T/2}.
\end{aligned}$$

The third inequality is based on the following claim.

Lemma B.3 *Define the distribution Q on $\mathbb{R}$ with density function $Q(s) = e^{-|s|}/2$. Then there exist a universal constant C such that for every $b \in \mathbb{R}$, it holds that*

$$\mathbb{E}_{s\sim Q}\Big[\Big(\frac{\mathcal{N}(s;b,1)}{Q(s)}\Big)^2\Big] \le Ce^{|b|}.$$

Thus, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\Big[\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\big[\Delta_{32}(\mathbf{x},\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T)\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big]\Big]\\
&\le \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\Big[\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\big[C'e^{\frac{1}{2}(|ax|+|c|)}\cdot e^{-T/2}\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big]\Big]\\
&= \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\big[C'e^{\frac{1}{2}(|ax|+|c|)}\cdot e^{-T/2}\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big]\Big]\\
&\le \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\big[C'e^{\frac{1}{2}|a'x|}\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\cdot e^{\frac{|c'|-T}{2}}\Big]\Big]\\
&\le \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Big[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\big[C''e^{(a'\lambda)^{1+\frac{1}{\alpha}}/2}\cdot e^{\frac{|c'|-T}{2}}\big]\Big]\Big]\\
&= C''e^{(a'\lambda)^{1+\frac{1}{\alpha}}/2}\cdot e^{\frac{|c'|-T}{2}}\\
&\le \epsilon,
\end{aligned}$$

when

$$T \ge (a'\lambda)^{1+\frac{1}{\alpha}} + |c'| + 2\log(C''/\epsilon).$$

By (B.3) and (B.4), we can take

$$T = O\Big((\rho\sigma)^{-\frac{1}{2}(1+\frac{1}{\alpha})}\lambda^{1+\frac{1}{\alpha}}\log(1/\epsilon)\Big). \tag{B.7}$$

The third last inequality is by Lemma C.6. Plugging this into the bound for k in (B.6), we can take

$$k \le \Big(C'a'\lambda\big(2c' + (a'\lambda)^{1+\frac{1}{\alpha}} + 2\log(C''/\epsilon)\big)\log(1/\epsilon)\Big)^{1+\frac{1}{\alpha}} + (1+\alpha)$$

for constant C', C''. By (B.3) and (B.4), we know that

$$k \le \Big(C'''\lambda^{2+\frac{1}{\alpha}}(\rho\sigma)^{-(1+\frac{1}{2\alpha})}\log^3(1/\epsilon)\Big)^{1+\frac{1}{\alpha}} + (1+\alpha), \tag{B.8}$$

where C''' is a large constant.

We now bound $\Delta_4(\mathbf{x})$. We have that

$$\begin{aligned}
&\Delta_4(\mathbf{x})\\
&\le \mathbb{E}_\mathbf{z}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\big[\big|\widetilde{p}_\mathbf{x}(\mathbf{y}_H,\mathbf{l}_H,\mathbf{m}_T) - \bar{p}_\mathbf{x}(\mathbf{y}_H,\mathbf{l}_H,\mathbf{m}_T)\big|\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big]\\
&= \frac{1}{\sqrt{2\pi}}\mathbb{E}_\mathbf{z}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\big[\big|e^{-\frac{1}{2}a^2x^2} - p_1(x)\big|\cdot\big|\mathbb{E}_{s\sim Q}[e^{-\frac{s^2}{2}-\log Q(s)+c(s-\frac{c}{2})}\cdot p_2(x,s)]\big|\\
&\qquad\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big]
\end{aligned}$$

Notice that

$$\begin{aligned}
&\big(\mathbb{E}_{s\sim Q}[e^{-\frac{s^2}{2}-\log Q(s)+c(s-\frac{c}{2})}\cdot p_2(x,s)]\big)^2\\
&\le \mathbb{E}_{s\sim Q}[e^{-s^2-2\log Q(s)+c(2s-c)}]\cdot\mathbb{E}_{s\sim Q}[p_2(x,s)^2]\\
&\le \mathbb{E}_{s\sim Q}\Big[\Big(\frac{\mathcal{N}(s;c,1)}{Q(s)}\Big)^2\Big]\cdot\mathbb{E}_{s\sim Q}\Big[\Big(\sum_{i=0}^{k-1}\frac{(a(s-c)x)^i}{i!}\Big)^2\cdot\mathbb{1}[|s|\le T]\Big]\\
&\le Ce^{|c|}\cdot\Big(\sum_{i=0}^{k-1}\frac{|a|^i(|T|+|c|)^i|x|^i}{i!}\Big)^2,
\end{aligned}$$

where the last inequality is by Lemma B.3. Thus, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\Delta_4(\mathbf{x})] \\
&\le \mathbb{E}_{\mathbf{z}}\Big[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\big[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\big[\big|e^{-\frac{1}{2}a^2x^2}-p_1(x)\big|\cdot\big|\mathbb{E}_{s\sim Q}[e^{-\frac{s^2}{2}-\log Q(s)+c(s-\frac{c}{2})}\cdot p_2(x,s)]\big|\big] \\
&\qquad\qquad \cdot \mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Big] \\
&\le \mathbb{E}_{\mathbf{z}}\Bigg[\mathbb{E}_{\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T}\Bigg[\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\big[\big|e^{-\frac{1}{2}a^2x^2}-p_1(x)\big|\cdot C^{\frac{1}{2}}e^{\frac{|c|}{2}}\cdot\Big(\sum_{i=0}^{k-1}\frac{|a|^i(|T|+|c|)^i|x|^i}{i!}\Big) \\
&\qquad\qquad \cdot \mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]\Bigg]\Bigg]
\end{aligned}
$$

Denote

$$
\begin{aligned}
\Delta_5(\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\Bigg[&\big|e^{-\frac{1}{2}a^2x^2}-p_1(x)\big|\cdot C^{\frac{1}{2}}e^{\frac{|c|}{2}}\cdot\Big(\sum_{i=0}^{k-1}\frac{|a|^i(|T|+|c|)^i|x|^i}{i!}\Big) \\
&\cdot \mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Bigg].
\end{aligned}
$$

We have

$$
\begin{aligned}
&\Delta_5(\mathbf{y}_H,\mathbf{l}_T,\mathbf{m}_T) \\
&\le C'e^{|c'|/2}\sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\big[\big(e^{-\frac{1}{2}a^2x^2}-p_1(x)\big)^2\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\big]} \\
&\qquad\cdot\sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\Big[\Big(\sum_{i=0}^{k-1}\frac{|a|^i(|T|+|c|)^i|x|^i}{i!}\Big)^2\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Big]} \\
&\le C'e^{|c'|/2}\cdot\delta\cdot\sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\Big[\Big(\sum_{i=0}^{k-1}\frac{|a|^i(|T|+|c|)^i|x|^i}{i!}\Big)^2\cdot\mathbb{1}[(\mathbf{l}_T,\mathbf{m}_T)\in\mathcal{E}_1]\cdot\mathbb{1}[\mathbf{y}_H\in\mathcal{E}^c]\Big]} \\
&\le C'e^{|c'|/2}\cdot\delta\cdot\sqrt{\Big(\sum_{i=0}^{k-1}\frac{|a'|^i(|T|+|c'|)^i}{i!}\Big)^2\cdot\max_{1\le i\le k-1}C\lambda^{2i}\cdot\Big(\frac{2i}{1+\alpha}\Big)^{\frac{2i}{1+\alpha}+\frac{1}{2}}e^{-\frac{2i}{1+\alpha}+\frac{1+\alpha}{24i}}} \\
&\le C''\delta e^{|c'|/2+|a'|(|T|+|c'|)}\lambda^k\Big(\frac{2k}{1+\alpha}\Big)^{\frac{k}{1+\alpha}+\frac{1}{4}}e^{\frac{1+\alpha}{48}} \\
&\le \epsilon,
\end{aligned}
$$

when δ is chosen accordingly. The third inequality is by Lemma C.5.

By Lemma B.2 and taking the exponent b as 2, the degree of $p_1(x)$ required to get the error is

$$
\begin{aligned}
&O\Big((a^2\lambda^2/2)^{1+1/\alpha}\big(C\log(a\lambda\sqrt{\log(1/\delta)})\big)^{2/\alpha}\big(\log(1/\delta)\big)^{\max\{1,\frac{1}{2}+\frac{1}{\alpha}\}}C^{1/\alpha^2}\Big) \\
&= O\Big((a'^2\lambda^2/2)^{1+1/\alpha}\big(C\log(a'\lambda)\big)^{2/\alpha} \\
&\qquad\cdot\big(\log(1/\epsilon)+|a'|(|T|+|c'|)+(k+1/4)\log(2k\lambda)\big)^{2/\alpha+1}C^{1/\alpha^2}\Big) \\
&= O\Big((a'^2\lambda^2/2)^{1+1/\alpha}\big(C\log(a'\lambda)\big)^{2/\alpha}\big(\log(1/\epsilon)+|a'|(|T|+|c'|)+k^{3/2}\lambda^{1/2}\big)^{2/\alpha+1}C^{1/\alpha^2}\Big).
\end{aligned}
$$

By using (B.3) and (B.4) and plugging in the order of k, T in (B.7) and (B.8), we know the degree of $p_1(x)$ is

$$
\begin{aligned}
&O\Big(((\rho\sigma)^{-1}\lambda^2/2)^{1+\frac{1}{\alpha}}\big(C\log((\rho\sigma)^{-\frac{1}{2}}\lambda)\big)^{\frac{2}{\alpha}} \\
&\qquad\cdot\big(C\lambda^{\frac{3}{2}(2+\frac{1}{\alpha})(1+\frac{1}{\alpha})+\frac{1}{2}}(\rho\sigma)^{-\frac{3}{2}(1+\frac{1}{2\alpha})(1+\frac{1}{\alpha})}\log^{\frac{9}{2}(1+\frac{1}{\alpha})}(1/\epsilon)\big)^{1+\frac{2}{\alpha}}C^{\frac{1}{\alpha^2}}\Big) \\
&= O\Big(\big(C(\rho\sigma)^{-\frac{1}{2}}\lambda\log(1/\epsilon)\big)^{6(1+\frac{1}{\alpha})^3}\Big).
\end{aligned}
$$

Putting everything together, we get that the degree of $p_{\mathbf{z}}(\mathbf{x})$ is at most $2H + \deg(p_1) + \deg(p_2)$ which is

$$O\big(\log(1+\lambda)/\alpha^2 + \log(1/\epsilon)\log(1/\alpha)/\rho\sigma\alpha^2\big) + O\Big(\big(C(\rho\sigma)^{-\frac{1}{2}}\lambda\log(1/\epsilon)\big)^{6(1+\frac{1}{\alpha})^3}\Big).$$

Plugging in the order of $\rho = \Omega(\epsilon^2)$ and $\alpha = \Omega(\rho\sigma/\sqrt{\log(1/\epsilon)})$, the degree can be bounded by

$$O\Big(\big(C\sigma^{-\frac{1}{2}}\lambda\log(1/\epsilon)/\epsilon\big)^{6(1+\frac{1}{\alpha})^3}\Big).$$

C Proofs of Auxiliary Lemmas

C.1 Proof of Theorem 4.2

Proof Let f^* be the optimal halfspace that achieves opt_σ. Let $p_{\mathbf{z}}$ be the polynomial of degree at most d such that

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_\sigma,\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p_{\mathbf{z}}(\mathbf{x}) - f^*(\mathbf{x}\odot\mathbf{z})|] \le \epsilon. \tag{C.1}$$

Consider the sample dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in a single run of Algorithm 1. Let p_S be the polynomial chosen by the algorithm and let h_S be the corresponding hypothesis that the algorithm outputs. By the proof of Theorem 4.1 in Kalai et al. (2008), we have that

$$\frac{1}{N}\sum_{i=1}^N \mathbb{1}[h_S(\mathbf{x}_i) \neq y_i] \le \frac{1}{2N}\sum_{i=1}^N |y_i - p_S(\mathbf{x}_i)|.$$

Notice that p_S is the minimizer of the error, and thus beats any polynomial $p_{\mathbf{z}}$ we choose, we have

$$\begin{aligned}\frac{1}{2N}\sum_{i=1}^N |y_i - p_S(\mathbf{x}_i)| &\le \mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\frac{1}{2N}\sum_{i=1}^N |y_i - p_{\mathbf{z}}(\mathbf{x}_i)|\right] \\ &\le \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\frac{1}{2N}\sum_{i=1}^N |f^*(\mathbf{x}_i\odot\mathbf{z}) - p_{\mathbf{z}}(\mathbf{x}_i)|\right]}_{\Delta_1(S)} + \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{N}_\sigma}\left[\frac{1}{2N}\sum_{i=1}^N |y_i - f^*(\mathbf{x}_i\odot\mathbf{z})|\right]}_{\Delta_2(S)}.\end{aligned}$$

By (C.1), we can bound $\Delta_1(S)$ as follows:

$$\mathbb{E}_{S\sim\mathcal{D}^{\otimes n}}[\Delta_1(S)] = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}},\mathbf{z}\sim\mathcal{D}_\sigma}[|f^*(\mathbf{x}\odot\mathbf{z}) - p_{\mathbf{z}}(\mathbf{x})|] \le \frac{1}{2}\epsilon.$$

By the optimality of f^*, we have

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^{\otimes n}}[\Delta_1(S)] &= \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D},\mathbf{z}\sim\mathcal{D}_\sigma}[|y - f^*(\mathbf{x}\odot\mathbf{z})|] \\ &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D},\mathbf{z}\sim\mathcal{D}_\sigma}[\mathbb{1}[y \neq f^*(\mathbf{x}\odot\mathbf{z})]] = \mathrm{opt}_\sigma.\end{aligned}$$

Thus, we obtain

$$\mathbb{E}_{S\sim\mathcal{D}^{\otimes n}}\left[\frac{1}{N}\sum_{i=1}^N \mathbb{1}[h_S(\mathbf{x}_i) \neq y_i]\right] \le \mathrm{opt}_\sigma + \frac{1}{2}\epsilon.$$

Since our hypothesis h_S is a polynomial threshold function of degree d on n variables, VC theory tells us that for $N = \mathrm{poly}(n^d/\epsilon)$, we have that

$$\mathbb{E}_{S\sim\mathcal{D}^{\otimes n}}\big[\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_S(\mathbf{x}) \neq y]\big] \le \mathrm{opt}_\sigma + \frac{3}{4}\epsilon.$$

By Markov's inequality, on any single repetition of the algorithm, we have that

$$\mathbb{P}_{S\sim\mathcal{D}^{\otimes n}}\Big[\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_S(\mathbf{x}) \neq y] \ge \mathrm{opt}_\sigma + \frac{7}{8}\epsilon\Big] \le \frac{\mathrm{opt}_\sigma + \frac{3}{4}\epsilon}{\mathrm{opt}_\sigma + \frac{7}{8}} \le 1 - \frac{\epsilon}{16}.$$

Hence, after $r = O(\log(1/\delta)/\epsilon)$ repetitions of the algorithm, with probability at least $1 - \delta/2$, one of them will have $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[h_S(\mathbf{x}) \neq y] \le \mathrm{opt}_\sigma + \frac{7}{8}\epsilon$. In this case, using an independent set of size $O(\log(1/\delta)/\epsilon^2)$, we probability at most $\delta/2$, we will choose one with error $> \mathrm{opt}_\sigma + \epsilon$. ■

C.2 Proof of Lemma B.3

Proof The proof below is straightforward calculation by completing the squares.

$$\begin{aligned}
\mathbb{E}_{s\sim Q}\left[\Big(\frac{\mathcal{N}(s;b,1)}{Q(s)}\Big)^2\right] &= \frac{1}{2}\int_{\mathbb{R}} \frac{\frac{1}{2\pi}e^{-(s-b)^2}}{\frac{1}{4}e^{-2|s|}}\cdot e^{-|s|}ds \\
&= \frac{1}{\pi}\int_{\mathbb{R}} e^{-(s-b)^2+|s|}ds \\
&\le \frac{1}{\pi}\int_{\mathbb{R}} e^{-(s-b)^2+s}ds + \frac{1}{\pi}\int_{\mathbb{R}} e^{-(s-b)^2-s}ds \\
&= \frac{1}{\pi}\int_{\mathbb{R}} e^{-(s-b+\frac{1}{2})^2+b+\frac{1}{4}}ds + \frac{1}{\pi}\int_{\mathbb{R}} e^{-(s-b-\frac{1}{2})^2-b+\frac{1}{4}}ds \\
&= \frac{1}{\sqrt{\pi}}e^{b+\frac{1}{4}} + \frac{1}{\sqrt{\pi}}e^{-b+\frac{1}{4}} \\
&\le \frac{2e^{\frac{1}{4}}}{\sqrt{\pi}}e^{|b|}
\end{aligned}$$

■

C.3 Proof of Lemma B.2

Proof Let $p(x) = \sum_{i=0}^{\deg(p)} c_i x^i$ be the polynomial obtained from Lemma C.4 with error $\epsilon/2$ and $T = \omega(\log(1/\epsilon))$ to be choosen later. Our final polynomial is $q(x) = p(\frac{1}{2}a^2x^2)$. Clearly, $\deg(q) = 2\cdot\deg(p) = O(\sqrt{T\log(1/\epsilon)})$. We now bound the error.

$$\begin{aligned}
&\mathbb{E}_x[(q(x) - e^{-\frac{1}{2}a^2x^2})^b] \\
&= \mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^b\cdot \mathbb{1}[a^2x^2/2 < T]\big] + \mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^b\cdot \mathbb{1}[a^2x^2/2 \ge T]\big] \\
&= \epsilon + \sqrt{\mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^{2b}\big]\cdot \mathbb{E}_x\big[\mathbb{1}[a^2x^2/2 \ge T]\big]} \\
&\le \epsilon + \sqrt{\mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^{2b}\big]\cdot 2e^{-(\frac{2T}{a^2\lambda^2})^{\frac{1+\alpha}{2}}}},
\end{aligned}$$

where the last inequality is by the tail bound of (α,λ)-strictly sub-exponential random variable.

We now bound $\mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^{2b}\big]$. We have that

$$\begin{aligned}
\mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^{2b}\big] &\le \mathbb{E}_x\big[(|q(x)|+1)^{2b}\big] \\
&\le \mathbb{E}_x\left[\left(1 + \sum_{i=0}^{\deg(p)}\frac{|c_i|}{2^i}a^{2i}x^{2i}\right)^{2b}\right] \\
&\le (\deg(p)+2)^{2b}\cdot \max_{i=0}^{\deg(p)}|c_i|^{2b}\cdot \max_{i=0}^{\deg(p)}\frac{a^{4bi}\mathbb{E}_x[x^{4bi}]}{2^{2bi}} \\
&\le e^{Cb\sqrt{T\log(1/\epsilon)}}\cdot \max_{i=0}^{\deg(p)}\frac{a^{4bi}\mathbb{E}_x[x^{4bi}]}{2^{2bi}} \\
&\le e^{Cb\sqrt{T\log(1/\epsilon)}}\cdot \max_{i=1}^{\deg(p)}\frac{a^{4bi}}{2^{2bi}}\cdot\frac{C'bi\lambda^{4bi}}{1+\alpha}\cdot\Big(\frac{4bi}{1+\alpha}\Big)^{\frac{4bi}{1+\alpha}-\frac{1}{2}}e^{-\frac{4bi}{1+\alpha}+\frac{1+\alpha}{48bi}} \\
&\le e^{Cb\sqrt{T\log(1/\epsilon)}+C'b\sqrt{T\log(1/\epsilon)}\cdot\log(ab\lambda\sqrt{T\log(1/\epsilon)})+\frac{1+\alpha}{48b}} \\
&\le e^{C''b(1+\alpha)\sqrt{T\log(1/\epsilon)}\cdot\log(ab\lambda\sqrt{T\log(1/\epsilon)})},
\end{aligned}$$

where the third last inequality is by Lemma C.5. Here C, C', C'' are large enough constant. Putting it all together, we get that

$$\mathbb{E}_x\big[(q(x) - e^{-\frac{1}{2}a^2x^2})^{2b}\big]\cdot 2e^{-(\frac{2T}{a^2\lambda^2})^{\frac{1+\alpha}{2}}} \le 2e^{C''b\sqrt{T\log(1/\epsilon)}\cdot\log(ab\lambda\sqrt{T\log(1/\epsilon)})-(\frac{2T}{a^2\lambda^2})^{\frac{1+\alpha}{2}}}.$$

Choosing

$$T = \Omega\Big((a^2\lambda^2/2)^{2(1+1/\alpha)}\big(C'''b^2\log^2(ab\lambda\sqrt{\log(1/\epsilon)})\big)^{2/\alpha}\big(\log(1/\epsilon)\big)^{\max\{1,2/\alpha\}}(C''')^{1/\alpha^2}\Big)$$

where C''' is a large constant makes the total error less than 2ϵ. Since T is $\omega(\log(1/\epsilon))$, the degree of the final polynomial is $O(\sqrt{T\log(1/\epsilon)})$ which is

$$T = O\Big((a^2\lambda^2/2)^{1+1/\alpha}\big(C'''b^2\log^2(ab\lambda\sqrt{\log(1/\epsilon)})\big)^{1/\alpha}\big(\log(1/\epsilon)\big)^{\max\{1,1/2+1/\alpha\}}(C''')^{1/2\alpha^2}\Big).$$

■

C.4 Other Auxiliary Lemmas

Lemma C.1 *Suppose* $\mathbf{x}$ *is a distribution on* $\{\pm1\}^n$ *that is* (α,λ)*-strictly subexponential and* $\mathbf{y}$ *distributed as* $\mathcal{N}_{1-\rho}(\mathbf{z})$*. For any fixed* $T\subseteq[n]$ *and fixed* $\mathbf{z}$*, with probability at least* $1-\epsilon$ *of* $\mathbf{x}$*, it holds that*

$$\mathbb{P}_{\mathbf{y}}\Big[|\langle\mathbf{u}_T,\mathbf{y}_T\rangle|\le C\cdot\|\mathbf{u}_T\|_2\Big]\ge 1-\epsilon,$$

where $\mathbf{u}=\mathbf{w}\odot\mathbf{x}$ *and* $C=(1-2\rho\sigma)\cdot\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)+\sqrt{2\log(2/\epsilon)}$.

Proof By Hoeffding's inequality, we have

$$\mathbb{P}_{\mathbf{y}}\Big[\big|\langle\mathbf{u}_T,\mathbf{y}_T\rangle-\mathbb{E}_{\mathbf{y}}[\langle\mathbf{u}_T,\mathbf{y}_T\rangle]\big|\ge t\Big]\le 2\exp\Big(-\frac{t^2}{2\|\mathbf{u}_T\|_2^2}\Big).$$

Therefore, with probability at least $1-\epsilon$ it holds that

$$\langle\mathbf{u}_T,\mathbf{y}_T\rangle\in\Big[\mathbb{E}_{\mathbf{y}}[\langle\mathbf{u}_T,\mathbf{y}_T\rangle]-\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2,\mathbb{E}_{\mathbf{y}}[\langle\mathbf{u}_T,\mathbf{y}_T\rangle]+\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2\Big].$$

Notice that

$$\begin{aligned}\mathbb{E}_{y_i}[u_iy_i]&=u_i\mathbb{E}[y_i]=u_i\big((1-\rho)z_i+\rho(1-2\sigma)\big),\\ \mathbb{E}_{\mathbf{y}}[\langle\mathbf{u}_T,\mathbf{y}_T\rangle]&=(1-\rho)\sum_{i\in T}u_iz_i+\rho(1-2\sigma)\sum_{i\in T}u_i\\ &=(1-\rho)\cdot\langle\mathbf{w}_T\odot\mathbf{z}_T,\mathbf{x}_T\rangle+\rho(1-2\sigma)\cdot\langle\mathbf{w}_T,\mathbf{x}_T\rangle,\end{aligned}$$

since $\mathbf{x}$ is (α,λ)-strictly subexponential, we have

$$\mathbb{P}_{\mathbf{x}}\Big[|\langle\mathbf{w}_T,\mathbf{x}_T\rangle|>t\cdot\|\mathbf{w}_T\|_2\Big]\le 2\exp\big(-(t/\lambda)^{1+\alpha}\big),$$

and

$$\mathbb{P}_{\mathbf{x}}\Big[|\langle\mathbf{w}_T\odot\mathbf{z}_T,\mathbf{x}_T\rangle|>t\cdot\|\mathbf{w}_T\|_2\Big]\le 2\exp\big(-(t/\lambda)^{1+\alpha}\big).$$

Thus, with probability at least $1-\epsilon$ of $\mathbf{x}$ it holds that

$$|\langle\mathbf{w}_T,\mathbf{x}_T\rangle|\le\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)\cdot\|\mathbf{w}_T\|_2,$$

and

$$|\langle\mathbf{w}_T\odot\mathbf{z}_T,\mathbf{x}_T\rangle|\le\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)\cdot\|\mathbf{w}_T\|_2.$$

Notice that $\mathbf{x}$ is distributed on $\{\pm1\}^n$ and hence $\|\mathbf{u}_T\|_2=\|\mathbf{w}_T\odot\mathbf{x}_T\|_2=\|\mathbf{w}_T\|_2$, then with probability at least $1-\epsilon$ of $\mathbf{y}$ it holds that

$$\begin{aligned}|\langle\mathbf{u}_T,\mathbf{y}_T\rangle|&\le\big|\mathbb{E}_{\mathbf{y}}[\langle\mathbf{u}_T,\mathbf{y}_T\rangle]\big|+\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2\\ &\le(1-\rho)\cdot|\langle\mathbf{w}_T\odot\mathbf{z}_T,\mathbf{x}_T\rangle|+\rho(1-2\sigma)\cdot|\langle\mathbf{w}_T,\mathbf{x}_T\rangle|+\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2\\ &\le\big((1-\rho)+\rho(1-2\sigma)\big)\cdot\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)\cdot\|\mathbf{w}_T\|_2+\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2\\ &\le(1-2\rho\sigma)\cdot\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)\cdot\|\mathbf{w}_T\|_2+\sqrt{2\log(2/\epsilon)}\cdot\|\mathbf{u}_T\|_2\\ &=C\cdot\|\mathbf{u}_T\|_2,\end{aligned}\tag{C.2}$$

where $C=(1-2\rho\sigma)\cdot\lambda\log^{\frac{1}{1+\alpha}}(4/\epsilon)+\sqrt{2\log(2/\epsilon)}$. ■

Lemma C.2 *Suppose that the vector $\mathbf{w} \in \mathbb{R}^n$ is α-regular, i.e., $\|\mathbf{w}\|_\infty \le \alpha \cdot \|\mathbf{w}\|_2$. Suppose $\mathbf{u}$ is a n-dimensional random vector where each coordinate is 1 with probability ρ and 0 with probability $1-\rho$ independently. If $\alpha \le \rho/\sqrt{\log(1/\delta)/2}$, then with probability at least $1-\delta$ over the randomness of $\mathbf{u}$ it holds that*

$$\|\mathbf{w}\|_\infty \le \alpha \cdot \|\mathbf{w}\|_2 \le c \cdot \alpha \cdot \|\mathbf{w} \odot \mathbf{u}\|_2,$$

where $c = (\rho - \sqrt{\log(1/\delta)/2} \cdot \alpha)^{-\frac{1}{2}}$. If condition $\alpha \le \rho/\sqrt{2\log(1/\delta)}$ holds, then with probability at least $1-\delta$ over the randomness of $\mathbf{u}$ it holds that

$$\|\mathbf{w}\|_\infty \le \alpha \cdot \|\mathbf{w}\|_2 \le (\rho/2)^{-\frac{1}{2}} \cdot \alpha \cdot \|\mathbf{w} \odot \mathbf{u}\|_2.$$

Proof By Hoeffding's inequality and notice that $\mathbb{E}[(w_i u_i)^2] = w_i^2 \cdot \mathbb{E}[u_i] = \rho w_i^2$ and $0 \le (w_i u_i)^2 \le w_i^2$, we obtain

$$\begin{aligned}
\mathbb{P}\Big(\|\mathbf{w} \odot \mathbf{u}\|_2^2 - \rho \cdot \|\mathbf{w}\|_2^2 \le -t\Big) &\le \exp\bigg(-\frac{2t^2}{\|\mathbf{w}\|_4^4}\bigg) \\
&\le \exp\bigg(-\frac{2t^2}{\|\mathbf{w}\|_\infty^2 \|\mathbf{w}\|_2^2}\bigg) \\
&\le \exp\bigg(-\frac{2t^2}{\alpha^2 \cdot \|\mathbf{w}\|_2^4}\bigg),
\end{aligned}$$

where the second inequality is by the definition of the infinity norm, and the third inequality is by the α-regularity condition. Therefore, with probability at least $1-\delta$, we can get

$$\|\mathbf{w} \odot \mathbf{u}\|_2^2 \ge \rho \cdot \|\mathbf{w}\|_2^2 - \sqrt{\frac{\log(1/\delta)}{2}} \cdot \alpha \cdot \|\mathbf{w}\|_2^2.$$

Then, with probability at least $1-\delta$, it holds that

$$\|\mathbf{w}\|_\infty \le \alpha \cdot \|\mathbf{w}\|_2 \le \alpha \cdot \bigg(\rho - \sqrt{\frac{\log(1/\delta)}{2}} \cdot \alpha\bigg)^{-\frac{1}{2}} \cdot \|\mathbf{w} \odot \mathbf{u}\|_2.$$

■

Theorem C.3 (Berry-Esseen CLT) *Let $X_1, X_2, ...,$ be independent random variables with $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma_i^2 > 0,$ and $\mathbb{E}[|X_i|^3] = \rho_i < \infty$. Also, let*

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}}$$

be the normalized n-th partial sum. Denote F_n the cdf of S_n, and Φ the cdf of the standard normal distribution. Then for all n there exists an absolute constant C such that

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le C \cdot \Big(\sum_{i=1}^n \sigma_i^2\Big)^{-\frac{3}{2}} \cdot \sum_{i=1}^n \rho_i.$$

Lemma C.4 (Lemma D.11 in Chandrasekaran et al. (2024)) *For $T > 0$ and error $\epsilon > 0$, there exists a polynomial p such that*

1. $\sup_{x \in [0,T]} |p(x) - e^{-x}| \le \epsilon$.
2. $\deg(p) \le O(\sqrt{T \log(1/\epsilon)})$, *if* $T = \omega(\log(1/\epsilon))$.
3. $p(x) = \sum_{i=1}^{\deg(p)} c_i x^i$ *where* $|c_i| \le e^{C\sqrt{T\log(1/\epsilon)}}$ *for all* $i \le \deg(p)$. *Here C is a large enough constant.*

Lemma C.5 *If x is (α, λ)-strictly sub-exponential random variable satisfying $\mathbb{P}_x[|x| > t] \le 2 \cdot e^{-(t/\lambda)^{1+\alpha}}$, then the k-th moment is upper bounded by:*

$$E_x[|x|^k] \le C\lambda^k \cdot \Big(\frac{k}{1+\alpha}\Big)^{\frac{k}{1+\alpha}+\frac{1}{2}} e^{-\frac{k}{1+\alpha}+\frac{1+\alpha}{12k}},$$

where C is a universal constant.

Proof By the layer cake representation and the tail bound, we have

$$E_x[|x|^k] = k\int_0^\infty t^{k-1}\cdot \mathbb{P}_x[|x|>t]dt \le 2k\int_0^\infty t^{k-1}\cdot e^{-(t/\lambda)^{1+\alpha}}dt.$$

Making the substitution $s=(t/\lambda)^{1+\alpha}$, i.e. $t=\lambda s^{\frac{1}{1+\alpha}}$ and $dt=\frac{\lambda}{1+\alpha}s^{-\frac{\alpha}{1+\alpha}}ds$, we get

$$\begin{aligned}E_x[|x|^k] &\le 2k\int_0^\infty \lambda^{k-1}s^{\frac{k-1}{1+\alpha}}\cdot e^{-s}\cdot\frac{\lambda}{1+\alpha}s^{-\frac{\alpha}{1+\alpha}}ds\\ &= \frac{2k\lambda^k}{1+\alpha}\int_0^\infty s^{\frac{k}{1+\alpha}-1}\cdot e^{-s}ds\\ &= \frac{2k\lambda^k}{1+\alpha}\cdot\Gamma\Big(\frac{k}{1+\alpha}\Big).\end{aligned}$$

Notice that $\Gamma(x)\le Cx^{x-1/2}e^{-x}e^{1/(12x)}$ for a positive constant C, then we have

$$E_x[|x|^k] \le \frac{2k\lambda^k}{1+\alpha}\cdot C\Big(\frac{k}{1+\alpha}\Big)^{\frac{k}{1+\alpha}-\frac{1}{2}}e^{-\frac{k}{1+\alpha}+\frac{1+\alpha}{12k}}.$$

■

Lemma C.6 *If x is (α,λ)-strictly sub-exponential random variable satisfying $\mathbb{P}_x[|x|>t]\le 2\cdot e^{-(t/\lambda)^{1+\alpha}}$, then for any $a>0$*

$$\mathbb{E}_x[e^{a|x|}] \le 3e^{2^{1/\alpha}(a\lambda)^{1+1/\alpha}}.$$

Proof We split the integral into two parts at some threshold T, which we choose later:

$$\begin{aligned}\mathbb{E}_x[e^{a|x|}] &= \mathbb{E}_x\big[e^{a|x|}\cdot\mathbb{1}[|x|<T]\big]+\mathbb{E}_x\big[e^{a|x|}\cdot\mathbb{1}[|x|\ge T]\big]\\ &\le e^{aT}+\mathbb{E}_x\big[e^{a|x|}\cdot\mathbb{1}[|x|\ge T]\big].\end{aligned}$$

By the layer cake representation and the tail bound, we have

$$\mathbb{E}_x\big[e^{a|x|}\cdot\mathbb{1}[|x|\ge T]\big] = \int_T^\infty ae^{at}\cdot\mathbb{P}_x[|x|>t]dt \le 2a\int_T^\infty e^{at}\cdot e^{-(t/\lambda)^{1+\alpha}}dt.$$

Choose T to be $(2a)^{1/\alpha}\lambda^{1+1/\alpha}$. Then, we have $(t/\lambda)^{1+\alpha}\ge 2at$ for $t\ge T$ and hence

$$\mathbb{E}_x\big[e^{a|x|}\cdot\mathbb{1}[|x|\ge T]\big] \le 2a\int_T^\infty e^{-at}dt = 2e^{-aT}.$$

Thus,

$$\mathbb{E}_x[e^{a|x|}] \le e^{aT}+2e^{-aT} \le 3e^{aT} = 3e^{2^{1/\alpha}(a\lambda)^{1+1/\alpha}}.$$

■

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contribution: a discrete analogue of smoothed agnostic learning for Boolean halfspaces under strictly sub-exponential distributions, along with a computationally efficient learning algorithm. These claims are supported by the theoretical results in the main body (Section 4.5).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of our work are discussed in the conclusion section of this paper (Section 5).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are clearly stated with assumptions, and full proofs are provided in the main paper or the supplementary material. Lemmas and tools from prior work are cited and contextualized.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

Justification: The paper is entirely theoretical and does not include empirical experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: No datasets or code are used or required, as the paper does not include experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experimental setting is involved; the paper is theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include empirical results or plots requiring error bars or statistical tests.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments were conducted; no compute resources were used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical, complies with NeurIPS ethical guidelines, and does not raise ethical concerns regarding data use, privacy, or fairness.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is purely theoretical.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release data, models, or tools with potential misuse risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No external datasets, models, or software assets are used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new datasets or software assets are introduced.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects or crowdsourcing are involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No IRB approval is necessary, as no human subjects are involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLMs were used in developing or supporting the core scientific contributions of this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.