Proceedings of Machine Learning Research vol 125:1–15, 2020 33rd Annual Conference on Learning Theory

Efficient Parameter Estimation of Truncated Boolean Product Distributions

Dimitris Fotakis FOTAKIS@CS.NTUA.GR
National Technical University of Athens

Alkis Kalavasis KALAVASIS.ALKIS@GMAIL.COM
National Technical University of Athens

Christos Tzamos TZAMOS@WISC.EDU
University of Wisconsin-Madison

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study the problem of estimating the parameters of a Boolean product distribution in d dimensions, when the samples are truncated by a set $S \subset \{0,1\}^d$ accessible through a membership oracle. This is the first time that the computational and statistical complexity of learning from truncated samples is considered in a discrete setting.

We introduce a natural notion of *fatness* of the truncation set S, under which truncated samples reveal enough information about the true distribution. We show that if the truncation set is sufficiently fat, samples from the true distribution can be generated from truncated samples. A stunning consequence is that virtually any statistical task (e.g., learning in total variation distance, parameter estimation, uniformity or identity testing) that can be performed efficiently for Boolean product distributions, can also be performed from truncated samples, with a small increase in sample complexity. We generalize our approach to ranking distributions over d alternatives, where we show how fatness implies efficient parameter estimation of Mallows models from truncated samples.

Exploring the limits of learning discrete models from truncated samples, we identify three natural conditions that are necessary for efficient identifiability: (i) the truncation set S should be rich enough; (ii) S should be accessible through membership queries; and (iii) the truncation by S should leave enough randomness in all directions. By carefully adapting the Stochastic Gradient Descent approach of (Daskalakis et al., FOCS 2018), we show that these conditions are also sufficient for efficient learning of truncated Boolean product distributions.

Keywords: Truncated Statistics, Boolean Product Distributions, Ranking Distributions, Stochastic Gradient Descent

1. Introduction

Parameter estimation and learning from truncated samples is an important and challenging problem in Statistics. The goal is to estimate the parameters of the true distribution based only on samples that fall within a (possibly small) subset S of the distribution's support.

Sample truncation occurs naturally in a variety of settings in science, engineering, economics, business and social sciences. Typical examples include selection bias in epidemiology and medical studies, and anecdotal "paradoxes" in damage and injury analysis explained by survivor bias. Statis-

0. The full version is available on arXiv with the same title and contains the proofs of all results discussed in this paper.

tical estimation from truncated samples goes back to at least Galton (1897), who analyzed truncated samples corresponding to speeds of trotting horses, and includes classical results on the use of the moments method (Pearson and Lee, 1908; Lee, 1914) and the maximum likelihood method (Fisher, 1931) for estimating a univariate Gaussian distribution from truncated samples.

In the last few years, there has been an increasing interest in computationally and statistically efficient algorithms for learning multivariate Gaussian distributions from truncated samples (when the truncation set is known (Daskalakis et al., 2018) or unknown (Kontonis et al., 2019)) and for training linear regression on models based on truncated (or censored) data (Daskalakis et al., 2019). In addition to the elegant and powerful application of Stochastic Gradient Descent to optimizing a seemingly unknown maximum likelihood function from truncated samples, a significant contribution of (Daskalakis et al., 2018; Kontonis et al., 2019; Daskalakis et al., 2019) concerns necessary conditions for efficient statistical estimation of multivariate Gaussian or regression models from truncated samples. More recently, Nagarajan and Panageas (2019) showed how to use Expectation-Maximization for learning mixtures of two Gaussian distributions from truncated samples.

Despite the strong results above for continuous settings, we are not aware of any previous work on learning discrete models from truncated samples. We note that certain elements of the prior approaches in inference from truncated data are inherently continuous and it is not clear if they can be adapted to a discrete setting. E.g., statistical estimation from truncated samples in a discrete setting should deal with a situation where the truncation removes virtually all randomness from certain directions, something that cannot happen naturally in a continuous setting.

Our Setting. Motivated by this gap in relevant literature, we investigate efficient parameter estimation of discrete models from truncated samples. We start with the fundamental setting of a Boolean product distribution $\mathcal{D}$ on the d-dimensional hypercube truncated by a set S, which is accessible through membership queries. The marginal of $\mathcal{D}$ in each direction i is an independent Bernoulli distribution with parameter $p_i \in (0,1)$. Our goal is to compute an estimation $\hat{\boldsymbol{p}}$ of the parameter vector $\boldsymbol{p}$ of $\mathcal{D}$ such that $\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_2 \leq \varepsilon$, with probability of at least $1-\delta$, with time and sample complexity polynomial in d, $1/\varepsilon$ and $\log(1/\delta)$. We note that such an estimation $\hat{\boldsymbol{p}}$ (or an estimation $\hat{\boldsymbol{z}}$ of the logit parameters $\boldsymbol{z} = (\log \frac{p_1}{1-p_1}, \ldots, \log \frac{p_d}{1-p_d})$ of similar accuracy) implies an estimation of the true distribution within total variation distance ε.

Our Contributions. Significantly departing from the maximum likelihood estimation approach of Daskalakis et al. (2018); Kontonis et al. (2019); Daskalakis et al. (2019), we introduce a natural notion of fatness of the truncation set S, under which samples from the truncated distribution $\mathcal{D}_S$ reveal enough information about the true distribution $\mathcal{D}$. Roughly speaking, a truncated Boolean product distribution $\mathcal{D}_S$ is α-*fat* in some direction i of the Boolean hypercube, if for an α probability mass of the truncated samples, the neighboring sample with its i-th coordinate flipped is also in S. Therefore, with probability α, conditional on the remaining coordinates, the i-th coordinate of a sample is distributed as the marginal of the true distribution $\mathcal{D}$ in direction i. So, if the truncated distribution $\mathcal{D}_S$ is α-fat in all directions (e.g., the halfspace of all vectors with L_1 norm at most k is a fat subset of the Boolean hypercube), a sample from $\mathcal{D}_S$ is quite likely to reveal significant information about the true distribution $\mathcal{D}$. Building on this intuition, we show how samples from the true distribution $\mathcal{D}$ can be generated from few truncated samples (see also Algorithm 1):

Informal Theorem 1 *With an expected number of $O(\log(d)/\alpha)$ samples from the α-fat truncation of a Boolean product distribution $\mathcal{D}$, we can generate a sample $\boldsymbol{x} \in \{0,1\}^d$ distributed as in $\mathcal{D}$.*

We show (Lemma 3) that fatness is also a necessary condition for Theorem 1. A stunning consequence of Theorem 1 is that virtually any statistical task (e.g., learning in total variation distance, parameter estimation, sparse recovery, uniformity or identity testing, differentially private uniformity testing) that can be performed efficiently for a Boolean product distribution $\mathcal{D}$, can also be performed using truncated samples from $\mathcal{D}$, at the expense of a factor $O(\log(d)/\alpha)$ increase in time and sample complexity. In Section 3, we obtain, as simple corollaries of Theorem 1, that the statistical tasks described in Acharya et al. (2015c); Diakonikolas et al. (2017b); Canonne et al. (2017, 2019b) for Boolean product distributions can be performed using only truncated samples!

To further demonstrate the power and the wide applicability of our approach, we extend the notion of fatness to the richer and more complex setting of ranking distributions on d alternatives. In Section 3.5, we show how to implement efficient statistical inference of Mallows models using samples from a fat truncated Mallows distribution (see Theorem 11).

Natural and powerful though, fatness is far from being necessary for efficient parameter estimation from truncated samples. Seeking a deeper understanding of the challenges of learning discrete models from truncated samples, we identify, in Section 4, three natural conditions that we show to be necessary for efficient parameter estimation in our setting:

Assumption 1: The support of the distribution $\mathcal{D}$ on S should be rich enough, in the sense that its truncation $\mathcal{D}_S$ should assign positive probability to a $\boldsymbol{x}^* \in S$ and d other vectors that remain linearly independent after we subtract $\boldsymbol{x}^*$ from them.

Assumption 2: S is accessible through a membership oracle that reveals whether $\boldsymbol{x} \in S$, for any $\boldsymbol{x}$ in the d-dimensional hypercube.

Assumption 3: The truncation of $\mathcal{D}$ by S leaves enough randomness in all directions. More precisely, we require that in any direction $\boldsymbol{w} \in \mathbb{R}^d$, any two samples from the truncated distribution $\mathcal{D}_S$ have sufficiently different projections on $\boldsymbol{w}$, with non-negligible probability.

Assumption 2 ensures that the learning algorithm has enough information about S and is also required in the continuous setting. Without oracle access to S, for any Boolean product distribution $\mathcal{D}$, we can construct a (possibly exponentially large) truncation set S such that sampling from the truncated distribution $\mathcal{D}_S$ appears identical to sampling from the uniform distribution, until the first duplicate sample appears (our construction is similar to (Daskalakis et al., 2018, Lemma 12)).

Similarly to Daskalakis et al. (2018), Assumption 2 is complemented by the additional natural requirement that the true distribution $\mathcal{D}$ should assign non-negligible probability mass to the truncation set S (Assumption 4). The reason is that the only way for a parameter estimation algorithm to evaluate the quality of its current estimation is by generating samples in S and comparing them with samples from $\mathcal{D}_S$. Assumptions 2 and 4 ensure that this can be performed efficiently.

Assumptions 1 and 3 are specific to the discrete setting of the Boolean hypercube. Assumption 1 requires that we should be able to normalize the truncation set S, by subtracting a vector $\boldsymbol{x}^*$, so that its dimension remains d. If this is true, we can recover the parameters of a Boolean product distribution $\mathcal{D}$ from truncated samples by solving a linear system with d equations and d unknowns, which we obtain after normalization. We prove, in Lemma 12, that Assumption 1 is both sufficient and necessary for parameter recovery from truncated samples in our setting.

Assumption 3 is a stronger version of Assumption 1 and is necessary for efficient parameter estimation from truncated samples in the Boolean hypercube. It essentially requires that with suf-

ficiently high probability, any set X of polynomially many samples from $\mathcal{D}_S$ can be normalized, subtracting a vector $\boldsymbol{x}^*$, so that X includes a well-conditioned $d \times d$ matrix, after normalization.

Beyond showing that these assumptions are necessary for efficient identifiability, we show that they are also sufficient and provide a computational efficient algorithm for learning Boolean product distributions. Our algorithm is based on a careful adaptation of the approach of Daskalakis et al. (2018) which uses Stochastic Gradient Descent on the negative log-likelihood. While the analysis consists of the same conceptual steps as that of Daskalakis et al. (2018), it requires dealing with a number of technical details that arise due to discreteness. One technical contribution of our work is using the necessary assumptions for identifiability to establish strong-convexity of the negative log-likelihood in a small ball around the true parameters. Our main result is that:

Informal Theorem 2 *Under Assumptions 1 - 4, Algorithm 2 computes an estimation $\hat{\boldsymbol{z}}$ of the logit vector $\boldsymbol{z}$ of the true distribution $\mathcal{D}$ such that $\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2 \leq \varepsilon$ with probability at least $1 - \delta$, and achieves time and sample complexity polynomial in d, $1/\varepsilon$ and $\log(1/\delta)$.*

Related Work. As aforementioned, there has been a large number of recent works dealing inference with truncated data from a Gaussian distribution (Daskalakis et al., 2018; Kontonis et al., 2019; Daskalakis et al., 2019) or mixtures of Gaussians (Nagarajan and Panageas, 2019) but to the best of our knowledge there is no work dealing with discrete distributions. An additional feature of our work compared to those results is that our methods are not limited to parameter estimation but enable any statistical task to be performed on truncated datasets by providing a sampler to the true underlying distribution. While this requires a mildly stronger than necessary but natural assumption on the truncation set, we show that the more complex SGD based methods developed in prior work can also be applied in the discrete settings we consider.

The field of robust statistics is also very related to our work as it also deals with biased datasets and aims to identify the distribution that generated the data. Truncation can be seen as an adversary erasing samples outside a certain set. Recently, there has been a lot of theoretical work for computationally-efficient robust estimation of high-dimensional distributions in the presence of arbitrary corruptions to a small ε fraction of the samples, allowing for both deletions of samples and additions of samples (Diakonikolas et al., 2016; Charikar et al., 2017; Lai et al., 2016; Diakonikolas et al., 2017a, 2018; Hopkins and Li, 2019). In particular, the work of Diakonikolas et al. (2016) deals with the problem of learning binary-product distributions.

Another line of related work concerns learning from positive examples. The work of De et al. (2014) considers a setting where samples are obtained from the uniform distribution over the hypercube truncated on a set S. However, their goal is somewhat orthogonal to ours. It aims to accurately learn the set S while the distribution is already known. In contrast, in our setting the truncation set is known and the goal is to learn the distribution. More recently, (Canonne et al., 2020) extend these results to learning the truncation set with truncated samples from continuous distributions.

Another related literature within learning theory aims to learn discrete distributions through conditional samples. In the conditional sampling model that was recently introduced concurrently by Chakraborty et al. (2013, 2016) and Canonne et al. (2014, 2015), the goal is again to learn an underlying discrete distribution through conditional/truncated samples but the learner can change the truncation set on demand. This is known to be a more powerful model for distribution learning and testing than standard sampling (Canonne, 2015; Falahatgar et al., 2015; Acharya et al., 2015b; Bhattacharyya and Chakraborty, 2018; Acharya et al., 2015a; Gouleakis et al., 2017; Kamath and Tzamos, 2019; Canonne et al., 2019a).

2. Preliminaries

We use lowercase bold letters $\boldsymbol{x}$ to denote d-dimensional vectors. We let $\|\boldsymbol{x}\|_p = (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ denote the L_p norm and $\|\boldsymbol{x}\|_\infty = \max_{i\in[d]}\{|x_i|\}$ denote the L_∞ norm of a vector $\boldsymbol{x}$. We let $[d] \stackrel{\text{def}}{=} \{1,\ldots,d\}$ and $\mathbb{F}_2 = \{0,1\}$. $\Pi_d = \{0,1\}^d$ denotes the d-dimensional Boolean hypercube.

For any vector $\boldsymbol{x}$, $\boldsymbol{x}_{-i}$ is the vector obtained from $\boldsymbol{x}$ by removing the i-th coordinate and $(\boldsymbol{x}_{-i}, y)$ is the vector obtained from $\boldsymbol{x}$ by replacing x_i by y. Similarly, given a set $S \subseteq \Pi_d$, we let $S_{-i} = \{\boldsymbol{x}_{-i} : (\boldsymbol{x}_{-i}, 0) \in S \vee (\boldsymbol{x}_{-i}, 1) \in S\}$ be the projection of S to $\Pi_{[d]\setminus\{i\}}$. For any $\boldsymbol{x} \in \Pi_d$ and any coordinate $i \in [d]$, we let $\textsc{flip}(\boldsymbol{x}, i) = (\boldsymbol{x}_{-i}, 1 - x_i)$ denote $\boldsymbol{x}$ with its i-th coordinated flipped.

Bernoulli Distribution. For any $p \in [0,1]$, we let $\mathcal{B}e(p)$ denote the Bernoulli distribution with parameter p. For any $x \in \mathbb{F}_2$, $\mathcal{B}e(p; x) = p^x(1-p)^{1-x}$ denotes the probability of value x under $\mathcal{B}e(p)$. The Bernoulli distribution is an exponential family[1], where the natural parameter, denoted z, is the logit $z = \log \frac{p}{1-p}$ of the parameter p[2]. The inverse parameter mapping is $p = \frac{1}{1+\exp(-z)}$. Also, the base measure is $h(x) = 1$, the sufficient statistic is the identity mapping $T(x) = x$ and the log-partition function with respect to p is $\alpha(p) = -\log(1-p)$.

Boolean Product Distribution. We mostly focus on a fundamental family of *Boolean product distributions* on the d-dimensional hypercube Π_d. A Boolean product distribution with parameter vector $\boldsymbol{p} = (p_1,\ldots,p_d)$, usually denoted by $\mathcal{D}(\boldsymbol{p})$, is the product of d independent Bernoulli distributions, i.e., $\mathcal{D}(\boldsymbol{p}) = \mathcal{B}e(p_1) \otimes \cdots \otimes \mathcal{B}e(p_d)$. The Boolean product distribution can be expressed in the form of an exponential family as follows:

$$\mathcal{D}(\boldsymbol{z}; \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{z})}{\prod_{i\in[d]}(1+\exp(z_i))}\,, \tag{1}$$

where $\boldsymbol{z} = (z_1,\ldots,z_d)$ is the natural parameter vector with $z_i = \log \frac{p_i}{1-p_i}$ for each $i \in [d]$.

In the following, we always let $\mathcal{D}$ (or $\mathcal{D}(\boldsymbol{p})$ or $\mathcal{D}(\boldsymbol{z})$, when we want to emphasize the parameter vector $\boldsymbol{p}$ or the natural parameter vector $\boldsymbol{z}$) denote a Boolean product distribution. We denote $\boldsymbol{z}(\boldsymbol{p})$ (or simply $\boldsymbol{z}$, when $\boldsymbol{p}$ is clear from the context) the vector of natural parameters of $\mathcal{D}$. We let $\mathcal{D}(\boldsymbol{p}; \boldsymbol{x})$ and $\mathcal{D}(\boldsymbol{z}; \boldsymbol{x})$ (or simply $\mathcal{D}(\boldsymbol{x})$, when $\boldsymbol{p}$ or $\boldsymbol{z}$ are clear from the context) denote the probability of $\boldsymbol{x} \in \Pi_d$ under $\mathcal{D}$. Given a subset $S \subset \Pi_d$ of the hypercube, the probability mass assigned to S by a distribution $\mathcal{D}(\boldsymbol{p})$, usually denoted $\mathcal{D}(\boldsymbol{p}; S)$ (or simply $\mathcal{D}(S)$, when $\boldsymbol{p}$ is clear from the context), $\mathcal{D}(\boldsymbol{p}; S) = \sum_{\boldsymbol{x}\in S} \mathcal{D}(\boldsymbol{p}; \boldsymbol{x})$.

Truncated Boolean Product Distribution. Given a Boolean product distribution $\mathcal{D}$, we define the *truncated Boolean product distribution* $\mathcal{D}_S$, for any fixed $S \subset \Pi_d$. $\mathcal{D}_S$ has $\mathcal{D}_S(\boldsymbol{x}) = \mathcal{D}(\boldsymbol{x})/\mathcal{D}(S)$, for all $\boldsymbol{x} \in S$, and $\mathcal{D}_S(\boldsymbol{x}) = 0$, otherwise. We often refer to $\mathcal{D}_S$ as the truncation of $\mathcal{D}$ (by S) and to S as the *truncation set*.

It is sometimes convenient (especially when we discuss assumptions 1 and 3, in Section 4), to refer to some fixed element of S. We observe that by swapping 1 with 0 (and p_i with $1-p_i$) in certain directions, we can *normalize* S so that $\mathbf{0} \in S$ and $\mathcal{D}_S(\mathbf{0}) > 0$. In the following, we always assume, without loss of generality, that S is normalized so that $\mathbf{0} \in S$ and $\mathcal{D}_S(\mathbf{0}) > 0$.

1. The exponential family $\mathcal{E}(\boldsymbol{T}, h)$ with sufficient statistics $\boldsymbol{T}$, carrier measure h and natural parameters $\boldsymbol{\eta}$ is the family of distributions $\mathcal{E}(\boldsymbol{T}, h) = \{\mathcal{P}_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H}_{\boldsymbol{T},h}\}$, where the probability distribution $\mathcal{P}_{\boldsymbol{\eta}}$ has density $p_{\boldsymbol{\eta}}(x) = h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x) - \alpha(\boldsymbol{\eta}))$.
2. The base of the logarithm function log used throughout the paper is insignificant.

Notions of Distance between Distributions. Let $\mathcal{P}, \mathcal{Q}$ be two probability measures in the discrete probability space $(\Omega, \mathcal{F})$. The *total variation distance* between $\mathcal{P}$ and $\mathcal{Q}$, denoted $D_{TV}(\mathcal{P}, \mathcal{Q})$, is defined as $D_{TV}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\sum_{x\in\Omega} |\mathcal{P}(x) - \mathcal{Q}(x)| = \max_{A\in\mathcal{F}} |\mathcal{P}(A) - \mathcal{Q}(A)|$. The *Kullback–Leibler divergence* (or simply, *KL divergence*), denoted $D_{KL}(\mathcal{P} \parallel \mathcal{Q})$, is defined as $D_{KL}(\mathcal{P} \parallel \mathcal{Q}) = \mathbb{E}_{x\sim\mathcal{P}}\left[\log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)}\right] = \sum_{x\in\Omega} \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)}$. The following summarizes some standard upper bounds on the total variation distance and the KL divergence of two Boolean product distributions.

Proposition 1 *Let $\mathcal{P}(\boldsymbol{p})$ and $\mathcal{Q}(\boldsymbol{q})$ be two Boolean product distributions, and let $\boldsymbol{z}(\boldsymbol{p})$ and $\boldsymbol{z}(\boldsymbol{q})$ be the vectors of their natural parameters. Then:*

(i) $D_{KL}(\mathcal{P} \parallel \mathcal{Q}) \leq \|\boldsymbol{z}(\boldsymbol{p}) - \boldsymbol{z}(\boldsymbol{q})\|_2^2$

(ii) $D_{TV}(\mathcal{P}, \mathcal{Q}) \leq \frac{\sqrt{2}}{2} \|\boldsymbol{z}(\boldsymbol{p}) - \boldsymbol{z}(\boldsymbol{q})\|_2$

(iii) $D_{TV}(\mathcal{P}, \mathcal{Q}) \leq \sqrt{\sum_{i=1}^{d} \frac{(p_i - q_i)^2}{p_i + q_i}}$

Identifiability and Learnability. A Boolean product distribution $\mathcal{D}(\boldsymbol{p})$ is *identifiable* from its truncation $\mathcal{D}_S(\boldsymbol{p})$, if given $\mathcal{D}_S(\boldsymbol{p}; \boldsymbol{x})$, for all $\boldsymbol{x} \in S$, we can recover the parameter vector $\boldsymbol{p}$.

A Boolean product distribution $\mathcal{D}(\boldsymbol{p})$ is *efficiently learnable* from its truncation $\mathcal{D}_S(\boldsymbol{p})$, if for any $\varepsilon, \delta > 0$, we can compute an estimation $\hat{\boldsymbol{p}}$ of the parameter vector $\boldsymbol{p}$ (or an estimation $\hat{\boldsymbol{z}}$ of the natural parameter vector $\boldsymbol{z}$) of $\mathcal{D}$ such that $\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_2 \leq \varepsilon$ (or $\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2 \leq \varepsilon$), with probability at least $1 - \delta$, with time and sample complexity polynomial in d, $1/\varepsilon$ and $\log(1/\delta)$ using truncated samples from $\mathcal{D}_S(\boldsymbol{p})$. By Proposition 1, an upper bound on the L_2 distance between $\hat{\boldsymbol{z}}$ and $\boldsymbol{z}$ (or between $\hat{\boldsymbol{p}}$ and $\boldsymbol{p}$) translates into an upper bound on the total variation distance between the true distribution and $\mathcal{D}(\hat{\boldsymbol{z}})$ (or $\mathcal{D}(\hat{\boldsymbol{p}})$).

3. Boolean Product Distributions Truncated by Fat Sets

In this section, we discuss *fatness* of the truncation set, a strong sufficient (and in a certain sense, necessary) condition, under which we can generate samples from a Boolean product distribution $\mathcal{D}$ using samples from its truncation $\mathcal{D}_S$ (and access to S through a membership oracle).

Definition 2 *A truncated Boolean product distribution $\mathcal{D}_S$ is α-fat in coordinate $i \in [d]$, for some $\alpha > 0$, if $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[\text{FLIP}(\boldsymbol{x}, i) \in S] \geq \alpha$. A truncated Boolean product distribution $\mathcal{D}_S$ is α-fat, for some $\alpha > 0$, if $\mathcal{D}_S$ is α-fat in every coordinate $i \in [d]$.*

If $\mathcal{D}_S$ is fat, it happens often that a sample $\boldsymbol{x} \sim \mathcal{D}_S$ has both $(\boldsymbol{x}_{-i}, 0), (\boldsymbol{x}_{-i}, 1) \in S$. Then, conditional on the remaining coordinates $\boldsymbol{x}_{-i}$, the i-th coordinate x_i of $\boldsymbol{x}$ is distributed as $\mathcal{B}e(p_i)$. We next focus on truncated Boolean product distributions $\mathcal{D}_S$ that are α-fat.

There are several natural classes of truncation subsets that give rise to fat truncated product distributions. E.g., for each $k \in [d]$, the halfspace $S_{\leq k} = \{\boldsymbol{x} \in \Pi_d : x_1 + \ldots + x_d \leq k\}$ results in an α-fat truncated distribution, if $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_{S_{\leq k}}}[x_i = 1] \geq \alpha$, for all $i \in [d]$. The same holds if S is any *downward closed*[3] subset of Π_d and $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[x_i = 1] \geq \alpha$, for all $i \in [d]$.

Fatness in coordinate $i \in [d]$ is necessary, if we want to distinguish between two truncated Boolean distributions based on their i-th parameter only, if the remaining coordinates are correlated.

3. A set $S \subseteq \Pi_d$ is downward closed if for any $\boldsymbol{x} \in S$ and any $\boldsymbol{y}$ with $y_i \leq x_i$, in all directions $i \in [d]$, $\boldsymbol{y} \in S$.

Algorithm 1 Sampling from $\mathcal{D}$ using samples from $\mathcal{D}_S$

procedure SAMPLER($\mathcal{D}_S$) ▷ $\mathcal{D}_S$ is α-fat.
  $\boldsymbol{y} \leftarrow (-1, \ldots -1)$
  while $\exists y_i = -1$ **do**
    Draw sample $\boldsymbol{x} \sim \mathcal{D}_S$
    for $i \leftarrow 1, \ldots, d$ **do**
      if FLIP$(\boldsymbol{x}, i) \in S$ **then** ▷ We assume oracle access to S
        $y_i \leftarrow x_i$
  return $\boldsymbol{y}$

Specifically, we can show that if $\mathcal{D}_S$ is 0-fat in some coordinate i, there exists a Boolean distribution with $q_i \neq p_i$ (and $|q_i - p_i|$ large enough) whose truncation by S appears identical to $\mathcal{D}_S$. Therefore, if the other coordinates are arbitrarily correlated, it is impossible to distinguish between the two distributions based on their i-th parameter alone. However, as we discuss in Section 4, if S is rich enough, but not necessarily fat, we can recover the entire parameter vector[4] of $\mathcal{D}$.

Lemma 3 *Let $i \in [d]$, let S be any subset of Π_d with FLIP$(\boldsymbol{x}, i) \notin S$, for all $\boldsymbol{x} \in S$, and consider any $0 < p < q < 1$. Then, for any Boolean distribution $\mathcal{D}_{-i}$ with $\mathcal{D}_{-i}(S_{-i}) \in (0, 1)$, there exists a distribution $\mathcal{D}'_{-i}$ such that $(\mathcal{B}e(p) \otimes \mathcal{D}_{-i})_S \equiv (\mathcal{B}e(q) \otimes \mathcal{D}'_{-i})_S$.*

3.1. Sampling from a Boolean Product Distribution using Samples from its Fat Truncation

An interesting consequence of fatness is that we can efficiently generate samples from a Boolean product distribution $\mathcal{D}$ using samples from any α-fat truncation of $\mathcal{D}$. The idea is described in Algorithm 1. Theorem 4 shows that for any sample $\boldsymbol{x}$ drawn from $\mathcal{D}_S$ and any $i \in [d]$ such that FLIP$(\boldsymbol{x}, i) \in S$, conditional on $\boldsymbol{x}_{-i}$, x_i is distributed as $\mathcal{B}e(p_i)$. So, we can generate a random sample $\boldsymbol{y} \sim \mathcal{D}$ by putting together d such values. α-fatness of the truncated distribution $\mathcal{D}_S$ implies that the expected number of samples $\boldsymbol{x} \sim \mathcal{D}_S$ required to generate a $\boldsymbol{y} \sim \mathcal{D}$ is $O(\log(d)/\alpha)$.

Theorem 4 *Let $\mathcal{D}$ be a Boolean product distribution over Π_d and let $\mathcal{D}_S$ be any α-fat truncation of $\mathcal{D}$. Then, (i) the distribution of the samples generated by Algorithm 1 is identical to $\mathcal{D}$; and (ii) the expected number of samples from $\mathcal{D}_S$ before a sample is returned by Algorithm 1 is $O(\log(d)/\alpha)$.*

3.2. Parameter Estimation and Learning in Total Variation Distance

Based on Algorithm 1, we can recover the parameters of any Boolean product distribution $\mathcal{D}$ using samples from any fat truncation of $\mathcal{D}$.

Theorem 5 *Let $\mathcal{D}(\boldsymbol{p})$ be a Boolean product distribution and let $\mathcal{D}_S(\boldsymbol{p})$ be a truncation of $\mathcal{D}$. If $\mathcal{D}_S$ is α-fat in any fixed coordinate i, then, for any $\varepsilon, \delta > 0$, we can compute an estimation $\hat{p}_i$ of the parameter p_i of $\mathcal{D}$ such that $|p_i - \hat{p}_i| \leq \varepsilon$, with probability at least $1 - \delta$, using an expected number of $O(\log(1/\delta)/(\varepsilon^2\alpha))$ samples from $\mathcal{D}_S$.*

4. For a concrete example, where we can recover the entire parameter vector of a truncated Boolean product distribution $\mathcal{D}_S$, we consider $S = \{000, 110, 011, 101\} \subset \Pi_3$, which is not fat in any coordinate, and let $p_{\boldsymbol{x}} = \mathcal{D}_S(\boldsymbol{x})$, for each $\boldsymbol{x} \in S$. Then, setting $z_i = \log \frac{p_i}{1-p_i}$, for each i, we can recover (p_1, p_2, p_3), by solving the following linear system: $z_1 + z_2 = \log \frac{p_{110}}{p_{000}}$, $z_2 + z_3 = \log \frac{p_{011}}{p_{000}}$, $z_1 + z_3 = \log \frac{p_{101}}{p_{000}}$. This is a special case of the more general identifiability condition discussed in Lemma 12.

Using $n = \log(2d/\delta)/\varepsilon^2$ samples $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ generated by Algorithm 1, we can estimate all the parameters $\boldsymbol{p}$ of $\mathcal{D}$, by taking $\hat{p}_i = \sum_{\ell=1}^{n} y_i^{(\ell)}/n$, for each $i \in [d]$. The following is an immediate consequence of theorems 4 and 5.

Corollary 6 *Let $\mathcal{D}(\boldsymbol{p})$ be a Boolean product distribution and $\mathcal{D}_S(\boldsymbol{p})$ be any α-fat truncation of $\mathcal{D}$. Then, for any $\varepsilon, \delta > 0$, we can compute an estimation $\hat{\boldsymbol{p}}$ such that $\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_\infty \leq \varepsilon$, with probability at least $1 - \delta$, using an expected number of $O(\log(d)\log(d/\delta)/(\varepsilon^2\alpha))$ samples from $\mathcal{D}_S$.*

3.3. Identity and Closeness Testing with Access to Truncated Samples

Theorem 4 implies that if we have sample access to an α-fat truncation $\mathcal{D}_S$ of a Boolean product distribution $\mathcal{D}$, we can pretend that we have sample access to the original distribution $\mathcal{D}$, at the expense of an increase in the sample complexity (from $\mathcal{D}_S$) by a factor of $O(\log(d)/\alpha)$. Therefore, we can extend virtually all known hypothesis testing and learning algorithms for Boolean product distributions to fat truncated Boolean product distributions.

For *identity testing* of Boolean product distributions, based on samples from fat truncated ones, we combine Algorithm 1 with the algorithm of (Canonne et al., 2017, Sec. 4.1). Combining Theorem 4 with (Canonne et al., 2017, Theorem 6), we obtain the following:

Corollary 7 (Identity Testing) *Let $\mathcal{Q}(\boldsymbol{q})$ be a Boolean product distribution described by its parameters $\boldsymbol{q}$, and let $\mathcal{D}$ be a Boolean product distribution for which we have sample access to its α-fat truncation $\mathcal{D}_S$. For any $\varepsilon > 0$, we can distinguish between $D_{TV}(\mathcal{Q}, \mathcal{D}) = 0$ and $D_{TV}(\mathcal{Q}, \mathcal{D}) > \varepsilon$, with probability $2/3$, using an expected number of $O(\log(d)\sqrt{d}/(\alpha\varepsilon^2))$ samples from $\mathcal{D}_S$.*

We can extend Corollary 7 to *closeness testing* of two Boolean product distributions, for which we only have sample access to their fat truncations. We combine Algorithm 1 with the algorithm of (Canonne et al., 2017, Sec. 5.1). The following is an immediate consequence of Theorem 4 and (Canonne et al., 2017, Theorem 9).

Corollary 8 (Closeness Testing) *Let $\mathcal{Q}$, $\mathcal{D}$ be two Boolean product distributions for which we have sample access to their α_1-fat truncation $\mathcal{Q}_{S_1}$ and α_2-fat truncation $\mathcal{D}_{S_2}$. For any $\varepsilon > 0$, we can distinguish between $D_{TV}(\mathcal{Q}, \mathcal{D}) = 0$ and $D_{TV}(\mathcal{Q}, \mathcal{D}) > \varepsilon$, with probability at least $2/3$, using an expected number of $O\Big((\frac{\log(d)}{\alpha_1} + \frac{\log(d)}{\alpha_2})\max\{\sqrt{d}/\varepsilon^2, d^{3/4}/\varepsilon\}\Big)$ samples from $\mathcal{Q}_{S_1}$ and $\mathcal{D}_{S_2}$.*

3.4. Learning in Total Variation Distance

Using Algorithm 1, we can learn a Boolean product distribution $\mathcal{D}(\boldsymbol{p})$, within ε in total variation distance, using samples from its fat truncation. The following uses a standard analysis of the sample complexity of learning a Boolean product distribution (see e.g., Kamath et al. (2018)).

Corollary 9 *Let $\mathcal{D}(\boldsymbol{p})$ be a Boolean product distribution and let $\mathcal{D}_S$ be any α-fat truncation of $\mathcal{D}$. Then, for any $\varepsilon, \delta > 0$, we can compute a Boolean product distribution $\hat{\mathcal{D}}(\hat{\boldsymbol{p}})$ such that $\mathcal{D}_{TV}(\mathcal{D}, \hat{\mathcal{D}}) \leq \varepsilon$, with probability at least $1 - \delta$, using $O(d\log(d/\delta)/(\varepsilon^2\alpha))$ samples from $\mathcal{D}_S$.*

We can improve the sample complexity in Corollary 9, if the original distribution $\mathcal{D}$ is sparse. We say that a Boolean product distribution $\mathcal{D}(\boldsymbol{p})$ is (k, c)-*sparse*, for some $k \in [d]$ and $c \in [0, 1]$, if there is an index set $I \subset [d]$, with $|I| = d - k$, such that for all $i \in I$, $p_i = c$. Namely, we

know that $d-k$ of $\mathcal{D}$'s parameters are equal to c (but we do not know which of them). Then, we first apply Corollary 6 and estimate all parameters of $\mathcal{D}$ within distance $\varepsilon/\sqrt{k}$. We set each p_i with $|p_i - c| \leq \varepsilon/\sqrt{k}$ to $p_i = c$. thus, we recover the index set I. For the remaining k parameters, we apply Corollary 9. The result is summarized by the following:

Corollary 10 *Let $\mathcal{D}(\boldsymbol{p})$ be a (k,c)-sparse Boolean product distribution and let $\mathcal{D}_S$ be any α-fat truncation of $\mathcal{D}$. Then, for any $\varepsilon, \delta > 0$, we can compute a Boolean product distribution $\hat{\mathcal{D}}(\hat{\boldsymbol{p}})$ such that $\mathcal{D}_{TV}(\mathcal{D}, \hat{\mathcal{D}}) \leq \varepsilon$, with probability at least $1-\delta$, using $O\left(\frac{k \log(d) \log(d/\delta)}{\varepsilon^2 \alpha}\right)$ samples from the truncate distribution $\mathcal{D}_S$.*

3.5. Learning Ranking Distributions from Truncated Samples

An interesting application of Theorem 4 is parameter estimation of ranking distributions from truncated samples. For clarity, we next focus on Mallows distributions. Our techniques imply similar results for other well known models of ranking distributions, such as Generalized Mallows distributions Fligner and Verducci (1986) and the models of Plackett (1975); Luce (1959), Bradley and Terry (1952) and B. Babington Smith (1950).

Definition and Notation. We start with some notation specific to this section. Let $\mathfrak{S}_d$ be the symmetric group over the finite set of items $[d]$. Given a ranking $\pi \in \mathfrak{S}_d$, we let $\pi(i)$ denote the position of item i in π. We say that i precedes j in π, denoted by $i \succ_\pi j$, if $\pi(i) < \pi(j)$. The Kendall tau distance of two rankings π and σ, denoted by $D_\tau(\pi, \sigma)$, is the number of discordant item pairs in π and σ. Formally, $D_\tau(\pi,\sigma) = \sum_{1 \leq i < j \leq d} \mathbb{1}\{(\pi(i) - \pi(j))(\sigma(i) - \sigma(j)) < 0\}$.

The *Mallows model* Mallows (1957) is a family of ranking distributions parameterized by the *central ranking* $\pi_0 \in \mathfrak{S}_d$ and the *spread parameter* $\phi \in [0,1]$. Assuming the Kendall tau distance between rankings, the probability mass function is $\mathcal{M}(\pi_0, \phi; \pi) = \phi^{D_\tau(\pi_0,\pi)}/Z(\phi)$, where the normalization factor is $Z(\phi) = \prod_{i=1}^{d} \frac{1-\phi^i}{1-\phi}$. For a given Mallows distribution $\mathcal{M}(\pi_0, \phi)$, we denote $p_{ij} = \mathbb{P}_{\pi \sim \mathcal{M}}[i \succ_\pi j]$ the probability that item i precedes item j in a random sample from $\mathcal{M}$.

Truncated Mallows Distributions. We consider parameter estimation for a Mallows distribution $\mathcal{M}(\pi_0, \phi)$ with sample access to its truncation $\mathcal{M}_S$ by a subset $S \subset \mathfrak{S}_d$. Then, $\mathcal{M}_S(\pi) = \mathcal{M}(\pi)/\mathcal{M}(S)$, for each $\pi \in S$, and $\mathcal{M}_S(\pi) = 0$, otherwise. Next, we generalize the notion of fatness to truncated ranking distributions and prove the equivalent of Theorem 5 and Corollary 6.

For a ranking π, we let $\textsc{flip}(\pi, i, j)$ denote the ranking π' obtained from π with the items i and j swapped. Formally, $\pi'(\ell) = \pi(\ell)$, for all items $\ell \in [d] \setminus \{i, j\}$, $\pi'(j) = \pi(i)$ and $\pi'(i) = \pi(j)$. We say that a truncated Mallows distribution $\mathcal{M}_S$ is *α-fat for pair* (i,j), if $\mathbb{P}_{\pi \sim \mathcal{M}_S}[\textsc{flip}(\pi, i, j) \in S] \geq \alpha$, for some $\alpha > 0$. A truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$ is α-fat, if $\mathcal{M}_S$ is α-fat for all pairs (i,j), and *neighboring α-fat*, if $\mathcal{M}_S$ is α-fat for all pairs (i,j) that occupy neighboring positions in the central ranking π_0, i.e., for all pairs (i,j) with $|\pi_0(i) - \pi_0(j)| = 1$.

Parameter Estimation and Learning of Mallows Distributions from Truncated Samples. We use a learning algorithm that draws a sample from the truncated Mallows distribution $\mathcal{M}_S$ and updates a vector $\boldsymbol{q}$ with estimations $\hat{p}_{ij} = q_{ij}/(q_{ij} + q_{ji})$ of the probability p_{ij} that item i precedes item j in a sample from the true Mallows distribution $\mathcal{M}$. Thus, we can show the following:

Theorem 11 *Let $\mathcal{M}(\pi_0, \phi)$ be a Mallows distribution with $\pi_0 \in \mathfrak{S}_d$ and $\phi \in [0, 1-\gamma]$, for some constant $\gamma > 0$, and let $\mathcal{M}_S$ be any neighboring α-fat truncation of $\mathcal{M}$. Then,*

(i) *For any $\delta > 0$, we can learn the central ranking π_0, with probability at least $1 - \delta$, using an expected number of $O(\log(d)\log(d/\delta)/(\gamma^2\alpha))$ samples from $\mathcal{M}_S$.*

(ii) *Assuming that the central ranking π_0 is known, for any $\varepsilon, \delta > 0$, we can compute an estimation $\hat{\phi}$ of the spread parameter such that $|\phi - \hat{\phi}| \leq O(\varepsilon)$, with probability at least $1-\delta$, using an expected number of $O(\log(1/\delta)/(\varepsilon^2\alpha))$ samples from $\mathcal{M}_S$.*

(iii) *For any $\varepsilon, \delta > 0$, we can compute a Mallows distribution $\hat{\mathcal{M}}(\pi_0, \hat{\phi})$ so that $\mathcal{D}_{TV}(\mathcal{M}, \hat{\mathcal{M}}) \leq O(\varepsilon)$, with probability at least $1-\delta$, using an expected number of $O(\log(d)\log(d/\delta)/(\gamma^2\alpha) + d\log(1/\delta)/(\varepsilon^2\alpha))$ samples from $\mathcal{M}_S$.*

4. Efficient Learnability from Truncated Samples: Necessary Conditions

We next discuss necessary conditions for identifiability and efficient learnability of a Boolean product distribution from truncated samples. For Assumption 1 and Lemma 12, we recall that we can assume without loss of generality that S is normalized so that $\mathcal{D}_S(\mathbf{0}) > 0$. The proof of Lemma 12 demonstrates that recovering $\boldsymbol{p}$ requires the solution to a linear system, similar to that in Footnote 4, which is solvable if and only if Assumption 1 holds.

Assumption 1 *For the truncated Boolean product distribution $\mathcal{D}_S$, $\mathcal{D}_S(\mathbf{0}) > 0$ (after possible normalization) and there are d linearly independent $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)} \in S$ with $\mathcal{D}_S(\boldsymbol{x}^{(j)}) > 0$, $j \in [d]$.*

Lemma 12 *A Boolean product distribution $\mathcal{D}(\boldsymbol{p})$ on Π_d is identifiable from its truncation $\mathcal{D}_S$ if and only if Assumption 1 holds.*

We proceed to show two necessary conditions for *efficient learnability*. Our first condition is that we have oracle access to the truncation set S. More formally, we assume that:

Assumption 2 *S is accessible through a membership oracle, which reveals whether $\boldsymbol{x} \in S$, for any $\boldsymbol{x} \in \Pi_d$.*

Based on the proof of (Daskalakis et al., 2018, Lemma 12), we show that if Assumption 2 does not hold, we can construct a (possibly exponentially large) truncation set S so that $\mathcal{D}_S$ appears identical to the uniform distribution $\mathcal{U}$ on Π_d as long as all the samples are distinct.

Lemma 13 *For any Boolean product distribution $\mathcal{D}(\boldsymbol{p})$, there is a truncation set S so that without additional information about S, we cannot distinguish between sampling from $\mathcal{D}_S$ and sampling from the uniform distribution $\mathcal{U}$ on Π_d, before an expected number of $\Omega(\sqrt{|S|})$ samples are drawn.*

Our second necessary condition for efficient learnability is that the truncated distribution is not extremely well concentrated in any direction. Intuitively, we need the Boolean product distribution $\mathcal{D}$, and its truncation $\mathcal{D}_S$, to behave well, so that we can get enough information about $\mathcal{D}$ based on few samples from $\mathcal{D}_S$. More formally, we quantify $\mathcal{D}_S$'s anticoncentration using λ^*, which is the maximum positive number so that for all unit vectors $\boldsymbol{w} \in \mathbb{R}^d$, $\|\boldsymbol{w}\|_2 = 1$, and all $c \in \mathbb{R}$, $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[\boldsymbol{w}^T\boldsymbol{x} \notin (c - \lambda^*, c + \lambda^*)] \geq \lambda^*$. Assumption 3 requires that λ^* is polynomially large.

Assumption 3 *There exists a $\lambda \geq 1/\mathrm{poly}(d)$ such that for all unit vectors $\boldsymbol{w} \in \mathbb{R}^d$, $\|\boldsymbol{w}\|_2 = 1$, and all $c \in \mathbb{R}$, $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[\boldsymbol{w}^T\boldsymbol{x} \notin (c - \lambda, c + \lambda)] \geq \lambda$.*

We note that Assumption 3 is a stronger version of Assumption 1. It also implies that all parameters $p_i \in (0,1)$ are bounded away from 0 and 1 by a safe margin. We next show that if $\mathcal{D}_S$ is well concentrated in some direction, estimating the parameter vector $\boldsymbol{p}$ requires a large number of samples from $\mathcal{D}_S$. More specifically, we show that either estimating $\mathcal{D}_S(\mathbf{0})$, which is needed for normalizing the linear system in Lemma 12, or sampling d vectors that result in a well-conditioned linear system, require $\Omega(1/\lambda^*)$ samples from $\mathcal{D}_S$. Therefore, if Assumption 3 does not hold, estimating $\boldsymbol{p}$ with truncated samples from $\mathcal{D}_S$ has superpolynomial sample complexity.

Lemma 14 *Let $\mathcal{D}(\boldsymbol{p})$ be a Boolean product distribution and let $\mathcal{D}_S$ be a truncation of $\mathcal{D}$. Then, computing an estimation $\hat{\boldsymbol{p}}$ of the parameter vector $\boldsymbol{p}$ of $\mathcal{D}$ such that $\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_2 \leq o(1)$ requires an expected number of $\Omega(1/\lambda^*)$ samples from $\mathcal{D}_S$.*

Proof (sketch) For a unit vector $\boldsymbol{w} \in \mathbb{R}^d$, we let $H_{\boldsymbol{w}} = \{\boldsymbol{x} \in S : \boldsymbol{w}^T\boldsymbol{x} \in (c-\lambda, c+\lambda)\}$. Intuitively, if λ^* is very small, there is a direction $\boldsymbol{w}$ such that virtually all samples $\boldsymbol{x} \sim \mathcal{D}_S$ lie in $H_{\boldsymbol{w}}$. Formally, by the definition of λ^*, for any $\lambda > \lambda^*$, there is a unit vector $\boldsymbol{w} \in \mathbb{R}^d$ and a $c \in \mathbb{R}$ such that $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[\boldsymbol{x} \notin H_{\boldsymbol{w}}] < \lambda$, or equivalently, $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_S}[\boldsymbol{x} \in H_{\boldsymbol{w}}] \geq 1-\lambda$.

Intuitively, recovering ($\boldsymbol{z}$ and) $\boldsymbol{p}$ boils down to the solution of a linear system as that in Footnote 4 and in Lemma 12. For that, we need d linearly independent vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)} \in S$ and an additional fixed element $\boldsymbol{x}^* \in S$ for the normalization of the probabilities in the right-hand side. With high probability, all $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)} \in H_{\boldsymbol{w}}$. If $\boldsymbol{x}^*$ is also in $H_{\boldsymbol{w}}$, normalizing the system by $\boldsymbol{x}^*$ results in an ill-conditioned system. The reason is that for any $\lambda > \lambda^*$ and any $\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \in H_{\boldsymbol{w}}$,

$$(\boldsymbol{x}^{(i)} - \boldsymbol{x}^*)^T(\boldsymbol{x}^{(j)} - \boldsymbol{x}^*) = (\boldsymbol{w}^T(\boldsymbol{x}^{(i)} - \boldsymbol{x}^*))^T(\boldsymbol{w}^T(\boldsymbol{x}^{(j)} - \boldsymbol{x}^*)) < 4\lambda^2 \,.$$

In fact, we can show that the condition number of the system is $\Omega(1/\lambda^*)$. Therefore, solving the linear system efficiently requires sampling a vector $\boldsymbol{x}^* \notin H_{\boldsymbol{w}}$ for normalization. However, the probability that we sample (and thus, can use for normalization) a vector $\boldsymbol{x}^* \notin H_{\boldsymbol{w}}$ is at most λ^*. ■

For the efficient estimation of $\boldsymbol{z}$, we also need to assume that the truncation set S is large enough.

Assumption 4 *For the truncation set S, there is an $\alpha > 0$ so that the Boolean product distribution $\mathcal{D}$ has $\mathcal{D}(S) \geq \alpha$.*

In the following section, we present a Projected Stochastic Gradient Descent algorithm and show that assumptions 2, 3 and 4 are sufficient for the efficient estimation of the natural parameter vector $\boldsymbol{z}$ of the Boolean product distribution $\mathcal{D}$ by sampling from its truncation $\mathcal{D}_S$.

5. Stochastic Gradient Descent for Learning Truncated Boolean Products

We next show how to estimate efficiently the natural parameter vector $\boldsymbol{z}^*$ of a Boolean product distribution $\mathcal{D}(\boldsymbol{z}^*)$ using samples from its truncation $\mathcal{D}_S(\boldsymbol{z}^*)$. Similarly to Daskalakis et al. (2018), we use Projected Stochastic Gradient Descent (SGD) on the negative log-likelihood of the truncated samples. Our SGD algorithm is described in Algorithm 2. We should highlight that Algorithm 2 runs in the space of the natural parameters $\boldsymbol{z}$ of the Boolean product distribution. Changing the parameters from $\boldsymbol{p}$ to $\boldsymbol{z}$ results in a linear system, similar to that in Footnote 4, and simplifies the analysis of the log-likelihood function. Furthermore, by Proposition 1, estimating $\boldsymbol{z}^*$ within error at most ε in L_2 norm results in a distribution within total variation distance at most ε to $\mathcal{D}(\boldsymbol{z}^*)$.

Algorithm 2 Projected Stochastic Gradient Descent with Samples from $\mathcal{D}_S(\boldsymbol{p}^*)$.

procedure SGD(M, η) ▷ M : number of steps, η : parameter
  $\boldsymbol{z}^{(0)} \leftarrow \hat{\boldsymbol{z}}$
  for $i = 1..M$ **do**
    Sample $\boldsymbol{x}^{(i)}$ from $\mathcal{D}_S$
    repeat
      Sample $\boldsymbol{y}$ from $\mathcal{D}(\boldsymbol{z}^{(i-1)})$
    until $\boldsymbol{y} \in S$ ▷ We assume oracle access to S
    $\boldsymbol{v}^{(i)} \leftarrow -\boldsymbol{x}^{(i)} + \boldsymbol{y}$
    $\boldsymbol{z}^{(i)} \leftarrow \Pi_{\mathcal{B}}(\boldsymbol{z}^{(i-1)} - \frac{1}{i\eta}\boldsymbol{v}^{(i)})$ ▷ $\eta_i = 1/(i\eta)$: step size
  return $\overline{\boldsymbol{z}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{z}^{(i)}$

Throughout the analysis of Algorithm 2, we make use of Assumptions 2 - 4. For the analysis, we first derive the negative log-likelihood function that Algorithm 2 optimizes. Since the truncation set S is only accessed through membership queries, we do not have a closed form of the log-likelihood. However, we can show that it is convex for any truncation set S. We prove that the natural parameter vector $\hat{\boldsymbol{z}}$ corresponding to the empirical estimator $\hat{\boldsymbol{p}}_S$ is a good initialization for Algorithm 2. Specifically, we show that $\hat{\boldsymbol{p}}_S$ is close to the true parameter vector $\boldsymbol{p}^*$ in L_2 distance, and that this proximity holds for the corresponding natural parameter vectors as well.

For the correctness of Algorithm 2, it is essential that it runs in a convex region. We can show that there exists a ball $\mathcal{B}$, centered at the initialization point $\hat{\boldsymbol{z}}$, which contains $\boldsymbol{z}^*$. The radius of the ball depends only on the lower bound α of $\mathcal{D}(S)$ (Assumption 4). We can prove that Assumptions 3 and 4 always hold inside $\mathcal{B}$. That is, for any vector $\boldsymbol{z} \in \mathcal{B}$ (and the corresponding parameter vector $\boldsymbol{p}$), the anti-concentration assumption holds for $\mathcal{D}_S(\boldsymbol{p})$ and the mass assigned to the truncation set S by $\mathcal{D}_S(\boldsymbol{p})$ can be lower bounded by a polynomial function of α. Under these two assumptions, we can prove that the negative log-likelihood is strongly-convex inside the ball $\mathcal{B}$. Hence, while Algorithm 2 iterates inside $\mathcal{B}$, the truncation set has always constant mass and the negative log-likelihood remains strongly-convex. Consequently, Algorithm 2 converges to the true vector of natural parameters $\boldsymbol{z}^*$. The following is a direct consequence of the steps described above:

Theorem 15 *Given oracle access to a measurable set $S \subset \Pi_d$ (Assumption 2), whose measure under some unknown Boolean product distribution $\mathcal{D}(\boldsymbol{z}^*)$ is at least $\alpha > 0$ (Assumption 4) and where the truncated distribution $\mathcal{D}_S(\boldsymbol{z}^*)$ satisfies Assumption 3 with parameter λ, and given samples from the truncation $\mathcal{D}_S(\boldsymbol{z}^*)$, there exists a polynomial-time algorithm that recovers an estimation $\overline{\boldsymbol{z}}$ of $\boldsymbol{z}^*$. For any $\varepsilon > 0$, the algorithm uses $\mathrm{poly}(1/\alpha, 1/\lambda)\tilde{O}(d/\varepsilon^2)$ truncated samples from $\mathcal{D}_S(\boldsymbol{z}^*)$ and membership queries to S and guarantees that $\|\boldsymbol{z}^* - \overline{\boldsymbol{z}}\|_2 \leq \varepsilon$, with probability 99%. Under these conditions, $\mathcal{D}_{TV}(\mathcal{D}(\boldsymbol{z}^*), \mathcal{D}(\overline{\boldsymbol{z}})) \leq O(\varepsilon)$ and the dependence of the sample complexity on d and ε is optimal (up to logarithmic factors), even when there is no truncation.*

Acknowledgments

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant", project BALSAM, HFRI-FM17-1424.

References

Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. Adaptive Estimation in Weighted Group Testing. In *Proceedings of the 2015 IEEE International Symposium on Information Theory*, ISIT '15, pages 2116–2120, 2015a.

Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. A Chasm Between Identity and Equivalence Testing with Conditional Queries. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, RANDOM '15, pages 449–466, 2015b.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal Testing for Properties of Distributions. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pages 3591–3599, 2015c. URL http://arxiv.org/abs/1507.05952.

B. Babington Smith. Discussion of Professor Ross's paper. *Journal of Royal Statistical Society B*, 12:53–56, 1950.

Rishiraj Bhattacharyya and Sourav Chakraborty. Property Testing of Joint Distributions using Conditional Samples. *Transactions on Computation Theory*, 10(4):16:1–16:20, 2018.

R.A. Bradley and M.E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39:324, 1952.

Clément L. Canonne. Big Data on the Rise? - Testing Monotonicity of Distributions. In *Proceedings of the 42nd International Colloquium on Automata, Languages, and Programming*, ICALP '15, pages 294–305, 2015.

Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1174–1192, 2014.

Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian Networks. In *Proceedings of the 30th Annual Conference on Learning Theory, (COLT)*, pages 370–448, 2017. URL http://arxiv.org/abs/1612.03156.

Clément L Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. Random Restrictions of High-Dimensional Distributions and Uniformity Testing with Subcube Conditioning. *CoRR*, abs/1911.07357, 2019a.

Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private Identity Testing for High-Dimensional Distributions. In *arXiv preprint arXiv:1905.11947*, 2019b. URL http://arxiv.org/abs/1905.11947.

Clément L Canonne, Anindya De, and Rocco A. Servedio. Learning from satisfying assignments under continuous distributions. In *14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 82–101. SIAM, 2020.

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the Power of Conditional Samples in Distribution Testing. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580. ACM, 2013.

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the Power of Conditional Samples in Distribution Testing. *SIAM Journal on Computing*, 45(4):1261–1296, 2016.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from Untrusted Data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017.

Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. Efficient Statistics, in High Dimensions, from Truncated Samples. In *59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018. URL https://arxiv.org/pdf/1809.03986.pdf.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and Statistically Efficient Truncated Regression. In *Conference on Learning Theory (COLT)*, pages 955–960, 2019.

Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. Learning from Satisfying Assignments. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 478–497. SIAM, 2014.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust Estimators in High Dimensions without the Computational Intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664, 2016. doi: 10.1109/FOCS.2016.85. URL https://doi.org/10.1109/FOCS.2016.85.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being Robust (in High Dimensions) Can Be Practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 999–1008, 2017a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical Query Lower Bounds for Robust Estimation of High-dimensional Gaussians and Gaussian Mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017b.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly Learning a Gaussian: Getting Optimal Error, Efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 2683–2702, 2018.

Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster Algorithms for Testing under Conditional Sampling. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 607–636, 2015.

RA Fisher. Properties and applications of Hh functions. *Mathematical tables*, 1:815–852, 1931.

Michael A Fligner and Joseph S Verducci. Distance Based Ranking Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

Francis Galton. An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1897.

Themistoklis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Faster Sublinear Algorithms Using Conditional Sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1743–1757, 2017.

Samuel B Hopkins and Jerry Li. How Hard is Robust Mean Estimation? In *Conference on Learning Theory*, pages 1649–1682, 2019.

Gautam Kamath and Christos Tzamos. Anaconda: A Non-Adaptive Conditional Sampling Algorithm for Distribution Testing. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 679–693. SIAM, 2019.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*, 2018. URL http://arxiv.org/abs/1805.00216.

Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient Truncated Statistics with Unknown Truncation. In *260th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic Estimation of Mean and Covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016.

Alice Lee. Table of the Gaussian "Tail" Functions; When the "Tail" is Larger than the Body. *Biometrika*, 10(2/3):208–214, 1914.

R.D. Luce. *Individual Choice Behavior*. Wiley, 1959.

Colin L Mallows. Non-Null Ranking Models. I. *Biometrika*, 44(1/2):114–130, 1957.

Sai Ganesh Nagarajan and Ioannis Panageas. On the Analysis of EM for truncated mixtures of two Gaussians. In *31st International Conference on Algorithmic Learning Theory (ALT)*, pages 955–960, 2019.

Karl Pearson and Alice Lee. On the Generalised Probable Error in Multiple Normal Correlation. *Biometrika*, 6(1):59–68, 1908.

R. Plackett. The Analysis of Permutations. *Applied Statistics*, 24:193–202, 1975.