Product distribution learning with imperfect advice

Arnab Bhattacharyya
Department of Computer Science
University of Warwick
arnab.bhattacharyya@warwick.ac.uk

Davin Choo
Harvard John A. Paulson School Of Engineering And Applied Sciences
Harvard University
davinchoo@seas.harvard.edu

Philips George John
CNRS@CREATE & Dept. of Computer Science
National University of Singapore
philips.george.john@u.nus.edu

Themis Gouleakis
College of Computing & Data Science
Nanyang Technological University
themis.gouleakis@ntu.edu.sg

Abstract

Given i.i.d. samples from an unknown distribution P, the goal of distribution learning is to recover the parameters of a distribution that is close to P. When P belongs to the class of product distributions on the Boolean hypercube $\{0,1\}^d$, it is known that $\Omega(d/\varepsilon^2)$ samples are necessary to learn P within total variation (TV) distance ϵ. We revisit this problem when the learner is also given as advice the parameters of a product distribution Q. We show that there is an efficient algorithm to learn P within TV distance ε that has sample complexity $\tilde{O}(d^{1-\eta}/\varepsilon^2)$, if $\|\mathbf{p}-\mathbf{q}\|_1 < \varepsilon d^{0.5-\Omega(\eta)}$. Here, $\mathbf{p}$ and $\mathbf{q}$ are the mean vectors of P and Q respectively, and no bound on $\|\mathbf{p}-\mathbf{q}\|_1$ is known to the algorithm a priori.

1 Introduction

Science fundamentally relies on the ability to learn models from data. In many real-world settings, the majority of available datasets consist of unlabeled examples – sample points drawn without corresponding labels, outputs or classifications. These unlabeled datasets are often modeled as samples from a joint probability distribution on a large domain. The goal of *distribution learning* is to output the description of a distribution that approximates the underlying distribution that generated the observed samples. See [Dia16] for a comprehensive survey.

In practice, distribution learning rarely occurs in isolation. While a given dataset may be new, one often has access to previously learned models from related datasets. Alternatively, the data may arise from an evolving process, motivating the reuse of information from past learning to guide current inference. This prior information can be viewed as a form of "advice" or "prediction" of some kind. In the framework of algorithms with predictions, the objective is to integrate such advice in a way that improves performance when the advice is accurate, while ensuring robustness: performance should not degrade beyond that an advice-free baseline algorithm, even when the predictions are inaccurate. Most previous works in this setting are in the context of online algorithms, e.g. for the ski-rental problem [GP19, WLW20, ADJ$^+$20], non-clairvoyant scheduling [PSK18], scheduling [LLMV20, BMRS20, AJS22], augmenting classical data structures with predictions (e.g. indexing [KBC$^+$18] and Bloom filters [Mit18]), online selection and matching problems [AGKK20,

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

DLPLV21, CGLB24, CJS25], online TSP [BLMS$^+$22, GLS23], and a more general framework of online primal-dual algorithms [BMS20]. However, there have been some recent applications to other areas, e.g. graph algorithms [CSVZ22, DIL$^+$21], causal learning [CGB23], mechanism design [GKST22, ABG$^+$22], and most relevantly to us, distribution learning [BCGG25].

In this work, we study the problem of learning *product distributions* over the d-dimensional Boolean hypercube $\{0,1\}^d$, arguably one of the most fundamental classes of discrete high-dimensional distributions. A product distribution P is fully specified by its *mean vector* $\mathbf{p} \in [0,1]^d$, where the i-th coordinate $\mathbf{p}_i$ represents the expectation of the i-th marginal of P, or equivalently, the probability that the i-th coordinate of a sample from P is 1. It is well-known that $\Theta(d/\varepsilon^2)$ samples from a product distribution P are both necessary and sufficient to learn a distribution $\widehat{P}$ such that $\mathrm{d_{TV}}(P,\widehat{P}) \le \varepsilon$ with probability at least $2/3$, where $\mathrm{d_{TV}}$ denotes the total variation distance. This optimal sample complexity is achieved by a simple, natural and efficient algorithm: computing the empirical mean of each coordinate. Motivated by the framework of algorithms with predictions, we investigate whether this sample complexity can be improved when, in addition to samples from P, the learner is given an advice mean vector $\mathbf{q} \in [0,1]^d$. Importantly, we make no assumption that $\mathbf{q}$ is close to the true mean $\mathbf{p}$. However, if we can detect that $\mathbf{q}$ is accurate – i.e., that $\|\mathbf{q}-\mathbf{p}\|$ is small in an appropriate norm – can this information be leveraged to constrain the search space and improve sample or computational efficiency? Our goal is to design algorithms that adapt to the quality of the advice: performing better when $\mathbf{q}$ is accurate, while remaining robust when it is not.

Our main result establishes that this is indeed possible. Specifically, we show that if $\|\mathbf{q}-\mathbf{p}\|_1 \ll \varepsilon\sqrt{d}$, then there exists a polynomial-time algorithm with sample complexity *sublinear in* d that outputs a distribution $\widehat{P}$ such that $\mathrm{d_{TV}}(P,\widehat{P}) \le \varepsilon$ with probability at least $2/3$. More precisely, under the regularity condition that no coordinate of P is too close to deterministic (i.e., bounded away from 0 and 1), we show that the sample complexity is:

$$\tilde{O}\left(\frac{d}{\varepsilon^2}\left(d^{-\eta} + \min\left(1, \frac{\|\mathbf{p}-\mathbf{q}\|_1^2}{d^{1-4\eta}\varepsilon^2}\right)\right)\right)$$

for any small enough constant η. In particular, when $\|\mathbf{p}-\mathbf{q}\|_1$ is small, the dependence on d becomes sublinear. We also prove that the non-determinism assumption is necessary: if coordinates of P can be arbitrarily close to 0 or 1, then sample complexity that is sublinear in d is impossible, even when $\|\mathbf{q}-\mathbf{p}\|_1 = O(1)$. Furthermore, we show that when $\|\mathbf{q}-\mathbf{p}\|_1 \gg \varepsilon\sqrt{d}$, no algorithm with sublinear sample complexity exists.

1.1 Technical Overview

We call a product distribution P *balanced* if no marginal of P is too biased. That is, each coordinate of $\mathbf{p}$ is bounded away from 0 and 1. It is known that, for balanced distributions, learning P in TV distance is equivalent to learning the mean vector $\mathbf{p}$ with ℓ_2-error. Hence, in this overview, we focus on the latter task.

To build intuition about how an advice vector $\mathbf{q}$ can be exploited, consider the following two situations:

1. **Exact advice**: Suppose $\mathbf{q} = \mathbf{p}$. Then, it suffices to verify that $\|\mathbf{q}-\mathbf{p}\|_2 \le \varepsilon$ and a learning algorithm can simply return $\mathbf{q}$. This is the classic *identity testing* problem, which has been extensively studied (see [Can20] for a detailed survey). For product distributions, Daskalakis and Pan [DP17] and Canonne, Diakonikolas, Kane and Stewart [CDKS17] independently showed that identity testing requires $\Theta(\sqrt{d}/\varepsilon^2)$ samples. This demonstrates that sublinear sample complexity is achievable when $\mathbf{q} = \mathbf{p}$. Morever, $\Omega(\sqrt{d}/\varepsilon^2)$ is a fundamental lower bound that applies even when $\mathbf{q} = \mathbf{p}$.
2. **Sparse disagreement**: Suppose $\mathbf{q}$ differs from $\mathbf{p}$ in at most t coordinates, i.e. $\|\mathbf{q}-\mathbf{p}\|_0 \le t$. In this case, one only needs to estimate $\mathbf{p}$ on those t coordinates, so the information-theoretic sample complexity should scale as $\sim \log\binom{d}{t}$. Unfortunately, since t is unknown a priori, we cannot directly exploit this sparsity. However, from the compressive sensing literature, it is known that closeness in ℓ_2 norm to a t-sparse vector can be certified by a small ℓ_1 norm (e.g., Theorem 2.5 of [FR13]).

Motivated by the above, we take the quality of the advice $\mathbf{q}$ to be governed by the ℓ_1-distance $\|\mathbf{p}-\mathbf{q}\|_1$, and aim for an algorithm whose sample complexity improves as this distance decreases.

Suppose we can certify that $\|\mathbf{p}-\mathbf{q}\|_1 \le \lambda$. Then, we may restrict attention to the ℓ_1-ball of radius λ centered at $\mathbf{q}$, and cover it with N ℓ_2-balls of radius ε, where the covering number N grows polynomially as $d^{O(\lambda^2/\varepsilon^2)}$. It is known (e.g. see Chapter 4 of [DL01]) that using the Scheffé tournament method, the sample complexity of learning $\mathbf{p}$ up to ℓ_2-norm ε scales as $(\log N)/\varepsilon^2$, which yields a bound of $O(\frac{\lambda^2}{\varepsilon^4}\log d)$. While the Scheffé tournament is computationally inefficient, the same sample complexity guarantee can be achieved efficiently by solving a constrained least squares problem. More precisely, given samples $\mathbf{x}_1,\dots,\mathbf{x}_n$ from P, we consider the estimator:

$$\operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^d:\|\mathbf{b}-\mathbf{q}\|_1\le\lambda} \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{b}\|_2^2$$

For $n = O(\lambda^2\varepsilon^{-4}\log d)$, this estimates achieves ℓ_2-error at most ε.

The key challenge that remains is then to approximate $\lambda \approx \|\mathbf{p}-\mathbf{q}\|_1$ using a sublinear number of samples from P. To this end, we devise a new identity testing algorithm that, using $O(\sqrt{d}/\varepsilon^2)$ samples, either (i) 2-approximates $\|\mathbf{p}-\mathbf{q}\|_2$, or (ii) certifies that $\|\mathbf{p}-\mathbf{q}\|_2 \le \varepsilon$, in which we simply return $\mathbf{q}$. If we are in case (i), we can upper bound $\|\mathbf{p}-\mathbf{q}\|_1$ with $\lambda = \|\mathbf{p}-\mathbf{q}\|_2\cdot\sqrt{d}$. However, this would make the sample complexity of the learning algorithm to be $O\left(\frac{\lambda^2}{\varepsilon^4}\log d\right) = O\left(\frac{d\log(d)\cdot\|\mathbf{p}-\mathbf{q}\|_2^2}{\varepsilon^4}\right) \gg \frac{d}{\varepsilon^2}$, i.e., exceeding the standard $O(d/\varepsilon^2)$ bound and defeating the purpose. To improve upon this, we can partition the d coordinates into d/k blocks of size k each. Then, within each block, the ratio between the ℓ_1 and ℓ_2 norms improves from $\sqrt{d}$ to $\sqrt{k}$. By appropriately choosing k, we can obtain a non-trivial reduction in overall sample complexity.

The structure of our algorithm described above and its analysis parallels the recent work by [BCGG25], which addresses the problem of learning Gaussian distributions with imperfect advice. However, our setting differs in several important ways:

- [BCGG25] used a well-known algorithm to approximate the ℓ_2 norm of a Gaussian's mean vector. In contrast, our ℓ_2-approximation algorithm for product distribution is new, to the best of our knowledge.
- We critically rely on the balancedness assumption to relate total variation distance and ℓ_2 error of the mean vector. No such assumption is needed in the Gaussian setting. In fact, we show that for product distributions, balancedness is essential: without it, no sublinear-sample algorithm exists, even when the advice vector is $O(1)$-close to the truth in ℓ_1 distance. We find this somewhat surprising since $O(\sqrt{d}/\varepsilon^2)$ samples suffice without any balancedness assumptions in identity testing [DP17, CDKS17].

2 Preliminaries

A distribution P on $\{0,1\}^d$ is said to be a *product distribution* if there exist distributions $P_1,\dots,P_d$ on $\{0,1\}$ such that $P(\mathbf{x}) = P_1(x_1)\cdot P_2(x_2)\cdots P_d(x_d)$ for every $x\in\{0,1\}^d$. In this case, we can write $P = P_1\otimes P_2\cdots\otimes P_d$.

Definition 2.1 (Mean vectors)**.** The mean vector of a distribution P is $\mathbf{p}\triangleq\mathbb{E}_{x\sim P}[x]$. In particular, if $P = P_1\otimes\cdots\otimes P_d$ is a product distribution, $\mathbf{p} = [p_1 \quad \cdots \quad p_d]$, where $p_i = P_i(1)$.

For a vector $\mathbf{p}\in[0,1]^d$, we denote by $\mathrm{Ber}(\mathbf{p})$ the product distribution with mean vector $\mathbf{p}$. We next define the notion of balancedness.

Definition 2.2. For $\tau\in[0,1/2]$, a product distribution P on $\{0,1\}^d$ is said to be *τ-balanced* if for every $i\in[d]$, the marginal P_i satisfies $\tau\le P_i(1)\le 1-\tau$.

Proposition 2.3 (e.g., [CDKS17], Lemma 1)**.** *Suppose P and Q are τ-balanced product distributions on $\{0,1\}^d$ with mean vectors $\mathbf{p}$ and $\mathbf{q}$ respectively. Then their KL divergence $\mathrm{d}_{\mathrm{KL}}(P\|Q)$ satisfies:*

$$2\|\mathbf{p}-\mathbf{q}\|_2^2 \le \mathrm{d}_{\mathrm{KL}}(P\|Q) \le \frac{2}{\tau}\|\mathbf{p}-\mathbf{q}\|_2^2.$$

Proposition 2.4. *Suppose P and Q are τ-balanced product distributions on $\{0,1\}^d$ with mean vectors $\mathbf{p}$ and $\mathbf{q}$ respectively. Then, for some constant $c<0.2$, the TV distance $\mathrm{d}_{\mathrm{TV}}(P,Q)$ satisfies:*

$$c\cdot\min\{1,\|\mathbf{p}-\mathbf{q}\|_2\} \le \mathrm{d}_{\mathrm{TV}}(P,Q) \le \frac{1}{\sqrt{\tau}}\|\mathbf{p}-\mathbf{q}\|_2.$$

Proof. The first inquality is the main result of [Kon25]. The second inequality follows from applying Pinsker to the upper bound on $\mathrm{d_{KL}}$ in Proposition 2.3. □

Note that the dependence on τ above is necessary (up to constant factors). For example, suppose $P = \mathrm{Ber}(\mathbf{0})$ and $Q = \mathrm{Ber}(\mathbf{u})$ where $\mathbf{0}$ is the all-zero vector and $\mathbf{u} = [\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]$. Then, $\|\mathbf{0} - \mathbf{u}\|_2 = 1/\sqrt{d}$, while $\mathrm{d_{TV}}(P, Q) \geq P(\mathbf{0}) - Q(\mathbf{0}) = 1 - (1 - 1/d)^d \approx 1 - 1/e$.

3 Algorithm

The goal of this section is to establish the following result.

Theorem 3.1. *There exists algorithm* TESTANDOPTIMIZEMEAN *that for any given $\varepsilon, \delta, \tau \in (0, 1)$, $\eta \geq 0$, and $\mathbf{q} \in [0, 1]^d$, and sample access to a τ-balanced product distribution $\mathrm{Ber}(\mathbf{p})$ on $\{0, 1\}^d$, it draws $n = \tilde{O}\left(\frac{d}{\varepsilon^2} \cdot (d^{-\eta} + \min\{1, f(\mathbf{p}, \mathbf{q}, d, \eta, \varepsilon)\})\right)$ i.i.d. samples from $\mathrm{Ber}(\mathbf{p})$, where:*

$$f(\mathbf{p}, \mathbf{q}, d, \eta, \varepsilon) = \frac{\|\mathbf{p} - \mathbf{q}\|_1^2}{d^{1-4\eta}\tau^6\varepsilon^2}.$$

The algorithm produces as output $\widehat{\mathbf{p}}$ in $\mathrm{poly}(n, d)$ time such that $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\widehat{\mathbf{p}})) \leq \varepsilon$ with success probability at least $1 - \delta$.

A basic component of the algorithm is a test to determine how close the advice $\mathbf{q}$ is to the true $\mathbf{p}$ in ℓ_2 norm.

Lemma 3.2 (Tolerant mean tester)**.** *Given $\varepsilon > 0$, $\delta \in (0, 1)$, d sufficiently large integer, and $\mathbf{q} \in [0, 1]^d$, there is a tolerant tester* TMT *that uses $O\left(\frac{\sqrt{d}}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$ i.i.d. samples from $\mathrm{Ber}(\mathbf{p})$ and satisfies both conditions below with probability at least $1 - \delta$:*

1. *If $\|\mathbf{p} - \mathbf{q}\|_2 \leq \varepsilon$, then the tester outputs* Accept
2. *If $\|\mathbf{p} - \mathbf{q}\|_2 \geq 2\varepsilon$, then the tester outputs* Reject

Proof. Notice that if $\sum p_i = \sum q_i = 1$, we could interpret $p_1, \ldots, p_d$ and $q_1, \ldots, q_d$ as distributions $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$ on $[d]$ that sample $i \in [d]$ with probability p_i and q_i respectively. Diakonikolas and Kane [DK16] showed that using $O(\|\tilde{\mathbf{p}}\|_2/\varepsilon^2)$ samples from $\tilde{\mathbf{p}}$, one can test whether $\|\tilde{\mathbf{p}} - \tilde{\mathbf{q}}\|_2 \leq \varepsilon$ or $\|\tilde{\mathbf{p}} - \tilde{\mathbf{q}}\|_2 \geq 2\varepsilon$. Inspired by this observation, we mimic the analysis of [DK16] to devise a tolerant tester for general product distributions.

Assume the desired failure probability to be $1/3$; we can reduce to any δ by repeating the test $O(\log 1/\delta)$ times and taking the majority vote. Set $m = c\sqrt{d}/\varepsilon^2$ for a large enough constant c, and let m_i be sampled independently from $\mathrm{Poi}(m)$ for each $i \in [d]$. Note that $\max_i m_i \leq 2em$ with high probability; we condition on this event and set the desired failure probability to $1/4$. Therefore, using $2em$ samples from $\mathrm{Ber}(\mathbf{p})$, for each i, we can obtain m_i samples from the ith coordinate, and let X_i be the number of times the i'th coordinate is sampled to be 1. Note that by standard properties of the Poisson distribution, the $X_1, \ldots, X_d$ are independent and each X_i is sampled from $\mathrm{Poi}(mp_i)$.

Define the statistic $Z = \sum_{i=1}^d Z_i$, where:

$$Z_i = (X_i - mq_i)^2 - X_i.$$

Using similar calculations as in [DK16], we can show that:

$$\mathbb{E}[Z_i] = m^2(p_i - q_i)^2 \qquad \text{and} \qquad \mathbb{E}[Z] = m^2\|\mathbf{p} - \mathbf{q}\|_2^2.$$

Also, we can calculate the variance to be:

$$\mathsf{Var}[Z] = 4m^3 \sum_{i=1}^d p_i(p_i - q_i)^2 + 2m^2 \sum_{i=1}^d p_i^2 \leq 4m^3\|\mathbf{p}\|_2\|\mathbf{p} - \mathbf{q}\|_4^2 + 2m^2\|\mathbf{p}\|_2^2,$$

where the inequality is by Cauchy-Schwarz.

If $\|\mathbf{p} - \mathbf{q}\|_2 \leq \varepsilon$, then $\mathbb{E}[Z] \leq c^2 d/\varepsilon^2$, and $\mathsf{Var}[Z] \leq (4c^3 + 2c^2)d^2/\varepsilon^4$. On the other hand, it always holds that $\mathbb{E}[Z] \geq c^2 d\|\mathbf{p} - \mathbf{q}\|_2^2/\varepsilon^4$, and $\mathsf{Var}[Z] \leq (4c^3 d^{1.5}\|\mathbf{p}\|_2\|\mathbf{p} - \mathbf{q}\|_2^2/\varepsilon^2 + 2c^2 d\|\mathbf{p}\|_2^2)/\varepsilon^4 \leq (4c^3\|\mathbf{p} - \mathbf{q}\|_2^2/\varepsilon^2 + 2c^2)d^2/\varepsilon^4$, since $\|\mathbf{p}\|_2 \leq \sqrt{d}$. Using Chebyshev's inequality, if c is large enough, when $\|\mathbf{p} - \mathbf{q}\|_2 \leq \varepsilon$, Z is at most $2c^2 d/\varepsilon^2$ with probability $3/4$, but when $\|\mathbf{p} - \mathbf{q}\|_2 \geq 2\varepsilon$, Z is at least $3c^2 d/\varepsilon^2$ with probability $3/4$. □

Algorithm 1 The APPROXL1 algorithm.

Input: Block size $k \in [d]$, lower bound $\alpha > 0$, upper bound $\zeta > 2\alpha$, failure rate $\delta \in (0,1)$, advice $\mathbf{q} \in [0,1]^d$, and i.i.d. samples $\mathcal{S}$ (multiset) from $\mathrm{Ber}(\mathbf{p})$.
Output: Fail or $\lambda \in \mathbb{R}$
Define $w = \lceil d/k \rceil$ and $\delta' = \frac{\delta}{w \cdot \lceil \log_2 \zeta/\alpha \rceil}$
Partition the index set $[d]$ into w blocks:

$$\mathbf{B}_1 = \{1, \ldots, k\}, \mathbf{B}_2 = \{k+1, \ldots, 2k\}, \ldots, \mathbf{B}_w = \{k(w-1)+1, \ldots, d\}$$

for $j \in \{1, \ldots, w\}$ **do**
  Define multiset $\mathcal{S}_j = \{\mathbf{x}_{\mathbf{B}_j} \in \mathbb{R}^{|\mathbf{B}_j|} : \mathbf{x} \in \mathcal{S}\}$ as the samples projected to $\mathbf{B}_j$
  Let $\mathbf{q}_{\mathbf{B}_j} \in \mathbb{R}^{|\mathbf{B}_j|}$ be the vector $\mathbf{q}$ projected to coordinates $\mathbf{B}_j$.
  Initialize $o_j =$ Fail
  for $i = 1, 2, \ldots, \lceil \log_2 \zeta/\alpha \rceil$ **do**
    Define $l_i = 2^{i-1} \cdot \alpha$
    Let `Outcome` be the output of the tolerant tester TMT of Lemma 3.2 using sample set $\mathcal{S}_j$
     with parameters $\varepsilon \leftarrow l_i, \delta \leftarrow \delta', d \leftarrow |\mathbf{B}_j|$ and $\mathbf{q} \leftarrow \mathbf{q}_{\mathbf{B}_j}$
    if `Outcome` is Accept **then**
      Set $o_j = l_i$ and **break** {Escape inner loop for block j}
    end if
  end for
end for
if there exists a Fail amongst $\{o_1, \ldots, o_w\}$ **then**
  return Fail
else
  return $\lambda = 2\sum_{j=1}^{w} \sqrt{|\mathbf{B}_j|} \cdot o_j$ {λ is an estimate for $\|\mathbf{p} - \mathbf{q}\|_1$}
end if

Lemma 3.3. *Let k, α, and ζ be the input parameters to the* APPROXL1 *algorithm. Given $q \in [0,1]^d$ and $m(k, \alpha, \delta) \triangleq \lceil \frac{16\sqrt{k}}{3\alpha^2} \rceil \cdot \left(1 + \lceil \log\left(\frac{12w \cdot \log_2 \lceil \zeta/\alpha \rceil}{\delta}\right) \rceil\right)$ i.i.d. samples from* $\mathrm{Ber}(\mathbf{p})$*,* APPROXL1 *succeeds with probability at least $1 - \delta$ and has the following properties:*
Property 1: If APPROXL1 *outputs* Fail*, then $\|\mathbf{p} - \mathbf{q}\|_2 > \zeta/2$.*
Property 2: If APPROXL1 *outputs $\lambda \in \mathbb{R}$, then $\|\mathbf{p} - \mathbf{q}\|_1 \le \lambda \le 2\sqrt{k} \cdot (\lceil d/k \rceil \cdot \alpha + 2\|\mathbf{p} - \mathbf{q}\|_1)$.*

Proof. We begin by stating some properties of $o_1, \ldots, o_w$. Fix an arbitrary index $j \in \{1, \ldots, w\}$ and suppose o_j is *not* a Fail, i.e. the tolerant tester of Lemma 3.2 outputs Accept for some $i^* \in \{1, 2, \ldots, \lceil \log_2 \zeta/\alpha \rceil\}$. Note that APPROXL1 sets $o_j = \ell_{i^*}$ and the tester outputs Reject for all smaller indices $i \in \{1, \ldots, i^* - 1\}$. Since the tester outputs Accept for i^*, we have that $\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \le 2\ell_{i^*} = 2o_j$. Meanwhile, if $i^* > 1$, then $\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > \ell_{i^*-1} = \ell_{i^*}/2 = o_j/2$ since the tester outputs Reject for $i^* - 1$. Thus, we see that

- When o_j is not Fail, we have $\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \le 2o_j$.
- When $\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \le 2\alpha$, we have $i^* = 1$ and $o_j = \ell_1 = \alpha$.
- When $\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha = 2\ell_1$, we have $i^* > 1$ and so $o_j < 2\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2$.

Success probability. Fix an arbitrary index $i \in \{1, 2, \ldots, \lceil \log_2 \zeta/\alpha \rceil\}$ with $\ell_i = 2^{i-1}\alpha$, where $\ell_i \le \ell_1 = \alpha$ for any i. We invoke the tolerant tester with $\varepsilon = \ell_i$ in the i^{th} invocation, so the i^{th} invocation uses at most $m_1(k, \varepsilon, \delta) \triangleq n_{k,\varepsilon} \cdot r_\delta$ i.i.d. samples to succeed with probability at least $1 - \delta$, where $n_{k,\varepsilon} \triangleq \lceil \frac{16\sqrt{k}}{3\varepsilon^2} \rceil$ and $r_\delta \triangleq 1 + \lceil \log(12/\delta) \rceil$.

So, with at most $m(k, \alpha, \delta) \triangleq m_1(k, \alpha, \delta') = n_{k,\alpha} \cdot r_{\delta'}$ samples, *any* call to the tolerant tester succeeds with probability at least $1 - \delta'$, where $\delta' \triangleq \frac{\delta}{w \cdot \lceil \log_2 \zeta/\alpha \rceil}$. By construction, there will be at most $w \cdot \lceil \log_2 \zeta/\alpha \rceil$ calls to the tolerant tester. Therefore, by union bound, *all* calls to the tolerant tester *jointly succeed* with probability at least $1 - \delta$.

Proof of Property 1. When APPROXL1 outputs Fail, there exists a Fail amongst $\{o_1, \ldots, o_w\}$. For any fixed index $j \in \{1, \ldots, w\}$, this can only happen when all calls to the tolerant tester outputs Reject. This means that $\|\boldsymbol{x}_{\mathbf{B}_j}\|_2 > \varepsilon_1 = \ell_i = 2^{i-1} \cdot \alpha$ for all $i \in \{1, 2, \ldots, \lceil \log_2 \zeta/\alpha \rceil\}$. In particular, this means that $\|\boldsymbol{x}_{\mathbf{B}_j}\|_2 > \zeta/2$.

Proof of Property 2. When APPROXL1 outputs $\lambda = 2\sum_{j=1}^{w} \sqrt{|\mathbf{B}_j|} \cdot o_j \in \mathbb{R}$, we can lower bound λ as follows:

$$\begin{aligned}
\lambda = 2\sum_{j=1}^{w} \sqrt{|\mathbf{B}_j|} \cdot o_j &\geq 2\sum_{j=1}^{w} \sqrt{|\mathbf{B}_j|} \cdot \frac{\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2}{2} && \text{(since } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2o_j\text{)} \\
&\geq \sum_{j=1}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 && \text{(since } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 \leq \sqrt{|\mathbf{B}_j|} \cdot \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2\text{)} \\
&= \|\mathbf{p} - \mathbf{q}\|_1 && \text{(since } \textstyle\sum_{j=1}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 = \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1\text{)}
\end{aligned}$$

That is, $\lambda \geq \|\mathbf{p} - \mathbf{q}\|_1$. Meanwhile, we can also upper bound λ as follows:

$$\begin{aligned}
\lambda = 2\sum_{j=1}^{w} \sqrt{|\mathbf{B}_j|} \cdot o_j &\leq 2\sqrt{k}\sum_{j=1}^{w} o_j && \text{(since } |\mathbf{B}_j| \leq k\text{)} \\
&= 2\sqrt{k} \cdot \left(\sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2\alpha}}^{w} o_j + \sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} o_j \right) \\
&&& \text{(partitioning the blocks based on } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \text{ versus } 2\alpha\text{)} \\
&= 2\sqrt{k} \cdot \left(\sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2\alpha}}^{w} \alpha + \sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} o_j \right) \\
&&& \text{(since } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2\alpha \text{ implies } o_j = \alpha\text{)} \\
&\leq 2\sqrt{k} \cdot \left(\sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2\alpha}}^{w} \alpha + \sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} 2\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \right) \\
&&& \text{(since } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha \text{ implies } o_j \leq 2\|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2\text{)} \\
&\leq 2\sqrt{k} \cdot \left(\sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq 2\alpha}}^{w} \alpha + 2 \sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 \right) \\
&&& \text{(since } \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 \leq \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1\text{)} \\
&\leq 2\sqrt{k} \cdot \left(\lceil d/k \rceil \cdot \alpha + 2 \sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 \right) \\
&&& \text{(since } |\{j \in [w] : \mathbf{p}_{\mathbf{B}_j}\|_2 \leq 2\alpha\}| \leq w\text{)} \\
&\leq 2\sqrt{k} \cdot (\lceil d/k \rceil \cdot \alpha + 2\|\mathbf{p} - \mathbf{q}\|_1) \\
&&& \text{(since } \textstyle\sum_{\substack{j=1 \\ \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_2 > 2\alpha}}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 \leq \sum_{j=1}^{w} \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1 = \|\mathbf{p}_{\mathbf{B}_j} - \mathbf{q}_{\mathbf{B}_j}\|_1\text{)}
\end{aligned}$$

That is, $\lambda \leq 2\sqrt{k} \cdot (\lceil d/k \rceil \cdot \alpha + 2\|\mathbf{p} - \mathbf{q}\|_1)$. The property follows by putting together both bounds. □

Now, suppose APPROXL1 tells us that $\|\mathbf{p}-\mathbf{q}\|_1 \le r$. We can then perform a constrained LASSO to search for a candidate $\widehat{\mathbf{p}} \in [0,1]^d$ using $O(\frac{r^2}{\varepsilon^4}\log\frac{d}{\delta})$ samples from $\mathrm{Ber}(\mathbf{p})$.

Lemma 3.4. *Fix $d \ge 1$, $r \ge 0$, $\varepsilon, \delta > 0$, and $\mathbf{q} \in [0,1]^d$. Given $O(\frac{r^2}{\varepsilon^4}\log\frac{d}{\delta})$ samples from $\mathrm{Ber}(\mathbf{p})$ for some unknown $\mathbf{p} \in [0,1]^d$ with $\|\mathbf{p}-\mathbf{q}\|_1 \le r$, one can produce an estimate $\widehat{\mathbf{p}} \in [0,1]^d$ in $\mathrm{poly}(n,d)$ time such that $\|\widehat{\mathbf{p}}-\mathbf{p}\|_2 \le \varepsilon$ with success probability at least $1-\delta$.*

Proof. Suppose we get n samples $\mathbf{y}_1,\dots,\mathbf{y}_n \sim \mathrm{Ber}(\mathbf{p})$. For $i \in [n]$, we can re-express each $\mathbf{y}_i$ as $\mathbf{y}_i = \mathbf{p} + \mathbf{z}_i$ for some $\mathbf{z}_i$ distributed as $\mathrm{Ber}(\mathbf{p}) - \mathbf{p}$. Let us define $\widehat{\mathbf{p}} \in [0,1]^d$ as follows:

$$\widehat{\mathbf{p}} = \operatorname*{argmin}_{\|\mathbf{b}-\mathbf{q}\|_1 \le r} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{b}\|_2^2 \tag{1}$$

By optimality of $\widehat{\mathbf{p}}$ in Equation (1), we have

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \widehat{\mathbf{p}}\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{p}\|_2^2 \tag{2}$$

By expanding and rearranging Equation (2), one can show:

$$\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2 \le \frac{2}{n}\left\langle \sum_{i=1}^{n} \mathbf{z}_i, \widehat{\mathbf{p}}-\mathbf{p} \right\rangle \tag{3}$$

Meanwhile, a standard Chernoff bound shows that $\Pr\left[\|\sum_{i=1}^n \mathbf{z}_i\|_\infty \ge \sqrt{2n\log\left(\frac{2d}{\delta}\right)}\right] \le \delta$. Therefore, using Hölder's inequality and triangle inequality with the above, we see that, with probability at least $1-\delta$,

$$\begin{aligned}
\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2 \le \frac{2}{n}\langle \sum_{i=1}^{n} \mathbf{z}_i, \widehat{\mathbf{p}}-\mathbf{p}\rangle &\le \frac{2}{n}\cdot \left\|\sum_{i=1}^{n}\mathbf{z}_i\right\|_\infty \cdot \|\widehat{\mathbf{p}}-\mathbf{p}\|_1 \\
&\le \frac{2}{n}\cdot \left\|\sum_{i=1}^{n}\mathbf{z}_i\right\|_\infty \cdot (\|\widehat{\mathbf{p}}-\mathbf{q}\|_1 + \|\mathbf{p}-\mathbf{q}\|_1) \\
&\le 4r\cdot\sqrt{\frac{2\log\left(\frac{2d}{\delta}\right)}{n}}
\end{aligned}$$

Finally, it is known that LASSO runs in $\mathrm{poly}(n,d)$ time. □

Using Lemma 3.4, we now ready to prove Theorem 3.1.

Proof of Theorem 3.1. **Correctness of $\widehat{\mathbf{p}}$ output.** TESTANDOPTIMIZEMEAN (Algorithm 2) has two possible outputs for $\widehat{\mathbf{p}}$:
Case 1: $\widehat{\mathbf{p}} = \mathrm{argmin}_{\|\mathbf{b}-\mathbf{q}\|_1 \le \lambda} \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{b}\|_2^2$, which can only happen when `Outcome` is $\lambda \in \mathbb{R}$ and $\lambda < \varepsilon\sqrt{d}$
Case 2: $\widehat{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i$

Conditioned on APPROXL1 succeeding, with probability at least $1-\delta$, we will show that $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\widehat{\mathbf{p}})) \le \varepsilon$ and failure probability at most δ in each of these cases, which implies the theorem statement.

Case 1: Using $r = \lambda$ as the upper bound, Lemma 3.4 tells us that $\|\widehat{\mathbf{p}}-\mathbf{p}\|_2 \le \varepsilon\sqrt{\tau(1-\tau)}/2$ with failure probability at most δ when $\widetilde{O}(\lambda^2/\tau^2\varepsilon^4)$ i.i.d. samples are used. Using Proposition 2.4, $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\widehat{\mathbf{p}})) \le \varepsilon$.

Case 2: With $\widetilde{O}(d/\varepsilon^2)$ samples, it is known that the empirical mean $\widehat{\mathbf{p}}$ achieves $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\widehat{\mathbf{p}})) \le \varepsilon$ with failure probability at most δ.

Algorithm 2 The TESTANDOPTIMIZEMEAN algorithm.

Input: Error rate $\varepsilon > 0$, failure rate $\delta \in (0,1)$, parameter $\eta \in [0, \frac{1}{4}]$, parameter $\tau \in [0, \frac{1}{2}]$, and sample access to $\mathrm{Ber}(\mathbf{p})$
Output: $\widehat{\mathbf{p}} \in \mathbb{R}^d$
Define $k = \min(\lceil d^{4\eta}/\tau^4 \rceil, d)$, $\alpha = \varepsilon d^{(3\eta-1)/2}/\tau$, $\zeta = 4\varepsilon \cdot \sqrt{d}$, and $\delta' = \frac{\delta}{\lceil d/k \rceil \cdot \lceil \log_2 \zeta/\alpha \rceil}$
Draw $O(\sqrt{k}\log(1/\delta')/\alpha^2)$ i.i.d. samples from $\mathrm{Ber}(\mathbf{p})$ and store it into a set $\mathcal{S}$
Let `Outcome` be the output of the APPROXL1 algorithm given k, α, ζ, and $\mathcal{S}$ as inputs
if `Outcome` is $\lambda \in \mathbb{R}$ and $\lambda < \varepsilon\sqrt{d}$ **then**
    Draw $n \in \widetilde{O}(\lambda^2/\varepsilon^4)$ i.i.d. samples $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \{0,1\}^d$
    return $\widehat{\mathbf{p}} = \mathrm{argmin}_{\|\mathbf{b}-\mathbf{q}\|_1 \le \lambda} \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{b}\|_2^2$
else
    Draw $n \in \widetilde{O}(d/\varepsilon^2)$ i.i.d. samples $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \{0,1\}^d$
    return $\widehat{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i$ {Empirical mean}
end if

Sample complexity used. APPROXL1 uses $|\mathbf{S}| = m(k, \alpha, \delta') \in \widetilde{O}(\sqrt{k}/\alpha^2)$ samples to produce `Outcome`. Then, APPROXL1 further uses $\widetilde{O}(\lambda^2/\tau^2\varepsilon^4)$ samples or $\widetilde{O}(d/\varepsilon^2)$ samples depending on whether $\lambda < \varepsilon\sqrt{d}$. So, TESTANDOPTIMIZEMEAN has a total sample complexity of $\widetilde{O}\left(\frac{\sqrt{k}}{\alpha^2} + \min\left\{\frac{\lambda^2}{\tau^2\varepsilon^4}, \frac{d}{\varepsilon^2}\right\}\right)$. Meanwhile, Lemma 3.3 states that $\|\mathbf{p}-\mathbf{q}\|_1 \le \lambda \le 2\sqrt{k} \cdot (\lceil d/k \rceil \cdot \alpha + 2\|\mathbf{p}-\mathbf{q}\|_1)$ whenever `Outcome` is $\lambda \in \mathbb{R}$. Since $(a+b)^2 \le 2a^2 + 2b^2$ for any two real numbers $a, b \in \mathbb{R}$, we see that $\frac{\lambda^2}{\tau^2\varepsilon^4} \in O\left(\frac{k}{\tau^2\varepsilon^4} \cdot \left(\frac{d^2\alpha^2}{k^2} + \|\mathbf{p}-\mathbf{q}\|_1^2\right)\right) \subseteq O\left(\frac{d}{\varepsilon^2} \cdot \frac{1}{\tau^2}\left(\frac{d\alpha^2}{\varepsilon^2 k} + \frac{k\cdot\|\mathbf{p}-\mathbf{q}\|_1^2}{d\varepsilon^2}\right)\right)$. Putting together the above observations, we see that the total sample complexity is

$$\widetilde{O}\left(\frac{\sqrt{k}}{\alpha^2} + \frac{d}{\varepsilon^2} \cdot \min\left\{1, \frac{d\alpha^2}{\varepsilon^2\tau^2 k} + \frac{k \cdot \|\mathbf{p}-\mathbf{q}\|_1^2}{d\tau^2\varepsilon^2}\right\}\right).$$

Recalling that TESTANDOPTIMIZEMEAN sets $k = \min(\lceil d^{4\eta}\tau^{-4} \rceil, d)$ and $\alpha = \varepsilon d^{(3\eta-1)/2}\tau^{-1}$, the above expression simplifies to $\widetilde{O}\left(\frac{d}{\varepsilon^2} \cdot \left(d^{-\eta} + \min\left(1, \frac{\|\mathbf{p}-\widetilde{\mathbf{p}}\|_1^2}{d^{1-4\eta}\tau^6\varepsilon^2}\right)\right)\right)$.. □

4 Lower Bounds

For proving of our lower bounds, we use the following corollary of Fano's inequality.

Lemma 4.1 (Lemma 6.1 of [ABDH+20]). *Let $\kappa : \mathbb{R} \to \mathbb{R}$ be a function and let $\mathcal{F}$ be a class of distributions such that, for all $\varepsilon > 0$, there exist distributions $f_1, \ldots, f_M \in \mathcal{F}$ such that*

$$\mathrm{d_{KL}}(f_i, f_j) \le \kappa(\varepsilon) \text{ and } \mathrm{d_{TV}}(f_i, f_j) > 2\varepsilon \ \forall i \ne j \in [M]$$

Then any method that learns $\mathcal{F}$ to within total variation distance ε with probability $\ge 2/3$ has sample complexity $\Omega\left(\frac{\log M}{\kappa(\varepsilon)\log(1/\varepsilon)}\right)$.

Lemma 4.2 (Learning unbalanced distributions requires linear samples). *Suppose ε is sufficiently small, and we are given sample access to a product distribution $\mathrm{Ber}(\mathbf{p})$ on $\{0,1\}^d$ with mean vector $\mathbf{p}$ having entries which are $O(1/d)$, along with an advice mean vector $\mathbf{q}$ such that $\|\mathbf{p}-\mathbf{q}\|_1 \le O(\varepsilon)$. Even in this case, learning $\widehat{\mathbf{p}}$ such that $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\widehat{\mathbf{p}})) \le \varepsilon$ requires $\widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$ samples*

Proof. Suppose that the advice distribution $\mathrm{Ber}(\mathbf{q})$ has mean vector $\mathbf{q} \triangleq \left[\frac{\varepsilon}{d} \ \cdots \ \frac{\varepsilon}{d}\right]$. If $S \subseteq [d]$, define $\mathbf{p}_S \in [0,1]^d$ with $\mathbf{p}_S[i] = \frac{2\varepsilon}{d}$ if $i \in S$ and $= \frac{\varepsilon}{d}$ otherwise. Then, we have $\|\mathbf{p}_S - \mathbf{q}\|_1 = |S|\frac{\varepsilon}{d}$ for all $S \subseteq [d]$. Also, for all $S, T \subseteq [d]$ we have $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p_S}), \mathrm{Ber}(\mathbf{p_T})) \ge \left|\Pr_{x\sim\mathrm{Ber}(\mathbf{p_S})}(x_{S\setminus T} = \mathbf{0}) - \Pr_{x\sim\mathrm{Ber}(\mathbf{p_T})}(x_{S\setminus T} = \mathbf{0})\right| = \left|\left(1-\frac{2\varepsilon}{d}\right)^{|S\setminus T|} - \left(1-\frac{\varepsilon}{d}\right)^{|S\setminus T|}\right| \ge 1 - \left(\frac{|S\setminus T|\varepsilon}{d} + \frac{1}{1+2\frac{|S\setminus T|\varepsilon}{d}}\right)$; this is using the inequalities $(1-x)^r \ge 1 - rx$ for $r \in \{0\} \cup [1, \infty)$,

$x \le 1$, and $(1-x)^r \le \frac{1}{1+rx}$ for $r \ge 0$, $x \in (-1/r, 1]$. Using the same argument with the set $T \setminus S$, we get $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p_S}), \mathrm{Ber}(\mathbf{p_T})) \ge 1 - \left(\frac{|T\setminus S|\varepsilon}{d} + \frac{1}{1+2\frac{|T\setminus S|\varepsilon}{d}}\right)$. Thus, we have $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p_S}), \mathrm{Ber}(\mathbf{p_T})) \ge \max_{\substack{H\in\{S\setminus T, T\setminus S\} \\ \xi \triangleq |H|\varepsilon/d}} 1 - \left(\xi + \frac{1}{1+2\xi}\right)$. Note that, by calculation, we can show that $1 - \left(\xi + \frac{1}{1+2\xi}\right) \ge \xi/2 - \xi^2$ for $\xi \ge 0$, which is $\Omega(\xi)$ for $\xi \in (0, 1/4)$.

Similarly, we have $\mathrm{d_{KL}}(\mathrm{Ber}(\mathbf{p_S})\|\mathrm{Ber}(\mathbf{p_T})) = \sum_{i\in[d]} \mathsf{kl}([\mathbf{p}_S]_i, [\mathbf{p}_T]_i) = \sum_{i\in S\setminus T} \mathsf{kl}(\frac{2\varepsilon}{d}, \frac{\varepsilon}{d}) + \sum_{i\in T\setminus S} \mathsf{kl}(\frac{\varepsilon}{d}, \frac{2\varepsilon}{d})$ (where $\mathsf{kl}(p,q) \triangleq \mathrm{d_{KL}}(\mathrm{Ber}(p)\|\mathrm{Ber}(q))$). We can see by simple calculations along with the logarithmic inequality $\ln(1+x) \le x$ for $x > -1$, that $\mathsf{kl}(\frac{\varepsilon}{d}, \frac{2\varepsilon}{d}) \le \frac{\varepsilon}{d}\left(1 - \ln(2) + \frac{\varepsilon}{d-2\varepsilon}\right) \le \frac{0.5\varepsilon}{d}$ (for $d \ge 10$ and $\varepsilon \le 1$), and $\mathsf{kl}\left(\frac{2\varepsilon}{d}, \frac{\varepsilon}{d}\right) \le \frac{\varepsilon}{d}\left(2\ln(2) - 1 + \frac{\varepsilon}{d-\varepsilon}\right) \le \frac{0.5}{d}$ (for $d \ge 10$ and $\varepsilon \le 1$). Thus $\mathrm{d_{KL}}(\mathrm{Ber}(\mathbf{p_S})\|\mathrm{Ber}(\mathbf{p_T})) \le \frac{\varepsilon}{2d}(|S \setminus T| + |T \setminus S|) = \frac{|S\oplus T|\varepsilon}{2d}$.

Viewing sets $S \subseteq [d]$ as vectors in $\mathbb{F}_2^d$ and using the Gilbert-Varshamov bound, we can say that, for any constant $c \in (0,1)$ and sufficiently large d, there exists a family of sets $\{S_1, \ldots, S_M\} \subseteq 2^{[d]}$ with $M \ge 2^{\Omega(cd)}$ such that $|S_i| = cd$ and $|S_i \oplus S_j| \ge \frac{cd}{4}$ for all $i, j \in [M]$. We use this family to instantiate distributions $f_i \triangleq \mathrm{Ber}(\mathbf{p_{S_i}})$ for each $i \in [M]$.

Suppose we take $S = S_i$, $T = S_j$, with $|S| = |T| = cd$ and $|S \oplus T| \in \left[\frac{cd}{4}, 2cd\right]$, so that $\mathrm{d_{KL}}(f_i\|f_j) \le \frac{|S\oplus T|\varepsilon}{2d} \le c\varepsilon$ for all $i, j \in [M]$. Since $|S_i \oplus S_j| \ge cd/4$, we will have at least one of $H \in \{S \setminus T, T \setminus S\}$ with $|H| \in \left[\frac{cd}{8}, cd\right]$. Thus, we will have $\frac{|H|\varepsilon}{d} \in \left[\frac{c\varepsilon}{8}, c\varepsilon\right] = \Theta(\varepsilon)$ (for constant $c > 0$), and $\mathrm{d_{TV}}(f_i, f_j) \ge \Omega(\varepsilon)$ (supposing $c\varepsilon < 1/4$).

By appropriately scaling ε and applying Lemma 4.1, we can show that we need $\tilde{\Omega}\left(\frac{d}{\varepsilon}\right)$ samples to learn a product distribution $f^* = \mathrm{Ber}(\mathbf{p_{S^*}}) \in \{f_1, \ldots, f_M\}$ to within ε in TV distance, even when given advice $\mathbf{q}$ with $\|\mathbf{p}_{S^*} - \mathbf{q}\|_1 < \varepsilon$, if the distribution mean vector $\mathbf{p}_{S^*}$ are allowed to be unbalanced (specifically, with entries $\le O(1/d)$). □

We also prove a sample complexity lower bound for learning product distributions balanced case given advice, which adapts the sample complexity lower bound in ([BCGG25], Lemma 32) for learning multivariate isotropic gaussians $\mathcal{N}(\mu^*, I_d)$ given an advice vector which is close to the true mean vector in ℓ_1 distance.

Lemma 4.3. *Let $\varepsilon > 0$ be sufficiently small. Suppose that we are given sample access to a distribution $\mathrm{Ber}(\mathbf{p})$ where $\mathbf{p}$ is $\frac{1}{4}$-balanced, and also an advice vector $\mathbf{q}$ with $\|\mathbf{p} - \mathbf{q}\|_1 = \lambda \ge 100\varepsilon$. Then, any algorithm that learns $\mathrm{Ber}(\mathbf{p})$ up to distance ε in total variation with constant failure probability requires $\widetilde{\Omega}\left(\min\left\{\|\mathbf{p} - \mathbf{q}\|_1^2/\varepsilon^4, d/\varepsilon^2\right\}\right)$ samples. In particular, when $\|\mathbf{p} - \mathbf{q}\|_1 \ge \varepsilon\sqrt{d}$, we need $\Omega(d/\varepsilon^2)$ samples.*

Proof. Suppose we want $\|\mathbf{p} - \mathbf{q}\|_1 = \lambda$ for λ sufficiently small. Fix $\mathbf{q} = \begin{bmatrix}\frac{1}{2} & \cdots & \frac{1}{2}\end{bmatrix}$ and suppose $\mathbf{p} = \mathbf{p}_S$ for some $S \subseteq [d]$ with $|S| = k$ such that $\mathbf{p}_S[i] = \frac{1}{2} + \frac{\lambda}{k}$ for $i \in S$ and $= \frac{1}{2}$ otherwise. Then $\|\mathbf{p}_S - \mathbf{q}\|_1 = \lambda$ and $\|\mathbf{p}_S - \mathbf{p}_T\|_2 = \frac{\lambda}{k}\sqrt{|S \oplus T|}$. If $\frac{\lambda}{k} < \frac{1}{4}$, the distributions $\mathbf{p}_S$ are τ-balanced for $\tau = \frac{1}{4}$. In that case, for any $S, T \subseteq [d]$, we can bound $\mathrm{d_{KL}}(\mathrm{Ber}(\mathbf{p_S})\|\mathrm{Ber}(\mathbf{p_T})) \le 8\left(\frac{\lambda}{k}\right)^2 |S \oplus T|$ (by Proposition 2.3), and the total variation distance by $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p}), \mathrm{Ber}(\mathbf{q})) \ge \Omega\left(\min\left\{1, \frac{\lambda}{k}\sqrt{|S \oplus T|}\right\}\right)$ (Proposition 2.4).

As in ([BCGG25], Lemma 32), we consider $\{S_1, \ldots, S_M\}$ and take $\mathbf{p}_i \triangleq \mathbf{p}_{S_i}$ for a family of k-subsets with $M \ge 2^{\Omega(k)}$ and $|S_i \oplus S_j| \ge k/4$ for all $i \ne j$, known to exist via the Gilbert-Varshamov bound. We can do this as long as, e.g. $k \ge 10$. This gives $\mathrm{d_{KL}}(\mathrm{Ber}(\mathbf{p_i})\|\mathrm{Ber}(\mathbf{p_j})) \le \frac{16\lambda^2}{k}$ (where λ is a function of ε) and $\mathrm{d_{TV}}(\mathrm{Ber}(\mathbf{p_i}), \mathrm{Ber}(\mathbf{p_j})) \ge c\frac{\lambda}{2\sqrt{k}}$ as long as $\frac{\lambda}{2\sqrt{k}} < 1$.

If we choose $k = \lceil\frac{\lambda^2}{\varepsilon^2}\rceil \ge 100$ such that $\lambda = \varepsilon\sqrt{k} < 2\sqrt{k}$, we will get pairwise TV $\ge c\varepsilon/2$ and pairwise KL $\le \frac{16\varepsilon^2 k}{k} \le O(\varepsilon^2)$. Finally, scaling ε before applying Lemma 4.1 gives the result. □

5 Conclusion

This work introduces an efficient algorithm for learning product distributions on the Boolean hypercube when provided with an imperfect advice distribution. The sample complexity of this algorithm is $O(d^{1-\eta}/\varepsilon^2)$ under specific conditions on the advice quality and a "balancedness" assumption on the true distribution. Note that the algorithm's sample complexity becomes sublinear in the dimension d if the advice is sufficiently accurate, and it remains robust even with poor advice. Key to this is a novel tolerant mean tester and techniques for approximating the ℓ_1-distance between the true and advice distributions. Future research could extend this learning-with-advice framework to other complex models like Bayesian networks and Ising models, aiming to understand how structural properties of these models interact with advice quality. It would also be interesting to investigate if advice can improve the sample complexity of learning an unstructured distribution over a discrete domain $[n]$ compared to the classical upper bound of $O(\frac{n}{\varepsilon^2})$ samples.

Acknowledgments and Disclosure of Funding

PGJ's research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. TG's research is supported by a start up grant at Nanyang Technological University.

References

[ABDH+20] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM*, 67(6), oct 2020.

[ABG+22] Priyank Agrawal, Eric Balkanski, Vasilis Gkatzelis, Tingting Ou, and Xizhi Tan. Learning-augmented mechanism design: Leveraging predictions for facility location. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 497–528, 2022.

[ADJ+20] Spyros Angelopoulos, Christoph Dürr, Shendan Jin, Shahin Kamali, and Marc Renault. Online Computation with Untrusted Advice. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[AGKK20] Antonios Antoniadis, Themis Gouleakis, Pieter Kleer, and Pavel Kolev. Secretary and online matching problems with machine learned advice. *Advances in Neural Information Processing Systems*, 33:7933–7944, 2020.

[AJS22] Antonios Antoniadis, Peyman Jabbarzade, and Golnoosh Shahkarami. A Novel Prediction Setup for Online Speed-Scaling. In *18th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[BCGG25] Arnab Bhattacharyya, Davin Choo, Philips George John, and Themis Gouleakis. Learning multivariate gaussians with imperfect advice. *International Conference on Machine Learning (ICML)*, 2025.

[BLMS+22] Giulia Bernardini, Alexander Lindermayr, Alberto Marchetti-Spaccamela, Nicole Megow, Leen Stougie, and Michelle Sweering. A Universal Error Measure for Input Predictions Applied to Online Graph Problems. In *Advances in Neural Information Processing Systems*, 2022.

[BMRS20] Étienne Bamas, Andreas Maggiori, Lars Rohwedder, and Ola Svensson. Learning Augmented Energy Minimization via Speed Scaling. *Advances in Neural Information Processing Systems*, 33:15350–15359, 2020.

[BMS20] Etienne Bamas, Andreas Maggiori, and Ola Svensson. The Primal-Dual method for Learning Augmented Algorithms. *Advances in Neural Information Processing Systems*, 33:20083–20094, 2020.

[Can20] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

[CDKS17] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. In *Conference on Learning Theory*, pages 370–448. PMLR, 2017.

[CGB23] Davin Choo, Themistoklis Gouleakis, and Arnab Bhattacharyya. Active causal structure learning with advice. In *International Conference on Machine Learning*, pages 5838–5867. PMLR, 2023.

[CGLB24] Davin Choo, Themistoklis Gouleakis, Chun Kai Ling, and Arnab Bhattacharyya. Online bipartite matching with imperfect advice. In *International Conference on Machine Learning*. PMLR, 2024.

[CJS25] Davin Choo, Billy Jin, and Yongho Shin. Learning-augmented online bipartite fractional matching. *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[CSVZ22] Justin Chen, Sandeep Silwal, Ali Vakilian, and Fred Zhang. Faster fundamental graph algorithms via learned predictions. In *International Conference on Machine Learning*, pages 3583–3602. PMLR, 2022.

[Dia16] Ilias Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267:10–1201, 2016.

[DIL+21] Michael Dinitz, Sungjin Im, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Faster matchings via learned duals. *Advances in neural information processing systems*, 34:10393–10406, 2021.

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

[DLPLV21] Paul Dütting, Silvio Lattanzi, Renato Paes Leme, and Sergei Vassilvitskii. Secretaries with Advice. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 409–429, 2021.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Conference on Learning Theory*, pages 697–703. PMLR, 2017.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[GKST22] Vasilis Gkatzelis, Kostas Kollias, Alkmini Sgouritsa, and Xizhi Tan. Improved price of anarchy via predictions. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 529–557, 2022.

[GLS23] Themis Gouleakis, Konstantinos Lakis, and Golnoosh Shahkarami. Learning-Augmented Algorithms for Online TSP on the Line. In *37th AAAI Conference on Artificial Intelligence*. AAAI, 2023.

[GP19] Sreenivas Gollapudi and Debmalya Panigrahi. Online Algorithms for Rent-or-Buy with Expert Advice. In *International Conference on Machine Learning*, pages 2319–2327. PMLR, 2019.

[KBC+18] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The Case for Learned Index Structures. In *Proceedings of the 2018 international conference on management of data*, pages 489–504, 2018.

[Kon25] Aryeh Kontorovich. On the tensorization of the variational distance. *Electronic Communications in Probability*, 30:1–10, 2025.

[LLMV20] Silvio Lattanzi, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Online Scheduling via Learned Weights. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1859–1877. SIAM, 2020.

[Mit18] Michael Mitzenmacher. A Model for Learned Bloom Filters, and Optimizing by Sandwiching. *Advances in Neural Information Processing Systems*, 31, 2018.

[PSK18] Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving Online Algorithms via ML Predictions. *Advances in Neural Information Processing Systems*, 31, 2018.

[WLW20] Shufan Wang, Jian Li, and Shiqiang Wang. Online Algorithms for Multi-shop Ski Rental with Machine Learned Advice. *Advances in Neural Information Processing Systems*, 33:8150–8160, 2020.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Formal theorems are stated and proved for every claim made.

   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: [NA]

   Justification: The paper is entirely theoretical. Theoretical limitations and assumptions etc are formally discussed. The claims are only made under the scope of these assumptions.

   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 assumptions are stated and full proofs are given in the paper itself.

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: This paper does not include 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: This paper does not include experiments, and does not provide code or data.

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: This paper does not include experiments.

   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: This paper does not include experiments.

   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: This paper does not include experiments.

   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 paper is entirely theoretical, and does not use human subjects or real-world datasets. The proposed algorithms are fully specified for reproducibility, and theorems/ideas from other papers have been acknowledged.

   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 work is theoretical and abstract; there is no direct societal impact.

    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: This paper does not introduce real world data or models, and poses no risks.

    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: This paper does not use existing assets.

    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: This paper does not introduce any assets other than the pseudocode and algorithms described.

    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: The paper is entirely theoretical and does not use crowdsourcing or human subjects.

    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: The paper is entirely theoretical and has no human subjects or study participants.

    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: LLMs were not used in the development 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.