Simple Binary Hypothesis Testing under Communication Constraints

Ankit Pensia
University of Wisconsin-Madison
ankitp@cs.wisc.edu

Po-Ling Loh
University of Cambridge
pll28@cam.ac.uk

Varun Jog
University of Cambridge
vj270@cam.ac.uk

***Abstract*—We study simple binary hypothesis testing under communication constraints, a.k.a. "decentralized detection". Here, each sample is mapped to a message from a finite set of messages via a channel before being revealed to a statistician. In the absence of communication constraints, it is well known that the sample complexity is characterized by the Hellinger distance between the distributions. We show that the sample complexity of hypothesis testing under communication constraints is at most a logarithmic factor larger than in the unconstrained setting, and demonstrate that distributions exist in which this characterization is tight. We also provide a polynomial-time algorithm which achieves the aforementioned sample complexity. Our proofs rely on a new reverse data processing inequality and a reverse Markov's inequality, which may be of independent interest.**

***Index Terms*—hypothesis testing, communication constraints, reverse data processing inequality, reverse Markov's inequality, f-divergences**

A full version of this paper is accessible here [1].

I. Introduction

Statistical inference has been studied under various constraints such as memory, privacy, and communication [2]–[10], designed to model physical or economical constraints. Our work focuses on communication constraints, where the statistician does not have access to the original sample, but only collects observations passed through a communication-constrained channel. For example, instead of observing the original sample $x \in \mathcal{X}$, the statistician might observe a single bit $f(x) \in \{0,1\}$, for some function $f : \mathcal{X} \to \{0,1\}$. The choice of the channel (here, the function f) crucially affects the quality of statistical inference and is the topic of study in our paper.

A recent line of work has established minimax optimal rates of communication-constrained channels [10]–[12] for a variety of problems, including distribution estimation and identity testing. However, previous work on communication constraints for simple hypothesis testing is fairly limited. Recall the simple hypothesis testing framework: Let $\mathcal{P}$ be a given finite set of distributions over the domain $\mathcal{X}$. Given i.i.d. samples $X_1, \dots, X_n$ from an unknown distribution $P \in \mathcal{P}$, the goal is to correctly identify P with high probability, with n as small as possible. We denote this problem as $\mathcal{B}(\mathcal{P})$ and use $n^*(\mathcal{P})$ to denote its sample complexity; i.e, the necessary and sufficient number of samples to solve $\mathcal{B}(\mathcal{P})$.

When $\mathcal{P} = \{p, q\}$, this is referred to as the binary hypothesis testing problem and has a rich history in statistics [13]–[16]. Given its historical and practical significance, we have a very good understanding of this problem (cf. Section II for details). In particular, it is known that $n^*(\mathcal{P}) = \Theta(1/d_h^2(p,q))$, where $d_h(p,q)$ refers to the Hellinger distance between p and q.

Hypothesis testing under communication constraints was studied in detail in the 1980s and 1990s under the name "decentralized detection" [2]. Briefly, the setup involves n users and a central server. Each user i observes an i.i.d. sample X_i from an unknown distribution $p' \in \mathcal{P}$, generates a D-valued message $Y_i \in \{0, 1, \dots, D-1\}$ using a channel $\mathbf{T}_i$ (chosen by the statistician), and transmits Y_i to the central server. The central server observes $(Y_1, \dots, Y_n)$ and produces an estimate $\widehat{p} \in \mathcal{P}$. The goal is to choose $(\mathbf{T}_1, \dots, \mathbf{T}_n)$ so that the central server can identify p' correctly with high probability, while keeping n as small as possible. We call this problem "simple hypothesis testing under communication constraints" and denote it by $\mathcal{B}(\mathcal{P}, D)$. We denote the corresponding sample complexity by $n^*(\mathcal{P}, D)$.

We will focus on the binary hypothesis testing problem, i.e., $\mathcal{P} = \{p, q\}$. It is known that the central server should perform a likelihood ratio test. Furthermore, an optimal choice of channels can be achieved using (deterministic) threshold tests, i.e., $Y_i = f_i(X_i)$ for some $f_i : \mathcal{X} \to \{0, 1, \dots, D-1\}$, such that f_i is characterized by D intervals that partition $\mathbb{R}_+$ and $f_i(x) = j$ if and only if $p(x)/q(x)$ lies in the j^{th} interval. The optimality of threshold tests crucially relies on the f_i's being possibly non-identical across users.

Despite such progress, fundamental statistical and computational questions have remained unanswered. We begin with the following statistical question:

For $\mathcal{P} = \{p, q\}$, what is the sample complexity of $\mathcal{B}(\mathcal{P}, D)$, and what is $\frac{n^*(\mathcal{P},D)}{n^*(\mathcal{P})}$?

Let $n^* = n^*(\mathcal{P})$ and $n^*_{\text{bin}} = n^*(\mathcal{P}, 2)$ for notational convenience. A folklore result using Scheffe's test implies that $n^*_{\text{bin}}/n^* \lesssim n^*$ (cf. Proposition A.3). Our main result is an exponential improvement in this guarantee, showing that $n^*_{\text{bin}}/n^* \lesssim \log(n^*)$, i.e., communication constraints only lead to at most a logarithmic increase in sample complexity. More specifically, we show the following sample complexity bound:

$$n^*(\mathcal{P}, D) \lesssim n^*(\mathcal{P}) \max\left(1, \frac{\log(n^*(\mathcal{P}))}{D}\right). \qquad (1)$$

AP was partially supported by NSF grants CCF-1841190 and CCF-2011255.

Furthermore, there exist cases where the bound (1) is tight (cf. Theorem IV.2). The bound can further be improved when the support size of p and q is smaller than $\log(n^*(\mathcal{P}))$.

Turning to computational considerations, let p and q be distributions over k elements. The optimality of threshold tests implies that each user can search over $k^{\Omega(D)}$ possible such channels, which is prohibitive for large D. This exponential-time barrier has been highlighted as a major computational bottleneck in decentralized detection [2], leading to the following question:

Is there a $\mathrm{poly}(k, D)$*-time algorithm to compute channels* $(\mathbf{T}_1, \ldots, \mathbf{T}_n)$ *that achieve the sample complexity bound* (1)*?*

We answer this question affirmatively by showing that it suffices to consider threshold tests parameterized by a single quantity (cf. equation (3)). In fact, we show that it suffices to use an identical channel across the users.

We summarize our main contributions as follows:

1) We establish the minimax optimal sample complexity (cf. inequality (1)) of binary simple hypothesis testing under communication constraints (Theorems IV.1 and IV.2).
2) We provide an efficient algorithm, running in $\mathrm{poly}(k, D)$ time, to find a channel that achieves the minimax optimal sample complexity.
3) Along the way, we prove the following two technical results which may be of independent interest: (i) a reverse data processing inequality for general f-divergences and communication-constrained channels (Theorem III.2), and (ii) a reverse Markov inequality for bounded random variables (Lemma III.6).

The remainder of the paper is organized as follows: we define notation, formally state the problem, and recall useful facts in Section II. Section III contains a reverse data processing inequality for f-divergences. Section IV then derives the statistical and computational guarantees for binary hypothesis testing. More technical proofs are deferred to the Appendix.

II. Preliminaries

Notation: Throughout this paper, we will focus on discrete distributions. For $n \in \mathbb{N}$, we use $[n]$ to denote $\{1, \ldots, n\}$ and $[0 : n]$ to denote $\{0, 1, \ldots, n\}$. We use Δ_k to denote the set of distributions over k elements. For a distribution $p \in \Delta_k$ and $i \in [k]$, we use both p_i and $p(i)$ to denote the probability of element i under p. For two distributions p and q, let $d_{\mathrm{TV}}(p, q)$ denote the total variation distance, and let $d_h(p, q) = \sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$ denote the Hellinger distance. Given n distributions $p_1, \ldots, p_n$, we use $\prod_{i=1}^n p_i$ to denote their product distribution. When each $p_i = p$, we use $p^{\otimes n}$ to denote the n-fold product distribution. For a set $A \subseteq \mathcal{X}$, we use $\mathbb{I}_A : \mathcal{X} \to \{0, 1\}$ to denote the indicator function of A. We consider $[a, b)$ to be an empty set when $b \leq a$. We will denote channels by bold capital letters, such as $\mathbf{T}$. For a channel $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ and a distribution p over $\mathcal{X}$, we use $\mathbf{T}p$ to denote the distribution over $\mathcal{Y}$ when $X \sim p$ passes through the channel $\mathbf{T}$. We use $c, c_1, c_2, \ldots$ to denote absolute positive constants, whose values might change from line to line, but with values which can be inferred by careful bookkeeping. We also use $C, C_1, C_2, \ldots$ to denote absolute positive constants that remain the same throughout the proof. We use $\lesssim$ and $\gtrsim$ to hide positive constants. We also use the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$. We use $\mathrm{poly}(\cdot)$ to denote a quantity that is polynomial in its arguments.

A. Definitions and basic facts

Definition II.1 (f-divergence)**.** For a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$, we use $I_f(p, q)$ to denote the f-divergence between p and q, defined as $I_f(p, q) := \sum_i q_i f(p_i/q_i)$.[1]

We now define the binary hypothesis testing problem:

Problem II.2 (Simple binary hypothesis testing)**.** Let p and q be two distributions over $\mathcal{X}$. Given p and q, we say a function (test) $\phi : \cup_{n=1}^{\infty} \mathcal{X}^n \to \{p, q\}$ solves the simple hypothesis testing problem with sample complexity n if it is the smallest n' that satisfies

$$\mathbb{P}_{x \sim p^{\otimes n'}} \{\phi(x) = q\} + \mathbb{P}_{x \sim q^{\otimes n'}} \{\phi(x) = p\} \leq 0.1.$$

We define the sample complexity of hypothesis testing to be the largest n such that for any $n' < n$, no test exists with sample complexity n'. We use $\mathcal{B}(p, q)$ to denote the binary hypothesis testing problem and $\mathrm{n}^*(p, q)$ to denote the sample complexity of $\mathcal{B}(p, q)$.

Fact II.3 (Hellinger distance and $\mathcal{B}(p, q)$ [18])**.** $\mathrm{n}^*(p, q) = \Theta(1/d_h^2(p, q))$.

We now define Scheffe's test.

Definition II.4 (Scheffe's test)**.** For two distributions p and q, consider the set $A = \{x : p(x) \geq q(x)\}$. Let p' and q' denote the distributions of $\mathbb{I}_A(X)$ when X is distributed as p and q, respectively. Given $(x_1, \ldots, x_n) \in \mathcal{X}^n$, Scheffe's test transforms each individual point x_i to $\mathbb{I}_A(x_i)$ and then applies the optimal test between p' and q' to the transformed points.[2]

It is easy to see that $d_{\mathrm{TV}}(p', q') = d_{\mathrm{TV}}(p, q)$, which implies that $d_h(p', q') \geq 0.5 d_h^2(p, q)$ (using Fact A.1), leading to an $O(1/d_h^4(p, q))$ sample complexity of Scheffe's test. This dependence is tight, see, for example, [19]. Formally, see Proposition A.3 in Appendix A.

B. Simple hypothesis testing under communication constraints

Definition II.5 (Channels with communication constraints)**.** Let $\mathcal{X}$ be the domain, $\mathcal{P}$ a family of distributions over $\mathcal{X}$, and $\mathcal{T}$ a family of channels from $\mathcal{X}$ to $\mathcal{Y}$. Let $\{U_i\}_{i=1}^n$ denote a set of n users who choose channels $\{\mathbf{T}_i\}_{i=1}^n \subseteq \mathcal{T}$ according to a rule $\mathcal{R} : [n] \to \mathcal{T}^n$.[3] Each user U_i then observes a

[1] We use the following convention [17]: $f(0) = \lim_{t \to 0^+} f(t)$, $0f(0/0) = 0$, and for $a > 0$, $0f(a/0) = a \lim_{u \to \infty} f(u)/u$.

[2] Note that p' and q' are Bernoulli distributions with probability of observing 1 equal to $p(A)$ and $q(A)$, respectively. The optimal test between p' and q' corresponds to a threshold on $\sum_i \mathbb{I}_A(x_i)$.

[3] We consider deterministic rules for simplicity.

random variable X_i i.i.d. from an (unknown) $p' \in \mathcal{P}$, and generates $Y_i = \mathbf{T}_i(X_i) \in \mathcal{Y}$. The central server U_0 observes $(Y_1, \ldots, Y_n)$ and constructs an estimate $\widehat{p} = \phi(Y_1, \ldots, Y_n)$.

Let $\mathcal{T}_D$ denote the set of all channels from $\mathcal{X}$ to $[0 : D-1]$. We now formally define the problem of simple binary hypothesis testing under communication constraints.

Definition II.6 (Simple hypothesis testing under communication constraints)**.** For $D \geq 2$, we define simple binary hypothesis testing under communication constraints of D-messages, denoted by $\mathcal{B}(p, q, \mathcal{T}_D)$, to be the problem in Definition II.5 when $\mathcal{P} = \{p, q\}$, $\mathcal{Y} = [0 : D-1]$, and $\mathcal{T} = \mathcal{T}_D$.

Definition II.7 (Sample complexity of $\mathcal{B}(p, q, \mathcal{T})$)**.** For a given test-rule pair $(\phi, \mathcal{R})$ with $\phi : \cup_{j=1}^{\infty} \mathcal{Y}^j \to \mathcal{P}$, we say that $(\phi, \mathcal{R})$ solves $\mathcal{B}(p, q, \mathcal{T}_D)$ with sample complexity n if it is the smallest n so that $\mathbb{P}_{(x_1,\ldots,x_n)\sim p^{\otimes n}}(\phi(y_1, \ldots, y_n) = q) + \mathbb{P}_{(x_1,\ldots,x_n)\sim q^{\otimes n}}(\phi(y_1, \ldots, y_n) = p) \leq 0.1$. We use $\mathrm{n}^*(p, q, \mathcal{T})$ to denote the sample complexity of this task, i.e., the smallest n so that there exists a $(\phi, \mathcal{R})$-pair that solves $\mathcal{B}(p, q, \mathcal{T})$. We use $\mathrm{n}^*_{\texttt{identical}}(p, q, \mathcal{T})$ to denote the setting where each channel is identical. We sometimes use $\mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T})$ to denote $\mathrm{n}^*(\mathcal{P}, \mathcal{T})$, to emphasize the setting where the channels need not be identical.

For a fixed rule $\mathcal{R}$, an optimal ϕ corresponds to the likelihood ratio test. Thus, our focus will be on designing the rule $\mathcal{R}$, while choosing the test ϕ implicitly, such that the test-rule pair $(\phi, \mathcal{R})$ has minimal sample complexity.

A subset of channels called *threshold channels* plays a key role in our theory. Consider a set $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ such that $0 < \gamma_1 \leq \cdots \leq \gamma_{D-1} < \infty$. Let $\gamma_0 := 0$ and $\gamma_D := \infty$. Define the function $w_\Gamma : [k] \to [0 : D-1]$ as follows[4]: if $q(x) = 0$, then $w_\Gamma(x) = D - 1$; otherwise,

$$w_\Gamma(x) = j \text{ if and only if } p(x)/q(x) \in [\gamma_j, \gamma_{j+1}). \tag{2}$$

We are now ready to define the threshold test:

Definition II.8 (Threshold test)**.** We say that a channel $\mathbf{T} \in \mathcal{T}_D$ corresponds to a threshold test for two distributions p and q over $[k]$ if there exists $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ such that $0 < \gamma_1 \leq \cdots \leq \gamma_{D-1} < \infty$ and $w_\Gamma(X) \sim \mathbf{T}p'$ when $X \sim p'$ (cf. equation (2)). Any such Γ is called the set of thresholds of the test $\mathbf{T}$. We use $\mathcal{T}_D^{\texttt{thresh}}$ to denote the set of all channels $\mathbf{T} \in \mathcal{T}_D$ that correspond to threshold tests.

Note that a priori, searching for an optimal channel over $\mathcal{T}_D^{\texttt{thresh}}$ seems to require $k^{\Omega(D)}$ time, as it requires searching over all possible values of Γ. By restricting our attention to a special class of thresholds parametrized by a single quantity, we will obtain a $\mathrm{poly}(k, D)$-time algorithm. In particular, we will focus on channels with thresholds in the following set:

$$\mathcal{C} := \{\Gamma = (\gamma_1, \ldots, \gamma_{D-1}) : \forall j \in [D-2],\ \gamma_{j+1}/\gamma_j = 2\}\,. \tag{3}$$

A classical result states that threshold tests (cf. Definition II.8) are optimal tests under communication constraints:

Theorem II.9. *[2, Proposition 2.4]* $\mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T}^{\texttt{thresh}}) = \mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D)$.

Our lower bounds on the sample complexity of hypothesis testing under communication constraints crucially rely on the optimality of threshold tests.

III. Reverse Data Processing Inequality for Quantized Channels

We first prove a reverse data processing inequality for a class of f-divergences for communication-constrained channels. We begin by defining a suitable family of f-divergences:

Definition III.1 (Well-behaved f-divergences)**.** We say $I_f(\cdot, \cdot)$ is a well-behaved f-divergence if it satisfies the following:

I.1 f is a convex non-negative function with $f(1) = 0$.
I.2 $xf(y/x) = yf(x/y)$.[5]
I.3 There exist $\alpha > 0, \kappa > 0, C_1 > 0$, and $C_2 > 0$ such that for all $x \in [0, \kappa]$, we have

$$C_1 x^\alpha \leq f(1 + x) \leq C_2 x^\alpha.$$

Some examples include the total variation distance, squared Hellinger distance, symmetrized χ^2-divergence, symmetrized KL-divergence, and triangular discrimination (see Claim D.1 for more details). If an f-divergence is symmetric, f is differentiable at 1, and $f'(1) = 0$, then f satisfies I.2 [20], [21]. Given an f-divergence that does not satisfy I.2, we can construct a new f-divergence with $\tilde{f}(x) := f(x) + xf(1/x)$, which is also a convex function[6] satisfying $\tilde{f}(1) = 0$ and I.2.

Theorem III.2. *Let I_f be a well-behaved f-divergence. Let p and q be two fixed distributions over $[k]$ such that for all $i \in [k], q_i \geq \nu p_i$ and $p_i \geq \nu q_i$, for some $\nu \in [0, 1]$. Then for any $D \geq 2$, there exists a channel $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ (and thus in $\mathcal{T}_D$) such that*

$$1 \leq \frac{I_f(p, q)}{I_f(\mathbf{T}^*p, \mathbf{T}^*q)} \leq 4\frac{f(\nu)}{f(1/(1+\kappa))} + \frac{52C_2}{C_1}\max\left(1, \frac{R}{D}\right), \tag{4}$$

where $R = \min(k, k')$ and $k' = 1 + \log\left(\frac{4C_2\kappa^\alpha}{I_f(p,q)}\right)$. Furthermore, given f, p, and q, there is a $\mathrm{poly}(k, D)$-time algorithm that finds a $\mathbf{T}^*$ achieving the rate in inequality (4).

We provide a brief proof sketch for the special case of $D = 2$ and Hellinger distance below, and defer the full proof to Appendix B-A. As our main focus will be on the Hellinger distance, we state the following corollary which will be used later (see Appendix D):

[4] When $q(x) = 0$ for some x and $p(x) \neq 0$, we take $p(x)/q(x) = \infty$. Without loss of generality, we can assume that for each $x \in [k]$, at least one of $p(x)$ or $q(x)$ is non-zero.

[5] This implies $I_f(p, q) = I_f(q, p)$.

[6] This can be checked by noting that $\tilde{f}''(x) = f''(x) + \frac{1}{x^3}f''(x)$, which is non-negative, as f is convex.

Corollary III.3 (Preservation of Hellinger distance)**.** *For any $p \in \Delta_k$, $q \in \Delta_k$, and $D \geq 2$, there exists a $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ such that the following holds:*

$$1 \leq \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}^* p, \mathbf{T}^* q)} \leq 1800 \max\left(1, \frac{\min(k',k)}{D}\right), \tag{5}$$

where $k' = \log(4/d_h^2(p,q))$. Given p and q, there is a $\operatorname{poly}(k,D)$*-time algorithm that finds $\mathbf{T}^*$ achieving the rate* (5)*.*

Remark III.4. Corollary III.3 can be interpreted as saying that the effective support size of p and q for the Hellinger distance is at most $k' := \log(4/d_h^2(p,q))$, because the distributions could be mapped to a k'-sized alphabet, with the pairwise Hellinger distance preserved up to constant terms.

The following result states that Corollary III.3 is also tight:

Lemma III.5 (Reverse data processing is tight)**.** *There exist positive constants c_1, c_2, c_3, and c such that for every $\rho \in (0, c_1)$ and $D \geq 2$, there exist (i) $k \in \mathbb{N}$ and (ii) two distributions p and q on $[k]$ such that the following hold: $d_h^2(p,q) \in [c_1\rho, c_2\rho]$, $k = \Theta(\log(1/\rho))$, and*

$$\inf_{\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}} \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}p, \mathbf{T}q)} \geq c\frac{R}{D}, \tag{6}$$

where $R = \max(k,k')$ and $k' = \log(1/\rho)$. Moreover, $k = R = \Theta(\log(1/\rho))$.

The proof of Lemma III.5 is given in Appendix B-B.

Proof of Theorem III.2 (sketch). We will focus on the case of the Hellinger distance and $D = 2$. We first establish the following result:

Lemma III.6 (Reverse Markov inequality)**.** *Let X be a random variable over $[0,1)$, supported on at most k points, with $\mathbb{E}[X] > 0$. Let $k' = 1 + \log(1/\mathbb{E}[X])$. Then*

$$\sup_{\delta \in [0,1)} \delta \mathbb{P}(X \geq \delta) \geq \frac{\mathbb{E}[X]}{13}\frac{1}{R}, \text{ where } R = \min(k,k'). \tag{7}$$

The generalized version of Lemma III.6 for the case $D > 2$, along with its proof, is given in Lemma B.1.

Remark III.7. Note that Lemma III.6 is tight, as shown in Claim B.4, which is crucially used in the proof of Lemma III.5.

Remark III.8 (Comparison with existing results). The guarantees of Lemma III.6 can be exponentially better than the Paley-Zygmund inequality and a standard version of the reverse Markov inequality. See Remark D.2 for details.

We now sketch the proof that there exists a channel $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ achieving $d_h^2(\mathbf{T}p, \mathbf{T}q) \gtrsim d_h^2(p,q)/R$. For simplicity of notation, we assume that for all $i \in [k]$, we have $p_i > 0$ and $q_i > 0$. We first define the sets

$$A_{l,u} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l_i, u_i) \right\},$$
$$A_{l,\infty} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l_i, \infty] \right\}. \tag{8}$$

Then $d_h^2(p,q)$ can be decomposed as follows:

$$d_h^2(p,q) = \sum_{i \in A_{0,1/2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{1/2,1}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{1,2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{2,\infty}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2.$$

We note that at least one of these terms must be at least $d_h^2(p,q)/4$. By symmetry, it suffices to consider the case when either the expression containing $A_{2,\infty}$ is at least $d_h^2(p,q)/4$, or the expression with $A_{1,2}$ is at least $d_h^2(p,q)/4$.[7]

a) Case 1: $\sum_{i \in A_{2,\infty}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \geq d_h^2(p,q)/4$: Let $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ be a threshold test with threshold $\Gamma = \{2\}$, i.e., $\mathbf{T}$ is a deterministic channel that corresponds to the function $i \mapsto \mathbb{I}_{p_i/q_i \geq 2}$. We note that $\mathbf{T}p$ and $\mathbf{T}q$ are binary distributions, characterized by $p' = \sum_{i \in A_{2,\infty}} p_i$ and $q' = \sum_{i \in A_{2,\infty}} q_i$, respectively. Then

$$d_h^2(p,q) \leq 4 \sum_{i \in A_{2,\infty}} (\sqrt{p_i} - \sqrt{q_i})^2 \leq 4 \sum_{i \in A_{2,\infty}} p_i \leq 4p'.$$

Using the fact that $p' \geq 2q'$, we also have

$$d_h^2(\mathbf{T}p, \mathbf{T}q) \geq (\sqrt{p'} - \sqrt{p'/2})^2 \geq 0.01p'.$$

Combining the two displayed equations, we obtain $d_h^2(\mathbf{T}p, \mathbf{T}q) \geq d_h^2(p,q)/400$. This completes the proof.

b) Case 2: $\sum_{i \in A_{1,2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \geq d_h^2(p,q)/4$: For $i \in A_{1,2}$, let $\delta_i := (p_i - q_i)/q_i$, which lies in $[0,1)$. Consider the random variable X over $[0,1)$ such that for $i \in A_{1,2}$, we define $\mathbb{P}(X = \delta_i) = q_i$ and $\mathbb{P}(X = 0) = 1 - \sum_{i \in A_{1,2}} q_i$. Let $\delta \in [0,1)$ be arbitrary (to be decided later). Consider the channel $\mathbf{T}$ that corresponds to the threshold $1 + \delta$. Suppose for now that the following inequalities hold:

$$d_h^2(p,q) \lesssim \mathbb{E}X^2 \quad \text{and} \quad d_h^2(\mathbf{T}p, \mathbf{T}q) \gtrsim \delta^2 \mathbb{P}(X \geq \delta), \tag{9}$$

which we will establish shortly using a Taylor approximation. Letting $Y = X^2$ and $\delta' = \delta^2$, we obtain the following inequality using the bounds (9):

$$\frac{d_h^2(\mathbf{T}p, \mathbf{T}q)}{d_h^2(p,q)} \gtrsim \frac{\delta^2 \mathbb{P}(X \geq \delta)}{\mathbb{E}[X^2]} = \frac{\delta' \mathbb{P}(Y \geq \delta')}{\mathbb{E}[Y]}, \tag{10}$$

which allows us to apply a reverse Markov inequality (Lemma III.6) to the random variable Y. Fix

$$R = \log(1/E[Y]) = \log(1/\mathbb{E}[X^2]) = \log(O(1/d_h^2(p,q)).$$

By Lemma III.6, we note that there exists δ' such that $\delta' \mathbb{P}(Y \geq \delta') \gtrsim \mathbb{E}[Y]/R$, which yields the desired lower bound $(d_h^2(\mathbf{T}p, \mathbf{T}q))/(d_h^2(p,q)) \gtrsim \frac{1}{R}$ using inequality (10).

We now provide a brief proof sketch of the bounds (9). We derive the first bound using the following arguments:

$$d_h^2(p,q) \leq 4 \sum_{i \in A_{1,2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 = 4 \sum_{i \in A_{1,2}} q_i(\sqrt{1 + \delta_i} - 1)^2$$
$$\leq 4 \sum_{i \in A_{1,2}} q_i \delta_i^2 = 4\mathbb{E}[X^2],$$

[7] An astute reader might note that the situation is not fully symmetric, as the set $A_{1,2}$ is right-open, while $A_{1/2,1}$ is left-closed. However, as shown in the full proof later, the desired conclusion still holds.

where the first inequality uses the assumption and the second inequality uses the fact that $\sqrt{1+x} \leq 1+x$ for $x \geq 0$.

We now turn our attention to the second bound (9). Recall that $\mathbf{T}$ is a channel corresponding to the threshold $1+\delta$. Let $p' = \sum_{i:\delta_i \in [\delta,1)} p_i$ and $q' = \sum_{i:\delta_i \in [\delta,1)} q_i$. Note that $q' = \mathbb{P}(X \geq \delta)$ and $p' - q' = \sum_{i:\delta_i \in [\delta,1)} \delta_i q_i = \mathbb{E}[X \mathbb{I}_{X \geq \delta}]$. Thus, $(p'-q')/q' = \mathbb{E}[X | X \geq \delta]$.

It can be shown that $d_h^2(\mathbf{T}p, \mathbf{T}q) \geq (\sqrt{p'} - \sqrt{q'})^2$ (cf. Appendix B), which leads to the following inequalities:

$$d_h^2(\mathbf{T}p, \mathbf{T}q) \geq (\sqrt{p'} - \sqrt{q'})^2 = q' \left(\sqrt{1 + \frac{p'-q'}{q'}} - 1 \right)^2$$
$$\gtrsim q' \left(\frac{p'-q'}{q'} \right)^2 \geq \delta^2 \mathbb{P}(X \geq \delta),$$

since $(\sqrt{1+x} - 1) \gtrsim x$ for $x \in [0,1)$. □

IV. Simple Binary hypothesis testing

We will now apply the results of previous sections to simple hypothesis testing under communication constraints. Let $\mathbf{T}$ be a fixed channel and suppose all users use the same channel $\mathbf{T}$. In this setting, Fact II.3 implies that the sample complexity would be $\Theta(1/(d_h^2(\mathbf{T}p, \mathbf{T}q)))$. Without any communication constraints, the sample complexity of the best test is known to be $\Theta(1/(d_h^2(p,q)))$. Thus, the additional (multiplicative) penalty of using the channel $\mathbf{T}$ is $d_h^2(p,q)/d_h^2(\mathbf{T}p, \mathbf{T}q)$, which is at least 1 by the data processing inequality. As we are allowed to choose any channel $\mathbf{T} \in \mathcal{T}_D$, we would like to choose the channel that minimizes this quantity, which was precisely studied in Section III.

A. Upper bound for simple hypothesis testing

Theorem IV.1. *There exists a constant $c > 0$ such that the following holds: Let p and q be any two distributions over Δ_k and define $\mathrm{n}^* := \mathrm{n}^*(\{p,q\})$.[8] Then for any $D \geq 2$,*

$$\mathrm{n}^*_{\texttt{identical}}(p, q, \mathcal{T}_D) \leq c \cdot \mathrm{n}^* \cdot \max\left(1, \frac{\min(k, \log \mathrm{n}^*)}{D}\right). \tag{11}$$

Furthermore, there is an algorithm which, given p, q, and D, finds a channel $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ in $\mathrm{poly}(k,D)$ time that achieves the rate in inequality (11).

Proof. As noted earlier, for a fixed $\mathbf{T}$, the sample complexity is $\Theta\left(\frac{1}{d_h^2(\mathbf{T}p, \mathbf{T}q)}\right)$. Our proof strategy will be to upper-bound the quantity $\inf_{\mathbf{T} \in \mathcal{T}_D} g(\mathbf{T})$, where $g(\mathbf{T}) := \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}p, \mathbf{T}q)}$. By Corollary III.3, there exists a $\mathbf{T}^*$ such that $g(\mathbf{T}^*) \lesssim \max(1, \min(k, n^*)/D)$, since $n^* = \Theta(\log(1/d_h^2(p,q)))$. Thus, the proof of Theorem IV.1 follows from Corollary III.3 by choosing the optimal $\mathbf{T}^*$ achieving the bound in Corollary III.3. As mentioned in Corollary III.3, the channel $\mathbf{T}^*$ can be found efficiently. □

B. Lower bounds

We now prove our lower bounds, showing that there exist distributions p and q such that the sample complexity must increase by a factor of $\log(n_*(p,q))$. As discussed in Section II-B, an optimal test that minimizes the probability of error under the communication constraints is one that thresholds based on $p(i)/q(i)$ [2]. However, this notion of optimality is conditioned on the fact that the channels are potentially non-identical; examples exist where this is necessary even for $D = 2, M = 2$, and $n = 2$ [22].

We will show that, up to constants in the sample complexity, it suffices to consider identical channels for simple hypothesis testing. In fact, we prove a much more general result, Lemma C.1 in Appendix C, that does not rely on restricting the function class to threshold tests. We have the following lower bound on $\mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T})$:

Theorem IV.2. *There exist positive constants c_1, c_2, and c_3 such that for every $n_0 \in \mathbb{N}$ and $D \geq 2$, there exist (i) $k \in \mathbb{N}$ and (ii) two distributions p and q on $[k]$ such that the following hold:*

1) $c_1 n_0 \leq \mathrm{n}^*(p,q) \leq c_2 n_0$, and
2) $\mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D) \geq n_0 \frac{\log(n_0)}{D}$.

Moreover, $k = \Theta(\log n_0)$, i.e., the domain of the two distributions is not too large.

Proof. (Sketch) Using Lemma C.1 and Theorem II.9, it suffices to consider the setting with identical threshold channels. With identical channels, say $\mathbf{T}$, the problem reduces to that of $\mathcal{B}(\mathbf{T}p, \mathbf{T}q)$, and thus to $d_h(\mathbf{T}p, \mathbf{T}q)$ using Fact II.3. Tightness of Lemma III.5 then gives the desired result. □

V. Discussion

In this paper, we studied the problem of simple binary hypothesis testing under communication constraints. We showed that communication constraints may lead to an at most logarithmic increase in the sample complexity of the test. At a technical level, our results rely on a reverse data processing inequality for communication-constrained channels.

Interesting questions for further study include characterizing the sample complexity for a robust version of simple binary hypothesis testing, i.e., the distribution p' in Definition II.5 may only be constrained to satisfy $\min\{d_{\mathrm{TV}}(p,p'), d_{\mathrm{TV}}(q,p')\} \leq \epsilon$, for some $\epsilon < d_{\mathrm{TV}}(p,q)/2$. Although Scheffe's test is both robust and communication-efficient, it is not clear whether the test is optimal. In a different research direction, it is not clear if adaptivity in the choice of channels, i.e., $\mathbf{T}_i$ is selected after observing $(Y_1, \ldots, Y_{i-1})$ in Definition II.5, would have any effect on the sample complexity of the problem. Finally, studying the simple hypothesis testing problem between $M > 2$ distributions would be a worthwhile endeavor.

References

[1] A. Pensia, V. Jog, and P. Loh, "Simple binary hypothesis testing under communication constraints," 2022. [Online]. Available: https://ankitp.net/research/hyp-testing.pdf

[8] Recall that Fact II.3 implies $n^* = \Theta(1/d_h^2(p,q))$.

[2] J. N. Tsitsiklis, "Decentralized detection," in *In Advances in Statistical Signal Processing*. JAI Press, 1993, pp. 297–344.

[3] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[4] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," *arXiv:1506.07216 [cs, math, stat]*, May 2016.

[5] J. Steinhardt, G. Valiant, and S. Wager, "Memory, communication, and statistical queries," in *Conference on Learning Theory, COLT 2016*, vol. 49. JMLR.org, 2016, pp. 1490–1516.

[6] V. Feldman, "A general characterization of the statistical query complexity," in *Conference on Learning Theory, COLT 2017*. PMLR, 2017.

[7] Y. Dagan and O. Shamir, "Detecting correlations with little memory and communication," in *Conference on Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, ser. Proceedings of Machine Learning Research, vol. 75. PMLR, 2018, pp. 1145–1198.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–201, Jan. 2018.

[9] Y. Han, A. Özgür, and T. Weissman, "Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints," *arXiv:1802.08417 [cs, math, stat]*, Jul. 2021.

[10] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7835–7855, Dec. 2020.

[11] ——, "Inference under information constraints II: Communication constraints and shared randomness," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7856–7877, 2020.

[12] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," *IEEE Transactions on Information Theory*, vol. 67, no. 12, pp. 8248–8263, Dec. 2021.

[13] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[14] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[15] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *The Annals of Statistics*, vol. 1, no. 2, Mar. 1973.

[16] L. L. Cam, *Asymptotic Methods in Statistical Decision Theory*, ser. Springer Series in Statistics. New York, NY: Springer New York, 1986.

[17] I. Sason, "On f-divergences: Integral representations, local behavior, and inequalities," *Entropy*, vol. 20, no. 5, p. 383, May 2018.

[18] C. L. Canonne, G. Kamath, A. McMillan, A. Smith, and J. Ullman, "The structure of optimal private tests for simple hypotheses," *arXiv:1811.11148 [cs, math, stat]*, Nov. 2018.

[19] A. T. Suresh, "Robust hypothesis testing and distribution estimation in hellinger distance," in *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, 2021.

[20] G. L. Gilardoni, "On the minimum f-divergence for given total variation," *Comptes Rendus Mathematique*, vol. 343, no. 11-12, pp. 763–766, Dec. 2006.

[21] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences," in *2015 IEEE Information Theory Workshop (ITW)*. Jerusalem, Israel: IEEE, Apr. 2015, pp. 1–5.

[22] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals, and Systems*, vol. 1, no. 2, pp. 167–182, Jun. 1988.

[23] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[24] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 275–283, May 1973.

[25] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge ; New York, NY: Cambridge University Press, 2019, no. 48.

[26] V. De la Peña and E. Giné, *Decoupling: From Dependence to Independence*, ser. Probability and Its Applications. New York: Springer, 1999.

[27] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014.

Additional Notation: Let $\beta_h(p,q)$ denote the Hellinger affinity, i.e., $\beta_h(p,q) := 1 - 0.5 d_h^2(p,q)$.

Appendix A
Additional Details from Section II

Fact A.1. *(see, e.g., [23], [24]) For any distributions $p, p_1, \dots, p_n$ and $q, q_1, \dots, q_n$ in Δ_k:*

1) $d_{TV}^2(p,q) \le d_h^2(p,q) \le 2 d_{TV}(p,q)$.
2) *(Hellinger tensorization)* $\beta_h\left(\prod_{i=1}^n p_i, \prod_{i=1}^n q_i\right) = \prod_{i=1}^n \beta_h(p_i, q_i)$.
3) *(Data processing) For any channel $\mathbf{T}$ and any distributions p and q, we have $I_f(\mathbf{T}p, \mathbf{T}q) \le I_f(p,q)$.*

Fact A.2. *(see, e.g., [18], [25])*

1) *(Total variation and hypothesis testing) For any random variable Z over $\mathcal{Z}$ and a test $\phi : \mathcal{Z} \to \{P, Q\}$, define the probability of error to be $\frac{1}{2}\mathbb{P}_P(\phi(Z) = Q) + \frac{1}{2}\mathbb{P}_Q(\phi(Z) = P)$. The minimum probability of error over all tests is $(1 - d_{TV}(P,Q))/2$ and is achieved by the following test: let $A^* \subseteq \mathcal{Z}$ be any set that maximizes $P(A) - Q(A)$ over $A \subseteq \mathcal{Z}$, and define $\phi(z) := P$ iff $z \in A^*$ and Q otherwise.*
2) *(Hellinger distance and $\mathcal{B}(p,q)$) The sample complexity for the simple binary hypothesis test between p and q is $\Theta(1/d_h^2(p,q))$, i.e., $\mathrm{n}^*(p,q) = \Theta(1/d_h^2(P,Q))$.*

Proposition A.3 (Folklore)**.** *The sample complexity of Scheffe's test is at most $O\left(1/d_h^4(p,q)\right)$. Furthermore, this is tight in the following sense: for any $\rho \in (0,1)$, there exist p and q such that $\mathrm{n}^*(p,q) = O(1/\rho^2)$, whereas the sample complexity of Scheffe's test is $\Omega\left(1/\rho^4\right)$.*

Proof. (Proof of Proposition A.3) Let p and q be the two given distributions and let $\rho^2 = d_h^2(p,q)$. Let $\mathbf{T}$ be the channel corresponding to the Scheffe's test. Since Scheffe's test preserves the total variation distance, we have that $d_{\mathrm{TV}}(p,q) = d_{\mathrm{TV}}(\mathbf{T}p, \mathbf{T}q)$. By using Fact A.1, we have that $d_h(\mathbf{T}p, \mathbf{T}q) \ge d_{\mathrm{TV}}(\mathbf{T}p, \mathbf{T}q) = d_{\mathrm{TV}}(p,q) \ge 0.5 d_h^2(p,q) \ge 0.5\rho^2$. By Fact A.2, we have that the sample complexity is at most $O(1/d_h^2(\mathbf{T}p, \mathbf{T}q)) = O(1/\rho^4)$.

Without loss of generality, we consider the setting when $\rho \le 0.01$. Consider the distributions $p = (\rho, 1/2 - 2\rho, 1/2 + \rho)$ and $q = (0, 1/2, 1/2)$. Let $\mathbf{T}$ be the channel corresponding to the Scheffe's test. Then we have $\mathbf{T}p = (1/2 + 2\rho, 1/2 - 2\rho)$ and $\mathbf{T}q = (1/2, 1/2)$. An elementary calculation shows that $d_h^2(p,q) = \Theta(\rho)$ and $d_h^2(\mathbf{T}p, \mathbf{T}q) = \Theta(\rho^2)$. Applying Fact A.2, we get the desired conclusion. □

Appendix B
Reverse Data Processing

Fix the distributions p and q over $[k]$. For $0 \le l < u < \infty$, we first define the following sets[9]:

$$A_{l,u} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, u) \right\} \text{ and}$$
$$A_{l,\infty} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, \infty] \right\}. \tag{12}$$

We will use the notation defined in Definition II.8.

A. Reverse data processing: Proof of Theorem III.2

Theorem III.2. *Let I_f be a well-behaved f-divergence. Let p and q be two fixed distributions over $[k]$ such that for all $i \in [k], q_i \ge \nu p_i$ and $p_i \ge \nu q_i$, for some $\nu \in [0,1]$. Then for any $D \ge 2$, there exists a channel $\mathbf{T}^* \in \mathcal{T}_D^{\mathtt{thresh}}$ (and thus in $\mathcal{T}_D$) such that*

$$1 \le \frac{I_f(p,q)}{I_f(\mathbf{T}^*p, \mathbf{T}^*q)} \le 4\frac{f(\nu)}{f(1/(1+\kappa))} + \frac{52C_2}{C_1}\max\left(1, \frac{R}{D}\right), \tag{4}$$

where $R = \min(k, k')$ and $k' = 1 + \log\left(\frac{4C_2\kappa^\alpha}{I_f(p,q)}\right)$. Furthermore, given f, p, and q, there is a $\mathrm{poly}(k, D)$-time algorithm that finds a $\mathbf{T}^*$ achieving the rate in inequality (4).

Proof. (Proof of Theorem III.2)

Let $\kappa > 0$ be as defined in Definition III.1. By definition of the f-divergence, we have the following:

$$I_f(p,q) = \sum_{i \in A_{1+\kappa,\infty}} q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i \in A_{1,1+\kappa}} q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i \in A_{1/(1+\kappa),1}} q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right), \tag{13}$$

where the sets $A_{l,u}$ are defined as in equation (8). Note that the sets $A_{1+\kappa,\infty}$ and $A_{0,1/(1+\kappa)}$ contain the elements that have a large ratio of probabilities under the two distributions. We will now consider a case-by-case basis.

a) Case 1: Main contribution by large ratio alphabets:: We first consider the case when $\sum_{i \in A_{1+\kappa,\infty} \cup A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \ge \frac{I_f(p,q)}{2}$. As we will show later, this is the simple case ($D = 2$ already achieves the claim). By symmetry of the I_f-divergence for the well-behaved f-divergence (I.2 in Definition III.1), it suffices to consider the case when $\sum_{i \in A_{1+\kappa,\infty}} q_i f\left(\frac{p_i}{q_i}\right) \ge \frac{I_f(p,q)}{4}$. [10]

[9] When $q(x) = 0$ for some x and $p(x) \ne 0$, then we think of $p(x)/q(x) = \infty$. Without loss of generality, we can assume that for each $x \in [k]$, at least one of $p(x)$ and $q(x)$ is non-zero.

[10] There is a slight asymmetry because of the corner cases but it suffices: if $\sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \ge I_f(p,q)/4$, then labeling $\tilde{p} := q$ and $\tilde{q} := p$, and defining $\tilde{A}$ as in Equation (12) with $\tilde{p}$ and $\tilde{q}$, we have that $\sum_{i \in \tilde{A}_{1+\kappa,\infty}} \tilde{q}_i f\left(\frac{\tilde{p}_i}{\tilde{q}_i}\right) \ge \sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \ge I_f(p,q)/4$ because $A_{0,1/(1+\kappa)} \subset \tilde{A}_{1+\kappa,\infty}$ since the interval in $A_{0,l}$ is left-open.

We will show that there exists a $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ such that $I_f(\mathbf{T}p, \mathbf{T}q) \geq 0.25(f(1/(1+\kappa))/f(0))I_f(p,q)$. Let $\mathbf{T}$ be the channel corresponding to the threshold $1+\kappa$, i.e., $\mathbf{T}$ corresponds to the function $i \mapsto \mathbb{I}_{A_{1+\kappa,\infty}}(i)$. Note that $\mathbf{T}p$ and $\mathbf{T}q$ are distribution on $\{0,1\}$ with $(\mathbf{T}p)_1 = \sum_{i \in A_{1+\kappa,\infty}} p_i$ and $(\mathbf{T}q)_1 = \sum_{i \in A_{1+\kappa,\infty}} q_i$, which we denote by p' and q' respectively. We have that $p' \geq (1+\kappa)q'$. Using convexity and non-negativity of f, and the fact that $f(1) = 0$ (see I.1), we have that $f(x) \leq f(y)$ for $0 \leq y \leq x \leq 1$. Using the non-negativity of f (I.1), symmetry of f (I.2), and monotonically decreasing property of f on $[0,1]$, we obtain the following:

$$\begin{aligned} I_f(\mathbf{T}p, \mathbf{T}q) &= p' f\left(\frac{q'}{p'}\right) + (1-p') f\left(\frac{1-q'}{1-p'}\right) \\ &\geq p' f\left(\frac{q'}{p'}\right) \\ &\geq p' f\left(\frac{1}{1+\kappa}\right). \end{aligned} \tag{14}$$

Moreover, by assumption that $\sum_{i \in A_{1+\kappa,\infty}} p_i f\left(\frac{p_i}{q_i}\right)$ is at least $0.25 I_f(p,q)$, we have that

$$\begin{aligned} 0.25 I_f(p,q) &\leq \sum_{i \in A_{1+\kappa,\infty}} p_i f\left(\frac{q_i}{p_i}\right) \\ &\leq \sum_{i \in A_{1+\kappa,\infty}} p_i f(\nu) = p' f(\nu), \end{aligned} \tag{15}$$

where we use $q_i/p_i \in [\nu, 1]$ and f is decreasing on $[\nu, 1]$. Combining (14) and (15), we get the following:

$$I_f(\mathbf{T}p, \mathbf{T}q) \geq \frac{f(1/(1+\kappa))}{4f(\nu)} I_f(p,q), \tag{16}$$

which implies $I_f(p,q)/I_f(\mathbf{T}p, \mathbf{T}q) \leq 4f(\nu)/f(1/(1+\kappa))$, proving the desired result.

We now comment on the computational complexity of finding a $\mathbf{T}^*$ that achieves the rate in (16). Since the channel $\mathbf{T}^*$ only depends on κ, the algorithm only needs to check whether the threshold should be $1+\kappa$ or $1/(1+\kappa)$, which requires at most $\mathrm{poly}(k)$ operations.

b) Case 2: Main contribution by small ratio alphabets:: We now consider the case when $\sum_{i \in A_{1,1+\kappa} \cup A_{1/(1+\kappa),1}} q_i f(p_i/q_i) \geq I_f(p,q)/2$. By symmetry (I.2), it suffices to consider the case when $\sum_{i \in A_{1,1+\kappa}} q_i f(p_i/q_i) \geq I_f(p,q)/4$[11]. This requires us to handle the elements where p_i and q_i are close, and the following arguments form the main technical core of this section.

We first state a reverse Markov inequality below, proved in Section B-C, whose objective will become clear later in the proof:

Lemma B.1. *Let Y be a random variable over $[0,\beta)$ with expectation $\mathbb{E}[Y] > 0$. Let $k' = 1 + \log(\beta/\mathbb{E}[Y])$. Then we have the following:*

$$\begin{aligned} \sup_{\beta = \nu_D \geq \ldots \geq \nu_1 \geq 0} \sum_{j=1}^{D-1} \nu_j \mathbb{P}\left(Y \in [\nu_j, \nu_{j+1})\right) \\ \geq \frac{1}{13}\mathbb{E}[Y] \min\left(1, \frac{D}{R}\right), \end{aligned} \tag{17}$$

where $R = k' := 1 + \log(\beta/\mathbb{E}[Y])$. Furthermore, the bound in (17) can be achieved by ν_j's such that $\nu_j = \min(\beta, x2^j)$ for an $x \in [0,\beta]$.

For the special setting when Y is supported on k points: we may set $R = \min(k, k')$, and there is a $\mathrm{poly}(k, D)$ algorithm to find ν_j's that achieve the right hand side of inequality (17).

For any $i \in A_{1,1+\kappa}$, we have that both q_i and p_i are positive. Let $\delta_i = (p_i/q_i) - 1$, which lies in $[0,\kappa)$ by definition. We thus have that $p_i = q_i(1+\delta_i)$. Let X be a random variable over $[0,\kappa)$ such that for $i \in A_{1,1+\kappa}$, define $\mathbb{P}(X = \delta_i) = q_i$, and $\mathbb{P}(X = 0) = 1 - \sum_{i \in A_{1,1+\kappa}} q_i$.

We now apply Lemma B.1 to the random variable $Y = X^\alpha$. Let $\beta = \kappa^\alpha$ and $R_2 = \min(k, 1 + \log(\kappa^\alpha/\mathbb{E}[X^\alpha]))$. Let $0 \leq \nu_1' \leq \ldots, \nu_D' = \beta$ be thresholds obtaining the bound in (17). Let $\nu_j = (\nu_j')^{1/\alpha}$ for all $j \in [D]$. We thus have that

$$\sum_{j=1}^{D-1} \nu_j^\alpha \mathbb{P}(X \in [\nu_j, \nu_{j+1})) \geq \frac{1}{13}\mathbb{E}[X^\alpha] \min\left(1, \frac{D}{R_2}\right). \tag{18}$$

We now define the thresholds $\Gamma = (\gamma_1, \ldots, \gamma_{D-1})$ such that $\gamma_j = 1 + \nu_j$ for $i \in [D-1]$. We set $\gamma_0 = 0$ and $\gamma_\infty = 0$. Note that for $j \in [D-1]$, we have $A_{\gamma_j, \gamma_{j+1}} = \{i : p_i/q_i \in [\gamma_i, \gamma_{i+1})\}$. Since $1 \leq \gamma_1 \leq \gamma_{D-1} \leq 1 + \kappa$, we have that for $j \in [D-2]$, $A_{\gamma_j,\gamma_{j+1}} = \{i \in A_{1,1+\kappa} : \delta_i \in [\nu_j, \nu_{j+1})\}$. Note that for any $j \in [D-2]$ and any function $g(\cdot)$, we have the following:

$$\begin{aligned} \sum_{i \in A_{\gamma_j,\gamma_{j+1}}} g(\delta_i) q_i &= \sum_{i \in A_{\gamma_j,\gamma_{j+1}}} g(\delta_i) \mathbb{P}(X = \delta_i) \\ &= \sum_{x \in [\nu_j, \nu_{j+1})} g(x) \mathbb{P}(X = x) \end{aligned} \tag{19}$$

Using the growth of $f(1+x)$ in $[0,\kappa]$ (I.3) and the fact that $0 \leq \delta_i \leq \kappa$, we have the following:

$$\begin{aligned} \sum_{i \in A_{1,1+\kappa}} q_i f\left(\frac{p_i}{q_i}\right) &= \sum_{i \in A_{1,1+\kappa}} q_i f(1+\delta_i) \\ &\leq \sum_{i \in A_{1,1+\kappa}} C_2 q_i \delta_i^\alpha = C_2 \mathbb{E}[X^\alpha], \end{aligned} \tag{20}$$

where the last equality uses the same arguments as (19). Finally we note that (20) and the assumption $I_f(p,q) \leq 4\sum_{i \in A_{1,1+\kappa}} q_i f(p_i/q_i)$ implies the following:

$$\begin{aligned} I_f(p,q) \leq 4C_2 \mathbb{E}[X^\alpha] \qquad \text{and} \\ R_2 \leq \min(k, 1 + \log(4C_2\kappa^\alpha / I_f(p,q))) = R. \end{aligned} \tag{21}$$

[11] There is a slight asymmetry here as well, but similar to the previous footnote, it suffices to consider this case

We use p' and q' to be denote the probability measures $\mathbf{T}p$ and $\mathbf{T}q$ respectively when $\mathbf{T}$ corresponds to thresholds Γ. Thus, we have that for $j \in [0:D-1]$, $p'(j) = \sum_{i \in A_{\gamma_j,\gamma_{j+1}}} p_i$; $q'(j)$ is defined similarly. We now define p'' to be the following positive measure: for $j \in [0:D-1]$, $p''_j = \sum_{i \in A_{\gamma_i,\gamma_{i+1}}} p_i$ and $p''_{D-1} = \sum_{i \in A_{\gamma_{D-1},1+\kappa}} p_i$; q'' is defined similarly. Recall that $\gamma_{D-1} = 1+\nu_{D-1} \leq 1+\kappa$. Note that p'' and q'' might not be probability measures as their sum might be smaller than 1, but they are equal to p' and q' respectively on all elements except the last. Moreover, we may define the "f-divergence" between p'' and q'' by mechanically applying the standard expression f-divergence but replacing the probability measures by p'' and q'' instead. The f-divergence between p'' and q'' so obtained is smaller than the f-divergence between p' and q' as follows:

$$q'_{D-1} f\left(\frac{p'_{D-1}}{q'_{D-1}}\right) \geq q''_{D-1} f\left(\frac{p''_{D-1}}{q''_{D-1}}\right),$$

which follows by noting that $q''_{D-1} \leq q'_{D-1}$, $p'_{D-1}/q'_{D-1} \geq p''_{D-1}/q''_{D-1} \geq 1$, and $f(x) \geq f(y) \geq 0$ for any $x \geq y \geq 1$. We thus get the following relation:

$$I_f(p',q') \geq \sum_{j=0}^{D-1} q''_j f\left(\frac{p''_j}{q''_j}\right). \tag{22}$$

Fix a $j \in [D-1]$. Using the fact that $0 \leq \frac{p''_j}{q''_j} - 1 \leq \kappa$ and that $f(1+x) \geq C_1 x^\alpha$ for $x \in [0,\kappa]$ (I.3), we have that for any j such that $q''_j > 0$,

$$\begin{aligned} q''_j f\left(\frac{p''_j}{q''_j}\right) &= q''_j f\left(1 + \frac{p''_j - q''_j}{q''_j}\right) \\ &\geq C_1 q''_j \left(\frac{p''_j - q''_j}{q''_j}\right)^\alpha \\ &= C_1 q''_j \left(\frac{\sum_{i \in A_{\gamma_j,\gamma_{j+1}}} q_i \delta_i}{\sum_{i \in A_{\gamma_j,\gamma_{j+1}}} q_i}\right)^\alpha \\ &\geq C_1 q''_j \nu_j^\alpha \qquad \text{(using } \delta_i \geq \nu_j\text{)} \\ &\geq C_1 \nu_j^\alpha \mathbb{P}(X \in [\nu_j, \nu_{j+1})) \end{aligned} \tag{23}$$

We note that this inequality is also true if $q''_j = 0$ because $q''_j = \mathbb{P}(X \in [\nu_j, \nu_{j+1}))$, and if the former is zero, then the expression in (23) is also zero.

Overall, we get the following series of inequalities:

$$\begin{aligned} I_f(p',q') &\geq \sum_{j=1}^{D-1} q''_j f\left(\frac{p''_j}{q''_j}\right) && \text{(using (22) and } f \geq 0\text{)} \\ &\geq C_1 \sum_{j=1}^{D-1} \nu_j^\alpha \mathbb{P}(X \in [\nu_j, \nu_{j+1})) && \text{(using (23))} \\ &\geq \frac{C_1}{13} \mathbb{E}[X^\alpha] \min\left(1, \frac{D}{R_2}\right) && \text{(using (18))} \\ &\geq \frac{C_1}{52C_2} I_f(p,q) \min\left(1, \frac{D}{R}\right) && \text{(using (21))}. \end{aligned}$$

This shows that there exists a $\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}$ such that

$$I_f(p,q)/I_f(\mathbf{T}p, \mathbf{T}q) \leq \frac{52C_2}{C_1} \max\left(1, \frac{R}{D}\right). \tag{24}$$

We now comment on the computational complexity of finding a $\mathbf{T}^*$ that achieves the rate in (24). Finding the thresholds Γ is equivalent to finding $(\nu'_1, \ldots, \nu'_{D-1})$. As noted in Lemma B.1 and its proof, the guarantee of (17) can be achieved by choosing ν'_j in one of the following ways:

- Setting $\nu'_j = \min(\kappa^\alpha, x2^j)$ for all j and optimizing over x that matter. As the random variable Y has support of at most k, this algorithm runs in $\operatorname{poly}(k, D)$-time.
- Choosing the top $D-1$ elements that maximize $\delta_i q_i$ and defining ν'_j appropriately.

□

B. Tightness of reverse data processing inequality: Proof of Lemma III.5

Lemma III.5 (Reverse data processing is tight)**.** *There exist positive constants c_1, c_2, c_3, and c such that for every $\rho \in (0, c_1)$ and $D \geq 2$, there exist (i) $k \in \mathbb{N}$ and (ii) two distributions p and q on $[k]$ such that the following hold: $d_h^2(p,q) \in [c_1\rho, c_2\rho]$, $k = \Theta(\log(1/\rho))$, and*

$$\inf_{\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}} \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}p, \mathbf{T}q)} \geq c\frac{R}{D}, \tag{6}$$

where $R = \max(k, k')$ and $k' = \log(1/\rho)$. Moreover, $k = R = \Theta(\log(1/\rho))$.

Proof. We will design p and q such that $p_i/q_i \in [0.5, 1.5]$. Fix any set of thresholds $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$, which without loss of generality lie in $[0.5, 1.5]$. Let $\mathbf{T}$ be the corresponding channel. Let p' and q' be the distributions after using the channel $\mathbf{T}$.

Note that k will depend on ρ, which will be decided later. For now, let k be even, equal to $2m$. Let $\tilde{q}$ be an arbitrary distribution on $[m]$ to be decided later. Let $\tilde{\delta} \in [0, 0.5]^m$ to be decided later. Using this $\tilde{q}$, we define a distribution on q as follows:

$$q_i = \begin{cases} 0.5\tilde{q}_i, & \text{if } i \in [m] \\ 0.5\tilde{q}_{i-m} & \text{if } i \in [k] \setminus [m] \end{cases}.$$

Using $\tilde{\delta}$, we define δ as follows:

$$\delta_i = \begin{cases} \tilde{\delta}_i, & \text{if } i \in [m] \\ -\tilde{\delta}_{i-m} & \text{if } i \in [k] \setminus [m] \end{cases}.$$

We now define p as follows: for $i \in [k]$, define $p_i = q_i(1+\delta_i)$. Equivalently,

$$p_i = \begin{cases} 0.5\tilde{q}_i(1+\tilde{\delta}_i), & \text{if } i \in [m] \\ 0.5\tilde{q}_{i-m}(1-\tilde{\delta}_{i-m}) & \text{if } i \in [k] \setminus [m] \end{cases}.$$

Thus, p is a valid distribution if q is a valid distribution. Let $\tilde{X}$ be the random variable such that $\mathbb{P}\{\tilde{X} = \tilde{\delta}_i\} = \tilde{q}_i$. We will need the following results, whose proofs are given at the end of this section:

Claim B.2. *We have the following inequality:*

$$0.02\mathbb{E}[\tilde{X}^2] \leq d_h^2(p,q) \leq \mathbb{E}[\tilde{X}^2].$$

Claim B.3. *Let $\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}$ be a channel corresponding to a threshold test. Then the following holds:*

$$d_h^2(\mathbf{T}p,\mathbf{T}q) \leq \sup_{0<\nu_1'<\cdots<\nu_D'=1} \sum_{j=1}^{D-1} \mathbb{P}\left\{\tilde{X} \geq \nu_j'\right\} \left(\mathbb{E}\left[\tilde{X}|\tilde{X} \geq \nu_j'\right]\right)^2 . \tag{25}$$

We will now show that there exists p and q (i.e., $\tilde{q} \in R^m$ and $\tilde{\delta} \in \mathbb{R}^m$), such that the desired conclusion holds. Defining $\tilde{q}$ and $\tilde{\delta}$ is equivalent to showing the existence of a random variable $\tilde{X}$ satisfying the desired properties. This is given in Claim B.4, showing that there exists a distribution $\tilde{X}$ such that the following holds: (i) $\mathbb{E}[\tilde{X}^2] = \Theta(\rho)$, and (ii) the expression on the right in (25), for any choice of thresholds Γ, is upper bounded by a constant multiple of $\frac{\mathbb{E}[\tilde{X}^2]D}{R}$, (iii) $R = \max(m,k') = \Theta(\log(1/\rho))$. See Claim B.4 for explicit constants. Using Claims B.2 to B.4, we get the following for any threshold channel $\mathbf{T}$:

$$d_h^2(\mathbf{T}p,\mathbf{T}q) \lesssim \mathbb{E}[\tilde{X}^2]\frac{D}{R} \lesssim d_h^2(p,q)\frac{D}{R},$$

completing the proof. □

The omitted proofs of Claim B.2 and Claim B.3 are given below.

Proof. (Proof of Claim B.2) We have the following:

$$\begin{aligned} d_h^2(p,q) &= \sum_{i\in[m]} \left(\sqrt{q_i\left(1+\delta_i\right)} - \sqrt{q_i}\right)^2 \\ &\quad + \sum_{i\in[k]\setminus[m]} \left(\sqrt{q_i} - \sqrt{q_i\left(1+\delta_i\right)}\right)^2 \\ &= 0.5\sum_{i\in[m]} \left(\sqrt{\tilde{q}_i\left(1+\tilde{\delta}_i\right)} - \sqrt{\tilde{q}_i}\right)^2 \\ &\quad + 0.5\sum_{i\in[m]} \left(\sqrt{\tilde{q}_i} - \sqrt{\tilde{q}_i\left(1-\tilde{\delta}_i\right)}\right)^2 \\ &= 0.5\sum_{i\in[m]} \tilde{q}_i\left(\left(\sqrt{1+\tilde{\delta}_i}-1\right)^2 + \left(1-\sqrt{1-\tilde{\delta}_i}\right)^2\right) \end{aligned}$$

Using that for $x \in [0,1]$: $\sqrt{1+x}-1 \geq 0.1x$, $1-\sqrt{1-x} \geq 0.1x$, $\sqrt{1+x} \leq 1+x$, $1-x \leq \sqrt{1-x}$, we have the following:

$$\mathbb{E}[\tilde{X}^2] \geq d_h^2(p,q) \geq 0.02\mathbb{E}[\tilde{X}^2].$$

□

Proof. (Proof of Claim B.3) Suppose $\mathbf{T}$ corresponds to a threshold test with thresholds $\Gamma = \{\gamma_1,\ldots,\gamma_{D-1}\}$ such that $\gamma_j < \gamma_{j+1}$. We define $\gamma_0 = \min_i p_i/q_i$ and $\gamma_D = \max_i p_i/q_i$. It suffices to consider the case when all $\gamma_j \in [0.5, 1.5]$ for $j \in [0:D]$. Let $p' = \mathbf{T}p$ and $q' = \mathbf{T}q$. Let $j^* \in [D-1]$ be such that $\gamma_{j^*-1} < 1$ and $\gamma_{j^*} \geq 1$.

We now define ν_j's as follows for $j \in [0:D-1]$:

$$\nu_j = \begin{cases} \gamma_j - 1, & \text{if } j \geq j^* \\ 1-\gamma_j, & \text{otherwise} \end{cases} .$$

Thus $\nu_j \in [0,1)$.

For $j \in [0:D-1]$, define $A_j := \{i \in [k] : (p_i/q_i) \in [\gamma_j,\gamma_{j+1})\} = \{i : 1+\delta_i \in [\gamma_j,\gamma_{j+1})\}$. For $j \geq j^*$, we have $A_j = \{i : \delta_i \in [\nu_j,\nu_{j+1})\}$. For $j < j^*$, we have $A_j = \{i \in [k] : -\delta_i \in (\nu_{j+1},\nu_j]\} = \left\{i \in [k] : \tilde{\delta}_{i-m} \in (\nu_{j+1},\nu_j]\right\}$. For $j \in [0:D-1]$, we have $p_j' = \sum_{i\in A_j} p_i$ and $q_j' = \sum_{i\in A_j} q_i$.

We have the following decomposition of the squared Hellinger distance between p' and q':

$$d_h^2(p',q') = \sum_{j<j^*}\left(\sqrt{p_j'} - \sqrt{q_j'}\right)^2 + \sum_{j\geq j^*}\left(\sqrt{p_j'} - \sqrt{q_j'}\right)^2 \tag{26}$$

We analyze these two terms separately:

a) Case 1: $j \geq j^*$: Let j be such that $q_j' > 0$. We have that $p_j' \in [q_j', 1.5q_j']$. We have the following using $\sqrt{1+x}-1 \leq x$ for $x \in [0,0.5]$:

$$\left(\sqrt{p_j'} - \sqrt{q_j'}\right)^2 = q_j'\left(\sqrt{1+\frac{p_\gamma' - q_j'}{q_j'}} - 1\right)^2 \leq \frac{(p_i' - q_j')^2}{q_j'}. \tag{27}$$

Since $\gamma_j \geq 1$, we note that

$$\begin{aligned} q_j' &= \sum_{i\in A_j} q_i = \sum_{i\in[m]:\tilde{\delta}_i\in[\nu_j,\nu_{j+1})} q_i \\ &= \sum_{i\in[m]:\tilde{\delta}_i\in[\nu_j,\nu_{j+1})} 0.5\tilde{q}_i = 0.5\mathbb{P}\{\tilde{X} \in [\nu_j,\nu_{j+1})\}. \end{aligned}$$

Similarly, we have the following:

$$\begin{aligned} p_j' - q_j' &= \sum_{i\in A_j} \delta_i q_i = \sum_{i\in[m]:\tilde{\delta}_i\in[\nu_j,\nu_{j+1})} \delta_i q_i \\ &= 0.5\sum_{i\in[m]:\tilde{\delta}_i\in[\nu_j,\nu_{j+1})} \tilde{\delta}_i\tilde{q}_i = 0.5\mathbb{E}\left[\tilde{X}\mathbb{I}_{\tilde{X}\in[\nu_j,\nu_{j+1})}\right]. \end{aligned}$$

Combining the last two displayed equations with (27) and using the definition of conditional expectation, we get the following:

$$\begin{aligned} &\sum_{j\geq j^*}\left(\sqrt{p_j'} - \sqrt{q_j'}\right)^2 \\ &\quad\leq 0.5\sum_{j\geq j^*} \mathbb{P}\left\{\tilde{X} \in [\nu_j,\nu_{j+1})\right\}\left(\mathbb{E}\left[\tilde{X}|\tilde{X} \in [\nu_j,\nu_{j+1})\right]\right)^2 \\ &\quad\leq 0.5\sum_{j\geq j^*} \mathbb{P}\left\{\tilde{X} \geq \nu_j\right\}\left(\mathbb{E}\left[\tilde{X}|\tilde{X} \geq \nu_j)\right]\right)^2 . \end{aligned} \tag{28}$$

Case $2 : j < j^*$**:** Let $j < j^*$ such that $q'_j > 0$. We have that $p'_j \in [q'_j/2, q'_j)$. Using $1 - \sqrt{1-x} \leq x$ for $x \in [0,1]$, we have

$$\left(\sqrt{q'_j} - \sqrt{p'_j}\right)^2 = q'_j \left(1 - \sqrt{1 - \frac{q'_j - p'_i}{q'_j}}\right)^2 \leq \frac{(q'_j - p'_i)^2}{q'_j}. \tag{29}$$

Since $\gamma_j < 1$, we note that

$$\begin{aligned} q'_j = \sum_{i \in A_j} q_i &= \sum_{i \in [k] \setminus [m] : \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} q_i \\ &= \sum_{i \in [k] \setminus [m] : \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} 0.5 \tilde{q}_{i-m} \\ &= 0.5 \mathbb{P}\{\tilde{X} \in (\nu_{j+1}, \nu_j]\}. \end{aligned}$$

Similarly, we have the following:

$$\begin{aligned} q'_j - p'_i = \sum_{i \in A_j} (-\delta_i q_i) &= \sum_{i \in [k] \setminus [m] : \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} \tilde{\delta}_i (0.5 \tilde{q}_{i-m}) \\ &= 0.5 \mathbb{E}\left[\tilde{X} \mathbb{I}_{\tilde{X} \in (\nu_{j+1}, \nu_j]}\right]. \end{aligned}$$

Combining the last two displayed equations with (29) and using the definition of conditional expectation, we get the following:

$$\begin{aligned} &\sum_{j < j^*} \left(\sqrt{p'_j} - \sqrt{q'_j}\right)^2 \\ &\leq 0.5 \sum_{j < j^*} \mathbb{P}\left\{\tilde{X} \in (\nu_{j+1}, \nu_j]\right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} \in (\nu_{j+1}, \nu_j]\right]\right)^2 \\ &\leq 0.5 \sum_{j < j^*} \mathbb{P}\left\{\tilde{X} > \nu_j]\right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} > \nu_j]\right]\right)^2 \end{aligned} \tag{30}$$

Combining (28) and (30), we get the proof by noting that $\tilde{X}$ is a discrete random variable and thus the distinction between $\tilde{X} \geq \nu_j$ (cf. (28)) and $\tilde{X} > \nu_j$ (cf. (30)) does not matter when taking the supremum.

□

C. Reverse Markov Inequality

Lemma B.1. *Let Y be a random variable over $[0, \beta)$ with expectation $\mathbb{E}[Y] > 0$. Let $k' = 1 + \log(\beta / \mathbb{E}[Y])$. Then we have the following:*

$$\begin{aligned} \sup_{\beta = \nu_D \geq \ldots \geq \nu_1 \geq 0} \sum_{j=1}^{D-1} \nu_j \mathbb{P}\left(Y \in [\nu_j, \nu_{j+1})\right) \\ \geq \frac{1}{13} \mathbb{E}[Y] \min\left(1, \frac{D}{R}\right), \end{aligned} \tag{17}$$

where $R = k' := 1 + \log(\beta / \mathbb{E}[Y])$. Furthermore, the bound in (17) can be achieved by ν_j's such that $\nu_j = \min(\beta, x 2^j)$ for an $x \in [0, \beta]$.

For the special setting when Y is supported on k points: we may set $R = \min(k, k')$, and there is a $\operatorname{poly}(k, D)$ algorithm to find ν_j's that achieve the right hand side of inequality (17).

Proof. We can safely assume that $D \leq R$. Under this assumption on D, we will show the desired expression is lower bounded by both of the following (up to constants): $\frac{\mathbb{E}[Y] D}{k}$ and $\frac{\mathbb{E}[Y] D}{k'}$. We will also assume that $\beta = 1$; otherwise, it suffices to apply the following argument to Y / β.

We first begin with dependence on k.

a) Dependence on k:: We have that Y has support size k. Let the support elements be $\{\delta'_i\}_{i=1}^k$ such that $\delta'_1 < \delta'_2 < \cdots < \delta'_k < 1$[12]. Let $\{p_i\}_{i=1}^k$ be such that $\mathbb{P}[Y = \delta'_i] = p_i$ and $\sum_{i=1}^k p_i = 1$.

It suffices to prove the following bound: there exists a labeling $\pi : [D-1] \to [k]$ such that $\pi(1) < \pi(2) < \cdots < \pi(D-1)$ and satisfies the following

$$\sum_{j=1}^{D-1} \delta'_{\pi(j)} p_{\pi(j)} \geq \mathbb{E}[Y] \left(\frac{D-1}{k}\right). \tag{31}$$

This is true because for $j \in [D-1]$, we have $p_{\pi(j)} = \mathbb{P}\left\{Y \in [\delta_{\pi(j)}, \delta_{\pi(j)+1})\right\} \leq \mathbb{P}\left\{Y \in [\delta_{\pi(j)}, \delta_{\pi(j+1)})\right\}$, where we define $\delta_{k+1} := 1$, and the desired conclusion follows by setting $\nu_j = \delta_{\pi(j)}$. In the rest of the proof, we will show that such a $\pi(\cdot)$ exists.

Let $\sigma : [k] \to [k]$ be the permutation such that $p_{\sigma(i)} \delta_{\sigma(i)} \geq p_{\sigma(i+1)} \delta_{\sigma(i+1)}$. Then we have that

$$\begin{aligned} \frac{\mathbb{E}[Y]}{k} &= \frac{\sum_{i=1}^k p_i \delta_i}{k} \\ &= \frac{\sum_{i=1}^k p_{\sigma(i)} \delta_{\sigma(i)}}{k} \leq \frac{\sum_{i=1}^{D-1} p_{\sigma(i)} \delta_{\sigma(i)}}{D-1} \\ &= \frac{\sum_{i=1}^{D-1} p_{\pi'(i)} \delta_{\pi'(i)}}{D-1}, \end{aligned}$$

for some $\pi' : [D-1] \to [k]$ such that $\pi'(1) < \pi'(2) < \cdots < \pi'(D-1)$. Thus we have established (31). Note that the desired bound is achieved by choosing ν_j's as follows: let S be the set of top $D-1$ elements among the support of Y that maximize $y \mathbb{P}(Y = y)$ and ν_j's have values in S such that they are increasing and distinct. It is clear that it can be implemented in $\operatorname{poly}(k, D)$ time.

b) Dependence on k':: We begin by noting that the desired expression can also be written as follows:

$$\sum_{j=1}^{D-1} (\nu_j - \nu_{j-1}) \mathbb{P}\left\{Y \geq \nu_j\right\},$$

where $\nu_0 := 0$. We need to obtain a lower bound on the supremum of this expression over ν_j's. In fact, we will show a stronger claim, where we fix ν_j in a particular way. We will take ν_j to be of the form $x 2^{j-1}$ for $j \in [D-1]$ and optimize over $x \in (0, 1)$. Note that we can allow $\nu_j \geq 1$ without loss of generality because their contribution to the desired expression

[12] It is easy to see that if the support size is strictly smaller than k, then we have a tighter bound.

is going to be 0. In the rest of the proof, we will show the following claim for $c = 1/13$:

$$\sup_{x\in(0,1)} x\mathbb{P}\left\{Y \ge x\right\} + \sum_{j=2}^{D-1}\left(x2^{j-1} - x2^{j-2}\right)\mathbb{P}\left\{Y \ge x2^{j-1}\right\} \ge c\mathbb{E}[Y]\frac{D}{k'} \tag{32}$$

Suppose that the desired conclusion does not hold. We will now derive a contradiction. Let $x \in (0,1)$ be arbitrary and set $\nu_j = x2^{j-1}$ for $j \in [D-1]$. Under the assumption that (32) is false, we have the following for each $x \in (0,1)$:

$$\begin{aligned} c\mathbb{E}[Y]\frac{D}{k'} &> x\mathbb{P}\left\{Y \ge x\right\} \\ &\quad + \sum_{j=2}^{D-1} x\left(x2^{j-1} - x2^{j-2}\right)\mathbb{P}\left\{Y \ge x2^{j-1}\right\} \\ &= x\left(\mathbb{P}\left\{Y \ge x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{P}\left\{2^{-j}Y \ge x\right\}\right). \end{aligned}$$

We thus get the following for all $x \in (0,1)$:

$$\mathbb{P}\left\{Y \ge x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{P}\left\{2^{-j}Y \ge x\right\} < c\mathbb{E}[Y]\frac{D}{k'}\frac{1}{x}. \tag{33}$$

Using that the probabilities are bounded by 1, we also have the following bound on the expression on the left in (33):

$$\mathbb{P}\left\{Y \ge x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{P}\left\{2^{-j}Y \ge x\right\} \le 1 + \sum_{j=1}^{D-2} 2^{j-1} = 2^{D-1}. \tag{34}$$

Combining the two bounds in (33) and (34), we have the following for every $x \in (0,1)$:

$$\mathbb{P}\left\{Y \ge x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{P}\left\{2^{-j}Y \ge x\right\} \le \min\left(2^{D-1}, \frac{cD\mathbb{E}[Y]}{k'x}\right). \tag{35}$$

Using that $Y \in [0,1]$, we also have the following for every $a > 1$:

$$\mathbb{E}[Y/a] = \int_0^{1/a}\mathbb{P}\left\{Y/a \ge t\right\}dt = \int_0^1 \mathbb{P}\left\{Y/a \ge t\right\}dt.$$

This leads to the following equality:

$$\begin{aligned} &\int_0^1\left(\mathbb{P}\left\{Y \ge t\right\} + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{P}\left\{Y2^{-j} \ge t\right\}\right)dt \\ &\qquad = \mathbb{E}[Y] + \sum_{j=1}^{D-2} 2^{j-1}\mathbb{E}[2^{-j}Y] = \frac{D\mathbb{E}[Y]}{2}. \end{aligned} \tag{36}$$

Combining (36) and (35), we get the following for an arbitrary $x_* \in (0,1)$:

$$\begin{aligned} \mathbb{E}[Y] &= \frac{2}{D}\int_0^1\left(\mathbb{P}\left\{Y \ge t\right\} + \sum_{i=1}^{D-2} 2^{i-1}\mathbb{P}\left\{Y2^{-i} \ge t\right\}\right)dt \\ &\le \frac{2}{D}\int_0^1 \min\left(2^{D-1}, \frac{cD\mathbb{E}[Y]}{k'x}\right)dt \\ &\le \frac{2^D x_*}{D} + \frac{2c\mathbb{E}[Y]}{k'}\log\left(\frac{1}{x_*}\right). \end{aligned} \tag{37}$$

Let $x_* = 2^{-D}\left(\mathbb{E}[Y]/k'\right)$. Then $\log(1/x_*) = D + \log(1/\mathbb{E}[Y]) + \log(1/k') \le D + 2k' \le 3k'$, where we use that $\max(D, \log(1/\mathbb{E}[Y])) \le k'$. Using $k' \ge 1$ and $D \ge 2$, the expression on the right in (37) can be further upper bounded to get the following inequality:

$$\mathbb{E}[Y] \le \left(\frac{\mathbb{E}[Y]}{k'D}\right) + \frac{2c\mathbb{E}[Y]}{k'}(3k') \le \frac{\mathbb{E}[Y]}{2} + 6c\mathbb{E}[Y],$$

which is a contradiction since $\mathbb{E}[Y] > 0$ and $c < 1/12$. Thus, we have that (32) is true.

□

D. Tightness of Reverse Markov Inequality

Claim B.4. *There exists constants c, c', c_1, c_2 such for every $\rho \in (0, c')$, there exists a $k \in \mathbb{N}$, where $k = \Theta(\log(1/\rho))$, and a probability distribution p, supported over k points in $(0, 0.5]$, such that the following holds:*

1) *$\mathbb{E}[X^2] \in [c_1\rho, c_2\rho]$ and for every $D \le 0.1k$,*

$$\sup_{1=\delta_D>\cdots>\delta_1>0}\sum_{j=1}^{D-1}\mathbb{P}\left\{X \ge \delta_j\right\}\left(\mathbb{E}\left[X|X \ge \delta_j\right]\right)^2 \le c\mathbb{E}X^2\frac{D}{R}, \tag{38}$$

where $R = \max(k, k')$ and $k' = \log(3/\mathbb{E}[X^2]))$.

2) *$\mathbb{E}[Y] = [c_1\rho, c_2\rho]$ and*

$$\sup_{1=\delta'_D>\ldots>\delta'_1>0}\sum_{j=1}^{D-1}\delta'_j\mathbb{P}\left(Y \in [\delta'_j, \delta'_{j+1})\right) \le c\mathbb{E}[Y]\frac{D}{R}, \tag{39}$$

where $R = \max(k, k')$ and $k' = \log(3/\mathbb{E}[Y]))$. Moreover, $R = \Theta((\log(1/\rho)))$.

Proof. For now, let $k \in \mathbb{N}$ be arbitrary; we will choose k so that $\mathbb{E}[X^2] \in [c_1\rho, c_2\rho]$. Consider the following discrete random variable Y supported on $\{2^{-i} : i \in [k]\}$:

$$\mathbb{P}\left\{Y = 2^{-i}\right\} = r2^i,$$

where r is chosen so that it is a valid distribution, i.e., r satisfies $1 = \sum_{i=1}^k r2^i = 2r(2^k - 1)$. Let $X = \sqrt{Y}$. We have the following:

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = \sum_{i=1}^k 2^{-i}\left(r2^i\right) = rk. \tag{40}$$

Consider a $\delta_i' \in (2^{-(j-1)}, 2^{-j}]$ for some $j \in [k]$ and $\delta_i = \sqrt{\delta_i'}$. For any such choice, we get the following:

$$\mathbb{P}(X \geq \delta_i) = \mathbb{P}(Y \geq \delta_i') = \mathbb{P}(Y \geq 2^{-j}) = \sum_{i \in [j]} \mathbb{P}\left\{Y = 2^{-i}\right\}$$

$$= \sum_{i \in [j]} r 2^i = 2r\left(2^j - 1\right) \leq 2r2^j. \tag{41}$$

Thus we get that for any δ', $\delta' \mathbb{P}\left\{Y \geq \delta'\right\} \leq 2r$, showing that the expression in (39) is upper bounded by $2(D-1)r$, which is equal to $2\frac{\mathbb{E}[Y](D-1)}{k}$ by (40). It remains to show that $R \leq ck/2$.

We first calculate bounds on k so that $\mathbb{E}[Y] = \Theta(\rho)$. Note that by construction $r = 1/\left(2(2^k-1)\right)$, implying that $r \in [2^{-k-1}, 2^{-k+1}]$. Thus it suffices to choose a k so that $f(k) := 2^{-k}k \in [2c_1\rho, 0.5c_2\rho]$. As $f(k+1)/f(k) \in (1/2, 1)$ for $k > 1$, $f(\lceil \ln(1/\rho) \rceil) \geq \rho \ln(1/\rho)$ and $f(\lfloor 2\ln(1/\rho) \rfloor) \geq 0.1\rho$, we know that such a k exists in $[\ln(1/\rho), 2\ln(1/\rho)]$ for $c_1 = 0.5$, $c_2 = 10$, $c' = 2^{-20}$.

We now calculate the quantity R. By definition of k', we have that

$$k' = \log\left(\frac{3}{\mathbb{E}[X^2]}\right) = \log\left(\frac{3}{rk}\right) \geq \log\left(\frac{32^{k-1}}{k}\right)$$
$$= k - \log k \pm \log 2.$$

As k is large enough, we have that $k' \in [0.5k, 2k]$. Since $R = \max(k, k')$, we have that $R \in [0.5k, 2k]$. This completes the proof of the claim in (39) with $c = 4$.

We now prove the claim in (38). We begin with the following:

$$\mathbb{E}[X\mathbb{I}_{X \geq 2^{-j/2}}] = \sum_{i \in [j]} 2^{-0.5i}\left(r2^i\right) = \sum_{i \in [j]} r2^{0.5i}$$
$$= (r\sqrt{2})\left(\frac{2^{0.5j}-1}{\sqrt{2}-1}\right) \leq 10r2^{0.5j}.$$

For $\delta_i \in (2^{-(j-1)/2}, 2^{-j/2}]$ for some j, we have the following

$$\mathbb{P}\left\{X \geq \delta_i\right\}\left(\mathbb{E}\Big[X | X \geq \delta_i\Big]\right)^2$$
$$= \mathbb{P}(X \geq 2^{-j/2})\left(\mathbb{E}[X | X \geq 2^{-j/2}]\right)^2$$
$$= \frac{\left(\mathbb{E}[X\mathbb{I}_{X \geq 2^{-j/2}}]\right)^2}{\mathbb{P}(X \geq 2^{-j/2})}$$
$$\leq \frac{100r^2 2^j}{2r2^j} = 50r.$$

We thus get that the supremum over any arbitrary δ_i is also upper bounded by $50r$. Hence, we get upper bound the expression on the left in (38) by $50(D-1)r$, which is equal to $50\mathbb{E}[X^2](D-1)/r$. Using the same calculations as in the first part of the proof, we prove (38) with $c = 50$.

□

Appendix C
Hypothesis Testing

Lemma C.1 (Equivalence between identical and non-identical channels for simple hypothesis testing)**.** *Let $\mathcal{T}$ be a collection of channels from $\mathcal{X} \to \mathcal{Y}$. Let p and q be two distributions on $\mathcal{X}$. Then*

$$\mathrm{n}^*_{\texttt{non-identical}}(p, q, \mathcal{T}) = \Theta\left(\mathrm{n}^*_{\texttt{identical}}(p, q, \mathcal{T})\right).$$

Proof. (Proof of Lemma C.1 Recall that we use $\beta_h(p,q)$ to denote the Hellinger-affinity between p and q. It suffices to consider the case that $\mathrm{n}^*_{\texttt{identical}}(p, q, \mathbf{T})$ is larger than a fixed constant. Define the following:

$$h_* = \sup_{\mathbf{T} \in \mathcal{T}} d_h(\mathbf{T}p, \mathbf{T}q), \qquad \beta_* := \inf_{\mathbf{T} \in \mathcal{T}} \beta_h(\mathbf{T}p, \mathbf{T}q)$$

and $\beta_* = 1 - 0.5h_*^2$. Let $n_* = \mathrm{n}^*_{\texttt{identical}}(p, q, \mathbf{T})$. Let $\mathbf{T}_*$ be any channel such that $\beta_h(\mathbf{T}_*p, \mathbf{T}_*q) \leq \beta_* + \epsilon\beta_*$, for some $\epsilon > 0$ satisfying $(1+\epsilon)^{n_*} \leq 2$. Let $p_* = \mathbf{T}_*p$ and $q_* = \mathbf{T}_*q$.

a) Identical channels: optimal $\mathbf{T}_*$: If each channel is identically $\mathbf{T}_*$, we have that the joint distributions of n_* samples would either be $p_*^{\otimes n_*}$ or $q_*^{\otimes n_*}$.

Let $f(n) = d_{\mathrm{TV}}(p_*^{\otimes n}, q_*^{\otimes n})$. Note that the probability of error for $p^{\otimes n}$ and $q^{\otimes n}$ is equal to $(1 - f(n))/2$ (Fact A.2). Since the sample complexity of $\mathcal{B}(p_*, q_*)$ is at least n_*, we must have that $f(n_* - 1) < 0.8$. Using Fact A.1, we have $d_h^2(p_*^{\otimes(n_*-1)}, q_*^{\otimes(n_*-1)}) \leq 1.6$ and consequently, $\beta_h(p_*^{\otimes(n_*-1)}, q_*^{\otimes(n_*-1)}) \geq 0.2$. Using tensorization of Hellinger affinity from Fact A.1 and relation between β_* and $\beta_h(p_*, q_*)$, we have $(\beta_*)^{n_*-1} \geq (\frac{\beta_h(p^*, q^*)}{1+\epsilon})^{n_*-1} \geq 0.1$.

b) Non-identical Channels: We will now show that even if n non-identical channels are allowed but $n \leq 0.1n_*$, then the probability of error is at least 0.2. For a choice of $\mathbf{T}_1, \ldots, \mathbf{T}_n$, let $P_n' := \prod_{i=1}^n \mathbf{T}_i p$ and $Q_n' := \prod_{i=1}^n \mathbf{T}_i q$ be the resulting joint probability distributions under p and q respectively. As the probability of error of the best test is $(1 - d_{\mathrm{TV}}\left(P_n', Q_n'\right)/2)$ (Fact A.2), it suffices to show that if $n \leq 0.1n_*$, then $d_{\mathrm{TV}}(P_n', Q_n') < 0.8$.

Using Fact A.1, it suffices to show that $d_h^2(P_n', Q_n') \leq 0.64$. Equivalently, it suffices to show that $\beta_h(P_n', Q_n') \geq 0.68$. Using tensorization of Hellinger affinity and optimality of β_*, we have the following:

$$\beta_h(P_n', Q_n') = \prod_{i=1}^n \beta_h(\mathbf{T}_i p, \mathbf{T}_i q) \geq \beta_*^n$$
$$= (\beta_*^{n_*-1})^{\frac{n}{n_*-1}} \geq \left((0.1)^{\frac{1}{10}}\right)^{\frac{10n}{n_*-1}} \geq 0.7^{\frac{10n}{n_*-1}}.$$

Thus if $n \leq 0.1(n_* - 1)$, then the Hellinger affinity is large, and thus total variation is small, and hence the probability of error is large.

□

Proof. (Proof of Theorem IV.2) We note that it suffices to consider $D \leq \log n_0$.

We will use the characterization of sample complexity of simple hypothesis testing in terms of Hellinger distance (Fact A.2). It thus suffices to show that there exists a constant

$c > 0$ such that for every $\rho \in (0, 0.01)$, there exist two distributions p and q on $[k]$ such that $d_h^2(p,q) = \rho$ and

$$\mathrm{n}^*_{\texttt{non-identical}}(p,q,\mathcal{T}_D) \geq c\frac{1}{d_h^2(p,q)} \min\left(1, \frac{\log\left(1/d_h^2(p,q)\right)}{D}\right). \tag{42}$$

This follows by noting that given any n_0, any two distributions p and q such that $d_h^2(p,q) = \rho$ would satisfy $\mathrm{n}^*(p,q) \in [c_1 n_0, c_2 no]$ for some absolute constants (see Fact A.2).

Fix any $\rho \in (0,1)$. Let p, q, and $k = O(\log(1/\rho))$ be from Lemma III.5 such that (i) $d_h^2(p,q) = \rho$ and (ii) inequality (6) holds.

We will use Theorem II.9, Lemma C.1, and Lemma III.5. For any p and q, we have the following:

$$\begin{aligned}
&\mathrm{n}^*_{\texttt{non-identical}}(\{p,q\},\mathcal{T}_D) \\
&\geq c'\mathrm{n}^*_{\texttt{non-identical}}(\{p,q\},\mathcal{T}_D^{\text{thresh}}) && \text{(using Theorem II.9)} \\
&\geq c'\mathrm{n}^*_{\texttt{identical}}(\{p,q\},\mathcal{T}_D^{\text{thresh}}) && \text{(using Lemma C.1)} \\
&= c' \inf_{\mathbf{T}\in\mathcal{T}_D^{\text{thresh}}} \mathrm{n}^*(p,q,\{\mathbf{T}\}) \\
&\geq c' \inf_{\mathbf{T}\in\mathcal{T}_D^{\text{thresh}}} \frac{1}{d_h^2(\mathbf{T}p,\mathbf{T}q)} && \text{(using Fact A.2)} \\
&= c'\frac{1}{d_h^2(p,q)} \inf_{\mathbf{T}\in\mathcal{T}_D^{\text{thresh}}} \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}p,\mathbf{T}q)} \\
&\geq c'\frac{1}{d_h^2(p,q)} \frac{d_h^2(p,q)}{d_h^2(\mathbf{T}p,\mathbf{T}q)} \\
&\geq c'\frac{1}{d_h^2(p,q)} \frac{\log(1/d_h^2(p,q))}{D}, && \text{(using Lemma III.5)}
\end{aligned}$$

which establishes the condition (42) and thus completes the proof.

□

Appendix D
Auxiliary Details

Claim D.1. *(Examples of well-behaved f-divergences) The following divergences are valid examples of well-behaved f-divergences, as defined in Definition III.1:*

1) *(Hellinger distance) $f(x) = (\sqrt{x}-1)^2$ with $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$.*
2) *(Total variation distance) $f(x) = 0.5|x-1|$ with $\kappa > 0$, $C_1 = 0.5$, $C_2 = 0.5$, $\alpha = 1$.*
3) *(Symmetrized KL-divergence) $f(x) = x\log x - \log x$ with $\kappa = 1$, $C_1 = 0.5$, $C_2 = 1$, $\alpha = 2$.*
4) *(Triangular Discrimination) $f(x) = \frac{(x-1)^2}{1+x}$ with $\kappa = 1$, $C_1 = 1/3$, $C_2 = 1/2$, $\alpha = 2$.*
5) *(Symmetrized χ^s divergence[13]) For $s \geq 1$, $f(x) = |x-1|^s + x^{1-s}|x-1|^s$ with $\kappa = 1$, $C_1 = 1$, $C_2 = 3$, $\alpha = s$.*

Proof. It is easy to see that these functions are non-negative, convex, and satisfy the symmetry property of Definition III.1. In the remainder of the proof, we outline how they satisfy Item I.3 property.

1) We will now show that we can take $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$. The upper bound $f(1+x) = (\sqrt{1+x}-1)^2 \leq x^2$ follows by noting that $\sqrt{1+x} \leq 1+x$ for any $x \geq 0$. For the lower bound, we define $g(x) := f(1+x) - C_1 x^2$. Note that $g(0) = 0$, $g'(x) = 1-(1+x)^{-0.5} - 2C_1 x$ and $g''(x) = 0.5(1+x)^{-1.5} - 2C_1$. We note that $g'(0) = 0$ and $g''(x) \geq g''(1) = 2^{-2.5} - 2C_1$ for all $x \in [0,1]$. Thus $g''(x) \geq 0$ for $x \in [0,1]$, and hence $g(x)$ is also non-negative on $x \in [0,1]$.
2) The result follows by noting that for $x \geq 0$, $f(1+x) = x$.
3) We have that $f(1+x) = x\log(1+x)$. We use the following result: for $x \geq 0$, $\frac{x}{1+x} \leq \log(1+x) \leq x$. This directly gives us that $f(1+x) = x\log(1+x) \leq x^2$. The lower bound follows by noting that for $x \in [0,1]$, $\log(1+x) \geq \frac{x}{2}$ and thus $f(1+x) \geq x^2/2$.
4) We have that $f(1+x) = x^2/(2+x)$, which lies between $x^2/3$ and $x^2/2$ for $x \in [0,1]$.
5) We have that $f(1+x) = |x|^s(1+(1+x)^{1-s})$, which is larger than x^s and less than $3x^s$ for $x \in [0,1]$.

□

We now provide the proof of Corollary III.3.

Proof. (Proof of Corollary III.3) The desired bound follows by noting that $f(x) = (\sqrt{x}-1)^2$ and taking $\nu = 0$. As shown in Appendix D (Claim D.1), we can take $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$. Note that $f(0) = 1$ and $f(1/(1+\kappa)) = (\sqrt{2}-1)^2/2 \geq 0.04$. This suffices to give a guarantee of $\frac{d_h^2(p,q)}{d_h^2(\mathbf{T}^*p,\mathbf{T}^*q)} \leq 100 + \frac{900}{D}(\min(k',k))$. □

Remark D.2. Lemma III.6 states that the truncation factor (compared to the usual upper bound in Markov's inequality) is at most $O(\log(1/\mathbb{E}[X]))$. It is instructive to compare the guarantee of Lemma III.6 with existing results in the literature:

1) The Paley-Zygmund inequality (see, e.g., [26, Corollary 3.3.2]) states that for any $\delta \in (0, \mathbb{E}[X])$, we have $\mathbb{P}(X \geq \delta) \geq \left(1 - \frac{\delta}{\mathbb{E}X}\right)^2 \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$. Multiplying both sides by δ and optimizing the lower bound over δ (achieved at $\delta = \mathbb{E}[X]/3$) yields

$$\sup_{\delta\geq 0} \delta\mathbb{P}(X \geq \delta) \gtrsim \mathbb{E}[X] \cdot \frac{1}{\mathbb{E}[X^2]/(\mathbb{E}[X])^2}.$$

Note that the truncation factor is $\mathbb{E}[X^2]/(\mathbb{E}[X])^2$, which is at most $1/\mathbb{E}[X]$ but could be exponentially larger than $\log(1/\mathbb{E}[X])$. (For example, consider a random variable with $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = 1/2) = p$: We have $\mathbb{E}[X^2]/(\mathbb{E}[X])^2 = 1/p$, whereas $\log(1/\mathbb{E}[X]) = \log(2/p)$).

2) A standard version of the reverse Markov inequality (see, e.g., [27, Lemma B.1]) for a random variable bounded in $[0,1]$ states the following for $\delta \in (0, \mathbb{E}[X])$: $\mathbb{P}(X \geq \delta) \geq \frac{\mathbb{E}[X]-\delta}{1-\delta}$. Multiplying both sides by δ and optimizing the bound over $\delta \in (0, \mathbb{E}[X])$, under the condition that $\mathbb{E}[X] \leq 0.1$, gives us following:

$$\sup_{\delta\geq 0} \delta\mathbb{P}(X \geq \delta) \gtrsim (\mathbb{E}[X])^2 = \mathbb{E}[X] \cdot \frac{1}{1/\mathbb{E}[X]},$$

[13] The usual χ^s-divergence corresponds to $f(x) = |x-1|^s$ for $s \geq 1$ [17]. We consider the symmeterized version with $\tilde{f}(x) = f(x) + xf(1/x)$.

i.e., the truncation factor is $1/\mathbb{E}[X]$, which is exponentially larger than $\log(1/\mathbb{E}[X])$.