Sample Amplification: Increasing Dataset Size even when Learning is Impossible

Brian Axelrod Shivam Garg Vatsal Sharan Gregory Valiant

{baxelrod, shivamgarg, vsharan, valiant}@stanford.edu

Stanford University[*]

Abstract

Given data drawn from an unknown distribution, D, to what extent is it possible to "amplify" this dataset and faithfully output an even larger set of samples that appear to have been drawn from D? We formalize this question as follows: an (n, m) *amplification procedure* takes as input n independent draws from an unknown distribution D, and outputs a set of $m > n$ "samples". An amplification procedure is *valid* if no algorithm can distinguish the set of m samples produced by the amplifier from a set of m independent draws from D, with probability greater than 2/3. Perhaps surprisingly, in many settings, a valid amplification procedure exists, even in the regime where the size of the input dataset, n, is significantly less than what would be necessary to *learn* distribution D to non-trivial accuracy. Specifically we consider two fundamental settings: the case where D is an arbitrary discrete distribution supported on $\leq k$ elements, and the case where D is a d-dimensional Gaussian with unknown mean, and fixed covariance matrix. In the first case, we show that an $\left(n, n + \Theta(\frac{n}{\sqrt{k}})\right)$ amplifier exists. In particular, given $n = O(\sqrt{k})$ samples from D, one can output a set of $m = n+1$ datapoints, whose total variation distance from the distribution of m i.i.d. draws from D is a small constant, despite the fact that one would need quadratically more data, $n = \Theta(k)$, to *learn* D up to small constant total variation distance. In the Gaussian case, we show that an $\left(n, n + \Theta(\frac{n}{\sqrt{d}})\right)$ amplifier exists, even though learning the distribution to small constant total variation distance requires $\Theta(d)$ samples. In both the discrete and Gaussian settings, we show that these results are tight, to constant factors. Beyond these results, we formalize a number of curious directions for future research along this vein.

1 Learning, Testing, and Sample Amplification

How much do you need to know about at distribution, D, in order to produce a dataset of size m that is indistinguishable from a set of independent draws from D? Do you need to *learn* D, to nontrivial accuracy in some natural metric, or does it suffice to have access to a smaller dataset of size $n < m$ drawn from D, and then "amplify" this dataset to create one of size m? In this work we formalize this question, and show that for two natural classes of distribution, discrete distributions with bounded support, and d-dimensional Gaussians, non-trivial data "amplification" is possible even in the regime in which you are given too few samples to learn.

From a theoretical perspective, this question is related to the meta-question underlying work on distributional property testing and estimation: *To answer basic hypothesis testing or property estimation questions regarding a distribution D, to what extent must one first learn D, and can such questions be reliably answered given a relatively modest amount of data drawn from D?* Much of the excitement surrounding distributional property testing and estimation stems from the fact that, for many such testing and estimation questions, a surprisingly small set of samples from D suffices—significantly fewer samples than would be required to learn D. These surprising answers have been revealed over the past two decades of work on property testing and estimation. The question posed in our work fits with this body of work, though instead of asking how

[*] This work was supported by NSF awards AF-1813049 and CCF-1704417, an ONR Young Investigator Award N00014-18-1-2295, and an NSF Graduate Fellowship.

much data is required to perform a hypothesis test, we are asking how much data is required to *fool* an optimal hypothesis test—in this case an "identity tester" which knows D and is trying to distinguish a set of m independent samples drawn from D, versus m datapoints constructed in some other fashion.

From a more practical perspective, the question we consider also seems timely. Deep neural network based systems, trained on a set of samples, can be designed to perform many tasks, including testing whether a given input was drawn from a distribution in question (i.e. "discrimination"), as well as sampling (often via the popular Generative Adversarial Network approach). There are many relevant questions regarding the extent to which current systems are successful in accomplishing these tasks, and the question of how to quantify the performance of these systems is still largely open. In this work, however, we ask a different question: Suppose a system *can* accomplish such a task—what would that actually mean? If a system can produce a dataset that is indistinguishable from a set of m independent draws from a distribution, D, does that mean the system knows D, or are there other ways of accomplishing this task?

1.1 Formal Problem Definition

We begin by formally stating two essentially equivalent definitions of sample amplification and then provide an illustrative example. Our first definition of sample amplification, states that a function f mapping a set of n datapoints to a set of m datapoints is a valid amplification procedure for a class of distributions $\mathcal{C}$, if for all $D \in \mathcal{C}$, letting X_n denote the random variable corresponding to n independent draws from D, the distribution of $f(X_n)$ has small total variation distance [1] to the distribution defined by m independent draws from D.

Definition 1. *A class $\mathcal{C}$ of distributions over domain S admits an (n, m)* amplification procedure *if there exists a (possibly randomized) function $f_{\mathcal{C},n,m} : S^n \to S^m$, mapping a dataset of size n to a dataset of size m, such that for every distribution $D \in \mathcal{C}$,*

$$D_{TV}\left(f_{\mathcal{C},n,m}(X_n), D^m\right) \leq 1/3,$$

where X_n is the random variable denoting n independent draws from D, and D^m denotes the distribution of m independent draws from D. If no such function $f_{\mathcal{C},n,m}$ exists, we say that $\mathcal{C}$ does not admit *an (n, m) amplification scheme.*

Crucially, in the above definition we are considering the random variable $f(X_n)$ whose randomness comes from the randomness of X_n, as well as any randomness in the function f itself. For example, every class of distributions admits an (n, n) amplification procedure, corresponding to taking the function f to be the identity function. If, instead, our definition had required that the *conditional* distribution of $f(X_n)$ given X_n be close to D^m, then the above definition would simply correspond to asking how well we can *learn* D, given the n samples denoted by X_n.

Definition 1 is also equivalent, up to the choice of constant $1/3$ in the bound on total variation distance, to the following intuitive formulation of sample amplification as a game between two parties: the amplifier who will produce a dataset of size m, and a "verifier" who knows D and will either accept or reject that dataset. The verifier's protocol, however, must satisfy the condition that given m independent draws from the true distribution in question, the verifier must accept with probability at least $3/4$, where the probability is with respect to both the randomness of the set of samples, and any internal randomness of the verifier. We briefly describe this formulation, as a number of natural directions for future work—such as if the verifier is computationally bounded, or only has sample access to D—are easier to articulate in this setting.

Definition 2. *The* sample amplification *game consists of two parties, an* amplifier *corresponding to a function $f_{n,m} : S^n \to S^m$ which maps a set of n datapoints in domain S to a set of m datapoints, and a* verifier *corresponding to a function $v : S^m \to \{ACCEPT, REJECT\}$. We say that a verifier v is* valid *for distribution D if, when given as input a set of m independent draws from D, the verifier accepts with probability at least $3/4$, where the probability is over both the randomness of the draws and any internal randomness of v:*

$$\Pr_{X_m \leftarrow D^m}[v(X_m) = ACCEPT] \geq 3/4.$$

[1] We overload the notation $D_{TV}(\cdot, \cdot)$ for total variation distance, and also use it when the argument is a random variable instead of the distribution of the random variable, whenever convenient.

A class $\mathcal{C}$ of distributions over domain S admits an (n, m) amplification procedure *if, and only if, there is an amplifier function $f_{\mathcal{C},n,m}$ that, for every $D \in \mathcal{C}$, can "win" the game with probability at least $2/3$; namely, such that for every $D \in \mathcal{C}$ and valid verifier v_D for D*

$$\Pr_{X_n \leftarrow D^n}[v_D(f_{\mathcal{C},n,m}(X_n)) = ACCEPT] \geq 2/3,$$

where the probability is with respect to the randomness of the choice of the n samples, X_n, and any internal randomness in the amplifier and verifier, f and v.

As was the case in Definition 1, in the above definition it is essential that the verifier only observes the output $f(X_n)$ produced by the amplifier. If the verifier sees both the amplified samples, $f(X_n)$ in addition to the original data, X_n, then the above definition also becomes equivalent to asking how well the class of distributions in question can be *learned* given n samples.

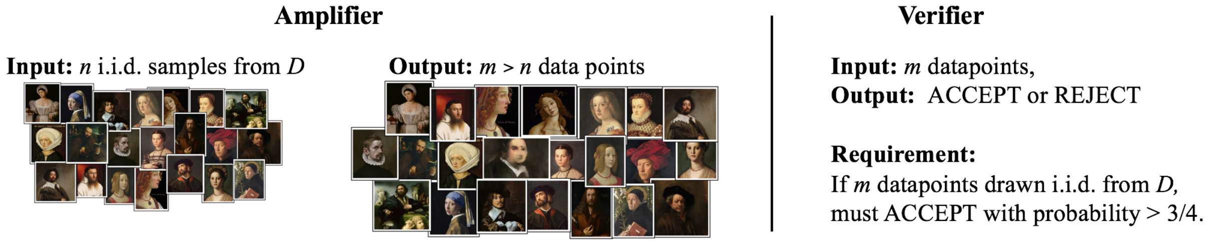
Amplifier
Input: n i.i.d. samples from D
Output: m > n data points
Verifier
Input: m datapoints,
Output: ACCEPT or REJECT
Requirement:
If m datapoints drawn i.i.d. from D,
must ACCEPT with probability > 3/4.

Figure 1: Sample amplification can be viewed as a game between an "amplifier" that obtains n independent draws from an unknown distribution D and must output a set of $m > n$ samples, and a "verifier" that receives the m samples and must ACCEPT or REJECT. The verifier knows the true distribution D and is computationally unbounded but does not know the amplifier's training set (the set of n input samples). An amplification scheme is successful if, for every verifier, with probability at least $2/3$ the verifier will accept the output of the amplifier. [In the setting illustrated above, observant readers might recognize that one of the images in the "Output" set is a painting which was sold in October, 2018 for over $400k by Christie's auction house, and which was "painted" by a Generative Adversarial Network (GAN) [10]].

Example 1. *Consider the class of distributions $\mathcal{C}$ corresponding to i.i.d. flips of a coin with unknown bias p. We claim that there are constants $c' \geq c > 0$ such that $(n, n + cn)$ sample amplification is possible, but $(n, n + c'n)$ amplification is not possible. To see this, consider the amplification strategy corresponding to returning a random permutation of the original samples together with cn additional tosses of a coin with bias $\hat{p}$, where $\hat{p}$ is the empirical bias of the n original samples. Because of the random permutation, the total variation distance between these samples and $n + cn$ i.i.d. tosses of the p-biased coin is a function of only the distribution of the total number of heads. Hence this is equivalent to the distance between $Binomial(n + cn, p)$, and the distribution corresponding to first drawing $h \leftarrow Binomial(n, p)$, and then returning $h + Binomial(cn, h/n)$. It is not hard to show that the total variation distance between these two can be bounded by any small constant by taking c to be a sufficiently small constant. Intuitively, this is because both distributions have the same mean, they are both unimodal, and have variances that differ by a small constant factor for small constant c. For the lower bound, to see that amplification by more than a constant factor is impossible, note that if it were possible, then one could learn p to error $o(1/\sqrt{n})$, with small constant probability of failure, by first amplifying the original samples and then returning the empirical estimate of p based on the amplified samples.*

In the setting, this constant factor amplification is not surprising, since the amplifier can learn the distribution to non-trivial accuracy. It is worth observing, however, that the above amplification scheme corresponding to a $(n, n+1)$ amplifier will return a set of $n+1$ samples, whose total variation distance from n i.i.d. samples is only $O(1/n)$; this is despite the fact that the amplifier can only learn the distribution to total variation distance $\Theta(1/\sqrt{n})$.

1.2 Summary of Results

Our main results provide tight bounds on the extent to which sample amplification is possible for two fundamental settings, unstructured discrete distributions, and d-dimensional Gaussians with unknown mean and fixed covariance. Our first result is for discrete distributions with support size at most k. In this case, we show that sample amplification is possible given only $O(\sqrt{k})$ samples from the distribution, and tightly characterize the extent to which amplification is possible. Note that learning the distribution to small total variation distance requires $\Theta(k)$ samples in this case.

Theorem 1. *Let $\mathcal{C}$ denote the class of discrete distributions with support size at most k. For sufficiently large k, and $m = n + O\left(\frac{n}{\sqrt{k}}\right)$, $\mathcal{C}$ admits an (n, m) amplification procedure.*

This bound is tight up to constants, i.e., there is a constant c, such that for every sufficiently large k, $\mathcal{C}$ does not admit an $\left(n, n + \frac{cn}{\sqrt{k}}\right)$ amplification procedure.

Our amplification procedure for discrete distributions is extremely simple: roughly, we generate additional samples from the empirical distribution of the initial set of n samples, and then randomly shuffle together the original and the new samples. For technical reasons, we do not exactly sample from the empirical distribution but from a suitable modification which facilitates the analysis.

Our second result concerns d-dimensional Gaussian distributions with unknown mean and fixed covariance. We show that we can amplify even with only $O(\sqrt{d})$ samples from the distribution. In contrast, learning the distribution to small constant total variation distance requires $\Theta(d)$ samples. As in the discrete case, we also tightly characterize the extent to which amplification is possible in this setting.

Theorem 2. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N(\mu, \Sigma)$ with unknown mean μ and fixed covariance Σ. For all $d, n > 0$ and $m = n + O\left(\frac{n}{\sqrt{d}}\right)$, $\mathcal{C}$ admits an (n, m) amplification procedure.*

This bound is tight up to constants, i.e., there is a fixed constant c such that for all $d, n > 0$, $\mathcal{C}$ does not admit an (n, m) amplification procedure for $m \geq n + \frac{cn}{\sqrt{d}}$.

The amplification algorithm in the Gaussian case first computes the empirical mean $\hat{\mu}$ of the original set X_n, and then draws $m - n$ new samples from $N(\hat{\mu}, \Sigma)$. We then shift the original n samples to "decorrelate" the original set and the new samples; intuitively, this step hides the fact that the $m - n$ new samples were generated based on the empirical mean of the original samples. The final set of returned samples consists of the shifted versions of the n original samples along with the $m - n$ freshly generated ones.

In contrast to the amplification procedure in the discrete setting, where the final set of returned samples contains the (unmodified) original set, in this Gaussian setting, none of the returned samples are contained in the original set of n samples. A natural question is whether this is necessary: *Does there exist a procedure which achieves optimal amplification in the Gaussian setting, that returns a superset of the original samples?* We show a lower bound proving that, for $n = o(d/\log d)$ there is no $(n, n+1)$ amplification procedure which always returns a superset of the original n samples. Curiously, however, there is such an amplification procedure in the regime where $n = \Omega(d/\log d)$, even though learning is not possible until $n = \Theta(d)$. Additionally, as n goes from $10\frac{d}{\log d}$ to $1000\frac{d}{\log d}$, such amplification procedures go from being unable to amplify at all, to being able to amplify by nearly $\sqrt{d}$ samples. This is formalized in the following proposition.

Proposition 1. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions with unknown mean μ and covariance Σ. There is an absolute constant, c, such that for sufficiently large d, if $n \leq \frac{cd}{\log d}$, there is no $(n, n+1)$ amplification procedure that always returns a superset of the original n points.*

On the other hand, there is a constant c' such that for any ϵ, for $n = \frac{d}{\epsilon \log d}$, and for sufficiently large d, there is an $\left(n, n + c' n^{\frac{1}{2} - 9\epsilon}\right)$ amplification protocol for $\mathcal{C}$ that returns a superset of the original n samples.

1.3 Open Directions

From a technical perspective, there are a number of natural open directions for future work, including establishing tight worst-case bounds on amplification for other natural distribution classes, such as d dimensional Gaussians with unknown means and covariance. More conceptually, it seems worth getting a broader understanding of the range of potential amplification algorithms, and the settings to which each can be applied.

In the discrete distribution setting, our amplification results are tight (to constant factors) in a worst-case sense, and our amplifier essentially just returns the original n samples, together with additional samples drawn from the empirical distribution of those n samples, and then randomly permutes the order of these datapoints. This begs the question: *In the case of discrete distributions, is there any benefit to considering more sophisticated amplification schemes?* Below we sketch one example motivating a more clever amplification approach.

Example 2. *Consider obtaining n samples corresponding to independent draws from a discrete distribution that puts probability $p \gg 1/n$ on a single domain element, and with probability $1-p$ draws a sample from the uniform distribution over some infinite discrete domain. If $p < 2/3$, then the amplification approach that adds samples from the empirical distribution of the data to the original set of samples, will fail. Indeed, with probability at least $1/3$ it will introduce a second sample of one of the "rare" elements, and such samples can be rejected by the verifier. For this setting, the optimal amplifier would always introduce extra samples corresponding to the element of probability p.*

The above example motivates a more sophisticated amplification strategy for the discrete distribution setting. Approaches such as Good-Turing frequency estimation, or more modern variants of it, adjust the empirical probabilities to more accurately reflect the true probabilities (see e.g. [14, 18, 22]). Indeed, in a setting such as Example 2, based on the fact that only one domain element is observed more than once, it is easy to conclude that the total probability mass of all the elements observed just once, is likely at most $O(1/n)$, which implies that a successful amplification scheme cannot duplicate any of them. While inserting samples from a Good-Turing adjusted empirical distribution will not improve the amplification in a worst-case sense for discrete distributions with a bounded support size, such schemes seem *strictly* better than the more naive schemes we analyze. The following question outlines one intriguing potential avenue for quantifying this, along the lines of the recent work on "instance optimal"/"competitive" distribution testing and estimation (see e.g. [1, 19, 22]):

> *Is there an "instance optimal" amplification scheme, which, for every distribution, D, amplifies as well as could be hoped? Specifically, to what extent is there an amplification scheme which performs nearly as well as a hypothetical optimal scheme that knows distribution D up to relabeling/permuting the domain?*

In a completely different direction, our results showing that non-trivial amplification is possible even in the regime in which learning is not possible, rely on the modeling assumption that the verifier gets no information about the amplifier's training set, X_n (the set of n i.i.d. samples). If this dataset is revealed to the verifier, then the question of amplification is equivalent to learning. This prompts the question about a middle ground, where the verifier has some information about the set X_n, but does not see the entire set; this middle ground also seems the most practically relevant (e.g. how much do I need to know about a GAN's training set to decide whether it actually understands a distribution of images?).

> *How does the power of the amplifier vary depending on how much information the verifier has about X_n? If the verifier is given a uniformly random subsample of X_n of size $n' \ll n$, how does the amount of possible amplification scale with n'?*

Finally, rather than considering how to increase the power of the verifier, as the above question asks, it might also be worth considering the consequences of decreasing either the computational power, or information theoretic power of the verifier.

> *If the verifier, instead of knowing distribution D, receives only a set of independent draws from D, how much more power does this give the amplifier? Alternately, if the verifier is constrained to be an efficiently computable function, does this provide additional power to the amplifier in any natural settings?*

1.4 Related Work

The question of deciding whether a set of samples consists of independent draws from a specified distribution is one of the fundamental problems at the core of distributional property testing. Interest in this problem was sparked by the seminal work of Goldreich and Ron [13], who considered the specific problem of determining whether a set of samples was drawn from a uniform distribution of support size k. This sparked a line of work on the slightly more general problem of "identity testing" whether a set of samples was drawn from a specified distribution, D, versus a distribution with distance at least ϵ from D. This includes the early work [2], and work of Paninski [20] who showed that if D is the uniform distribution over k elements, $n = \Theta(\sqrt{k}/\epsilon^2)$ samples are necessary and sufficient, and the more recent work [23] who gave tight bounds on the sample complexity as a function of the distribution in question (which also established that $n = \Theta(\sqrt{k}/\epsilon^2)$ samples suffice for any distribution supported on at most k elements). While the identity testing problem is clearly related to the amplification problem we consider, these appear to be quite distinct problems. In particular, in the identity testing setting, the main technical challenge is understanding what statistics of a set of i.i.d. samples are capable of distinguishing samples drawn from the prescribed distribution, versus samples drawn from any distribution that is at least ϵ-far from that distribution. In contrast, in the amplification setting, the core of the question is how the amplifier can leverage a set of independent samples from D to generate a larger set of (presumably) non-independent samples that can successfully masquerade as a set of independent samples drawn from D; of course, the catch is that the amplifier must do this in the data regime in which it is impossible for them to learn much about D.

Beyond the specific question of identity testing, there is an enormous body of work on other distributional property testing questions, including the "tolerant" version of identity testing where one wishes to distinguish whether samples were drawn independently from a distribution that is close to a specified distribution, D, versus far from D, as well as the multi-distribution analogs where one obtains two (or more) sets of i.i.d. samples, drawn respectively from unknown distributions D_1, D_2, and wishes to distinguish the case that the two distributions are identical (or close) versus have significant total variation distance (see e.g. [3, 24, 8, 19, 5, 16, 12]). In the majority of these works, the assumption is that given samples consist of independent draws from some fixed distribution, and the common theme in these results is the punchline that such tests can typically be accomplished with far less data than would be required to learn the distribution in question.

Within this line of work on distributional property testing and estimation, there is also a recent thread of work on designing estimators for specific properties (such as entropy, or distance to uniformity), whose performance given n independent draws from the distribution in question is comparable to the expected performance of a naive "plugin" estimator (which returns the property value of the empirical distribution) based on $m > n$ independent draws [22, 27]. The term "data amplification" has been applied to this line of work, although it is a different problem from the one we consider. In particular, we are considering the extent to which the samples can be used to create a larger set of samples; the work on property estimation is asking to what extent one can craft superior estimators whose performance is comparable to the performance that a more basic estimator would achieve with a larger sample size.

The recent work on *sampling correctors* [7] also considers the question of how to produce a "good" set of draws from a given distribution. That work assumes access to independent draws from a distribution, D, which is close to having some desired structural property, such as monotonicity or uniformity, and considers how to "correct" or "improve" those samples to produce a set of samples that appear to have been drawn from a different distribution D' that possesses the desired property (or is closer to possessing the property). Part of that work also considers the question of whether such a protocol requires access to additional randomness.

Our formulation of sample amplification as a game between an amplifier and a verifier, closely resembles the setup for *pseudo-randomness* (see [21] for a relatively recent survey). There, the pseudo-random generator takes a set of n independent fair coin flips, and outputs a longer string of $m > n$ outcomes. The verifier's job is to distinguish the output of the generator from a set of m independent tosses of the fair coin (i.e. truly random bits). In contrast to our setting, in pseudo-randomness, both players know that the distribution in question is the uniform distribution, the catch is that the generator does not have access to randomness, and the verifier is computationally bounded. Beyond the superficial similarity in setup, we are not aware of any deeper connections between our notion of amplification and pseudorandomness.

Finally, it is also worth mentioning the work of Viola on the complexity of sampling from distributions [25]. That work also considers the challenge of generating samples from a specified distribution, though the problem is posed as the computational challenge of producing samples from a specified distribution, given access to uniformly random bits. One of the punchlines of that work is that there are distributions, such as the distribution over pairs (x, y) where x is a uniformly random length-n string, and $y = parity(x)$, where small circuits can sample from the distribution, yet no small circuit can compute $y = parity(x)$ given x. A different way of phrasing that punchline is that there are distributions that are easy to sample from, for which it is much harder to sample from their conditional distributions (e.g. in the parity case, sampling (x, y) given x is hard).

2 Algorithms and Proof Overview

In this section, we describe our algorithms for data amplification for discrete and Gaussian distributions. We also give an intuitive overview of the proofs of both the upper and lower bounds.

2.1 Discrete Distributions with Bounded Support

We begin by providing some intuition for amplification in the discrete distribution setting, by considering the simple case where the distribution in question is a uniform distribution over an unknown support. We then describe how our more general amplification algorithm extends this intuition.

2.1.1 Intuition via the Uniform Distribution

Consider the problem of generating $(n + 1)$ samples from a uniform distribution over k unknown elements, given a set of n samples from the distribution. Suppose $n \ll \sqrt{k}$. Then with high probability, no element appears more than once in a set of $(n + 1)$ samples from Unif$[k]$. Therefore, as the amplifier only knows n elements of the support with n samples, it cannot produce a set of $(n + 1)$ samples such that each element only appears once in the set. Hence, no amplification is possible in this regime. Now consider the case when $n = c\sqrt{k}$ for a large constant c. By the birthday paradox, we now expect some elements to appear more than once, and the number of elements appearing twice has expectation $\approx \frac{c^2}{2}$ and standard deviation $\Theta(c)$. In light of this fact, consider an amplification procedure which takes any element that appears only once in the set X_n, adds an additional copy of this element to the set X_n, and then randomly shuffles these $n + 1$ samples to produce the final set Z_{n+1}. It is easy to verify that the distribution of Z_{n+1} will be close in total variation distance to a set X_{n+1} of $(n + 1)$ i.i.d. samples drawn from the original uniform distribution. Since the standard deviation of the number of elements in X_{n+1} that appear twice is $\Theta(c)$, intuitively, we should be able to amplify by an additional $\Theta(c)$ samples, by taking $\Theta(c)$ elements which appear only once and repeating them, and then randomly permuting these $n + \Theta(c)$ samples. Note that with high probability, most elements only appear once in the set X_n, and hence the previous amplifier is almost equivalent to an amplifier which generates new samples by sampling from the empirical distribution of the original n samples, and then randomly shuffles together the original and new samples. Our amplification procedure for general discrete distributions is based on this sampling-from-empirical procedure.

2.1.2 Algorithm and Upper Bound

To facilitate the analysis, our general amplification procedure, which applies to any discrete distribution D, deviates from the sampling-from-empirical-then-shuffle scheme in two ways. First, we use the "Poissonization" trick and go from working with the multinomial distribution to the Poisson distribution—making the element counts independent for all $\leq k$ elements. And second, instead of generating samples from the empirical distribution and shuffling, we (i) divide the input samples into two sets, (ii) use the first set to estimate the empirical distribution, (iii) generate new samples using this empirical distribution, and (iv) randomly shuffle these new samples with the samples in the second set. More precisely, we simulate two sets X_{N_1} and X_{N_2}, of Poisson$(n/4)$ samples from the distribution D, using the original set X_n of n samples from D. This is straightforward to do, as a Poisson$(n/4)$ random variable is $\leq n/2$ with high probability. We then estimate the probabilities of the elements using the first set X_{N_1}, and use these estimated probabilities to generate $R \approx m - n$ more samples from a Poisson distribution, which are then randomly shuffled with the samples in X_{N_2} to produce Z_{N_2+R}. Then the set of output samples Z_m just consist of the samples in X_{N_1}

concatenated with those in Z_{N_2+R}. This describes the main steps in the procedure, more technical details can be found in the full description in Algorithm 3. We show that this procedure achieves a (n, m) amplifier for sufficiently large k and $m = n + O\left(\frac{n}{\sqrt{k}}\right)$.

To prove this upper bound, first note that the counts of each element in a shuffled set Z_m are a sufficient statistics for the probability of observing Z_m, as the ordering of the elements is uniformly random. Hence we only need to show that the distribution of the counts in the set Z_m is close in total variation distance to the distribution of counts in a set X_m of m elements drawn i.i.d. from D. Note that as the first set X_{N_1} is independent of the second set X_{N_2}, the additional samples added to X_{N_2} are independent of the samples originally in X_{N_2}, which avoids additional dependencies in the analysis. Using this independence, we show a technical lemma that with high probability over the first set X_{N_1}, the KL-divergence between the distribution of the set Z_{N_2+R} and D^{N_2+R} of $N_2 + R$ i.i.d. samples from D is small. Then using Pinsker's inequality, it follows that the total variation distance is also small. The final result then follows by a coupling argument, and showing that the Poissonization steps are successful with high probability.

2.1.3 Lower Bound

We now describe the intuition for showing our lower bound that the class of discrete distributions with support at most k does not admit an (n, m) amplification scheme for $m \geq n + \frac{cn}{\sqrt{k}}$, where c is a fixed constant. For $n \leq \frac{k}{4}$, we show this lower bound for the class of uniform distributions $D = \text{Unif}[k]$ on some unknown k elements. In this case, a verifier can distinguish between true samples from D and a set of amplified samples by counting the number of unique samples in the set. Note that as the support of D is unknown, the number of unique samples in the amplified set is at most the number of unique samples in the original set X_n, unless the amplifier includes samples that are outside the support of D, in which case the verifier will trivially reject this set. The expected number of unique samples in n and m draws from D differs by $\frac{c_1 n}{\sqrt{k}}$, for some fixed constant c_1. We use a Doob martingale and martingale concentration bounds to show that the number of unique samples in n samples from D concentrates within a $\frac{c_2 n}{\sqrt{k}}$ margin of its expectation with high probability, for some fixed constant $c_2 \ll c_1$. This implies that there will be a large gap between the number of unique samples in n and m draws from D. The verifier uses this to distinguish between true samples from D and an amplified set, which cannot have sufficiently many unique samples.

Finally, we show that for $n > \frac{k}{4}$, a $\left(n, n + \frac{c'k}{\sqrt{k}}\right)$ amplification procedure for discrete distributions on k elements implies a $(\frac{k}{4}, \frac{k}{4} + c'\sqrt{k})$ amplification procedure for the uniform distribution on $(k-1)$ elements, and for sufficiently large c' this is a contradiction to the previous part. This reduction follows by considering the distribution which has $1 - \frac{k}{4n}$ mass on one element and $\frac{k}{4n}$ mass uniformly distributed on the remaining $(k-1)$ elements. With sufficiently large probability, the number of samples in the uniform section will be $\approx \frac{k}{4}$, and hence we can apply the previous result.

2.2 Gaussian Distributions with Unknown Mean and Fixed Covariance

Given the success of the simple sampling-from-empirical scheme for the discrete case, it is natural to consider the analogous algorithm for d-dimensional Gaussian distributions with unknown mean and fixed covariance. In this section, we first show that this analogous procedure achieves non-trivial amplification for $n = \Omega(d/\log d)$. We then describe the idea behind the lower bound that any procedure which does not modify the input samples does not work for $n = o(d/\log d)$. Inspired by the insights from this lower bound, we then discuss a more sophisticated procedure, which is optimal and achieves non-trivial amplification for n as small as $\Omega(\sqrt{d})$.

2.2.1 Upper Bound for Algorithm which Samples from the Empirical Distribution

Let $\hat{\mu}$ be the empirical mean of the original set X_n. Consider the (n, m) amplification scheme which draws $(m - n)$ new samples from $N(\hat{\mu}, \Sigma)$ and then randomly shuffles together the original samples and the new samples. We show that for any ϵ, this procedure—with a small modification to facilitate the analysis—achieves $\left(n, n + O\left(n^{\frac{1}{2}-9\epsilon}\right)\right)$ amplification for $n = \frac{d}{\epsilon \log d}$. This is despite the empirical distribution $N(\hat{\mu}, \Sigma)$ being $1 - o(1)$ far in total variation distance from the true distribution $N(\mu, \Sigma)$, for $n = o(d)$.

We now provide the proof intuition for this result. First, note that it is sufficient to prove the result for $\Sigma = I$. This is because all the operations performed by our amplification procedure are invariant under linear transformations. The intuition for the result in the identity covariance case is as follows. Consider $n = \Theta(d/\log d)$. In this case, with high probability, the empirical mean $\hat{\mu}$ satisfies $\|\mu - \hat{\mu}\| = O(\sqrt{\log d}) \leq \sqrt{c \log n}$ for a fixed constant c. If we center and rotate the coordinate system, such that $\hat{\mu}$ has the coordinates $(\|\mu - \hat{\mu}\|, 0, \ldots, 0)$, then the distribution of samples from $N(\hat{\mu}, I)$ and $N(\mu, I)$ only differs along the first axis, and is independent across different axes. Hence, with some technical work, our problem reduces to the following univariate problem: what is the total variation distance between $(n+1)$ samples from the univariate distributions $N(0,1)$ and $\tilde{D}$, where $\tilde{D}$ is a mixture distribution where each sample is drawn from $N(0,1)$ with probability $1 - \frac{1}{n+1}$ and from $N(\sqrt{c \log n}, 1)$ with probability $\frac{1}{n+1}$? We show that the total variation distance between these distributions is small, by bounding the squared Hellinger distance between them. Intuitively, the reason for the total variation distance being small is that, even though one sample from $N(\sqrt{c \log n}, 1)$ is easy to distinguish from one sample from $N(0,1)$, for sufficiently small c it is difficult to distinguish between these two samples in the presence of n other samples from $N(0,1)$. This is because for n draws from $N(0,1)$, with high probability there are $O(n^{1-c})$ samples in a constant length interval around $\sqrt{c \log n}$, and hence it is difficult to detect the presence or absence of one extra sample in this interval.

2.2.2 Lower Bound for any Procedure which Returns a Superset of the Input Samples

We show that procedures which return a superset of the input samples are inherently limited in this Gaussian setting, in the sense that they cannot achieve $(n, n+1)$ amplification for $n \leq \frac{cd}{\log d}$, where c is a fixed constant.

The idea behind the lower bound is as follows. If we consider any arbitrary direction and project a true sample from $N(\mu, I)$ along that direction, then with high probability, the projection lies close to the projection of the mean. However, for input set X_n with mean $\hat{\mu}$, the projection of an extra sample added by any amplification procedure along the direction $\mu - \hat{\mu}$ will be far from the projection of the mean μ. This is because after seeing just $\frac{cd}{\log d}$ samples, any amplification procedure will have high uncertainty about the location of μ relative to $\hat{\mu}$. Based on this, we construct a verifier which can distinguish between a set of true samples and a set of amplified samples, for $n \leq \frac{cd}{\log d}$.

We now explain this more formally. Let x'_i be the i-th sample returned by the procedure, and let $\hat{\mu}_{-i}$ be the mean of all except the i-th sample. Let "new" be the index of the additional point added by the amplifier to the original set X_n, hence the amplifier returns the set $\{x'_{\text{new}}, X_n\}$. Note that $\hat{\mu} \leftarrow N(\mu, \frac{I}{n})$, hence $\|\mu - \hat{\mu}\|^2 \approx \frac{d}{n}$ with high probability. Suppose the verifier evaluates the following inner product for the additional point x'_{new},

$$\langle x'_{\text{new}} - \hat{\mu}_{-\text{new}}, \mu - \hat{\mu}_{-\text{new}} \rangle. \tag{1}$$

Note that $\hat{\mu}_{-\text{new}} = \hat{\mu}$ as the amplifier has not modified any of the original samples in X_n. For a point x'_{new} drawn from $N(\mu, I)$, this inner product concentrates around $\|\mu - \hat{\mu}\|^2 \approx \frac{d}{n}$. We now argue that if the true mean μ is drawn from the distribution $N(0, \sqrt{d} I)$, then the above inner product is much smaller than $\frac{d}{n}$ with high probability over μ. The reason for this is as follows. After seeing the samples in X_n, the amplification algorithm knows that μ lies in a ball of radius $\approx \sqrt{\frac{d}{n}}$ centered at $\hat{\mu}$, but μ could lie along any direction in that ball. Formally, we can show that if μ is drawn from the distribution $N(0, \sqrt{d} I)$, then the posterior distribution of $\mu \mid X_n$ is a Gaussian $N(\bar{\mu}, \bar{\sigma} I)$ with $\bar{\mu} \approx \hat{\mu}$ and $\bar{\sigma} \approx \frac{1}{n}$. As $\mu - \hat{\mu}$ is a random direction, for any x'_{new} that the algorithm returns, the inner product in (1) is $\approx \|x'_{\text{new}} - \hat{\mu}\| \|\mu - \hat{\mu}\| \left(\frac{1}{\sqrt{d}}\right)$ with high probability over the randomness in $\mu \mid X_n$. The verifier checks and ensures that $\|x'_{\text{new}} - \hat{\mu}_{-\text{new}}\| = \|x'_{\text{new}} - \hat{\mu}\| \approx \sqrt{d}$. Hence for any $(n, n+1)$ amplification scheme, the inner product in (1) is at most $\approx \sqrt{\frac{d}{n}}$ with high probability over $\mu \mid X_n$. In contrast, we argued before that this inner product is $\approx \frac{d}{n}$ for a true sample from $N(\mu, I)$.

Finally, note that the algorithm can randomly shuffle the samples, and hence the verifier needs to do the above inner product test for every returned sample x'_i, for a total of $(n+1)$ tests. If $(n+1)$ tests are performed, then the inner product is expected to deviate by $\sqrt{\frac{d \log n}{n}}$ around its expected value of $\frac{d}{n}$, even for $(n+1)$ true samples drawn for the distribution. But if $n \ll \frac{d}{\log d}$, then $\sqrt{\frac{d}{n}} \ll \frac{d}{n} - \sqrt{\frac{d \log n}{n}}$, and hence any

$(n, n+1)$ amplification scheme in this regime fails at least one of the following tests with high probability over μ:

1. $\forall\, i \in [n+1], \langle x'_i - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i}\rangle \geq \frac{d}{n} - \sqrt{\frac{d \log n}{n}}$,
2. $\forall\, i \in [n+1], \|x'_i - \hat{\mu}_{-i}\| \approx \sqrt{d}$.

As true samples pass all the tests with high probability, this shows that $(n, n+1)$ amplification without modifying the provided samples is impossible for $n \ll \frac{d}{\log d}$.

2.2.3 Optimal Amplification Procedure for Gaussians: Algorithm and Upper Bound

The above lower bound shows that it is necessary to modify the input samples X_n to achieve amplification for $n = o(d/\log d)$. It also shows what a modification to cross the threshold must achieve—the inner product in (1) should be driven towards its expected value of $\frac{d}{n}$ for a true sample drawn from the distribution. Note that the inner product is too small for the algorithm which samples from the empirical distribution $N(\hat{\mu}, I)$ as the generated point x'_{new} is too correlated with the mean $\hat{\mu}_{-\text{new}} = \hat{\mu}$ of the remaining points. We can fix this by shifting the original points in X_n themselves, to hide the correlation between x'_{new} and the original mean $\hat{\mu}$ of X_n. The full procedure is quite simple to state, and is described in Algorithm 1. Note that unlike our other amplification procedures, this procedure does not involve any random shuffling of the samples. We show that this procedure achieves (n, m) amplification for all $d > 0$ and $m = n + O\left(\frac{n}{\sqrt{d}}\right)$.

Algorithm 1 Sample Amplification for Gaussian with Unknown Mean and Fixed Covariance

Input: $X_n = (x_1, x_2, \dots, x_n)$, where $x_i \leftarrow N(\mu, \Sigma_{d\times d})$.
Output: $Z_m = (x'_1, x'_2, \dots, x'_m)$, such that $D_{TV}(D^m, Z_m) \leq \frac{1}{3}$, where D is $N(\mu, \Sigma_{d\times d})$

```
procedure AmplifyGaussian($X_n$)
    $\hat{\mu} := \sum_{i=1}^{n} \frac{x_i}{n}$
    $\epsilon_i \leftarrow N(0, \Sigma_{d\times d})$, for $i \in \{n+1, n+2, \dots, m\}$        ▷ Draw $m - n$ i.i.d samples from $N(0, \Sigma_{d\times d})$
    $x'_i := \hat{\mu} + \epsilon_i$, for $i \in \{n+1, n+2, \dots, m\}$
    $x'_i := x_i - \sum_{j=n+1}^{m} \frac{\epsilon_j}{n}$, for $i \in \{1, 2, \dots, n\}$        ▷ Remove correlations between old and new samples
    return $Z_m := (x'_1, x'_2, \dots, x'_m)$
```

We now provide a brief proof sketch for this upper bound, for the case when $m = n+1$. First, we show that the returned samples in Z_m can also be thought of as a single sample from a $(m \times d)$-dimensional Gaussian distribution $N\Big(\underbrace{(\mu, \mu, \dots, \mu)}_{m \text{ times}}, \tilde{\Sigma}_{md\times md}\Big)$, as the returned samples are linear combinations of Gaussian random variables. Hence, it is sufficient to find their mean and covariance, and use that to bound their total variation distance to true samples from the distribution (which can also be though of as a single sample from a $(d \times m)$-dimensional Gaussian distribution $N\Big(\underbrace{(\mu, \mu, \dots, \mu)}_{m \text{ times}}, I_{md\times md}\Big)$). Therefore, our problem reduces to ensuring that the total variation distance between these two Gaussian distributions is small, and this distance is proportional to $\|\tilde{\Sigma}_{md\times md} - I_{md\times md}\|_{\text{F}}$. Our modification procedure removes the correlations between the original samples and the generated samples to ensure that the non-diagonal entries of $\tilde{\Sigma}_{md\times md}$ are small, and hence the total variation distance is also small. For example, the original correlation between the first coordinates of the original sample x_1 and the generated sample x'_{n+1} is too large, but it is easy to verify that the correlation between the first coordinates of the modified sample $x'_1 = x_1 - \frac{x'_{n+1} - \hat{\mu}}{n}$ and the generated sample x'_{n+1} is zero.

2.2.4 General Lower Bound for Gaussians

We show a lower bound that there is no (n, m) amplification procedure for Gaussian distibutions with unknown mean for $m \geq n + \frac{cn}{\sqrt{d}}$, where c is a fixed constant. The idea behind the lower bound is similar to the lower bound for procedures which return a superset of the input samples in Section 2.2.2. We define a verifier such that for $\mu \leftarrow N(0, \sqrt{d}I)$ and $m = n + \frac{cn}{\sqrt{d}}$, m true samples from $N(\mu, I)$ are accepted by the verifier with high probability over the randomness in the samples, but m samples generated by any (n, m) amplification scheme are rejected by the verifier with high probability over the randomness in the samples and μ. As before, let x'_i be the i-th sample returned by the procedure, and let $\hat{\mu}_{-i}$ be the mean of all except the i-th sample. Our verifier for this setting evaluates the inner product and the distance of the sample from the mean as in Section 2.2.2, and in addition, also checks $\|\mu - \hat{\mu}_{-m}\|^2$. In total, it evaluates the following,

1. $\langle x'_m - \hat{\mu}_{-m}, \mu - \hat{\mu}_{-m} \rangle$,
2. $\|x'_m - \hat{\mu}_{-m}\|$,
3. $\|\mu - \hat{\mu}_{-m}\|$.

Note that unlike Section 2.2.2, we show that the verifier only needs to do the above tests for any one index i—chosen to be m above—instead of $\forall\, i \in [m]$. We now explain why these tests are sufficient to prove the lower bound. Note that for m true samples drawn from $N(\mu, I)$, $\|\mu - \hat{\mu}_{-m}\|^2 \approx \frac{d}{m-1}$. Also, the squared distance $\|\mu - \hat{\mu}^2\|$ of the mean $\hat{\mu}$ of the original set X_n from μ is concentrated around $\frac{d}{n}$. Using this, for $m = n + 1 + \frac{cn}{\sqrt{d}}$, we can show that no algorithm can find a $\hat{\mu}_{-m}$ which satisfies $\|\mu - \hat{\mu}_{-m}\|^2 \approx \frac{d}{m-1} \ll \frac{d}{n}$ with decent probability over $\mu \leftarrow N(0, \sqrt{d}I)$. This is because such an algorithm could be used to find the true mean μ with much smaller error than $\frac{d}{n}$, which is not possible with n samples. This argument works for $m = n + 1 + \frac{cn}{\sqrt{d}}$, but that does not rule out $(n, n+1)$ amplification schemes for $n \ll \sqrt{d}$. To show that $(n, n+1)$ amplification is not possible for $n \ll \sqrt{d}$, we use the inner product test along with the test for $\|\mu - \hat{\mu}_{-m}\|^2$ to distinguish between $(n+1)$ true samples from $N(\mu, I)$ and those produced by an $(n, n+1)$ amplification scheme, with high probability over μ and the samples. The analysis is similar to the analysis in Section 2.2.2.

3 Proofs: Gaussian with Unknown Mean and Fixed Covariance

3.1 Upper Bound

In this section, we prove the upper bound in Theorem 2 by showing that Algorithm 1 can be used as a $(n, n + \frac{n}{\sqrt{d}})$ amplification procedure.

First, note that it is sufficient to prove the theorem for the case when input samples come from an identity covariance Gaussian. This is because, for the purpose of analysis we can transform our samples to those coming from indentity covariance Gaussian, as our amplification procedure is invariant to linear transformations to samples. In particular, let f_Σ denote our amplification procedure for samples coming from $N(\mu, \Sigma)$, and, $Y_n = (y_1, y_2, \ldots, y_n)$ denote the random variable corresponding to n samples from $N(\mu, \Sigma)$. Let $X_n = (x_1, x_2, \ldots, x_n)$ denote n samples from $N(\mu, I)$, such that $Y_n = \Sigma^{\frac{1}{2}}(X_n - \mu) + \mu = (\Sigma^{\frac{1}{2}}(x_1 - \mu) + \mu, \Sigma^{\frac{1}{2}}(x_2 - \mu) + \mu, \ldots, \Sigma^{\frac{1}{2}}(x_n - \mu) + \mu)$. Due to invariance of our amplification procedure to linear transformations, we get that $\Sigma^{\frac{1}{2}}(f_I(X_n) - \mu) + \mu$ is equal in distribution to $f_\Sigma(\Sigma^{\frac{1}{2}}(X_n - \mu) + \mu) = f_\Sigma(Y_n)$. This gives us

$$\begin{aligned} D_{TV}(f_\Sigma(Y_n), Y_m) &= D_{TV}(f_\Sigma(\Sigma^{\frac{1}{2}}(X_n - \mu) + \mu), \Sigma^{\frac{1}{2}}(X_m - \mu) + \mu) \\ &= D_{TV}(\Sigma^{\frac{1}{2}}(f_I(X_n) - \mu) + \mu, \Sigma^{\frac{1}{2}}(X_m - \mu) + \mu) \\ &\leq D_{TV}(f_I(X_n), X_m), \end{aligned}$$

where the last inequality is true because the total variation distance between two distributions can't increase if we apply the same transformation to both the distributions. Hence, we can conclude that it is sufficient to prove our results for identity covariance case. This is true for both the amplification procedures for Gaussians that we have discussed. So in this whole section, we will work with identity covariance Gaussian distributions.

Proposition 2. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N(\mu, I)$ with unknown mean μ. For all $d, n > 0$ and $m = n + O\left(\frac{n}{\sqrt{d}}\right)$, $\mathcal{C}$ admits an (n, m) amplification procedure.*

Proof. The amplification procedure consists of two parts. The first uses the provided samples to learn the empirical mean $\hat{\mu}$ and generate $m-n$ new samples from $\mathcal{N}(\hat{\mu}, I)$. The second part adjusts the first n samples to "hide" the correlations that would otherwise arise from using the empirical mean to generate additional samples.

Let $\epsilon_{n+1}, \epsilon_{n+2}, \ldots, \epsilon_m$ be $m-n$ i.i.d. samples generated from $N(0, I)$, and let $\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n}$. The amplification procedure will return $x'_1, \ldots, x'_m$ with:

$$x'_i = \begin{cases} x_i - \frac{\sum_{j=n+1}^m \epsilon_j}{n}, & \text{for } i \in \{1, 2, \ldots, n\} \\ \hat{\mu} + \epsilon_i, & \text{for } i \in \{n+1, n+2, \ldots, m\}. \end{cases} \tag{2}$$

We will show later in this proof that subtracting $\frac{\sum_{j=n+1}^m \epsilon_j}{n}$ will serve to decorrelate the first n samples from the remaining samples.

Let $f_{\mathcal{C},n,m} : S^n \to S^m$ be the random function denoting the map from X_n to Z_m as described above, where $S = \mathbb{R}^d$. We need to show

$$D_{TV}\left(Z_m = f_{\mathcal{C},n,m}\left(X_n\right), X_m\right) \leq 1/3,$$

where X_n and X_m denote n and m independent samples from $N(\mu, I)$ respectively.

For ease of understanding, we first prove this result for univariate case, and then extend it to the general setting.

So consider the setting where $d = 1$. In this case, X_m corresponds to m i.i.d. samples from a gaussian with mean μ, and variance 1. X_m can also be thought of as a single sample from an $m-$dimensional Gaussian $N\Big(\underbrace{(\mu, \mu, \ldots, \mu)}_{m \text{ times}}, I_{m\times m}\Big)$. Now, $f_{\mathcal{C},n,m}$ is a map that takes n i.i.d samples from $N(\mu, 1)$, $m-n$ i.i.d samples (ϵ_i) from $N(0, 1)$, and outputs m samples that are a linear combination of the m input samples. So, $f_{\mathcal{C},n,m}(X_n)$ can be thought of as a $m-$dimensional random variable obtained by applying a linear transformation to a sample drawn from $N\Big(\big(\underbrace{\mu, \mu, \ldots, \mu}_{n \text{ times}}, \underbrace{0, 0, \ldots, 0}_{m-n \text{ times}}\big), I_{m\times m}\Big)$. As a linear transformation applied to gaussian random variable outputs a gaussian random variable, we get that $Z_m = (x'_1, x'_2, \ldots, x'_m)$ is distributed according to $N(\tilde{\mu}, \Sigma_{m\times m})$, where $\tilde{\mu}$ and $\Sigma_{m\times m}$ denote the mean and covariance. Note that $\tilde{\mu} = \underbrace{(\mu, \mu, \ldots, \mu)}_{m \text{ times}}$ as

$$\mathbb{E}[x'_i] = \begin{cases} \mathbb{E}[x_i] - \mathbb{E}\left[\frac{\sum_{j=n+1}^m \epsilon_j}{n}\right] = \mu - 0 = \mu, & \text{for } i \in \{1, 2, \ldots, n\} \\ \mathbb{E}[\hat{\mu}] + \mathbb{E}[\epsilon_i] = \mu + 0 = \mu, & \text{for } i \in \{n+1, n+2, \ldots, m\} \end{cases} \tag{3}$$

Next, we compute the covariance matrix $\Sigma_{m\times m}$.

For $i = j$, and $i \in \{1, 2, \ldots, n\}$, we get

$$\begin{aligned}\Sigma_{ii} &= \mathbb{E}[(x'_i - \mu)^2] \\ &= \mathbb{E}\left[(x_i - \mu)^2\right] + \mathbb{E}\left[\left(\frac{\sum_{j=n+1}^m \epsilon_j}{n}\right)^2\right] \\ &= 1 + \frac{m-n}{n^2}.\end{aligned}$$

For $i = j$, and $i \in \{n+1, n+2, \ldots, n+m\}$, we get

$$\begin{aligned}\Sigma_{ii} &= \mathbb{E}\left[(x'_i - \mu)^2\right] \\ &= \mathbb{E}\left[(\hat{\mu} - \mu)^2\right] + \mathbb{E}\left[\epsilon_i^2\right] \\ &= \frac{1}{n} + 1.\end{aligned}$$

For $i \in \{1, 2, \ldots, n\}, j \in \{n+1, n+2, \ldots, n+m\}$, we get

$$\begin{aligned}
\Sigma_{ij} &= \mathbb{E}\left[\left(x_i' - \mu\right)\left(x_j' - \mu\right)\right] \\
&= \mathbb{E}\left[\left(x_i - \frac{\sum_{k=n+1}^{m} \epsilon_k}{n} - \mu\right)\left(\hat{\mu} + \epsilon_j - \mu\right)\right] \\
&= \mathbb{E}[(x_i - \mu)\left(\hat{\mu} - \mu\right)] - \mathbb{E}\left[\left(\frac{\sum_{k=n+1}^{m} \epsilon_k}{n}\right)\left(\epsilon_j\right)\right] \\
&= \frac{1}{n} - \frac{1}{n} \\
&= 0.
\end{aligned}$$

For $i, j \in \{1, 2, \ldots, n\}, i \neq j$, we get

$$\begin{aligned}
\Sigma_{ij} &= \mathbb{E}\Big[\left(x_i' - \mu\right)\left(x_j' - \mu\right)\Big] \\
&= \mathbb{E}\left[\left(x_i - \frac{\sum_{k=n+1}^{m} \epsilon_k}{n} - \mu\right)\left(x_j - \frac{\sum_{k=n+1}^{m} \epsilon_k}{n} - \mu\right)\right] \\
&= \mathbb{E}\left[(x_i - \mu)(x_j - \mu)\right] + \mathbb{E}\left[\left(\frac{\sum_{k=n+1}^{m} \epsilon_k}{n}\right)^2\right] \\
&= \frac{m-n}{n^2}.
\end{aligned}$$

For $i, j \in \{n+1, n+2, \ldots, m\}, i \neq j$, we get

$$\begin{aligned}
\Sigma_{ij} &= \mathbb{E}[(x_i' - \mu)\left(x_j' - \mu\right)] \\
&= \mathbb{E}[(\hat{\mu} + \epsilon_i - \mu)\left(\hat{\mu} + \epsilon_j - \mu\right)] \\
&= \mathbb{E}\left[\left(\hat{\mu} - \mu\right)^2\right] \\
&= \frac{1}{n}.
\end{aligned}$$

This gives us

$$\Sigma_{m \times m} = \begin{bmatrix}
1 + \frac{m-n}{n^2} & \frac{m-n}{n^2} & \cdots & \frac{m-n}{n^2} & 0 & 0 & \cdots & 0 \\
\frac{m-n}{n^2} & 1 + \frac{m-n}{n^2} & \cdots & \frac{m-n}{n^2} & 0 & 0 & \cdots & 0 \\
\vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \frac{m-n}{n^2} & \vdots & \cdots & \cdots & \vdots \\
\frac{m-n}{n^2} & \cdots & \frac{m-n}{n^2} & 1 + \frac{m-n}{n^2} & 0 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & 0 & 1 + \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\
0 & \cdots & \cdots & 0 & \frac{1}{n} & 1 + \frac{1}{n} & \cdots & \frac{1}{n} \\
\vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \frac{1}{n} \\
0 & \cdots & \cdots & 0 & \frac{1}{n} & \cdots & \frac{1}{n} & 1 + \frac{1}{n}
\end{bmatrix}.$$

Now, finding $D_{TV}\left(Z_m, X_m\right)$ reduces to computing $D_{TV}\left(N\left(\tilde{\mu}, I_{m \times m}\right), N\left(\tilde{\mu}, \Sigma_{m \times m}\right)\right)$. From [11, Theorem 1.1], we know that $D_{TV}\left(N\left(\tilde{\mu}, I_{m \times m}\right), N\left(\tilde{\mu}, \Sigma_{m \times m}\right)\right) \leq \min\left(1, \frac{3}{2}||\Sigma - I||_F\right)$. This gives us

$$
\begin{aligned}
D_{TV}\left(N\left(\tilde{\mu}, I_{m\times m}\right), N\left(\tilde{\mu}, \Sigma_{m\times m}\right)\right) &\leq \min\left(1, \frac{3}{2}||\Sigma - I||_F\right) \\
&\leq \sqrt{\frac{3}{2}\left(\left(\frac{m-n}{n^2}\right)^2 n^2 + \frac{1}{n^2}\left(m-n\right)^2\right)} \\
&= \frac{\sqrt{3}\left(m-n\right)}{n}.
\end{aligned}
\tag{4}
$$

Now, for $d > 1$, by a similar argument as above, X_m can be thought of as d independent samples from the following d distributions: $N\Big(\underbrace{(\mu_1, \mu_1, \ldots, \mu_1)}_{m \text{ times}}, I_{m\times m}\Big), \ldots, N\Big(\underbrace{(\mu_d, \mu_d, \ldots, \mu_d)}_{m \text{ times}}, I_{m\times m}\Big)$. Or equivalently, as a single sample from $N\Big(\big(\underbrace{\mu_1, \mu_1, \ldots, \mu_1}_{m \text{ times}}, \ldots, \underbrace{\mu_d, \mu_d, \ldots, \mu_d}_{m \text{ times}}\big), I_{md\times md}\Big)$. Similarly, Z_m can be thought of as d independent samples from $N\Big(\underbrace{(\mu_i, \mu_i, \ldots, \mu_i)}_{m \text{ times}}, \Sigma_{m\times m}\Big)$, or equivalently, a single sample from $N\Big(\big(\underbrace{\mu_1, \mu_1, \ldots, \mu_1}_{m \text{ times}}, \ldots, \underbrace{\mu_d, \mu_d, \ldots, \mu_d}_{m \text{ times}}\big), \tilde{\Sigma}_{md\times md}\Big)$ where $\tilde{\Sigma}_{md\times md}$ is a block diagonal matrix with block diagonal entries equal to $\Sigma_{m\times m}$ (denoted as Σ in the figure).

$$
\tilde{\Sigma}_{md\times md} = \begin{bmatrix} \Sigma & 0 & \cdots & \cdots & 0 \\ 0 & \Sigma & 0 & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \Sigma \end{bmatrix}.
$$

Similar to (4), we get

$$
\begin{aligned}
&D_{TV}\Big(N\Big(\big(\underbrace{\mu_1, \mu_1, \ldots, \mu_1}_{m \text{ times}}, \ldots, \underbrace{\mu_d, \mu_d, \ldots, \mu_d}_{m \text{ times}}\big), I_{md\times md}\Big), N\Big(\big(\underbrace{\mu_1, \mu_1, \ldots, \mu_1}_{m \text{ times}}, \ldots, \underbrace{\mu_d, \mu_d, \ldots, \mu_d}_{m \text{ times}}\big), \tilde{\Sigma}_{md\times md}\Big)\Big) \\
&\qquad \leq \min\left(1, \frac{3}{2}||\tilde{\Sigma} - I||_F\right) \\
&\qquad \leq \sqrt{d\left(\frac{3}{2}\left(\left(\frac{m-n}{n^2}\right)^2 n^2 + \frac{1}{n^2}\left(m-n\right)^2\right)\right)} \\
&\qquad = \frac{\sqrt{3d}\left(m-n\right)}{n}.
\end{aligned}
$$

If we want the total variation distance to be less than δ, we get $m - n = O\left(\frac{n\delta}{\sqrt{d}}\right)$. Setting $\delta = \frac{1}{3}$, we get $m = n + O\left(\frac{n}{\sqrt{d}}\right)$, which completes the proof. □

3.2 Lower Bound

In this section we prove the lower bound from Theorem 2 and show that it is impossible to amplify beyond $O\left(\frac{n}{\sqrt{d}}\right)$ samples.

Proposition 3. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N\left(\mu, I\right)$ with unknown mean μ. There is a fixed constant c such that for all sufficiently large $d, n > 0$, $\mathcal{C}$ does not admit an (n, m) amplification procedure for $m \geq n + \frac{cn}{\sqrt{d}}$.*

Proof. We prove the theorem in two parts. For a fixed constant C, we show that,

1. For $n < \sqrt{d}/C$, $\mathcal{C}$ does not admit an (n, m) amplification procedure for $m > n$.
2. For $n \geq \sqrt{d}/C$, $\mathcal{C}$ does not admit an (n, m) amplification procedure for $m \geq n + 1 + Cn/\sqrt{d}$.

Note that these together imply the lower bound in Theorem 3 with $c = 2C$. We begin with the proof of the first part of the theorem.

Part 1 of Theorem 3: Note that it is sufficient to prove the theorem for $m = n + 1$, as an amplification procedure for $m > n+1$ implies an amplification procedure for $m = n+1$ by discarding the residual samples. We define a verifier $v(Z_{n+1})$ for the distribution $N(\mu, I)$ which takes as input a set Z_{n+1} of $(n + 1)$ samples $\{x'_i \in \mathbb{R}^d, i \in [n+1]\}$, and a distribution D_μ over μ, such that for a fixed constant C; (i) for all μ, the verifier will accept with probability $1 - 1/e^2$ when given as input a set Z_{n+1} of $(n + 1)$ i.i.d. samples from $N(\mu, I)$, (ii) but will reject any $(n, n + 1)$ amplification procedure for $n < \sqrt{d}/C$ with probability $1 - 1/e^2$, where the probability is with respect to the randomness in $\mu \leftarrow D_\mu$, the set X_n and in any internal randomness of the amplifier. Note that by Definition 2 of an amplification procedure, this implies that there is no $(n, n + 1)$ amplification procedure for $n < \sqrt{d}/C$. We choose D_μ to be $N(0, \sqrt{d}I)$. Let $\hat{\mu}_n$ be the mean of the first n samples returned by the amplification procedure. The verifier performs the following three tests, accepts if all tests pass, and rejects otherwise—

1. $\|x'_{n+1} - \mu\|^2 \leq 15d$.
2. $\left|\|\hat{\mu}_n - \mu\|^2 - d/n\right| \leq 10\sqrt{d}/n$.
3. $\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n\rangle \geq d/(4n)$.

We first show that for a fixed constant C, $n < \sqrt{d}/C$ and any μ, $(n + 1)$ i.i.d. samples from $N(\mu, I)$ pass the above tests with probability $1 - 1/e^2$. We will use the following concentration bounds for a χ^2 random variable Z with d degrees of freedom [15, 26],

$$\Pr\left[Z - d \geq 2\sqrt{dt} + 2t\right] \leq e^{-t},\ \forall\, t > 0, \tag{5}$$

$$\Pr\left[|Z - d| \geq dt\right] \leq 2e^{-dt^2/8},\ \forall\, t \in (0, 1). \tag{6}$$

As $\|x'_{n+1} - \mu\|^2$ is a χ^2 random variable with d degrees of freedom, by using (5) and setting $t = 3d$, a true sample x'_{n+1} passes the first test with failure probability e^{-3d}. Note that $\hat{\mu}_n \leftarrow N(\mu, \frac{I}{n})$. Hence by using (6) and setting $t = 10/\sqrt{d}$,

$$\Pr\left[\left|\|\hat{\mu}_n - \mu\|^2 - d/n\right| > 10\sqrt{d}/n\right] \leq 1/e^3.$$

Hence $\hat{\mu}_n$ passes the second test with probability $1 - 1/e^3$. Conditioned on $\hat{\mu}_n$ passing the second test, we show that x'_{n+1} passes the third test with probability $1 - 1/e^3$. If $\hat{\mu}_n$ passes the second test,

$$\|\hat{\mu}_n - \mu\|^2 \geq d/n - 10\sqrt{d}/n \geq d/(2n),$$
$$\|\hat{\mu}_n - \mu\|^2 \leq d/n + 10\sqrt{d}/n \leq 2d/n.$$

Note that as $x'_{n+1} \leftarrow N(\mu, I)$, for a fixed $\hat{\mu}_n$, $\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n\rangle \leftarrow N(\|\hat{\mu}_n - \mu\|^2, \|\hat{\mu}_n - \mu\|^2)$. Hence conditioned on passing the second test, by standard Gaussian tail bounds,

$$\Pr\left[\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n\rangle \leq d/(2n) - 10\sqrt{2d/n}\right] \leq 1/e^3,$$
$$\implies \Pr\left[\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n\rangle \leq d/(4n)\right] \leq 1/e^3,$$

where in the last step $10\sqrt{2d/n} \leq d/(4n)$ as $n < \sqrt{d}/C$. Hence by a union bound, $(n + 1)$ samples drawn from $N(\mu, I)$ will satisfy all 3 tests with failure probability $e^{-3d} + 2/e^3 \leq 1/e^2$. Hence for any μ, the verifier accepts $(n + 1)$ i.i.d. samples from $N(\mu, I)$ with probability $1 - 1/e^2$.

We now show that for μ sampled from $D_\mu = N(0, \sqrt{d}I)$, the verifier rejects any $(n, n+1)$ amplification procedure for $n < \sqrt{d}/C$ with high probability over the randomness in μ. Let $D_{\mu|X_n}$ be the posterior distribution of μ conditioned on the set X_n. We will show that for any set X_n received by the amplifier, the amplified set Z_{n+1} is accepted by the verifier with probability at most $1/e^2$ over $\mu \leftarrow D_{\mu|X_n}$. This implies that with probability $1 - 1/e^2$ over the randomness in $\mu \leftarrow D_\mu$, the set X_n and any internal randomness in the amplifier, the amplifier cannot output a set Z_{n+1} which is accepted by the verifier, completing the proof of the first part of Proposition 3.

To show the above claim, we first find the posterior distribution $D_{\mu|X_n}$ of μ conditioned on the amplifier's set X_n. Let μ_0 be the mean of the set X_n. By standard Bayesian analysis (see, for instance, [17]), the posterior distribution $D_{\mu|X_n} = N(\bar{\mu}, \bar{\sigma}^2 I)$, where,

$$\bar{\mu} = \frac{n}{n + 1/\sqrt{d}}\mu_0, \quad \bar{\sigma}^2 = \frac{1}{n + 1/\sqrt{d}}.$$

We now break into three cases to show that with probability $1 - 1/e^2$ over $\mu \mid X_n$, the set Z_{n+1} cannot satisfy all 3 tests.

Case 1: $\|x'_{n+1} - \bar{\mu}\|^2 \geq 100d$.

We show that the first test is not satisfied with high probability in this case. As $\mu \mid X_n \leftarrow N(\bar{\mu}, \bar{\sigma}^2)$, hence by (5), $\|\mu - \bar{\mu}\|^2 \leq 15d/n$ with probability $1 - e^{-3d}$. Therefore, if $\|x'_{n+1} - \bar{\mu}\|^2 \geq 100d$, then with probability e^{-d},

$$\|x'_{n+1} - \mu\|^2 \geq (\sqrt{100d} - \sqrt{15d/n})^2 > 15d,$$

in which case the first test is not satisfied. Hence in the first case, the amplifier succeeds with probability at most e^{-3d}.

Case 2: $\|\hat{\mu}_n - \bar{\mu}\|^2 \geq 30\sqrt{d}/n$.

We expand $\|\hat{\mu}_n - \mu\|^2$ in the second test as follows,

$$\begin{aligned}\|\hat{\mu}_n - \mu\|^2 &= \|\hat{\mu}_n - \bar{\mu} - (\mu - \bar{\mu})\|^2 \\ &= \|\hat{\mu}_n - \bar{\mu}\|^2 - 2\langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu}\rangle + \|\mu - \bar{\mu}\|^2.\end{aligned}$$

By using (6) and setting $t = 10/\sqrt{d}$, with probability $1 - 1/e^3$,

$$\begin{aligned}\|\mu - \bar{\mu}\|^2 &\geq \frac{d}{n + 1/\sqrt{d}} - \frac{10\sqrt{d}}{n + 1/\sqrt{d}} \\ &\geq \Big(\frac{d}{n}\Big)\Big(1 - \frac{1}{n\sqrt{d}}\Big) - \frac{10\sqrt{d}}{n} \\ &= d/n - \sqrt{d}/n^2 - 10\sqrt{d}/n \\ &\geq d/n - 12\sqrt{d}/n.\end{aligned}$$

As $\mu \mid X_n \leftarrow N(\bar{\mu}, \bar{\sigma}^2)$, $\langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu}\rangle$ is distributed as $N(0, \bar{\sigma}^2\|\hat{\mu}_n - \bar{\mu}\|^2)$. Hence with probability $1 - 1/e^3$, $\langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu}\rangle \leq 10\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n + 1/\sqrt{d}} \leq 10\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n}$. Therefore, with probability $1 - 2/e^3$,

$$\begin{aligned}\|\hat{\mu}_n - \mu\|^2 &\geq \|\hat{\mu}_n - \bar{\mu}\|^2 - (20/\sqrt{n})\|\hat{\mu}_n - \bar{\mu}\| + d/n - 12\sqrt{d}/n, \\ &= \|\hat{\mu}_n - \bar{\mu}\|^2\Big(1 - \frac{20}{\|\hat{\mu}_n - \bar{\mu}\|\sqrt{n}}\Big) + d/n - 12\sqrt{d}/n, \\ &\geq 0.9\|\hat{\mu}_n - \bar{\mu}\|^2 + d/n - 12\sqrt{d}/n, \\ &\geq d/n + 15\sqrt{d}/n,\end{aligned}$$

which implies that the second test is not satisfied. Hence in this case, the amplifier succeeds with probability at most $2/e^3$.

Case 3: $\|x'_{n+1} - \bar{\mu}\|^2 < 100d$ and $\|\hat{\mu}_n - \bar{\mu}\|^2 < 30\sqrt{d}/n$.

We expand $\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n \rangle$ in the third test as follows,

$$\begin{aligned}\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n \rangle &= \langle x'_{n+1} - \bar{\mu}, \mu - \bar{\mu} \rangle - \langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu} \rangle - \langle x'_{n+1} - \bar{\mu}, \hat{\mu}_n - \bar{\mu} \rangle + \|\hat{\mu}_n - \bar{\mu}\|^2, \\ &\leq \langle x'_{n+1} - \bar{\mu}, \mu - \bar{\mu} \rangle - \langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu} \rangle + \|x'_{n+1} - \bar{\mu}\| \|\hat{\mu}_n - \bar{\mu}\| + \|\hat{\mu}_n - \bar{\mu}\|^2.\end{aligned}$$

As in the previous case, $\langle \hat{\mu}_n - \bar{\mu}, \mu - \bar{\mu} \rangle$ is distributed as $N(0, \bar{\sigma}^2 \|\hat{\mu}_n - \bar{\mu}\|^2)$ and hence with probability $1 - 1/e^3$ it is at most $10\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n}$. Similarly, with probability $1 - 1/e^3$, $\langle x'_{n+1} - \bar{\mu}, \mu - \bar{\mu} \rangle$ is at most $10\|x'_{n+1} - \bar{\mu}\|/\sqrt{n}$. Therefore, with probability $1 - 2/e^3$,

$$\begin{aligned}\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n \rangle &\leq 10\|x'_{n+1} - \bar{\mu}\|/\sqrt{n} + 10\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n} + \|x'_{n+1} - \bar{\mu}\| \|\hat{\mu}_n - \bar{\mu}\| + \|\hat{\mu}_n - \bar{\mu}\|^2, \\ &\leq 100\sqrt{\frac{d}{n}} + 100\frac{d^{1/4}}{n} + 60\frac{d^{3/4}}{\sqrt{n}} + 30\frac{\sqrt{d}}{n} \\ &\leq 300\frac{d^{3/4}}{\sqrt{n}} = 300\Big(\frac{d}{n}\Big)\Big(\frac{\sqrt{n}}{d^{1/4}}\Big).\end{aligned}$$

Hence for sufficiently large constant C and $n < \sqrt{d}/C$, with probability $1 - 4/e^3$,

$$\langle x'_{n+1} - \hat{\mu}_n, \mu - \hat{\mu}_n \rangle < \frac{d}{5n},$$

which implies that the third test is not satisfied. Hence the amplifier succeeds in this case with probability at most $2/e^3$.

The overall probability of the amplifier succeeding is the maximum probability of success across the three cases, hence the verifier accepts Z_{n+1} with probability at most $1/e^2$ for $n < \sqrt{d}/C$. This completes the proof of the first part of the theorem.

Part 2 of Proposition 3: In this case, the verifier just uses an analog of the second test of the previous verifier and tests the following,

$$\Big| \|\hat{\mu}_{m-1} - \mu\|^2 - d/(m-1) \Big| \leq 10\sqrt{d}/(m-1). \tag{7}$$

By the same argument as in part 1, m samples drawn i.i.d. from $N(\mu, I)$ will be accepted by the verifier with probability $1 - 1/e^2$. We now show that for $\mu \leftarrow N(0, \sqrt{d}I)$, the verifier will reject the samples produced by a (n, m) amplification scheme for $m = n + 1 + 100n/\sqrt{d}$ with probability $1 - 1/e^2$ over the randomness in μ. As before, this implies that there is no (n, m) amplification scheme for $m = n + 1 + 100n/\sqrt{d}$.

The posterior distribution $D_{\mu|X_n}$ of the mean given the samples X_n is the same as in the previous case. We show that any set Z_m returned by the amplifier fails (7) with probability $1 - 1/e^2$ over the randomness in $\mu \mid X_n$. By the same analysis as in case 2 of the previous part, we get that with probability at least $1 - 2/e^3$,

$$\|\hat{\mu}_{m-1} - \mu\|^2 \geq \|\hat{\mu}_n - \bar{\mu}\|^2 - 20\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n} + d/n - \sqrt{d}/n^2 - 10\sqrt{d}/n.$$

Note that $\|\hat{\mu}_n - \bar{\mu}\|^2 - 20\|\hat{\mu}_n - \bar{\mu}\|/\sqrt{n} \geq -100/n$. Therefore, with probability at least $1 - 2/e^3$,

$$\|\hat{\mu}_{m-1} - \mu\|^2 \geq -100/n + d/n - \sqrt{d}/n^2 - 10\sqrt{d}/n \geq d/n - 20\sqrt{d}/n.$$

To pass (7), $\|\hat{\mu}_{m-1} - \mu\|^2 \leq d/(m-1) + 10\sqrt{d}/(m-1)$. Therefore, if an amplifier passes the test with probability greater than $1 - 2/e^3$ over the randomness in $\mu \mid X_n$ for $m > n + 1 + 100n/\sqrt{d}$, then,

$$\begin{aligned}
&d/n - 20\sqrt{d}/n \leq \|\hat{\mu}_{m-1} - \mu\|^2 \leq d/(m-1) + 10\sqrt{d}/(m-1),\\
\implies &d/n - 20\sqrt{d}/n \leq d/(m-1) + 10\sqrt{d}/(m-1),\\
\implies &d/n - 20\sqrt{d}/n \leq d/(n + 100n/\sqrt{d}) + 10\sqrt{d}/(n + 100n/\sqrt{d}),\\
\implies &d/n - 20\sqrt{d}/n \leq d/n(1 + 100/\sqrt{d})^{-1} + 10\sqrt{d}/n(1 + 100/\sqrt{d})^{-1},\\
\implies &d/n - 20\sqrt{d}/n \leq d/n(1 - 50/\sqrt{d}) + 10\sqrt{d}/n(1 - 50/\sqrt{d}),\\
\implies &- 20\sqrt{d}/n \leq -40\sqrt{d}/n - 1000/n,\\
\implies &- 20\sqrt{d}/n \leq -30\sqrt{d}/n,
\end{aligned}$$

which is a contradiction. Hence for $m > n + 1 + 100n/\sqrt{d}$, every (n, m) amplifier is rejected by the verifier with probability greater than $1 - 1/e^2$ over the randomness in μ, the set X_n, and any internal randomness of the amplifier.

□

3.3 Upper Bound for Procedures which Returns a Superset of the Input Samples

In this section we prove the upper bound in Proposition 1. The algorithm itself is presented in Algorithm 2. Before we proceed with the proof we prove a brief lemma that will be useful for bounding the total variation distance.

Lemma 1. *Let X, Y_1, Y_2 be random variables such that with probability at least $1-\epsilon$ over X, $D_{TV}(Y_1|X, Y_2|X) \leq \epsilon'$, then $D_{TV}((X, Y_1), (X, Y_2)) \leq \epsilon + \epsilon'$.*

Proof. From the definition of total variation distance, we know

$$\begin{aligned}
D_{TV}((X, Y_1), (X, Y_2)) &= \frac{1}{2}\sum_{x,y} |\Pr((X, Y_1) = (x, y)) - \Pr((X, Y_2) = (x, y)))|\\
&= \frac{1}{2}\sum_{x,y} \Pr(X = x) \left|\Pr\left(Y_1 = y \mid X = x\right) - \Pr(Y_2 = y \mid X = x)\right|\\
&= \sum_{x} \Pr(X = x)\ \frac{1}{2}\sum_{y} |\Pr(Y_1 = y \mid X = x) - Pr(Y_2 = y \mid X = x)|\\
&= \sum_{x} \Pr(X = x)\ d_{TV}(Y_1 \mid X = x, Y_2 \mid X = x).
\end{aligned}$$

Since with probability $(1-\epsilon)$ over X, $d_{TV}(Y_1 \mid X, Y_2 \mid X)$ is at most ϵ', and total variation distance is always bounded by 1, we get $\sum_x Pr(X = x)\ d_{TV}(Y_1 \mid X = x, Y_2 \mid X = x) \leq (1-\epsilon)\epsilon' + \epsilon \leq \epsilon' + \epsilon$.

This same proof with summations appropriately replaced with integrals will go through when the random variables in consideration are defined over continuous domains. □

Now we prove the upper bound from Proposition 1. As in Proposition 2, it is sufficient to prove this bound only for the case of identity covariance gaussians as our algorithm in this case is also invariant to linear transformation.

Proposition 4. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N(\mu, I)$ with unknown mean μ. There is a constant c' such that for any ϵ, and $n = \frac{d}{\epsilon \log d}$, and for sufficiently large d, there is an $\left(n, n + c'n^{\frac{1}{2}-9\epsilon}\right)$ amplification protocol for $\mathcal{C}$ that returns a superset of the original n samples.*

Algorithm 2 Sample Amplification for Gaussian with Unknown Mean and Fixed Covariance Without Modifying Input Samples

Input: $X_n = (x_1, x_2, \dots, x_n)$, where $x_i \leftarrow N(\mu, \Sigma_{d\times d})$.
Output: $Z_m = (x'_1, x'_2, \dots, x'_m)$, such that $D_{TV}(D^m, Z_m) \le \frac{1}{3}$, where D is $N(\mu, \Sigma_{d\times d})$

```
procedure AmplifyGaussian2($X_n$)
    $r := m - n$
    $\hat{\mu} := \sum_{i=1}^{\frac{n}{2}} \frac{x_i}{n/2}$
    $x'_i := x_i$, for $i \in \{1, 2, \dots, \frac{n}{2}\}$
    $X_{\text{remaining}} := (x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \dots, x_n)$
    for $i = \frac{n}{2} + 1$ to m do
        $T \leftarrow$ Bernoulli($\frac{2r}{r+n/2}$)        ▷ Set $T = 1$ with probability $\frac{2r}{r+n/2}$, and 0 otherwise
        if $T$ equals 1 then
            $x'_i \leftarrow N(\hat{\mu}, \Sigma_{d\times d})$
        else
            if $X_{\text{remaining}}$ is not empty then
                $x'_i :=$ Random Element Drawn without Replacement from $X_{\text{remaining}}$
            else
                $x'_i := x_1$        ▷ Happens with small probability
    $Z_m := (x'_1, x'_2, \dots, x'_m)$
    return $Z_m$
```

Proof. Let $m = n + r$,where $r = O\left(n^{\frac{1}{2}-9\epsilon}\right)$. We begin by describing the procedure to generate m samples $Z_m = (x'_1, x'_2, \dots, x'_m)$, given n i.i.d. samples $X_n = (x_1, x_2, \dots, x_n)$ drawn from $N(\mu, I)$. Let $\tilde{\mu} = \sum_{i=1}^{n/2} \frac{x_i}{n/2}$ denote the mean of first $\frac{n}{2}$ samples in X_n. For distributions P and Q, let $(1-\alpha)P + \alpha Q$ denote the mixture distribution where $(1-\alpha)$ and α are the respective mixture weights.

We first describe how to generate $Z'_m = (x''_1, x''_2, \dots, x''_m)$, given n i.i.d samples X_n. For $i \in \{1, 2, \dots, \frac{n}{2}\}$, we set $x''_i = x_i$. For $i \in \{\frac{n}{2}+1, \frac{n}{2}+2, \dots, m\}$, we set x''_i to a random independent draw from the mixture distribution $\left(1 - \frac{10r}{r+\frac{n}{2}}\right) N(\mu, I_{d\times d}) + \frac{10r}{r+\frac{n}{2}} N(\tilde{\mu}, I_{d\times d})$.

Now, the construction of Z_m is very similar to Z'_m except that we don't have access to $N(\mu, I_{d\times d})$ to sample points from the mixture distribution. So, for Z_m, set $x'_i = x_i$ for $i \in \{1, 2, \dots, \frac{n}{2}\}$. For $i \in \{\frac{n}{2}+1, \frac{n}{2}+2, \dots, m\}$, we use samples from $(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \dots, x_n)$ instead of producing new samples from $N(\mu, I_{d\times d})$. With probability $\left(1 - \frac{10r}{r+\frac{n}{2}}\right)$, we generate a random sample without replacement from $\left(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \dots, x_n\right)$, and with probability $\frac{10r}{r+\frac{n}{2}}$ we generate a sample from $N(\tilde{\mu}, I)$, and set x'_i equal to that sample. As we sample from $(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \dots, x_n)$ without replacement, we can generate only $\frac{n}{2}$ samples this way. The expected number of samples needed is $(\frac{n}{2} + r)(1 - \frac{10r}{r+\frac{n}{2}}) = \frac{n}{2} - 9r$, and with high probability, we won't need more than $\frac{n}{2}$ samples. If the total number of required samples from $\left(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \dots, x_n\right)$ turns out to be more than $\frac{n}{2}$, we set x_i to an arbitrary $d-$dimensional vector (say x_1) but this happens with low probability, leading to insignificant loss in total variation distance.

Let X_m denote the random variable corresponding to m i.i.d. samples from $N(\mu, I)$. We want to show that $D_{TV}(X_m, Z_m)$ is small. By triangle inequality, $D_{TV}(X_m, Z_m) \le D_{TV}(X_m, Z'_m) + D_{TV}(Z'_m, Z_m)$.

We first bound $D_{TV}(Z_m, Z'_m)$. Let $Y, Y' \leftarrow$ Binomial $\left(r + \frac{n}{2}, 1 - \frac{10r}{r+\frac{n}{2}}\right)$ be random variables that denotes the number of samples from $(1 - \frac{10r}{r+\frac{n}{2}})$ mixture component in Z_m and Z'_m respectively. Let Ω denote the sample space of Z_m and Z'_m.

$$\begin{aligned} D_{TV}(Z_m, Z'_m) &= \max_{E\subseteq\Omega} |\Pr(Z_m \in E) - \Pr(Z'_m \in E)| \\ &= \max_{E\subseteq\Omega} |\Pr\left(Z_m \in E \mid Y \le \frac{n}{2}\right) \Pr\left(Y \le \frac{n}{2}\right) + \Pr\left(Z_m \in E \mid Y > \frac{n}{2}\right) \Pr\left(Y > \frac{n}{2}\right) \\ &\qquad - \Pr\left(Z'_m \in E \mid Y' \le \frac{n}{2}\right) \Pr\left(Y' \le \frac{n}{2}\right) - \Pr\left(Z'_m \in E \mid Y' > \frac{n}{2}\right) \Pr\left(Y' > \frac{n}{2}\right)| \end{aligned}$$

Since Y and Y' have the same distribution, we have $\Pr\left(Y' \le \frac{n}{2}\right) = \Pr\left(Y \le \frac{n}{2}\right)$, and $\Pr\left(Y' > \frac{n}{2}\right) = \Pr\left(Y > \frac{n}{2}\right)$. This gives us

$$\begin{aligned} D_{TV}(Z_m, Z'_m) = \max_{E \subseteq \Omega} |&\Pr\left(Z_m \in E \mid Y \le \frac{n}{2}\right) \Pr\left(Y \le \frac{n}{2}\right) + \Pr\left(Z_m \in E \mid Y > \frac{n}{2}\right) \Pr\left(Y > \frac{n}{2}\right) \\ &- \Pr\left(Z'_m \in E \mid Y' \le \frac{n}{2}\right) \Pr\left(Y \le \frac{n}{2}\right) - \Pr\left(Z'_m \in E \mid Y' > \frac{n}{2}\right) \Pr\left(Y > \frac{n}{2}\right)| \\ \le \max_{E \subseteq \Omega} &\Pr\left(Y \le \frac{n}{2}\right) \mid \Pr\left(Z_m \in E | Y \le \frac{n}{2}\right) - \Pr\left(Z'_m \in E | Y' \le \frac{n}{2}\right) \mid \\ &+ \Pr\left(Y > \frac{n}{2}\right) \mid \Pr\left(Z_m \in E | Y > \frac{n}{2}\right) - \Pr\left(Z'_m \in E | Y' > \frac{n}{2}\right) \mid . \end{aligned}$$

where the last inequality holds because of the triangle inequality. Now, note that $\Pr(Z_m \in E | Y \le \frac{n}{2}) = \Pr(Z'_m \in E | Y' \le \frac{n}{2})$ for all E, and $|\Pr(Z_m \in E | Y > \frac{n}{2}) - \Pr(Z'_m \in E | Y' > \frac{n}{2})| \le 1$. This gives us

$$D_{TV}(Z_m, Z'_m) \le \Pr\left(Y > \frac{n}{2}\right).$$

We know $\mathbb{E}[Y] = \frac{n}{2} - 9r$, and $\mathrm{Var}[Y] = \left(\frac{n}{2} + r\right)\left(1 - \frac{10r}{\frac{n}{2}+r}\right)\left(\frac{10r}{\frac{n}{2}+r}\right) \le 10r$. Using Bernstein's inequality, we get

$$\begin{aligned} \Pr\left[Y > \frac{n}{2}\right] &= \Pr(Y - \mathbb{E}[Y] > 9r) \\ &\le \exp\left(\frac{-(9r)^2}{2(10r + 9r/3)}\right) \\ &\le \exp\left(\frac{-81r}{26}\right). \end{aligned}$$

So we get $D_{TV}(Z_m, Z'_m) \le \exp\left(\frac{-81r}{26}\right)$.

Next, we calculate $D_{TV}(X_m, Z'_m)$. We write $X_m = (X_m^1, X_m^2)$ and $Z'_m = (Z_m^{1'}, Z_m^{2'})$ where X_m^1 and $Z_m^{1'}$ denote the first $\frac{n}{2}$ samples of X_m and Z'_m , and X_m^2 and $Z_m^{2'}$ denote rest of their samples. Since X_m^1 and $Z_m^{1'}$ are drawn from the same distribution, $\Pi_{i=1}^{\frac{n}{2}} N(\mu, I)$, and $Z_m^{1'}, X_m^1, X_m^2$ are independent, we get $(Z_m^{1'}, X_m^2)$ and (X_m^1, X_m^2) are equal in distribution. This gives us

$$D_{TV}(X_m, Z'_m) = D_{TV}((X_m^1, X_m^2), (Z_m^{1'}, Z_m^{2'})) = D_{TV}((Z_m^{1'}, X_m^2), (Z_m^{1'}, Z_m^{2'})).$$

From Lemma 1, we know that, if with probability at least $1 - \epsilon_1$ over $Z_m^{1'}$, $D_{TV}(X_m^2 | Z_m^{1'}, Z_m^{2'} | Z_m^{1'}) \le \epsilon_2$, then $D_{TV}((Z_m^{1'}, X_m^2), (Z_m^{1'}, Z_m^{2'})) \le \epsilon_1 + \epsilon_2$. Here, $Z_m^{1'}$ and X_m^2 are independent, and the only dependency between $Z_m^{1'}$ and $Z_m^{2'}$ is via the mean $\tilde{\mu}$ of the elements of $Z_m^{1'}$. So $D_{TV}(X_m^2 | Z_m^{1'}, Z_m^{2'} | Z_m^{1'}) = D_{TV}(X_m^2, Z_m^{2'} | \tilde{\mu})$. We will show that with high probability over $\tilde{\mu}$, this total variation distance is small.

We first estimate $\|\tilde{\mu} - \mu\|$. Note that $\mathbb{E}_{Z_m^{1'}}[\|\tilde{\mu} - \mu\|^2] = \frac{2d}{n}$, and $\frac{n}{2}\|\tilde{\mu} - \mu\|^2$ is a χ^2 random variable with d degrees of freedom. To bound the deviation of $\|\tilde{\mu} - \mu\|^2$ around it's mean, we will use the following concentration bound for a χ^2 random variable R with d degrees of freedom [26, Example 2.5].

$$\Pr[|R - d| \ge dt] \le 2e^{-dt^2/8}, \text{ for all } t \in (0, 1).$$

This gives us $\Pr(|\frac{n}{2}\|\tilde{\mu} - \mu\|^2 - d| \ge 0.5d) \le 2e^{-d/32}$, that is, $\|\tilde{\mu} - \mu\| \le \sqrt{\frac{3d}{n}} \le \sqrt{3\epsilon \log d}$ with probability at least $1 - 2e^{-d/32}$.

X_m^2 is distributed as the product of $\frac{n}{2} + r$ gaussiaus $\Pi_{i=1}^{\frac{n}{2}+r} N(\mu, I_{d\times d})$ and $Z_m^{2'} | \tilde{\mu}$ is distributed as the product of $\frac{n}{2} + r$ mixture distributions $\Pi_{i=1}^{\frac{n}{2}+r}(1 - \frac{10r}{\frac{n}{2}+r})N(\mu, I_{d\times d}) + \frac{10r}{\frac{n}{2}+r}N(\tilde{\mu}, I_{d\times d})$. We evaluate the total variation distance between these two distributions by bounding their squared Hellinger distance, since squared Hellinger distance is easy to bound for product distributions and is within a quadratic factor of the total variation distance for any distribution. By the subadditivity of the squared Hellinger distance, we get

$$\begin{aligned} &H\left(\Pi_{i=1}^{\frac{n}{2}+r} N(\mu, I_{d\times d}), \Pi_{i=1}^{\frac{n}{2}+r}\left(1-\frac{10r}{\frac{n}{2}+r}\right) N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r} N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2 \\ &\le \left(\frac{n}{2}+r\right)\; H\left(N\left(\mu, I_{d\times d}\right), \left(1-\frac{10r}{\frac{n}{2}+r}\right) N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r} N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2 . \end{aligned} \tag{8}$$

For sufficiently large d, r and n satisfy $r \le \frac{n}{18}$, so we can use Lemma 2 to get

$$\begin{aligned} H\left(N(\mu, I_{d\times d}), \left(1-\frac{10r}{\frac{n}{2}+r}\right) N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r} N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2 &\le \frac{576r^2}{n^2} e^{3\|\tilde{\mu}-\mu\|^2} \\ &\le \frac{576r^2 d^{9\epsilon}}{n^2}, \end{aligned} \tag{9}$$

with probability at least $1-2e^{-d/32}$ over $\tilde{\mu}$. From (8) and (9), we get that with probability at least $1-2e^{-d/32}$ over $\tilde{\mu}$,

$$H\left(\Pi_{i=1}^{\frac{n}{2}+r} N(\mu, I_{d\times d}), \Pi_{i=1}^{\frac{n}{2}+r}\left(1-\frac{10r}{\frac{n}{2}+r}\right) N(\mu, I_{d\times d}) + \frac{10r}{\frac{n}{2}+r} N(\tilde{\mu}, I_{d\times d})\right)^2 \le (\frac{n}{2}+r)\frac{576r^2 d^{9\epsilon}}{n^2} \le \frac{576r^2 d^{9\epsilon}}{n},$$

where the last inequality holds because $r < \frac{n}{2}$. As the total variation distance between two distributions is upper bounded by $\sqrt{2}$ times their Hellinger distance, we get that with probability at least $1-2e^{-d/32}$ over $\tilde{\mu}$,

$$\begin{aligned} &D_{TV}\left(\Pi_{i=1}^{\frac{n}{2}+r} N(\mu, I_{d\times d}), \Pi_{i=1}^{\frac{n}{2}+r}\left(1-\frac{10r}{\frac{n}{2}+r}\right) N(\mu, I_{d\times d}) + \frac{10r}{\frac{n}{2}+r} N(\tilde{\mu}, I_{d\times d})\right) \\ &\le \frac{24\sqrt{2}rd^{9\epsilon/2}}{\sqrt{n}} \le \frac{24\sqrt{2}rn^{9\epsilon}}{\sqrt{n}}, \end{aligned}$$

where the last inequality is true because $n > \sqrt{d}$.

Now, from Lemma 1, we know that if with probability at least $1-\epsilon_1$ over $Z_m^{1'}$, $D_{TV}(X_m^2|Z_m^{1'}, Z_m^{2'}|Z_m^{1'}) \le \epsilon_2$, then $D_{TV}((Z_m^{1'}, X_m^2), (Z_m^{1'}, Z_m^{2'})) \le \epsilon_1 + \epsilon_2$. In this case, $\epsilon_1 = 2e^{-d/32}$ and $\epsilon_2 = \frac{24\sqrt{2}rn^{9\epsilon}}{\sqrt{n}}$, so we get $D_{TV}((Z_m^{1'}, X_m^2), (Z_m^{1'}, Z_m^{2'})) = D_{TV}(X_m, Z_m') \le 2e^{-d/32} + \frac{24\sqrt{2}rn^{9\epsilon}}{\sqrt{n}}$. We also know that $D_{TV}(Z_m, Z_m') \le e^{-81r/26}$. Using triangle inequality, we get

$$D_{TV}(X_m, Z_m) \le 2e^{-d/32} + \frac{24\sqrt{2}rn^{9\epsilon}}{\sqrt{n}} + e^{-81r/26}.$$

For $\delta > 2(2e^{-d/32} + e^{-81r/26})$, and for $r \le \frac{n^{\frac{1}{2}-9\epsilon}\delta}{48\sqrt{2}}$, we get $D_{TV}(X_m, Z_m) \le \delta$. For d large enough, setting $\delta = \frac{1}{3}$ and $r \le \frac{n^{\frac{1}{2}-9\epsilon}}{144\sqrt{2}}$, we get the desired result. Note that we haven't tried to optimize the constants in this proof.

Lemma 2. *Let $P = N(0, I_{d\times d})$ and $Q = N(\hat{\mu}, I_{d\times d})$ be d-dimensional gaussian distributions. For $r \le \frac{n}{18}$, $H\left(P, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P + \frac{10r}{r+\frac{n}{2}}Q\right) \le \frac{24r}{n}e^{\frac{3\|\hat{\mu}\|^2}{2}}$.*

Proof. We work in the rotated basis where $Q = N((\|\hat{\mu}\|, \underbrace{0, 0, \ldots, 0}_{d-1 \text{ times}}), I_{d\times d})$ and $P = N(0, I_{d\times d})$. Let $P_1 = N(0,1)$ and $Q_1 = N(\|\hat{\mu}\|, 1)$ denote the projection of P and Q along the first coordinate axis respectively. Note that the mixture distribution in question is the product of $\left(\left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1 + \frac{10r}{r+\frac{n}{2}}Q_1\right)$ and

$N(0, I_{d-1\times d-1})$, and P is the product of P_1 and $N(0, I_{d-1\times d-1})$. Since the squared Hellinger distance is subadditive for product distributions, we get,

$$\begin{aligned} H\left(P, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P+\frac{10r}{r+\frac{n}{2}}Q\right)^2 &\leq H\left(P_1, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)^2 + H(N(0,I_{d-1\times d-1}), N(0,I_{d-1\times d-1}))^2 \\ &= H\left(P_1, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)^2. \end{aligned}$$

Therefore, to bound the required Hellinger distance, we just need to bound $H\left(P_1, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)$. Let p_1 and q_1 denote the probability densities of P_1 and $\left(\left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)$ respectively. We get $H\left(P_1, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)^2 = \int_{-\infty}^{\infty}\left(\sqrt{p_1}-\sqrt{q_1}\right)^2 dx$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\left(\sqrt{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} - \sqrt{\left(1-\frac{10r}{r+\frac{n}{2}}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}+\frac{10r}{r+\frac{n}{2}}\frac{1}{\sqrt{2\pi}}e^{-(x-\|\hat{\mu}\|)^2/2}}\right)^2 dx \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1-\sqrt{1-\frac{10r}{r+\frac{n}{2}}+\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}}\right)^2 dx. \end{aligned}$$

We will evaluate this integral as a sum of integral in two regions.

1. From $-\infty$ to $\|\hat{\mu}\|/2$:

   $$\int_{-\infty}^{\|\hat{\mu}\|/2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1-\sqrt{1-\frac{10r}{r+\frac{n}{2}}+\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}}\right)^2 dx \leq \int_{-\infty}^{\|\hat{\mu}\|/2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1-\sqrt{1-\frac{10r}{r+\frac{n}{2}}}\right)^2 dx.$$

   Since $r \leq \frac{n}{18}$, we get $\frac{10r}{r+\frac{n}{2}} \leq 1$. Using $1-y \leq \sqrt{1-y}$ for $0 \leq y \leq 1$, we get

   $$\begin{aligned} \int_{-\infty}^{\|\hat{\mu}\|/2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1-\sqrt{1-\frac{10r}{r+\frac{n}{2}}}\right)^2 dx &\leq \int_{-\infty}^{\|\hat{\mu}\|/2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(\frac{10r}{r+\frac{n}{2}}\right)^2 dx \\ &\leq \frac{400r^2}{n^2}. \end{aligned}$$

2. From $\frac{\|\hat{\mu}\|}{2}$ to ∞, we get $\int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1-\sqrt{1-\frac{10r}{r+\frac{n}{2}}+\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}}\right)^2 dx.$

   $$\leq \int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(\sqrt{1+\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}}-1\right)^2 dx.$$

   This is because $x \geq \|\hat{\mu}\|/2$, and therefore $\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}} \geq \frac{10r}{r+\frac{n}{2}}$. Now, using $\sqrt{1+y} \leq 1+\frac{y}{2}$, we get

   $$\begin{aligned} &\int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(\sqrt{1+\frac{10r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}}-1\right)^2 dx \\ &\leq \int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(1+\frac{5r}{r+\frac{n}{2}}e^{\frac{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}{2}}-1\right)^2 dx \\ &\leq \frac{100r^2}{n^2}\int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\|\hat{\mu}\|^2+2\|\hat{\mu}\|x}e^{-x^2/2}dx \\ &= \frac{100r^2}{n^2}e^{-\|\hat{\mu}\|^2}\int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{2\|\hat{\mu}\|x-x^2/4}e^{-x^2/4}dx. \end{aligned}$$

Since $2\|\hat{\mu}\|x - x^2/4 \leq 4\|\hat{\mu}\|^2$, we get

$$\begin{aligned}\frac{100r^2}{n^2}e^{-\|\hat{\mu}\|^2}\left(\int_{\|\hat{\mu}\|/2}^{\infty} e^{2\|\hat{\mu}\|x-x^2/4}\frac{1}{\sqrt{2\pi}}e^{-x^2/4}\right)dx &\leq \frac{100r^2}{n^2}e^{3\|\hat{\mu}\|^2}\left(\int_{\|\hat{\mu}\|/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/4}\right)dx\\ &\leq \frac{100\sqrt{2}r^2}{n^2}e^{3\|\hat{\mu}\|^2}\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi}}e^{-x^2/4}\right)dx\\ &\leq \frac{100\sqrt{2}r^2}{n^2}e^{3\|\hat{\mu}\|^2}.\end{aligned}$$

Adding the two integrals, we get

$$\begin{aligned}H\left(P_1, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1 + \frac{10r}{r+\frac{n}{2}}Q_1\right)^2 &\leq \frac{400r^2}{n^2} + \frac{100\sqrt{2}r^2}{n^2}e^{3\|\hat{\mu}\|^2}\\ &\leq \frac{576r^2}{n^2}e^{3\|\hat{\mu}\|^2}.\end{aligned}$$

This gives us $H(P, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P + \frac{10r}{r+\frac{n}{2}}Q) \leq \frac{24r}{n}e^{3\|\hat{\mu}\|^2/2}$ which completes the proof. □

□

3.4 Lower Bound for Procedures which Return a Superset of the Input Samples

In this section we prove the lower bound from Proposition 1.

Proposition 5. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N(\mu, I)$ with unknown mean μ. There is an absolute constant, c, such that for sufficiently large d, if $n \leq \frac{cd}{\log d}$, there is no $(n, n+1)$ amplification procedure that always returns a superset of the original n points.*

Proof. The outline of the proof is very similar to the proof of Proposition 3. As in the proof of Proposition 3, we define a verifier $v(Z_{n+1})$ for the distribution $N(\mu, I)$ which takes as input $(n+1)$ samples $\{x_i' \in \mathbb{R}^d, i \in [n+1]\}$, and a distribution D_μ over μ, such that if $n < O(d/\log(d))$; (i) for all μ, the verifier will accept with probability $1 - 1/e^2$ when given as input a set Z_{n+1} of $(n+1)$ i.i.d. samples from $N(\mu, I)$, (ii) but will reject any $(n, n+1)$ amplification procedure which does not modify the input samples with probability $1-1/e^2$, where the probability is with respect to the randomness in $\mu \leftarrow D_\mu$, the set X_n and in any internal randomness of the amplifier. Note that by Definition 2 of an amplification procedure, this implies that there is no $(n, n+1)$ amplification procedure which does not modify the input samples for $n < O(d/\log(d))$. We choose D_μ to be $N(0, \sqrt{d}I)$. Let $\hat{\mu}_{-i}$ be the mean of the all except the i-th sample returned by the amplification procedure. The verifier performs the following tests, and accepts if all tests pass, and rejects otherwise—

1. $\forall\, i \in [n+1], \|x_i' - \mu\|^2 \leq 15d$.
2. $\forall\, i \in [n+1], \langle x_i' - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i}\rangle \geq d/(4n)$.

We first show that for a sufficiently large constant C and $n < O(d/\log(d))$, $(n+1)$ i.i.d. samples from $N(\mu, I)$ pass the above tests with probability at least $1 - 1/e^2$. As $\|x_i' - \mu\|^2$ is a χ^2 random variable with d degrees of freedom, by the concentration bound for a χ^2 random variable (5), a true sample x_i' passes the first test with failure probability e^{-3d}. Hence by a union bound, all samples $\{x_i, i \in [n+1]\}$ pass the first test with probability at least $1 - de^{-3d} \geq 1 - 1/e^3$. Let E denote the following event,

$$\begin{aligned}&\forall\, i \in [n+1], \|\hat{\mu}_n - \mu\|^2 \geq d/n - \sqrt{20d\log d}/n \geq d/(2n),\\ &\forall\, i \in [n+1], \|\hat{\mu}_n - \mu\|^2 \leq d/n + \sqrt{20d\log d}/n \leq 2d/n.\end{aligned}$$

Note that $\hat{\mu}_{-i} \leftarrow N(\mu, \frac{I}{n})$. Hence, by using (6) with $t = 20\sqrt{\frac{\log d}{d}}$, and a union bound over all $i \in [n+1]$,

$$\Pr[E] \geq 1 - 1/e^3.$$

Note that as $x'_i \leftarrow N(\mu, I)$, for a fixed $\hat{\mu}_{-i}$, $\langle x'_i - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i}\rangle \leftarrow N(\|\hat{\mu}_{-i} - \mu\|^2, \|\hat{\mu}_{-i} - \mu\|^2)$. Hence conditioned on E, by standard Gaussian tail bounds,

$$\Pr\left[\langle x'_i - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i}\rangle \leq d/(2n) - \sqrt{20d\log d/n}\right] \leq 1/n^2,$$
$$\implies \Pr\left[\langle x'_i - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i}\rangle \leq d/(4n)\right] \leq 1/n^2,$$

where in the last step we use the fact that $n < \frac{d}{C\log d}$ for a large constant C. Therefore, conditioned on E, $\{x_i, i \in [n+1]\}$ pass the third test with probability at least $1 - 1/e^3$. Hence by a union bound, $(n+1)$ samples drawn from $N(\mu, I)$ will satisfy all 3 tests with failure probability at most $1/e^2$. Hence for any μ, the verifier accepts $n+1$ i.i.d. samples from $N(\mu, I)$ with probability at least $1 - 1/e^2$.

We now show that for $n < \frac{d}{C\log d}$ and μ sampled from $D_\mu = N(0, \sqrt{d}I)$, the verifier rejects any $(n, n+1)$ amplification procedure which does not modify the input samples with high probability over the randomness in μ and the set X_n. Let $D_{\mu|X_n}$ be the posterior distribution of μ conditioned on the set X_n. As in Proposition 3, $D_{\mu|X_n} = N(\bar{\mu}, \bar{\sigma}^2 I)$, where,

$$\bar{\mu} = \frac{n}{n + 1/\sqrt{d}}\mu_0, \quad \bar{\sigma}^2 = \frac{1}{n + 1/\sqrt{d}}.$$

We will show that with probability $1 - e^{-3d}$ over the randomness in the set X_n received by the amplifier and with probability $1 - 1/e^2$ over $\mu \leftarrow D_{\mu|X_n}$ and any internal randomness of the amplifier, the amplifier cannot output a set Z_{n+1} which contains the set X_n as a subset and which is accepted by the verifier. To show this, we first claim that $\|\mu_0\| \leq 30d^{3/4}$ with probability $1 - e^d$. Note that $\mu_0 \leftarrow N(\mu, \frac{I}{n})$, where $\mu \leftarrow N(0, \sqrt{d}I)$. By (5), with probability at least $1 - e^{-3d}$, $\|\mu\| \leq 15d^{3/4}$ and $\|\mu - \mu_0\| \leq 15\sqrt{d}$. Hence by the triangle inequality, $\|\mu_0\| \leq 30d^{3/4}$ with probability at least $1 - e^{-3d}$. We now show that for sets X_n such that $\|\mu_0\| \leq 30d^{3/4}$, Z_{n+1} cannot pass the verifier with probability more than $1 - e^2$ over the randomness in $\mu|X_n$. The proof consists of two cases, and the analysis of the cases is similar to the proof of Proposition 3. Without loss of generality, assume that $Z_{n+1} = \{x'_1, X_n\}$, hence x'_1 is the only sample not present in the set. We will show that either x'_1 or $\hat{\mu}_{-1}$ fail one of the three tests performed by the verifier with high probability.

Case 1: $\|x'_1 - \bar{\mu}\|^2 \geq 100d$.

We show that the first test is not satisfied with high probability in this case. As $\mu|X_n \leftarrow N(\bar{\mu}, \bar{\sigma}^2)$, hence by (5), $\|\mu - \bar{\mu}\|^2 \leq 15d/n$ with probability $1 - e^{-3d}$. Therefore, if $\|x'_1 - \bar{\mu}\|^2 \geq 100d$, then with probability e^{-3d},

$$\|x'_1 - \mu\|^2 \geq (\sqrt{100d} - \sqrt{15d/n})^2 > 15d,$$

in which case the first test is not satisfied. Hence in the first case, the amplifier succeeds with probability at most e^{-3d}.

Case 2: $\|x'_1 - \bar{\mu}\|^2 < 100d$.

Note that for the sample x'_1, $\mu_{-1} = \mu_0$ as the last n samples are the same as the original set X_n. We now bound $\|\hat{\mu}_{-1} - \bar{\mu}\|$ as follows,

$$\|\hat{\mu}_{-1} - \bar{\mu}\| = \left\|\mu_0 - \frac{n}{n + 1/\sqrt{d}}\mu_0\right\| \leq \frac{\|\mu_0\|}{n\sqrt{d}} \leq \frac{30d^{1/4}}{n}.$$

We now expand $\langle x'_1 - \hat{\mu}_{-1}, \mu - \hat{\mu}_{-1}\rangle$ in the third test as follows,

$$\begin{aligned}\langle x'_1 - \hat{\mu}_{-1}, \mu - \hat{\mu}_{-1}\rangle &= \langle x'_1 - \bar{\mu}, \mu - \bar{\mu}\rangle - \langle \hat{\mu}_{-1} - \bar{\mu}, \mu - \bar{\mu}\rangle - \langle x'_1 - \bar{\mu}, \hat{\mu}_{-1} - \bar{\mu}\rangle + \|\hat{\mu}_{-1} - \bar{\mu}\|^2,\\ &\leq \langle x'_1 - \bar{\mu}, \mu - \bar{\mu}\rangle - \langle \hat{\mu}_{-1} - \bar{\mu}, \mu - \bar{\mu}\rangle + \|x'_1 - \bar{\mu}\|\|\hat{\mu}_{-1} - \bar{\mu}\| + \|\hat{\mu}_{-1} - \bar{\mu}\|^2.\end{aligned}$$

Note that $\langle \hat{\mu}_{-1} - \bar{\mu}, \mu - \bar{\mu} \rangle$ is distributed as $N(0, \bar{\sigma}^2 \|\hat{\mu}_{-1} - \bar{\mu}\|^2)$ and hence with probability $1 - 1/e^3$ it is at most $10\|\hat{\mu}_{-1} - \bar{\mu}\|/\sqrt{n}$. Similarly, with probability $1 - 1/e^3$, $\langle x'_1 - \bar{\mu}, \mu - \bar{\mu} \rangle$ is at most $10\|x'_1 - \bar{\mu}\|/\sqrt{n}$. Therefore, with probability $1 - 2/e^3$,

$$\begin{aligned}
\langle x'_1 - \hat{\mu}_{-1}, \mu - \hat{\mu}_{-1} \rangle &\leq 10\|x'_1 - \bar{\mu}\|/\sqrt{n} + 10\|\hat{\mu}_{-1} - \bar{\mu}\|/\sqrt{n} + \|x'_1 - \bar{\mu}\|\|\hat{\mu}_{-1} - \bar{\mu}\| + \|\hat{\mu}_{-1} - \bar{\mu}\|^2, \\
&\leq 100\sqrt{\frac{d}{n}} + 300\frac{d^{3/4}}{n^2} + 300\frac{d^{3/4}}{n} + 900\frac{\sqrt{d}}{n^2} \\
&\leq 100\sqrt{\frac{d}{n}} + 1500\frac{d^{3/4}}{n} \\
&= 100\sqrt{\frac{n}{d}}\Big(\frac{d}{n}\Big) + \frac{1500}{d^{1/4}}\Big(\frac{d}{n}\Big).
\end{aligned}$$

Hence for a sufficiently large constant C, $n < \frac{d}{C \log d}$ and d sufficiently large, with probability $1 - 2/e^3$,

$$\langle x'_1 - \hat{\mu}_{-1}, \mu - \hat{\mu}_{-1} \rangle \leq \frac{d}{5n},$$

which implies that the second test is not satisfied. Hence the amplifier succeeds in this case with probability at most $2/e^3$.

The overall success probability of the amplifier is the maximum success probability across the two cases, hence for sets X_n such that the $\|\mu_0\| \leq 30d^{3/4}$, the verifier accepts the amplified set Z_{n+1} with probability at most $2/e^3$. As $\Pr\Big[\|\mu_0\| \leq 30d^{3/4}\Big] \geq 1 - e^{-3d}$, the overall success probability of the amplifier over the randomness in μ, X_n and any internal randomness of the amplifier is at most $1/e^2$. □

4 Proofs: Discrete Distributions with Bounded Support

4.1 Upper Bound

In this section we prove the upper bound from Theorem 1. The algorithm itself is presented in Algorithm 3. For clarity of writing, we assume that the number of input samples is $4n$, instead of n.

Proposition 6. *Let $\mathcal{C}$ denote the class of discrete distributions with support size at most k. For sufficiently large k, and $m = 4n + O\left(\frac{n}{\sqrt{k}}\right)$, $\mathcal{C}$ admits an $(4n, m)$ amplification procedure.*

Proof. To avoid dependencies between the count of different elements, we first prove our results in a poissonized setting, and then in lemma 4, we describe how to use the amplifier for poissonized setting to get an amplifier for the original multinomial setting. Let $D \in \mathcal{C}$ be an unknown probability distribution over $[k]$, and let p_i denote the probability mass associated with $i \in [k]$. Throughout the proof, we use random variable X_q to denote q independent samples from D, where q can also be a random variable. Suppose we are given $N = N_1 + N_2$ independent samples from D, denoted by X_{N_1} and X_{N_2}, where N_1 and N_2 are drawn from Poisson(n). We show how to amplify them to $\tilde{M} = N + R$ samples, denoted by $Z_{\tilde{M}}$, such that $D_{TV}(Z_{\tilde{M}}, X_M)$ is small, where $M \leftarrow$ Poisson($2n + r$).

Our amplifying procedure involves estimating the probability of each element using X_{N_1}, generating R independent samples using these estimates, and randomly shuffling these samples with X_{N_2}. Let u_i be the count of element i in X_{N_1} and y_i be the count of i in X_{N_2} noting they are both distributed as Poisson(np_i). The amplification procedure proceeds through the following steps:

1. Estimate the frequency $\hat{p}_i$ of each element using u_i, that is, $\hat{p}_i = \frac{u_i}{n}$.
2. Draw $\hat{z}_i \leftarrow$ Poisson($r\hat{p}_i$) additional samples of element i for all $i \in [k]$.
3. Append these generated samples to X_{N_2} to get Z_{N_2+R}.
4. Randomly permute the elements of Z_{N_2+R}, and append them to X_{N_1} to get $Z_{\tilde{M}}$.

Algorithm 3 Sample Amplification for Discrete Distributions

Input: $X_{4n} = (x_1, x_2, \ldots, x_{4n})$, where $x_i \leftarrow D$, for any discrete distribution D over $[k]$.
Output: $Z_m = (x'_1, x'_2, \ldots, x'_m)$, such that $D_{TV}(D^m, Z_m) \leq \frac{1}{3}$.

```
procedure AmplifyDiscrete($X_{4n}$)
    $N_1, N_2 \leftarrow$ Poisson($n$)        ▷ Draw two i.i.d samples $N_1$ and $N_2$ from Poisson($n$)
    $N := N_1 + N_2$
    if $N \leq 4n$ then
        $X_{N_1} := (x_1, x_2, \ldots, x_{N_1})$
        $X_{N_2} := (x_{N_1+1}, x_{N_1+2}, \ldots, x_{N_1+N_2})$
    else        ▷ Uninteresting case: happens with low probability
        $X_{N_1} := \underbrace{(x_1, x_1, \ldots, x_1)}_{N_1 \text{ times}}$
        $X_{N_2} := \underbrace{(x_1, x_1, \ldots, x_1)}_{N_2 \text{ times}}$
    $r := 8(m - n)$
    $(x'_1, x'_2, \ldots, x'_{N+R})$ = AmplifyDiscretePoissonized($X_{N_1}, X_{N_2}, r, n$)
      ▷ Amplify first $N_1 + N_2$ samples to $N_1 + N_2 + R$ samples, for $R$ roughly distributed as Poisson($r$)

    $R_1 := max(R, r/8)$
    if $R < r/8$ then        ▷ Uninteresting case: happens with low probability
        $(x'_1, x'_2, \ldots, x'_{N+R_1}) := (x'_1, x'_2, \ldots, x'_{N+R}, \underbrace{x_1, x_1, \ldots, x_1}_{\frac{r}{8}-R \text{ times}})$
    $(x'_{N+R_1+1}, x'_{N+R_1+2}, \ldots, x'_m) := (x_{N+1}, x_{N+2}, \ldots, x_{4n-(R_1-\frac{r}{8})})$
      ▷ Add the remaining samples to get $4n + r/8$ samples in total
    $Z_m := (x'_1, x'_2, \ldots, x'_m)$
    return $Z_M$

procedure AmplifyDiscretePoissonized($X_{N_1}, X_{N_2}, r, n$)
      ▷ Generates approximately Poisson($r$) more samples given $N_1 + N_2$ input samples
      ▷ $X_{N_1} = (x_1, x_2, \ldots, x_{N_1})$, $X_{N_2} = (x_{N_1+1}, x_{N_1+2}, \ldots, x_{N_1+N_2})$, and $r = 8(m - n)$
    count$_j := \sum_{i=1}^{N_1} \mathbb{1}(x_i = j)$, for $j \in [k]$        ▷ Find the count of each element in first $N_1$ samples
    $\hat{p}_j := \frac{count_j}{n}$, for $j \in [k]$
    $\hat{z}_j \leftarrow$ Poisson($\hat{p}_j r$), for $j \in [k]$
    $R := \sum_{j=1}^{k} \hat{z}_j$
    $(x'_1, x'_2, \ldots, x'_{N_1}) := (x_1, x_2, \ldots, x_{N_1})$
    $(x'_{N_1+1}, \ldots, x'_{N_1+N_2+R}) :=$ RandomPermute($(x_{N_1+1}, x_{N_1+2}, \ldots, x_{N_1+N_2}, \underbrace{1, 1, \ldots, 1}_{\hat{z}_1 \text{ times}}, \ldots, \underbrace{k, k, \ldots, k}_{\hat{z}_k \text{ times}})$)
    return $(x'_1, x'_2, \ldots, x'_{N_1+N_2+R})$
```

We first show that $Z_{\tilde{M}}$ is close in total variation distance, to Poisson($2n + r$) samples generated from D. We will prove this by showing that with high probability over the choice of X_{N_1}, the distribution of Z_{N_2+R} is close to Poisson($n + r$) samples generated from D. After this, we can use lemma 1 to show that appending Z_{N_2+R} to the samples in X_{N_1} results in a sequence with low total variation distance to X_M. Since our amplification procedure randomly permutes the last $N_2 + R$ elements, we can argue this using only the count of each element. Recall y_i is the count of element i in X_{N_2}, and $\hat{z}_i$ is the number of additional samples of element i added by our amplification procedure. Let $z_i \leftarrow$ Poisson(rp_i), and let $v_i = y_i + z_i$ and $\hat{v}_i = y_i + \hat{z}_i$. Here, v_i denotes the count of element i in Poisson($n + r$) samples drawn from D, and $\hat{v}_i$ denotes the corresponding count in samples generated using our amplification procedure. We use P_v to denote the distribution associated with random variable v.

Lemma 3. *For $r \leq n\epsilon^{1.5}/(4\sqrt{k})$, with probability $1-\epsilon$ over the randomness in $\{u_i, i \in [k]\}$,*

$$d_{TV}\left(\prod_{i=1}^{k} v_i, \prod_{i=1}^{k} \hat{v}_i\right) \leq \epsilon/2.$$

where $\prod$ refers to the product distribution.

Proof. We partition the support $[k]$ into two sets. Let $S = \{i : p_i \geq \epsilon/(2nk)\}$ and $S^c = [k]\backslash S$. Let $|S| = k'$. Without loss of generality, assume that $S = \{i : 1 \leq i \leq k'\}$ and $S^c = \{i : k'+1 \leq i \leq k\}$. We will separately bound the contribution of the variables in the set S and S^c to the total variation distance. For the first set S, we will upper bound $\sum_{i=1}^{k'} D_{KL}(v_i \parallel \hat{v}_i)$, and use Pinsker's inequality to then bound the total variation distance. For the second set S^c, we will directly bound $\sum_{i=k'+1}^{k} d_{TV}(v_i, \hat{v}_i)$. All our bounds will be with high probability over the randomness in the first set $\{u_i, i \in [k]\}$.

We first bound the total variation distance for the variables in the first set S. Note that because the sum of two Poisson random variables is a Poisson random variable, v_i is distributed as Poisson$(np_i + rp_i)$ and $\hat{v}_i$ is distributed as Poisson$(np_i + ru_i/n)$. We will use the following expression for the KL divergence $D_{KL}(P \parallel Q)$ between two Poisson distributions P and Q with means λ_1 and λ_2 respectively—

$$D_{KL}(P \parallel Q) = \lambda_1 \log\left(\frac{\lambda_1}{\lambda_2}\right) + \lambda_2 - \lambda_1. \tag{10}$$

Using this expression, we can write the KL divergence between the distributions of v_i and $\hat{v}_i$ as follows,

$$D_{KL}(v_i \parallel \hat{v}_i) = p_i(n+r)\log\left(\frac{p_i(n+r)}{p_i n + ru_i/n}\right) + (ru_i/n - rp_i).$$

Let $\delta_i = u_i - np_i$. We can rewrite the above expression as follows,

$$\begin{aligned} D_{KL}(v_i \parallel \hat{v}_i) &= p_i(n+r)\log\left(\frac{p_i(n+r)}{p_i(n+r) + r\delta_i/n}\right) + r\delta_i/n, \\ &= p_i(n+r)\log\left(\frac{1}{1 + r\delta_i/(np_i(n+r))}\right) + r\delta_i/n. \end{aligned}$$

Note that $\log(1+x) \geq x - 2x^2$ for $x \geq 0.8$. As $\delta_i \geq -np_i$, therefore $r\delta_i/(np_i(n+r)) \geq -0.8$ for $r \leq n$. Therefore,

$$\begin{aligned} p_i(n+r)\log\left(\frac{1}{1 + r\delta_i/(np_i(n+r))}\right) &\leq -r\delta_i/n + \frac{2r^2\delta_i^2}{n^2 p_i(n+r)}, \\ \implies D_{KL}(v_i \parallel \hat{v}_i) &\leq \frac{2r^2\delta_i^2}{n^2 p_i(n+r)}, \\ \implies \sum_{i=1}^{k'} D_{KL}(v_i \parallel \hat{v}_i) &\leq \frac{2r^2}{n^2}\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i}. \end{aligned} \tag{11}$$

We will now bound $\sum_{i=1}^{k'} \frac{\delta_i^2}{np_i}$. As a Poisson$(\lambda)$ random variable has variance λ and $\delta_i = u_i - np_i$ where $u_i \leftarrow$ Poisson(np_i), therefore,

$$\mathbb{E}\left[\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i}\right] = k'.$$

Also, the fourth central moment of a Poisson(λ) random variable is $\lambda(1+3\lambda)$, hence

$$\begin{aligned} \text{Var}[\delta_i^2] &= \mathbb{E}\left[\delta_i^4\right] - \mathbb{E}\left[\delta_i^2\right]^2, \\ &= np_i(1+3np_i) - (np_i)^2 = np_i(1+2np_i), \\ \implies \text{Var}\left[\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i}\right] &= \sum_{i=1}^{k'}\frac{1+2np_i}{np_i}. \end{aligned}$$

As $p_i \geq \epsilon/(2nk)$ for $i \in S$ and $k' \leq k$, therefore,

$$\mathrm{Var}\left[\sum_{i=1}^{k'} \frac{\delta_i^2}{np_i}\right] \leq 2k^2/\epsilon + 2k \leq 4k^2/\epsilon.$$

Hence by Chebyshev's inequality,

$$\begin{aligned} &\Pr\left[\sum_{i=1}^{k'} \frac{\delta_i^2}{np_i} - k' \geq 4k/\epsilon\right] \leq \epsilon/4, \\ \implies &\Pr\left[\sum_{i=1}^{k'} \frac{\delta_i^2}{np_i} \geq 4k/\epsilon\right] \leq \epsilon/4. \end{aligned} \tag{12}$$

Let E_1 be the event that $\sum_{i=1}^{k'} \frac{\delta_i^2}{np_i} \leq 4k/\epsilon$. By (12), $\Pr(E_1) \geq 1 - \epsilon/4$. Conditioned on the event E_1 and using (11), we can bound the KL divergence as follows,

$$D_{KL}\Big(\prod_{i\in S} v_i \;\Big\|\; \prod_{i \in S} \hat{v}_i\Big) = \sum_{i=1}^{k'} D_{KL}(v_i \parallel \hat{v}_i) \leq \frac{8r^2k}{n^2\epsilon}.$$

Hence for $r \leq n\epsilon^{1.5}/(4\sqrt{k})$ and conditioned on the event E_1,

$$D_{KL}\Big(\prod_{i\in S} v_i \;\Big\|\; \prod_{i \in S} \hat{v}_i\Big) \leq \epsilon^2/2.$$

Hence using Pinsker's inequality, conditioned on the event E_1,

$$d_{TV}\left(\prod_{i\in S} v_i, \prod_{i\in S} \hat{v}_i\right) \leq \epsilon/2.$$

We will now bound the total variation distance for the variables in the set S^c. Let E_2 be the event that $u_i = 0,\ \forall\ i \in S^c$. Note that as $u_i \sim$ Poisson(np_i) where $p_i < \epsilon/(2nk)$, $u_i = 0$ with probability at least $e^{-\epsilon/(2k)}$, hence $\Pr(E_2) \geq e^{-\epsilon/2} \geq 1 - \epsilon/2$. We now condition on the event E_2. Recall that $v_i = y_i + z_i$, where $z_i \sim$ Poisson(rp_i) and $\hat{v}_i = y_i + \hat{z}_i$, where $\hat{z}_i = 0$ conditioned on E_2. By a coupling argument on y_i, the total variation distance between the distributions of v_i and $\hat{v}_i$ equals the total variation distance between the distributions of z_i and $\hat{z}_i$. As $\hat{z}_i = 0$, conditioned on the event E_2,

$$\begin{aligned} d_{TV}(v_i, \hat{v}_i) &= \Pr[z_i \neq 0] = 1 - e^{-rp_i} \leq 1 - e^{-r\epsilon/(2nk)} \\ &\leq \frac{r\epsilon}{2nk} \leq \frac{\epsilon}{2k}, \quad \text{as } r \leq n. \end{aligned}$$

Hence conditioned on E_2,

$$d_{TV}\left(\prod_{i\in S^c} v_i, \prod_{i\in S^c} \hat{v}_i\right) \leq \sum_{i=k'+1}^{k} d_{TV}\left(v_i, \hat{v}_i\right) \leq \epsilon/2.$$

Hence conditioned on the events E_1 and E_2,

$$d_{TV}\left(\prod_{i=1}^{k} v_i, \prod_{i=1}^{k} \hat{v}_i\right) \leq d_{TV}\left(\prod_{i\in S} v_i, \prod_{i\in S} \hat{v}_i\right) + d_{TV}\left(\prod_{i\in S^c} v_i, \prod_{i\in S^c} \hat{v}_i\right) \leq \epsilon.$$

As $\Pr(E_1) \geq 1 - \epsilon/4$ and $\Pr(E_2) \geq 1 - \epsilon/2$, by a union bound $\Pr(E_1 \cup E_2) \geq 1 - \epsilon$. Hence with probability $1 - \epsilon$ over the randomness in $\{u_i, i \in [k]\}$,

$$d_{TV}\left(\prod_{i=1}^{k} v_i, \prod_{i=1}^{k} \hat{v}_i\right) \leq \epsilon.$$

□

Lemma 3 says that with high probability over the first N_1 samples, the $N_2 + R$ samples are close in total variation distance to Poisson$(n + r)$ samples drawn from D. Using lemma 3 and lemma 1, we can conclude that for $r \leq n\epsilon^{1.5}/(4\sqrt{k})$, $D_{TV}(X_M, Z_{\tilde{M}}) \leq \epsilon + \epsilon/2 = 3\epsilon/2$.

Next, we show how to use the above amplification procedure to amplify samples in the non-poissonized setting. Given $N = N_1 + N_2$ samples from D, we have shown how to amplify them to get $\widetilde{M} = N + R$ samples. Given such an amplifier as a black box, and $4n$ samples from D, one can use the first N samples to generate M samples. Then append these M samples with the remaining $4n - N$ samples to get an amplifier in our original non-poissonized setting.

Lemma 4. *Let $N = N_1 + N_2$ where $N_1, N_2 \leftarrow$ Poisson(n), and let $M \leftarrow$ Poisson$(2n + r)$. Suppose we are given an (N, M) amplifier f (as described above) satisfying $D_{TV}(f(X_N), X_M) \leq \frac{3\epsilon}{2}$, for all $D \in \mathcal{C}$. Then there exists an amplifier $f' : [k]^{4n} \to [k]^{4n+\frac{r}{8}}$, such that $D_{TV}(f'(X_{4n}), X_{4n+\frac{r}{8}}) \leq \frac{5\epsilon}{2}$, for $\epsilon \geq 2e^{-\frac{n}{20}} + e^{-\frac{25r}{88}}$, and for $r \leq n\epsilon^{1.5}/(4\sqrt{k})$.*

Proof. We divide the proof into three steps:

- **Step 1:** f takes as input X_{N_1} and X_{N_2}, samples of size N_1 and N_2 drawn from D. To simulate these samples, we use the 4n samples available to us from D. We draw $N'_1, N'_2 \leftarrow$ Poisson(n), and let $N' = N'_1 + N'_2$. If $N' \leq 4n$, we set $X_{N'_1} = (x_1, x_2, \ldots, x_{N'_1})$ and $X_{N'_2} = (x_{N'_1+1}, x_{N'_1+2}, \ldots, x_{N'_2})$. Otherwise, we set $X_{N'_1} = \underbrace{(x_1, x_1, \ldots, x_1)}_{N'_1 \text{ times}}$, and $X_{N'_2} = \underbrace{(x_1, x_1, \ldots, x_1)}_{N'_2 \text{ times}}$, but this happens with very small probability leading to small total variation distance between $f(X_{N_1}, X_{N_2})$ and $f(X_{N'_1}, X_{N'_2})$, and by triangle inequality, small TV distance between $f(X_{N'_1}, X_{N'_2})$ and X_M. We denote (X_{N_1}, X_{N_2}) by X_N and $(X_{N'_1}, X_{N'_2})$ by $X_{N'}$.

- **Step 2:** We would like to finally output $\frac{r}{8}$ more samples. Let us denote the number of samples in $f(X_{N'})$ by M'. If $M' < N' + \frac{r}{8}$, we append $N' + \frac{r}{8} - M'$ arbitrary samples to it (say x_1) so that the total sample size is equal to $N' + \frac{r}{8}$. If $M' \geq N' + \frac{r}{8}$, we don't do anything in this step. Let $t_1(f(X_{N'}))$ denote the samples outputted in this step. Since the number of new samples added by f is roughly distributed as Poisson(r), the probability that the number of new samples is less than $r/8$ is small, leading to small TV distance between $t_1(f(X_{N'}))$ and $f(X_{N'})$, and by triangle inequality, small TV distance between $t_1(f(X_{N'}))$ and X_M.

- **Step 3:** Let M'_1 denote the number of samples in $t_1(f(X_{N'}))$, and let $Q'_1 = 4n + \frac{r}{8} - M'_1$ denote the number of extra samples needed to output $4n + \frac{r}{8}$ samples in total. If $Q'_1 \geq 0$, we append Q'_1 i.i.d. samples from D to $t_1(f(X_{N'}))$, and if $Q'_1 < 0$, we remove last $|Q'_1|$ samples from $t_1(f(X_{N'}))$. We use $t_2(t_1(f(X_{N'})))$ to denote the output of this step. Step 2 ensures $M'_1 \geq N' + \frac{r}{8}$, which implies $Q'_1 \leq 4n - N'$. Let $X_{4n-N'} = (x_{N'+1}, x_{N'+2}, \ldots, x_{4n})$ denote the leftover samples in X_{4n} after removing the first N' samples. When $Q'_1 \geq 0$, we use the first Q'_1 samples from $X_{4n-N'}$ to simulate i.i.d. samples from D, that is, $t_2(t_1(f(X_{N'}))) = \text{append}(t_1(f(X_{N'})), (x_{N'+1}, x_{N'+2}, \ldots, x_{N'+Q'_1}))$. $t_2(t_1(f(X_{N'})))$ is the final output of our amplifier f'.

  Similarly, let $Q_1 = 4n + \frac{r}{8} - M$ denote the number of extra samples needed to be appended to X_M to output $4n + \frac{r}{8}$ samples in total. If $Q_1 \geq 0$, $t_2(X_M)$ correspond to appending Q_1 samples from D to X_M, and otherwise, it corresponds to removing last $|Q_1|$ samples from X_M. Since applying the same transformation to two random variables can't increase their total variation distance, and from step 2, we know that $D_{TV}(t_1(f(X_{N'})), X_M)$ is small, we get $D_{TV}(t_2(t_1(f(X_{N'}))), t_2(X_M))$ is small.

  As $t_2(X_M)$ corresponds to $4n + \frac{r}{8}$ i.i.d. samples from D, $D_{TV}(X_{4n+\frac{r}{8}}, t_2(X_M)) = 0$. Using triangle inequality, we get $D_{TV}(t_2(t_1(f(X_{N'}))), X_{4n+\frac{r}{8}})$ is small which is the desired result.

Next, we prove that the total variation distances involved in each of these steps are small.

- **Step 1:** We first bound $D_{TV}(f(X_N), f(X_{N'}))$.

  $$\begin{aligned}D_{TV}(f(X_N), f(X_{N'})) &\le D_{TV}(X_N, X_{N'})\\ &= \frac{1}{2}\sum_x |\Pr(X_N = x) - \Pr(X_{N'} = x)|\\ &= \frac{1}{2}\sum_x |\Pr(X_N = x \mid N \le 4n)\Pr(N \le 4n) - \Pr(X_{N'} = x \mid N' \le 4n)\Pr(N' \le 4n)\\ &+ \Pr(X_N = x \mid N > 4n)\Pr(N > 4n) - \Pr(X_{N'} = x \mid N' > 4n)\Pr(N' > 4n)|\end{aligned}$$

  where the first inequality holds as applying the same transformation to two random variables can't increase their total variation distance. Now, note that X_N and $X_{N'}$ have the same distribution conditioned on $N \le 4n$ and $N' \le 4n$. Also, $\Pr(N \le 4n) = \Pr(N' \le 4n)$ and $\Pr(N > 4n) = \Pr(N' > 4n)$, as both N and N' are drawn from Poisson$(2n)$ distribution. This gives us

  $$\begin{aligned}D_{TV}(f(X_N), f(X_{N'})) &= \frac{1}{2}\sum_x \Pr(N > 4n)|\Pr(X_N = x \mid N > 4n) - \Pr(X_{N'} = x \mid N' > 4n)|\\ &\le \Pr(N > 4n)\end{aligned}$$

  Using traingle inequality, we get $D_{TV}(X_M, f(X'_N)) \le \Pr(N > 4n) + 3\epsilon/2$. To bound $\Pr(N > 4n)$, we use the following poisson tail bound [6]: for $X \leftarrow \text{Poisson}(\lambda)$,

  $$\Pr[X \ge \lambda + x], \Pr[X \le \lambda - x] \le e^{\frac{-x^2}{\lambda + x}}. \tag{13}$$

  As N is distributed as Poisson$(2n)$, we get $\Pr(N > 4n) \le e^{-n}$, which implies $D_{TV}(X_M, f(X'_N)) \le e^{-n} + \frac{3\epsilon}{2}$.

- **Step 2:** In this step, we need to show $D_{TV}(t_1(f(X_{N'})), X_M)$ is small. Note that $t_1(f(X_{N'}))$ is equal to $f(X_{N'})$ except when $M' < N' + \frac{r}{8}$. From step 1, we know that $D_{TV}(f(X_{N'}), X_M)$ is small. If we show $D_{TV}(f(X_{N'}), t_1(f(X_{N'})))$ is small, then by traingle inequality, we get $D_{TV}(X_M, t_1(f(X_{N'})))$ is small. Let $M' = N' + R'$ where R' denote the number of new samples added by the amplification procedure f to $X_{N'}$.

  $$\begin{aligned}&D_{TV}\left(t_1\left(f\left(X_{N'}\right)\right), f\left(X_{N'}\right)\right)\\ &= \frac{1}{2}\sum_x |\Pr\left(t_1\left(f\left(X_{N'}\right)\right) = x\right) - \Pr\left(f\left(X_{N'}\right) = x\right)|\\ &= \frac{1}{2}\sum_x |\Pr\left(R' < \frac{r}{8}\right)\left(\Pr\left(t_1\left(f\left(X_{N'}\right)\right) = x \mid R' < \frac{r}{8}\right) - \Pr\left(f\left(X_{N'}\right) = x \mid R' < \frac{r}{8}\right)\right)\\ &+ \Pr\left(R' \ge \frac{r}{8}\right)\left(\Pr\left(t_1\left(f\left(X_{N'}\right)\right) = x \mid R' \ge \frac{r}{8}\right) - \Pr\left(f\left(X_{N'}\right) = x \mid R' \ge \frac{r}{8}\right)\right)|\end{aligned}$$

  We know $\Pr\left(t_1\left(f\left(X_{N'}\right)\right) = x \mid R' \ge \frac{r}{8}\right) = \Pr\left(f\left(X_{N'}\right) = x \mid R' \ge \frac{r}{8}\right)$. This gives

  $$\begin{aligned}&D_{TV}\left(t_1\left(f\left(X_{N'}\right)\right), f\left(X_{N'}\right)\right)\\ &= \frac{1}{2}\sum_x |\Pr\left(R' < \frac{r}{8}\right)\left(\Pr\left(t_1\left(f\left(X_{N'}\right)\right) = x \mid R' < \frac{r}{8}\right) - \Pr\left(f\left(X_{N'}\right) = x \mid R' < \frac{r}{8}\right)\right)|\\ &\le \Pr\left(R' < \frac{r}{8}\right)\end{aligned}$$

  Now, we need to bound $\Pr\left(R' < \frac{r}{8}\right)$. From the description of f, we know that the number of new copies of element i added by f is distributed as Poisson$\left(r\hat{p}_i\right)$. Here, $\hat{p}_i = \frac{u_i}{n}$ where u_i denotes the number of occurrences of element i in $X_{N'_1}$. Since the total number of samples in $X_{N'_1}$ is N'_1, we get

$\sum_{i=1}^k \hat{p}_i = \frac{\sum_{i=1}^k u_i}{n} = \frac{N_1'}{n}$. Note that R' is equal to the sum of number of new copies of each element, and as the sum of Poisson random variables is Poisson, we get R' is distributed as Poisson $\left(r\frac{N_1'}{n}\right)$.

$$\begin{aligned}\Pr\left(R' < \frac{r}{8}\right) &= \Pr\left(R' < \frac{r}{8} \mid N_1' \geq \frac{3n}{4}\right)\Pr\left(N_1' \geq \frac{3n}{4}\right) + \Pr\left(R' < \frac{r}{8} \mid N_1' < \frac{3n}{4}\right)\Pr\left(N_1' < \frac{3n}{4}\right)\\ &\leq \Pr\left(R' < \frac{r}{8} \mid N_1' \geq \frac{3n}{4}\right) + \Pr\left(N_1' < \frac{3n}{4}\right)\end{aligned}$$

Using Poisson tail bound (13), we get

$$\begin{aligned}\Pr\left(R' < \frac{r}{8} \mid N_1' \geq \frac{3n}{4}\right) &\leq exp\left(-\frac{(5r/8)^2}{3r/4 + 5r/8}\right) = e^{-25r/88}\\ \Pr\left(N_1' < \frac{3n}{4}\right) &\leq exp\left(-\frac{(n/4)^2}{n + n/4}\right) = e^{-n/20}\end{aligned}$$

This gives us $D_{TV}(f(X_{N'}), t_1(f(X_{N'}))) \leq e^{-25r/88} + e^{-n/20}$. By triangle inequality, we get $D_{TV}(X_M, t_1(f(X_{N'}))) \leq \frac{3\epsilon}{2} + e^{-n} + e^{-25r/88} + e^{-n/20}$.

- **Step 3:** For this step, we need to show $D_{TV}(t_2(t_1(f(X_{N'}))), t_2(X_M))$ is small. Since applying the same transformation to two random variables doesn't increase their TV distance, we get

  $$\begin{aligned}D_{TV}(t_2(t_1(f(X_{N'}))), t_2(X_M)) &\leq D_{TV}(t_1(f(X_{N'})), X_M)\\ &\leq \frac{3\epsilon}{2} + e^{-n} + +e^{-25r/88} + e^{-n/20}\end{aligned}$$

  As $D_{TV}(X_{4n+\frac{r}{8}}, t_2(X_M)) = 0$, using triangle inequality, we get

  $$D_{TV}(t_2(t_1(f(X_{N'}))), X_{4n+\frac{r}{2}}) \leq \frac{3\epsilon}{2} + e^{-n} + e^{-25r/88} + e^{-n/20}$$

For $\epsilon \geq 2e^{-n/20} + e^{-25r/88}$, this gives us $D_{TV}(f'(X_{4n}), X_{4n+\frac{r}{8}}) = D_{TV}(t_2(t_1(f(X_{N'}))), X_{4n+\frac{r}{8}}) \leq \frac{5\epsilon}{2}$. □

From lemma 4, we get that for $\epsilon \geq 2e^{-n/20} + e^{-25r/88}$, and for $r \leq n\epsilon^{1.5}/(4\sqrt{k})$, $D_{TV}(f'(X_{4n}), X_{4n+\frac{r}{8}}) \leq \frac{5\epsilon}{2}$. We can assume n is at least $\sqrt{k}$, and r is at least 8, as otherwise the theorem is trivially true. So for k large enough (implying large n), we can put $\epsilon = \frac{2}{15}$, to get $D_{TV}(t_2(t_1(f(X_{N'}))), X_{4n+\frac{r}{8}}) \leq \frac{1}{3}$, which finishes the proof! □

4.2 Lower Bound

In this section we show that the above procedure is optimal, up to constant factors for amplifying samples from discrete distributions. The proof is constructive and shows that a simple verifier can distinguish any amplifier when $m > \alpha\frac{n}{\sqrt{k}}$ for a fixed α. The proof relies on the fact that the amplifier cannot add samples beyond the support of the samples it has already seen. When m is sufficiently larger than n, we can show there are distributions for which large regions of the support are below the threshold required for the birthday paradox meaning that with high probability every new sample will reveal additional information about the support. The amplifier will not be able to add samples in that region.

Proposition 7. *There is a constant c, such that for every sufficiently large k, $\mathcal{C}$ does not admit an $\left(n, n + \frac{cn}{\sqrt{k}}\right)$ amplification procedure.*

The proposition follows by constructing a verifier and class of discrete distributions over k elements, $\mathcal{C}$ with the following property: for a universal constant c and $p \leftarrow Uniform[\mathcal{C}]$, the verifier can detect *any* $(n, n + \frac{cn}{\sqrt{d}})$ amplifier from with sufficiently high probability.

Before we prove Proposition 7, we introduce some additional notation and a basic martingale inequality. Let C^k be the set of discrete uniform distributions over k integers in $0, \ldots, 8k$. Let C^k_l be the set of discrete distributions with mass $1-l$ on one element and uniform mass over $k-1$ remaining integers in $0, \ldots, 8k$. We also rely on some martingale inequalities which can be found in [9].

Fact 1. *Let X be the martingale associated with a filter $\mathcal{F}$ satisfying:*

1. $\mathrm{Var}[X_i \mid \mathcal{F}_{i-1}] \le \sigma_i^2$ *for* $1 \le i \le n$
2. $0 \le X_i \le 1$ *almost surely.*

Then, we have

$$Pr(X - \mathbb{E}[x] \ge \lambda) \le e^{-\frac{\lambda^2}{2(\sum \sigma_i^2 + \lambda/3)}}.$$

Similarly the following holds (though not simultaneously):

$$Pr(X - \mathbb{E}[x] \le -\lambda) \le e^{-\frac{\lambda^2}{2(\sum \sigma_i^2 + \lambda/3)}}.$$

Finally we rely on slight generalization of the birthday paradox which can be found in [4].

Fact 2. *Let n samples be drawn from a uniform distribution over k elements. Then the probability of the samples containing a duplicate is less than $\frac{n^2}{2k}$.*

The proof proceeds in two parts. First we prove a lemma that shows the desired result for $n \le \frac{k}{2}$. We then show show a class of distributions that allows us to reduce the general case to the result shown in the lemma.

Lemma 5. *For sufficiently large k, fixed c and $m = n + 30\frac{n}{\sqrt{k}} \le \frac{k}{4}$ the following holds:*

There exists a verifier that for $p \sim Uniform[C^k]$ the following holds true:

1. *For all p, it accepts X_m with probability at least $\frac{3}{4}$ over the randomness in X_m.*
2. *It rejects $f(X_n)$ with probability at least $\frac{3}{4}$ for any amplifier f over the randomness in X_n, p and the amplifier.*

Proof. First we consider the case when $n \le \frac{\sqrt{k}}{2}$. Consider the verifier that takes $\frac{\sqrt{k}}{2} + 1 < \sqrt{\frac{k}{2}}$ samples from the given samples uniformly at random and accepts if there are no repeats by Fact 2 and the support is correct. The probability of a duplicate with the real distribution is less than $\frac{1}{4}$ by fact 2 so the verifier will accept samples from the true distribution with at least probability $\frac{3}{4}$.

An amplified set, on the other hand, must have repeats outside of the original elements it saw. This is because if the amplifier expanded the support of the set, the verifier would catch it with probability $\frac{7}{8}$. To show this, consider a sample added by the amplifier outside of the seen support. Conditioned on the at most $\frac{k}{4}$ unique samples seen so far (which implies that $\frac{3}{4}$ of the support is still unseen), the probability, over the choice of p, of said sample being in the set is at most $\frac{(3/4)k}{8k-n} \le \frac{(3/4)k}{7.5k} \le \frac{1}{8}$. Hence if the amplified set has any element outside the original support then it is rejected with probability $\frac{7}{8}$. Note that if the amplified set has at most $\frac{\sqrt{k}}{2}$ unique elements, then it can be immediately distinguished for having too many repeats.

We now examine the case when $n > \frac{\sqrt{k}}{2}$. Since the verifier can identify when the amplifier introduces unseen elements with probabiltiy at least $\frac{7}{8}$, we condition on the event that the verifier identifies such elements for the remainder of this proof. The proof proceeds by showing that a set the size of the amplified set must have significantly more unique elements than the original set. Before we proceed with the details of the proof we define the martingale that is central to the argument. Consider the scenario where the n samples are drawn in sequence, and let $\mathcal{F}_i$ denote the filtration corresponding to the i-th draw (i.e., information in the first i draws). Let U_i be the indicator that is the ith sample was previously unseen. Let $U^n = \sum_{i=1}^{n} U_i$. Note that $B_i = \mathbb{E}\left[\sum_{j=1}^{n} U_j \mid \mathcal{F}_i\right]$ is a Doob martingale with respect to the filtration $\mathcal{F}_i$ and $B_n = U$. Also, B_i has differences bounded by 1 as U_i is an indicator random variable. If j is the count of previously seen elements then $\mathrm{Var}\left[B_i \mid \mathcal{F}_i\right] \leq \mathrm{Var}[U_i \mid \mathcal{F}_i] \leq \frac{(k-j)j}{k^2}$. Since $n < \frac{k}{2}$, the variance is upper bounded by $\frac{i}{k} \leq \frac{n}{k}$.

The verifier will accept only if all elements are within the support of the distribution and the number unique elements is greater than $\mathbb{E}[U^n] + 7\frac{n}{\sqrt{k}}$ under X_n.

The remainder of the proof will show the following:

1. U^n concentrates around its expectation within a $O\left(\frac{n}{\sqrt{k}}\right)$ margin for X_n (this shows the amplifier gets too few unique samples to be accepted by the verifier).

2. The expectation $\mathbb{E}[U^m - U^n]$ increases by at least $\Omega\left(\frac{n}{\sqrt{k}}\right)$ from X_n to X_m (which shows the number of unique items is sufficiently different in expectation between X_n and X_m).

3. U^m concentrates around its expectation within a $O\left(\frac{n}{\sqrt{k}}\right)$ margin for X_m (this combined with the previous statement shows the verifier accepts real samples with sufficiently high probability).

The upper tail bound follows via Fact 1. Recall that $\frac{n}{\sqrt{k}} < 4\frac{n^2}{k}$ since $n > \frac{\sqrt{k}}{2}$.

$$\begin{aligned}
\Pr\left(U^n - \mathbb{E}[U^n] \geq 7\frac{n}{\sqrt{k}}\right) &\leq \exp\left(-\frac{7^2\frac{n^2}{k}}{2\left(\sum \sigma_i^2 + \frac{7n}{3\sqrt{k}}\right)}\right)\\
&\leq \exp\left(-\frac{7^2\frac{n^2}{k}}{2\left(\frac{n^2}{k} + \frac{7n}{3\sqrt{k}}\right)}\right)\\
&\leq \exp\left(-\frac{7^2\frac{n^2}{k}}{2\left(\frac{n^2}{k} + 7\frac{4n^2}{3k}\right)}\right)\\
&= \exp\left(-\frac{7^2\frac{n^2}{k}}{2\left(1 + \frac{4}{3}7\right)\frac{n^2}{k}}\right)\\
&\leq \frac{1}{8}.
\end{aligned}$$

Note that this suffices to show that the verifier can distinguish any amplifier with sufficiently many unique samples.

Let k be sufficiently large that the following conditions hold for both k and $k-1$:

1. $n + 30\frac{n}{\sqrt{k}} < \frac{k}{2}$

2. The samples increased by at most a factor of 2

Now we note that the $\mathbb{E}[U^n]$ and $\mathbb{E}[U^m]$ must differ by at least $\frac{15n}{\sqrt{k}}$, since $m < \frac{k}{2}$ implying that every new sample has at least a $\frac{1}{2}$ probability of being unique. Now all the remains to show that the verifier will accept X_m is to show concentration of U within $\frac{8n}{\sqrt{k}}$ of its mean.

Since the number of samples increased by at most a factor of two, the bound on the σ_i^2 increased by at most a factor of two. This suffices for the lower tail bound on U for X_m—

$$\begin{aligned}\Pr\left(U^m - \mathbb{E}[U^m] \leq -8\frac{n}{\sqrt{k}}\right) &\leq \exp\left(-\frac{8^2\frac{n^2}{k}}{2\left(\sum \sigma_i^2 + \frac{8n}{3\sqrt{k}}\right)}\right) \\ &\leq \exp\left(-\frac{8^2\frac{n^2}{k}}{2\left(4\frac{n^2}{k} + \frac{8n}{3\sqrt{k}}\right)}\right) \\ &\leq \exp\left(-\frac{8^2\frac{n^2}{k}}{2\left(4\frac{n^2}{k} + 8\frac{4n^2}{3k}\right)}\right) \\ &= \exp\left(-\frac{8^2\frac{n^2}{k}}{2\left(4 + \frac{4}{3}8\right)\frac{n^2}{k}}\right) \\ &< \frac{1}{8}.\end{aligned}$$

Thus X_m will have sufficiently many unique elements to be accepted by the verifier with probability at least $\frac{7}{8}$. A success probability of $\frac{3}{4}$ follows from subtracting the probabiltiy that the verifier did not properly identify unseen samples.

□

We are now ready to prove Proposition 7.

Proof. If $n \leq \frac{k}{4}$, then Lemma 5 applies directly. If not, we use the set of distributions $\mathcal{C}^k_{\frac{k}{4}}$ with the intention of applying Lemma 5 on samples that land in the uniform region.

The verifier will check that the samples are within the support of the distribution, more than $n + 7\frac{n}{\sqrt{k}}$ samples are in the uniform region and the verifier from Lemma 5 accepts on the uniform region.

First note that after n samples, at most $\frac{k}{4} + \frac{\sqrt{k}}{4}$ samples will be in the uniform region with at least probability $\frac{15}{16}$ by a Chebyshev bound. Conditioned on this event, Lemma 5 shows that the amplifier cannot output more than $\frac{k}{4} + O(\sqrt{k})$ samples in the uniform region and will be rejected by our verifier.

Now we show that the verifier will accept real samples with good probability. Note that the expected number of samples to receive in the uniform region for X_m is $\frac{k}{4} + c\sqrt{k}$. The variance on this quantity is $\frac{k}{4} + c\sqrt{k}$. An application of Chebyshev's inequality shows that with probability at least $\frac{15}{16}$ sufficiently many samples will land in the uniform region.

$$\begin{aligned}\frac{k}{4} + c\sqrt{k} - 4\sqrt{\frac{k}{4} + c\sqrt{k}} &\geq \frac{k}{4} + c\sqrt{k} - 2\sqrt{k} - 4\sqrt{c\sqrt{k}} \\ &\geq \frac{k}{4} + c\sqrt{k} - 2\sqrt{k} - 4\sqrt{ck}.\end{aligned}$$

Since the expression above is increasing with c, we can choose a c sufficiently large so that the verifier will accept with sufficiently high probability. □

References

[1] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1, 2012.

[2] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2001.

[3] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

[4] Mihir Bellare, Joe Kilian, and Phillip Rogaway. The security of the cipher block chaining message authentication code. In *Advances in Cryptology–CRYPTO*, volume 94, pages 341–358. Citeseer, 1994.

[5] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, pages 2611–2619, 2015.

[6] Clément L Canonne. A short note on poisson tail bounds, 2017.

[7] Clément L Canonne, Themis Gouleakis, and Ronitt Rubinfeld. Sampling correctors. *SIAM Journal on Computing*, 47(4):1373–1423, 2018.

[8] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

[9] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

[10] Gabe Cohn. Ai art at christie's sells for $432,500. *The New York Times* *https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html*, Oct 2018.

[11] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[12] 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.

[13] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity*, 2000.

[14] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[15] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[16] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(1):295–347, 2013.

[17] Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1($2\sigma 2$):16, 2007.

[18] Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always good turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

[19] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

[20] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[21] Salil P Vadhan et al. Pseudorandomness. *Foundations and Trends® in Theoretical Computer Science*, 7(1–3):1–336, 2012.

[22] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155. ACM, 2016.

[23] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[24] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[25] Emanuele Viola. The complexity of distributions. *SIAM Journal on Computing*, 41(1):191–218, 2012.

[26] Martin Wainwright. Basic tail and concentration bounds. *https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf (visited on 12/31/2017)*, 2015.

[27] Hao Yi, Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pages 8848–8857, 2018.