Sample Amplification: Increasing Dataset Size even when Learning is Impossible

Brian Axelrod [*1] **Shivam Garg** [*1] **Vatsal Sharan** [*1] **Gregory Valiant** [*1]

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" which must be indistinguishable from m samples drawn iid from D. We consider this sample amplification problem in two fundamental settings: the case where D is an arbitrary discrete distribution supported on k elements, and the case where D is a d-dimensional Gaussian with unknown mean, and fixed covariance matrix. Perhaps surprisingly, we show a valid amplification procedure exists for both of these settings, 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. We also show that our procedures are optimal up to constant factors. Beyond these results, we describe potential applications of sample amplification, and formalize a number of curious directions for future research.

1. Learning, Testing, and Sample Amplification

How much do you need to know about a 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. The question posed in our work fits with this body of work, though instead of asking how 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 (GAN) 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

[*]Alphabetical ordering [1]Stanford University, United States. Correspondence to: Brian Axelrod <baxelrod@stanford.edu>, Shivam Garg <shivamgarg@stanford.edu>, Vatsal Sharan <vsharan@cs.stanford.edu>, Gregory Valiant <valiant@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

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 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.*[2]

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 it parallels the pseudo-randomness framework, and 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 here.

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:*

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

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.

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*

[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.

[2] The constant in the definition is chosen for ease of exposition, and we prove the theorems for general tolerance parameter.

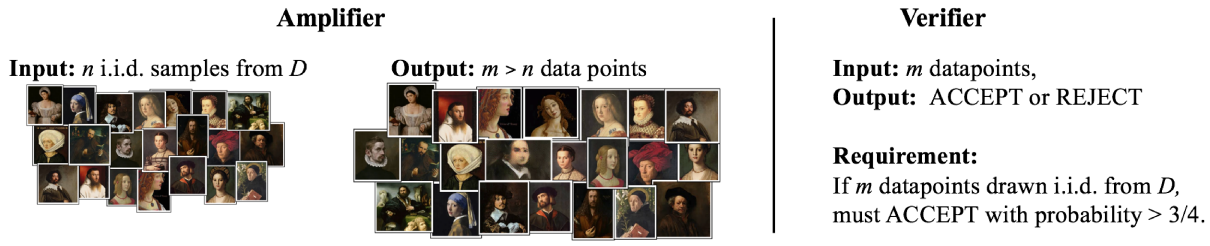

Figure 1: Sample amplification can be viewed as a game between an "amplifier" that obtains n independent draws from an unknown distribution D and must output a set of $m > n$ samples, and a "verifier" that receives the m samples and must ACCEPT or REJECT. The verifier knows the true distribution D and 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) (Cohn, 2018)].

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.

This constant factor amplification above 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 $n+1$ i.i.d. samples is only $O(1/n)$; this is despite the fact that the amplifier can only learn the distribution to TV 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 to small constant total variation distance requires $\Theta(d)$ samples. Unlike the discrete setting, however, we do not get optimal amplification in this setting by generating additional samples from the empirical distribution of the initial set of n samples, and then randomly shuffling together the original and new samples. Here, by empirical distribution, we refer to the Gaussian distribution centered at the empirical mean of the n input samples. Moreover, 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, the procedure that generates new samples from the empirical distribution, and then randomly shuffles together the new and old samples, is able to amplify at $n = \Omega(d/\log d)$, even though learning is not possible until $n = \Theta(d)$. Additionally, as n goes from $10\frac{d}{\log d}$ to $1000\frac{d}{\log d}$, this amplification procedure goes 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 $(n, n+1)$ am-*

plification 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.

The above proposition suggests that to be able to amplify at input size $n = o(d/\log d)$, one must modify the input samples. A naive way to modify the input samples is to discard all the original n samples and generate m new samples from the distribution $N(\hat{\mu}, \Sigma)$, where $\hat{\mu}$ is empirical mean $\hat{\mu}$ of the original set X_n. However this does not even give an (n, n) amplification procedure for any value of n. To achieve optimal amplification in the Gaussian case, the amplifier 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. This procedure gives $(n, n + O(\frac{n}{\sqrt{d}}))$ amplification, and we also show that this is tight up to constant factors.

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 $m \geq n + \frac{cn}{\sqrt{d}}$.

1.3. Open Directions

From a technical perspective, there are a number of natural open directions for future work, including establishing tight bounds on amplification for other natural distribution classes, such as d dimensional Gaussians with unknown mean 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.

Weaker or More Powerful Verifiers? 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 we 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'?*

Rather than considering how to increase the power of the verifier, 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?*

Potential Applications of Sample Amplification. An interesting future direction is to examine if amplification is a useful primitive in settings where the samples are given as input to downstream analysis. Amplification does not add any new information to the original data, but it could still make the original information more easily accessible to certain types of algorithms which interact with the data in limited ways. For example, many popular algorithms and heuristics are not information theoretically optimal, despite their widespread use. It seems worth examining if amplification schemes could improve the statistical efficiency of these commonly used methods. Since the amplified samples are "good" in an information theoretic sense (they are indistinguishable from true samples), the performance of downstream algorithms *cannot* be significantly hurt. Below, we provide a toy example where amplification improves the accuracy of a standard downstream estimator.

Example 2. *Given labeled examples, $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from a distribution, D, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, a natural quantity to estimate is the fraction of variance in y explainable as a linear function of x: $\inf_{\theta \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim D}[(\theta^T x - y)^2]$. The standard unbiased estimator for this quantity is the training error of the least-squares linear model, scaled by a factor of $\frac{1}{n-d}$. This scaling factor makes this estimate unbiased, although the variance is large when n is not much larger than d. Figure 2 shows the expected squared error of this estimator on raw samples, and on $(n, n + 2)$ amplified samples, in*

the case where $x_i \sim N(0, I_d)$, and $y_i = \theta^T x_i + \eta$ for some model $\|\theta\|_2 = 1$ and independent noise $\eta \sim N(0, \frac{1}{4})$—hence the true value for the "unexplainable variance" is $1/4$. Here, the amplification procedure draws two additional datapoints, x from the isotropic Gaussian with mean equal to the empirical mean, and labels them according to the learned least-squares regression model $\hat{\theta}$ with independent noise of variance $5/n$ times the empirical estimate of the unexplained variance.

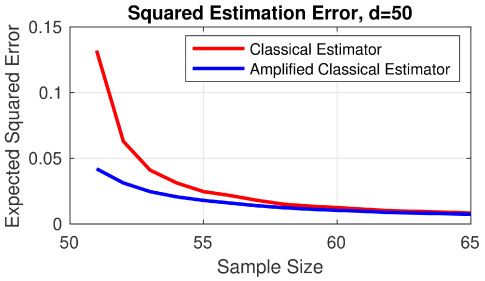

Figure 2: Performance improvement by amplification. See example 2 for a description of the setup.

One potential limitation to applications of amplification is that our existing results show that it is only possible to amplify the sample size by sub-constant factors (for the settings considered). If the algorithm using the amplified data is limited, however, then we could hope for much larger amplification factors. This is reminiscent of the open problem in the previous section on whether larger amplification is possible against weaker classes of verifiers.

In practice, there is already growing interest in using generative models for data augmentation to improve classification accuracy (Antoniou et al., 2017; Frid-Adar et al., 2018; Wang et al., 2018; Yi et al., 2019). Given our results which show that amplification is significantly easier than learning, such pipelines might be more effective than one would initially suspect. It is also worth thinking more generally about how to design modular data analysis or learning pipelines, where a first component of the pipeline could be an amplifier tailored to the specific data distribution, followed by more generic learning algorithms that do not attempt to leverage structural properties of the data distribution. Such modular pipelines might prove to be significantly easier to develop and maintain, in practice.

Implications for Generative Models. The sample amplification framework has some connections to generative modelling. Generative models such as GANs aim to produce new samples from an unknown distribution D given a training set drawn from D. It is tempting to try to relate the amplification setting to GANs by viewing the amplifier and verifier as analogs of the generator and discriminator, respectively. This is *not* an accurate correspondence: For GANs, the discriminator typically evaluates examples individually (or in small batches), and often has seen the same training set as the generator, whereas our verifier explicitly evaluates a full set of samples without knowledge of the training samples. The samples generated by a generative model are often evaluated by humans (either manually or algorithmically). This evaluation is usually aimed at understanding the quality of output samples *conditioned on the training data*—if some of the output samples are copies of the training set, this is not satisfactory—which again corresponds to learning rather than sample amplification.

Despite these differences, some ways that generative models are actually used, do closely mirror the amplification setting. For example, when generative models are used to augment a training set that is used to learn a classifier, both the generated samples and the original dataset are fed into the learning algorithm. The learning algorithm does not necessarily distinguish between "new" and "old" samples. In this setting, it does make sense to evaluate the set of "new" and "old" samples together, as a single set, rather than evaluating the "new" samples conditioned on the "old" ones. This exactly corresponds to our amplification formulation. As amplification is often easier than learning, it might be worthwhile trying to develop more techniques that are explicitly trying to amplify, rather than learn.

A second, distinct connection between amplification and GANs, relates to the question of how humans can evaluate the samples produced by a GAN. The gap between learning (evaluating the generated samples conditioned on the training set), and amplification (evaluating the generated samples without knowing the training set), suggests that in order to truly evaluate the samples produced by a GAN, we would need to closely inspect the training data used by the GAN. This is clearly impractical in many settings, and motivates some of the questions described above concerning how much access a verifier needs to the input data in order for there to be a gap between learning, and amplifying.

1.4. Related Work

The question of deciding whether a set of samples consists of independent draws from a specified distribution D—known as *identity testing*—is one of the fundamental problems at the core of distributional property testing (Goldreich & Ron, 2000; Batu et al., 2001; Paninski, 2008; Valiant & Valiant, 2017; Diakonikolas & Kane, 2016; Batu et al., 2013; Valiant, 2011; Chan et al., 2014; Orlitsky & Suresh, 2015; Bhattacharya & Valiant, 2015; Levi et al., 2013; Diakonikolas & Kane, 2016)). In the majority of these works, the assumption is that the samples are i.i.d. draws from some fixed distribution, and the common theme in these results is that these hypothesis tests can be accomplished with far less data than would be required to learn the distribution. While the identity testing problem is clearly related to the amplification problem we consider, these appear to

be quite distinct problems. In the amplification setting, the core 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 i.i.d. draws from D.

Within this line of work on distributional property testing and estimation, there is also a recent thread of work on designing estimators or tests, whose performance given n i.i.d. samples is comparable to the expected performance of a naive "plugin" estimator (which returns the property value of the empirical distribution) based on $m > n$ i.i.d. draws (Valiant & Valiant, 2016; Yi et al., 2018). The term "data amplification" has been applied to this line of work, although it is a different problem from the one we consider.

The recent work on *sampling correctors* (Canonne et al., 2018) also considers the question of how to produce a "good" set of draws from a given distribution. That work assumes access to 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).

Our formulation of sample amplification as a game between an amplifier and a verifier, closely resembles the setup for *pseudo-randomness* (see (Vadhan et al., 2012) 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. 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.

Finally, it is also worth mentioning the work of Viola on the complexity of sampling from distributions (Viola, 2012). 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 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.

2. Algorithms and Proof Overview

In this section, we describe our sample amplification algorithms for the discrete and Gaussian settings, and give an overview of their analyses. The full proofs of the upper and lower bounds are provided in the supplementary material.

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 extend this intuition to general discrete distributions.

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. 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 it 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 them with the original samples. Our amplification procedure for general discrete distributions is based on this sample-from-empirical procedure.

Algorithm and Upper Bound. To facilitate the analysis, our general amplification procedure which applies to any discrete distribution D, deviates from the sample-from-empirical-then-shuffle scheme in two ways. The modifications avoid two sources of dependencies in the sample-from-empirical-then-shuffle schemes. 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 sec-

ond, note that the new samples are dependent on the old samples if we generate the new samples from the empirical distribution. To leverage independence, we instead (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 the supplementary. We show that this procedure achieves $(n, n + O\left(\frac{n}{\sqrt{k}}\right))$ amplification.

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. Since 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.

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. 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 $n = \Omega(\sqrt{d})$.

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 am-

plification 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.

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 $n \leq \frac{cd}{\log d}$.

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 know 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 does 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) $\forall\ i \in [n+1], \langle x'_i - \hat{\mu}_{-i}, \mu - \hat{\mu}_{-i} \rangle \geq \frac{d}{n} - \sqrt{\frac{d \log n}{n}}$, and (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 $n \ll \frac{d}{\log d}$.

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 $n = o(d/\log d)$. What would be the most naive amplification scheme which does not output a superset of the input samples? One candidate could be an amplifier which first estimates the sample mean $\hat{\mu}$ of X_n, and then just outputs m samples from $N(\hat{\mu}, I)$. It is not hard to see that this scheme does not even give a valid (n, n) amplification procedure. The verifier in this case could check the distance between the true mean and

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

Input: $X_n = (x_1, x_2, \ldots, x_n)$, where $x_i \leftarrow N(\mu, \Sigma_{d\times d})$.
Output: $Z_m = (x'_1, x'_2, \ldots, x'_m)$, 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 $i \in \{n+1, n+2, \ldots, m\}$ ▷ Draw $m-n$ i.i.d samples from $N(0, \Sigma_{d\times d})$
$x'_i := \hat{\mu} + \epsilon_i$, for $i \in \{n+1, n+2, \ldots, m\}$
$x'_i := x_i - \sum_{j=n+1}^{m} \frac{\epsilon_j}{n}$, for $i \in \{1, 2, \ldots, n\}$ ▷ Remove correlations between old and new samples
return $Z_m := (x'_1, x'_2, \ldots, x'_m)$

the mean of the returned samples, which would be significantly more than expected, with high probability.

How should one modify the input samples then? The above lower bound also shows what such an amplification procedure 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 $m = n + O\left(\frac{n}{\sqrt{d}}\right)$.

We now provide a brief proof sketch for this upper bound, for the case when $m = n + 1$. Note that the returned samples in Z_m can also be thought of as a single sample from a $(m \times d)$-dimensional Gaussian distribution $N\Big(\underbrace{(\mu, \mu, \ldots, \mu)}_{m \text{ times}}, \tilde{\Sigma}_{md\times md}\Big)$, 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 $N\Big((\mu, \mu, \ldots, \mu), I_{md\times md}\Big)$). The TV distance between the two distributions 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.

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 intuition behind the lower bound is that any such amplification procedure could be used to find the true mean μ with much smaller error than what is possible with n samples.

To show this formally, 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. In this case, the verifier only needs to evaluate the squared distance $\|\mu - \hat{\mu}_m\|^2$ of the empirical mean $\hat{\mu}_m$ of the returned samples from the true mean μ, and accept the samples if and only if this squared distance is less than $\frac{d}{m} + \frac{c_1\sqrt{d}}{m}$ for some fixed constant c_1. It is not difficult to see why this test is sufficient. Note that for m true samples drawn from $N(\mu, I)$, $\|\mu - \hat{\mu}_m\|^2 = \frac{d}{m} \pm O\left(\frac{\sqrt{d}}{m}\right)$. Also, the squared distance $\|\mu - \hat{\mu}^2\|$ of the mean $\hat{\mu}$ of the original set X_n from the true mean μ is concentrated around $\frac{d}{n} \pm O\left(\frac{\sqrt{d}}{n}\right)$. Using this, for $m > n + \frac{cn}{\sqrt{d}}$, we can show that no algorithm can find a $\hat{\mu}_m$ which satisfies $\|\mu - \hat{\mu}_m\|^2 \leq \frac{d}{m} \pm O\left(\frac{\sqrt{d}}{m}\right)$ with decent probability over $\mu \leftarrow N(0, \sqrt{d}I)$. This is because the algorithm only knows μ up to squared error $\frac{d}{n} \pm O\left(\frac{\sqrt{d}}{n}\right)$ based on the original set X_n.

3. Conclusion

We introduce the notion of sample amplification, which formalizes what it means to enlarge a dataset that consists of independent draws from an unknown distribution, D. For two fundamental classes of distributions—discrete distributions and high dimensional Gaussians—we show that non-trivial amplification is possible even when one does not have enough data to learn D. Beyond these results, we present a toy example illustrating one potential application of sample amplification, and outline several intriguing directions of future work in this vein. We believe that further exploration of sample amplification may inform how we view and evaluate generative models.

Acknowledgements

We would like to thank Clément Canonne and the anonymous reviewers for their comments and feedback. This work was supported by an NSF fellowship, NSF awards 1804222, 1813049, IIS1908774 and 1704417, DOE award DE-SC0019205 and an ONR Young Investigator Award.

References

Antoniou, A., Storkey, A., and Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

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

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., and White, P. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

Bhattacharya, B. and Valiant, G. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, pp. 2611–2619, 2015.

Canonne, C. L., Gouleakis, T., and Rubinfeld, R. Sampling correctors. *SIAM Journal on Computing*, 47(4):1373–1423, 2018.

Chan, S.-O., Diakonikolas, I., Valiant, P., and Valiant, G. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1193–1203. SIAM, 2014.

Cohn, G. AI art at Christie's sells for $432,500. *The New York Times*, Oct 2018.

Diakonikolas, I. and Kane, D. M. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 685–694. IEEE, 2016.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

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

Levi, R., Ron, D., and Rubinfeld, R. Testing properties of collections of distributions. *Theory of Computing*, 9(1): 295–347, 2013.

Orlitsky, A. and Suresh, A. T. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pp. 2143–2151, 2015.

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

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

Valiant, G. and Valiant, P. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 142–155. ACM, 2016.

Valiant, G. and Valiant, P. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

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

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

Wang, Y.-X., Girshick, R., Hebert, M., and Hariharan, B. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7278–7286, 2018.

Yi, H., Orlitsky, A., Suresh, A. T., and Wu, Y. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pp. 8848–8857, 2018.

Yi, X., Walia, E., and Babyn, P. Generative adversarial network in medical imaging: A review. *Medical image analysis*, pp. 101552, 2019.