Sample Amplification: Increasing Dataset Size even when Learning is Impossible (Supplementary Material)

A. Proofs: Gaussian with Unknown Mean and Fixed Covariance

A.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, $Y_n = (y_1, y_2, \dots, y_n)$ denote the random variable corresponding to n samples from $N(\mu, \Sigma)$. Let $X_n = (x_1, x_2, \dots, x_n)$ denote n samples from $N(\mu, I)$, such that $Y_n = \Sigma^{\frac{1}{2}}(X_n - \mu) + \mu = (\Sigma^{\frac{1}{2}}(x_1 - \mu) + \mu, \Sigma^{\frac{1}{2}}(x_2 - \mu) + \mu, \dots, \Sigma^{\frac{1}{2}}(x_n - \mu) + \mu)$. 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 1. *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 $\epsilon_{n+1}, \epsilon_{n+2}, \dots, \epsilon_m$ 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 $x'_1, \dots, x'_m$ with:

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

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 the 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

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

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

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

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

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

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 (**?**)Theorem 1.1]devroye2018total, 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

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

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 (3), we get

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

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. $\square$

A.2. Lower Bound

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

Proposition 2. *Let $\mathcal{C}$ denote the class of $d-$dimensional Gaussian distributions $N(\mu, I)$ 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. Note that it is sufficient to prove the theorem for $m = n + cn/\sqrt{d}$ for a fixed constant c, as an amplification procedure for $m > n + cn/\sqrt{d}$ implies an amplification procedure for $m = n + cn/\sqrt{d}$ by discarding the residual samples. To prove the theorem for $m = n + cn/\sqrt{d}$, we will define a distribution D_μ over μ and a verifier $v(Z_m)$ for the distribution $N(\mu, I)$ which takes as input a set Z_m of m samples, such that: (i) for all μ, the verifier $v(Z_m)$ will accept with probability $1 - 1/e^2$ when given as input a set Z_m of m i.i.d. samples from $N(\mu, I)$, (ii) but will reject any (n, m) amplification procedure for $m = n + cn/\sqrt{d}$ 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, m) amplification procedure for $m = n+cn/\sqrt{d}$.

We now define the distribution D_μ and the verifier $v(Z_m)$. We choose D_μ to be $N(0, \sqrt{d}I)$. Let $\hat{\mu}_m$ be the mean of the samples Z_m returned by the amplification procedure. The verifier $v(Z_m)$ performs the following test, accepts if $\hat{\mu}_m$ passes the test, and rejects otherwise—

$$\Big| \|\hat{\mu}_m - \mu\|^2 - d/m \Big| \leq 10\sqrt{d}/m. \tag{4}$$

We first show that m i.i.d. samples from $N(\mu, I)$ pass the above test 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 (**??**),

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

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

Note that $\hat{\mu}_m \leftarrow N(\mu, \frac{I}{m})$ for m i.i.d. samples from $N(\mu, I)$. Hence by using (6) and setting $t = 10/\sqrt{d}$,

$$\Pr\Big[\Big|\|\hat{\mu}_m - \mu\|^2 - d/m\Big| > 10\sqrt{d}/m\Big] \leq 1/e^3.$$

Hence m i.i.d. samples from $N(\mu, I)$ pass the test with probability at least $1 - 1/e^2$.

We now show that for μ sampled from $D_\mu = N(0, \sqrt{d}I)$, the verifier rejects any (n, m) amplification procedure for $m = n + cn/\sqrt{d}$ 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_m 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_m which is accepted by the verifier, completing the proof of Proposition 2.

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, (**?**)), 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 show that any set Z_m returned by the amplifier for $m = n + 100n/\sqrt{d}$ fails the test (4) with probability $1 - 1/e^2$ over the randomness in $\mu \mid X_n$. We expand $\|\hat{\mu}_m - \mu\|^2$ in the test as follows,

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

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

$$\|\hat{\mu}_m - \mu\|^2 \geq \|\hat{\mu}_m - \bar{\mu}\|^2 - (20/\sqrt{n})\|\hat{\mu}_m - \bar{\mu}\| + d/n - 12\sqrt{d}/n.$$

We claim that $\|\hat{\mu}_m - \bar{\mu}\|^2 - 20\|\hat{\mu}_m - \bar{\mu}\|/\sqrt{n} \geq -100/n$. To verify, note that $\|\hat{\mu}_m - \bar{\mu}\|^2 - 20\|\hat{\mu}_m - \bar{\mu}\|/\sqrt{n} + 100/n$ is a non-negative quadratic function in $\|\hat{\mu}_m - \bar{\mu}\|$. Therefore, with probability at least $1 - 2/e^3$,

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

To pass (4), $\|\hat{\mu}_m - \mu\|^2 \leq d/m + 10\sqrt{d}/m$. Therefore, if an amplifier passes the test with probability greater than $1 - 2/e^3$ over the randomness in $\mu \mid X_n$ for $m = n + 100n/\sqrt{d}$, then,

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

which is a contradiction. Hence for $m = n + 100n/\sqrt{d}$, 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.

□

A.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 1. 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

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

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 1, 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 3. *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 1 Sample Amplification for Gaussian with Unknown Mean and Fixed Covariance Without Modifying Input Samples

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 AMPLIFYGAUSSIAN2(X_n)
  $r := m - n$
  $\hat{\mu} := \sum_{i=1}^{\frac{n}{2}} \frac{x_i}{n/2}$
  $x'_i := x_i$, for $i \in \{1, 2, \ldots, \frac{n}{2}\}$
  $X_{\text{remaining}} := (x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \ldots, x_n)$
  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
  $Z_m := (x'_1, x'_2, \ldots, x'_m)$
  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 $Z_m = (x'_1, x'_2, \ldots, x'_m)$, given n i.i.d. samples $X_n = (x_1, x_2, \ldots, x_n)$ 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 $Z'_m = (x''_1, x''_2, \ldots, x''_m)$, given n i.i.d samples X_n. For $i \in \{1, 2, \ldots, \frac{n}{2}\}$, we set $x''_i = x_i$. For $i \in \{\frac{n}{2} + 1, \frac{n}{2} + 2, \ldots, m\}$, 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 $i \in \{1, 2, \ldots, \frac{n}{2}\}$. For $i \in \{\frac{n}{2} + 1, \frac{n}{2} + 2, \ldots, m\}$, we use samples from $(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \ldots, x_n)$ 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 $\left(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \ldots, x_n\right)$, 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 $(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \ldots, x_n)$ 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 $\left(x_{\frac{n}{2}+1}, x_{\frac{n}{2}+2}, \ldots, x_n\right)$ 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, $D_{TV}(X_m, Z_m) \leq D_{TV}(X_m, Z'_m) + D_{TV}(Z'_m, Z_m)$.

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.

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

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

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

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

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

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

We first estimate $\|\tilde{\mu} - \mu\|$. Note that $\mathbb{E}_{Z^{1'}_m}[\|\tilde{\mu} - \mu\|^2] = \frac{2d}{n}$, 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 (**?**, Example 2.5).

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

This gives us $\Pr(|\frac{n}{2}\|\tilde{\mu}-\mu\|^2 - d| \geq 0.5d) \leq 2e^{-d/32}$, that is, $\|\tilde{\mu}-\mu\| \leq \sqrt{\frac{3d}{n}} \leq \sqrt{3\epsilon\log d}$ 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

$$
\begin{aligned}
&H\left(\Pi_{i=1}^{\frac{n}{2}+r} N(\mu, I_{d\times d}), \Pi_{i=1}^{\frac{n}{2}+r}\left(1-\frac{10r}{\frac{n}{2}+r}\right)N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r}N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2 \\
&\leq \left(\frac{n}{2}+r\right)\ H\left(N\left(\mu, I_{d\times d}\right), \left(1-\frac{10r}{\frac{n}{2}+r}\right)N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r}N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2.
\end{aligned} \tag{7}
$$

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

$$
\begin{aligned}
H\left(N(\mu, I_{d\times d}), \left(1-\frac{10r}{\frac{n}{2}+r}\right)N\left(\mu, I_{d\times d}\right) + \frac{10r}{\frac{n}{2}+r}N\left(\tilde{\mu}, I_{d\times d}\right)\right)^2 &\leq \frac{576r^2}{n^2}e^{3\|\tilde{\mu}-\mu\|^2} \\
&\leq \frac{576r^2d^{9\epsilon}}{n^2},
\end{aligned} \tag{8}
$$

with probability at least $1-2e^{-d/32}$ over $\tilde{\mu}$. From (7) and (8), we get that with probability at least $1-2e^{-d/32}$ over $\tilde{\mu}$,

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

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 $\tilde{\mu}$,

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

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

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

For $\delta > 2(2e^{-d/32} + e^{-81r/26})$, and for $r \leq \frac{n^{\frac{1}{2}-9\epsilon}\delta}{48\sqrt{2}}$, we get $D_{TV}(X_m, Z_m) \leq \delta$. For d large enough, setting $\delta = \frac{1}{3}$ and $r \leq \frac{n^{\frac{1}{2}-9\epsilon}}{144\sqrt{2}}$, 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 $r \leq \frac{n}{18}$, $H\left(P, \left(1-\frac{10r}{r+\frac{n}{2}}\right)P + \frac{10r}{r+\frac{n}{2}}Q\right) \leq \frac{24r}{n}e^{\frac{3\|\hat{\mu}\|^2}{2}}$.*

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, $H\left(P,\left(1-\frac{10r}{r+\frac{n}{2}}\right)P+\frac{10r}{r+\frac{n}{2}}Q\right)^2$

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

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 $H\left(P_1,\left(1-\frac{10r}{r+\frac{n}{2}}\right)P_1+\frac{10r}{r+\frac{n}{2}}Q_1\right)^2=\int_{-\infty}^{\infty}\left(\sqrt{p_1}-\sqrt{q_1}\right)^2dx$

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

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

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

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

   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

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

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

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

   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. □

□

A.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 4. *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 2. As in the proof of Proposition 2, 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,

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

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 2, $D_{\mu|X_n} = N(\bar{\mu}, \bar{\sigma}^2 I)$, 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 2. 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 $1 - 2/e^3$,

$$\begin{aligned}\langle x_1' - \hat{\mu}_{-1}, \mu - \hat{\mu}_{-1}\rangle &\leq 10\|x_1' - \bar{\mu}\|/\sqrt{n} + 10\|\hat{\mu}_{-1} - \bar{\mu}\|/\sqrt{n} + \|x_1' - \bar{\mu}\|\|\hat{\mu}_{-1} - \bar{\mu}\| + \|\hat{\mu}_{-1} - \bar{\mu}\|^2, \\ &\leq 100\sqrt{\frac{d}{n}} + 300\frac{d^{3/4}}{n^2} + 300\frac{d^{3/4}}{n} + 900\frac{\sqrt{d}}{n^2} \\ &\leq 100\sqrt{\frac{d}{n}} + 1500\frac{d^{3/4}}{n} \\ &= 100\sqrt{\frac{n}{d}}\left(\frac{d}{n}\right) + \frac{1500}{d^{1/4}}\left(\frac{d}{n}\right).\end{aligned}$$

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 $\Pr\left[\|\mu_0\| \leq 30d^{3/4}\right] \geq 1 - e^{-3d}$, the overall success probability of the amplifier over the randomness in μ, X_n and any internal randomness of the amplifier is at most $1/e^2$. □

B. Proofs: Discrete Distributions with Bounded Support

B.1. Upper Bound

In this section we prove the upper bound from Theorem 1. The algorithm itself is presented in Algorithm 2. For clarity of writing, we assume that the number of input samples is $4n$, instead of n.

Proposition 5. *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$(2n + r)$.

Algorithm 2 Sample Amplification for Discrete Distributions

Input: $X_{4n} = (x_1, x_2, \dots, x_{4n})$, where $x_i \leftarrow D$, for any discrete distribution D over $[k]$.
Output: $Z_m = (x'_1, x'_2, \dots, x'_m)$, 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**
    $X_{N_1} := (x_1, x_2, \dots, x_{N_1})$
    $X_{N_2} := (x_{N_1+1}, x_{N_1+2}, \dots, x_{N_1+N_2})$
  else ▷ Uninteresting case: happens with low probability
    $X_{N_1} := \underbrace{(x_1, x_1, \dots, x_1)}_{N_1 \text{ times}}$
    $X_{N_2} := \underbrace{(x_1, x_1, \dots, x_1)}_{N_2 \text{ times}}$
  $r := 8(m-n)$
  $(x'_1, x'_2, \dots, x'_{N+R})$ = 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
    $(x'_1, x'_2, \dots, x'_{N+R_1}) := (x'_1, x'_2, \dots, x'_{N+R}, \underbrace{x_1, x_1, \dots, x_1}_{\frac{r}{8} - R \text{ times}})$
  $(x'_{N+R_1+1}, x'_{N+R_1+2}, \dots, x'_m) := (x_{N+1}, x_{N+2}, \dots, x_{4n-(R_1-\frac{r}{8})})$
  Add the remaining samples to get $4n + r/8$ samples in total
  $Z_m := (x'_1, x'_2, \dots, x'_m)$
  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
  $X_{N_1} = (x_1, x_2, \dots, x_{N_1})$, $X_{N_2} = (x_{N_1+1}, x_{N_1+2}, \dots, x_{N_1+N_2})$, 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 $j \in [k]$
  $R := \sum_{j=1}^{k} \hat{z}_j$
  $(x'_1, x'_2, \dots, x'_{N_1}) := (x_1, x_2, \dots, x_{N_1})$
  $(x'_{N_1+1}, \dots, x'_{N_1+N_2+R}) :=$ RandomPermute$((x_{N_1+1}, x_{N_1+2}, \dots, x_{N_1+N_2}, \underbrace{1, 1, \dots, 1}_{\hat{z}_1 \text{ times}}, \dots, \underbrace{k, k, \dots, k}_{\hat{z}_k \text{ times}}))$
  return $(x'_1, x'_2, \dots, x'_{N_1+N_2+R})$

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 $Z_{\tilde{M}}$.

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—

$$D_{KL}(P \parallel Q) = \lambda_1 \log\left(\frac{\lambda_1}{\lambda_2}\right) + \lambda_2 - \lambda_1. \tag{9}$$

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,

$$p_i(n+r)\log\left(\frac{1}{1+r\delta_i/(np_i(n+r))}\right) \leq -r\delta_i/n + \frac{2r^2\delta_i^2}{n^2p_i(n+r)},$$
$$\implies D_{KL}(v_i \parallel \hat{v}_i) \leq \frac{2r^2\delta_i^2}{n^2p_i(n+r)},$$
$$\implies \sum_{i=1}^{k'} D_{KL}(v_i \parallel \hat{v}_i) \leq \frac{2r^2}{n^2}\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i}. \tag{10}$$

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

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

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

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

Hence by Chebyshev's inequality,

$$\Pr\left[\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i} - k' \geq 4k/\epsilon\right] \leq \epsilon/4,$$
$$\implies \Pr\left[\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i} \geq 4k/\epsilon\right] \leq \epsilon/4. \tag{11}$$

Let E_1 be the event that $\sum_{i=1}^{k'}\frac{\delta_i^2}{np_i} \leq 4k/\epsilon$. By (11), $\Pr(E_1) \geq 1-\epsilon/4$. Conditioned on the event E_1 and using (10), 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 $u_i \sim \text{Poisson}(np_i)$ 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 $z_i \sim \text{Poisson}(rp_i)$ 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,

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

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 $\text{Poisson}(n + r)$ 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 $\tilde{M} = N + R$ 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 $N'_1, N'_2 \leftarrow \text{Poisson}(n)$, 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 $D_{TV}(f(X_N), f(X_{N'}))$.

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

  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 $N \leq 4n$ and $N' \leq 4n$. Also, $\Pr(N \leq 4n) = \Pr(N' \leq 4n)$ and $\Pr(N > 4n) = \Pr(N' > 4n)$, as both N and N' are drawn from Poisson$(2n)$ distribution. This gives us

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

  Using the triangle inequality, we get $D_{TV}(X_M, f(X_N')) \leq \Pr(N > 4n) + 3\epsilon/2$. To bound $\Pr(N > 4n)$, we use the following Poisson tail bound (**?**): for $X \leftarrow \text{Poisson}(\lambda)$,

$$
\Pr[X \geq \lambda + x], \Pr[X \leq \lambda - x] \leq e^{\frac{-x^2}{\lambda + x}}. \tag{12}
$$

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

- **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 triangle 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 $X_{N'}$.

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

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

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

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 $(r\hat{p}_i)$. 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 (12), we get

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

$$\Pr\left(N_1' < \frac{3n}{4}\right) \leq exp\left(-\frac{(n/4)^2}{n + n/4}\right) = e^{-n/20}$$

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! □

B.2. Lower Bound

In this section we show that the above procedure is optimal, up to constant factors for amplifying samples from discrete distributions. We first 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.

Proposition 6. *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 6, 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 (**?**).

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

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

Then, we have

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

Similarly the following holds (though not simultaneously):

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

Finally we rely on slight generalization of the birthday paradox which can be found in (**?**).

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 $n \leq \frac{k}{2}$. 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 $m = n + 30\frac{n}{\sqrt{k}} \leq \frac{k}{4}$ 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 $n \leq \frac{\sqrt{k}}{2}$. 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 $\frac{(3/4)k}{8k-n} \leq \frac{(3/4)k}{7.5k} \leq \frac{1}{8}$. 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 6.

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. $\square$