ON THE STATISTICAL COMPLEXITY OF SAMPLE AMPLIFICATION

BY BRIAN AXELROD[1,a], SHIVAM GARG[2,c], YANJUN HAN[3,d], VATSAL SHARAN[4,e], AND GREGORY VALIANT[1,b]

[1]*Department of Computer Science, Stanford University;* [a]*baxelrod@cs.stanford.edu;* [b]*valiant@cs.stanford.edu*

[2]*Microsoft Research;* [c]*shigarg@microsoft.com*

[3]*Courannt Institute of Mathematical Sciences and Center for Data Science, New York University;* [d]*yanjunhan@nyu.edu*

[4]*Department of Computer Science, University of Southern California;* [e]*vsharan@usc.edu*

The "sample amplification" problem formalizes the following question: Given n i.i.d. samples drawn from an unknown distribution P, when is it possible to produce a larger set of $n + m$ samples which cannot be distinguished from $n + m$ i.i.d. samples drawn from P? In this work, we provide a firm statistical foundation for this problem by deriving generally applicable amplification procedures, lower bound techniques and connections to existing statistical notions. Our techniques apply to a large class of distributions including the exponential family, and establish a rigorous connection between sample amplification and distribution learning.

1. Introduction. Consider the following problem of manufacturing more data: an amplifier is given n samples drawn i.i.d. from an unknown distribution P, and the goal is to generate a larger set of $n + m$ samples which are indistinguishable from $n + m$ i.i.d. samples from P. How large can m be as a function of n and the distribution class in question? Are there sound and systematic ways to generate a larger set of samples? Is this task strictly *easier* than the learning task, in the sense that the number of samples required for generating $n + 1$ samples is smaller than that required for learning P?

In our preliminary work [2], we formalized this problem as the *sample amplification* problem, considering total variation (TV) as the measure for indistinguishability.

DEFINITION 1.1 (Sample Amplification). *Let $\mathcal{P}$ be a class of probability distributions over a domain $\mathcal{X}$. We say that $\mathcal{P}$ admits an $(n, n + m, \epsilon)$ sample amplification procedure if there exists a (possibly randomized) map $T_{\mathcal{P},n,m,\epsilon} : \mathcal{X}^n \to \mathcal{X}^{n+m}$ such that*

$$(1.1) \qquad \sup_{P \in \mathcal{P}} \|P^{\otimes n} \circ T_{\mathcal{P},n,m,\epsilon}^{-1} - P^{\otimes(n+m)}\|_{\mathrm{TV}} \le \epsilon.$$

An equivalent formulation to view Definition 1.1 is as a game between two parties: an amplifier and a verifier. The amplifier gets n samples drawn i.i.d. from the unknown distribution P in the class $\mathcal{P}$, and her goal is to generate a larger dataset of $n + m$ samples which must be accepted with good probability by any verifier that also accepts a real dataset of $n + m$ i.i.d. samples from P with good probability. Here, the verifier is computationally unbounded and knows the distribution P, but does not observe the amplifier's original set of n samples.

Along with being a natural statistical task, the sample amplification framework is also relevant from a practical standpoint. Currently, there is an enormous trend in the machine learning community to train models on datasets that have been enlarged in various ways. There are relatively transparent and classical approaches to achieve this, such as leveraging

Distribution Class	Amplification	Procedure
Gaussian with unknown mean and fixed covariance (UB: Example 4.1, 4.2, A.8; LB: Theorem 6.2, 6.5)	$(n, n + \Theta(n\epsilon/\sqrt{d}))$	Sufficiency/Learning
Gaussian with unknown mean and covariance (UB: Example A.1, A.3; LB: Example A.20)	$(n, n + \Theta(n\epsilon/d))$	Sufficiency
Gaussian with s-sparse mean and identity covariance (UB: Example A.12; LB: Example A.18)	$(n, n + \Theta(n\epsilon/\sqrt{s\log d}))$	Learning
Discrete distributions with support size at most k (UB: Example A.9; LB: [2, Theorem 1])	$(n, n + \Theta(n\epsilon/\sqrt{k}))$	Learning
Poissonized discrete distributions with support at most k (UB: Example A.16; LB: Example A.16)	$(n, n + \Theta(\sqrt{n}\epsilon + n\epsilon/\sqrt{k}))$	Learning
d-dim. product of Exponential distributions (UB: Example A.5, A.11; LB: Theorem 6.5)	$(n, n + \Theta(n\epsilon/\sqrt{d}))$	Sufficiency/Learning
Uniform distribution on d-dim. rectangle (UB: Example A.6, A.10; LB: Theorem 6.5)	$(n, n + \Theta(n\epsilon/\sqrt{d}))$	Sufficiency/Learning
d-dim. product of Poisson distributions (UB: Example A.14; LB: Theorem 6.5)	$(n, n + \Theta(n\epsilon/\sqrt{d}))$	Sufficiency+Learning

TABLE 1

Sample amplification achieved by the presented procedures. Results include matching upper bounds (UB) and lower bounds (LB), with appropriate pointers to specific examples or theorems for details.

known invariances such as rotation or translation invariance to augment the dataset by including transformed versions of each original datapoint [22, 23, 29, 44, 45]. More recently, deep generative models have been used to both directly enlarge training data and construct larger datasets consisting of samples with properties that are rare in naturally occurring datasets [1, 3, 18, 19, 20, 21, 25, 26, 35, 36, 38, 43, 53, 54]. More opaque approaches such as MixUp [56] and related techniques [28, 47, 50, 55] which add a significant fraction of new datapoints that are explicitly *not* supported in the true data distribution are also very popular since they seem to improve the performance of the trained models. Given this current zoo of widely implemented approaches to enlarging datasets, there is a clear motivation for bringing a more principled statistical understanding to such approaches. One natural starting point is the statistical setting we consider that asks the extent to which datasets can be enlarged in a perfect sense—where it is not possible to distinguish the enlarged dataset from a set of i.i.d. draws. Moreover, this work lays a foundation for the ambitious broader goal of understanding how various amplification techniques interact with downstream learning algorithms and statistical estimators, and developing amplification techniques that are optimal for certain classes of such algorithms and estimators.

In [2], a subset of the authors introduced the sample amplification problem, and studied two classes of distributions: the Gaussian location model and discrete distribution model. For these examples, they characterized the statistical complexity of sample amplification and showed that it is strictly smaller than that of learning. In this paper, we work towards a general understanding of the statistical complexity of the sample amplification problem, and its relationship with learning. The main contributions of this paper are as follows:

1. **Amplification via sufficiency.** Our first contribution is a simple yet powerful procedure for sample amplification, i.e. apply the sample amplification map only to sufficient statistics. Specifically, the learner computes a sufficient statistic T_n from X^n, maps T_n properly to some T_{n+m}, and generates new samples $\widehat{X}^{n+m}$ from some conditional distribution conditioned on T_{n+m}. Surprisingly, this simple idea leads to a much cleaner pro-

cedure than [2] under Gaussian location models which is also exactly optimal (cf. Theorem 6.2). The range of applicability also extends to general exponential families, with rate-optimal sample amplification performances. Specifically, for general d-dimensional exponential families with a mild moment condition, the sufficiency-based procedure achieves an $(n, n + O(n\epsilon/\sqrt{d}), \epsilon)$ sample amplification for large enough n, which by our matching lower bounds in Section 6 is asymptotically minimax rate-optimal.

2. **Amplification via learning.** Our second contribution is another general sample amplification procedure that does not require the existence of a sufficient statistic, and also sheds light on the relationship between learning and sample amplification. This procedure essentially draws new samples from a rate-optimal estimate of the true distribution, and outputs a random permutation of the old and new samples. The procedure achieves an $(n, n + O(\epsilon\sqrt{n/r_{\chi^2}(\mathcal{P}, n)}), \epsilon)$ sample amplification, where $r_{\chi^2}(\mathcal{P}, n)$ denotes the minimax risk for learning $P \in \mathcal{P}$ under the expected χ^2 divergence given n samples. This shows that learning P to χ^2 divergence $O(n/\epsilon^2)$ is sufficient for non-trivial sample amplification.

In addition, we show that for the special case of product distributions, it is important that the permutation step be applied coordinatewise to achieve the optimal sample amplification. Specifically, if $\mathcal{P} = \prod_{j=1}^{d} \mathcal{P}_j$, this procedure achieves a better sample amplification

$$\left(n, n + O\left(\epsilon\sqrt{\frac{n}{\sum_{j=1}^{d} r_{\chi^2}(\mathcal{P}_j, n)}}\right), \epsilon\right).$$

We have summarized several examples in Table 1 where the sufficiency and/or learning based sample amplification procedures are optimal. Note that there is no golden rule for choosing one idea over the other, and there exists an example where the above two ideas must be combined.

3. **Minimax lower bounds.** Complementing our sample amplification procedures, we provide a general recipe for proving lower bounds for sample amplification. This recipe is intrinsically different from the standard techniques of proving lower bounds for hypothesis testing, for the sample amplification problem differs significantly from an estimation problem. In particular, specializing our recipe to product models shows that, for essentially *all* d-dimensional product models, an $(n, n + Cn\epsilon/\sqrt{d}, \epsilon)$ sample amplification is impossible for some absolute constant $C < \infty$ *independent of* the product model.

For non-product models, the above powerful result does not directly apply, but proper applications of the general recipe could still lead to tight lower bounds for sample amplification. Specifically, we obtain matching lower bounds for all examples listed in Table 1, including the non-product examples.

We now provide several numerical simulations to suggest the potential utility of sample amplification. Recall that a practical motivation of sample amplification is to produce an enlarged dataset that can be fed into a *distribution-agnostic* algorithm in downstream applications. Here, we consider the following basic experiments in that vein:

- Fourth moment estimation for one-dimensional Gaussian: here we observe $X_1, \cdots, X_n \sim \mathcal{N}(\mu, 1)$ with $n = 100$ and $\mu = 1$, and we consider three estimators. The empirical estimator operates in a distribution-agnostic fashion and is simply the empirical fourth moment $n^{-1}\sum_{i=1}^{n} X_i^4$. The plug-in estimator uses the knowledge of Gaussianity: it first estimates $\widehat{\mu} = \bar{X}$ and then uses $\mathbb{E}_{X \sim \mathcal{N}(\widehat{\mu}, 1)}[X^4] = \widehat{\mu}^4 + 6\widehat{\mu}^2 + 3$. The amplified estimator first amplifies the sample X^n into Y^{n+m} via sufficiency (cf. Example 4.1), and then uses the empirical estimator $(n + m)^{-1}\sum_{j=1}^{n+m} Y_j^4$ based on the enlarged sample Y^{n+m}. The plots of the mean absolute errors (MAEs) are displayed in Figure 1a. We observe that although
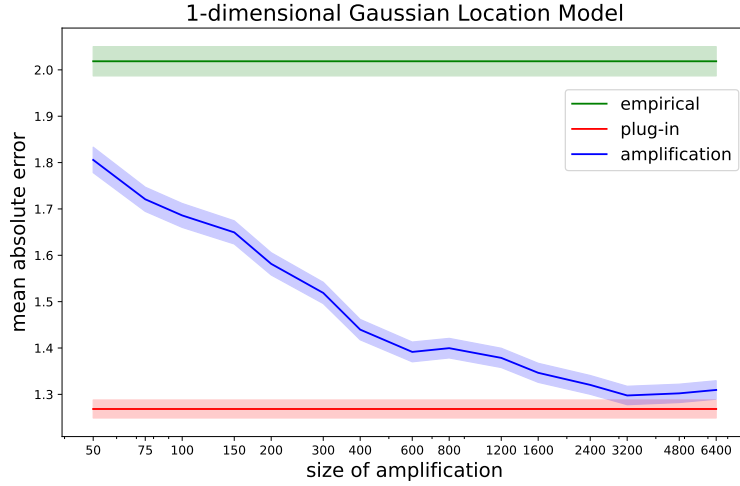
(a) Mean absolute errors for estimation of the fourth moment $\mathbb{E}[X^4]$ in a one-dimensional Gaussian location model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, 1)$ with $n = 100, \mu = 1$.
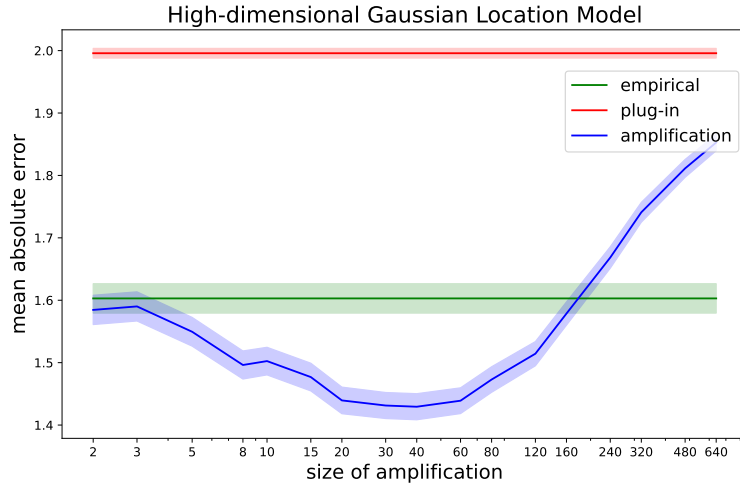
(b) Mean absolute errors for estimation of the squared L_2 norm $\mathbb{E}[\|X\|_2^2]$ in a d-dimensional Gaussian location model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, I_d)$ with $n = 50, d = 100, \mu = \mathbf{1}/\sqrt{d}$.

(c) Classification errors for binary classification between two clusters of covariates $X_1, \cdots, X_{n/2} \sim \mathcal{N}(e_1, I_d)$ and $X_{n/2+1}, \cdots, X_n \sim \mathcal{N}(-e_1, I_d)$, with $n = 50$ and $d = 10$.

Fig 1: Sample amplification experiments. The x-axis corresponds to the amount of amplification, m, and the shaded area depicts the 95% confidence interval based on 5,000 simulations.

the empirical estimator based on the original sample X^n has a large MAE, its performance is improved based on the amplified sample Y^{n+m}.
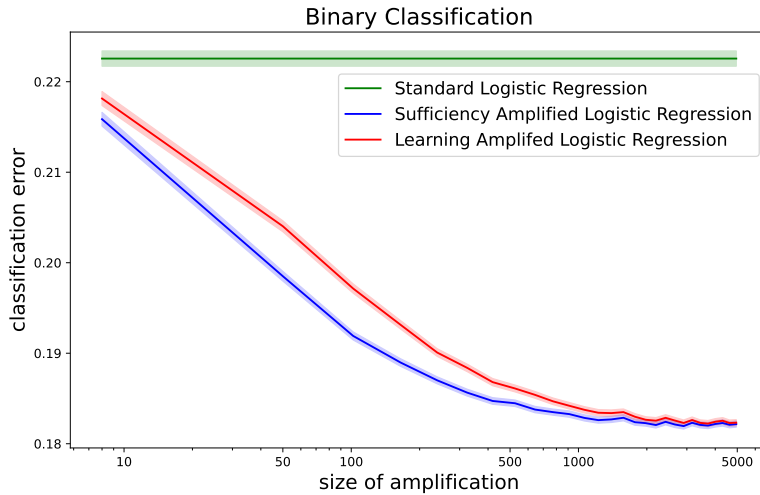
- Squared L_2 norm estimation for high-dimensional Gaussian: here we observe $X_1, \cdots, X_n \sim \mathcal{N}(\mu, I_d)$ with $n = 50, d = 100$ and $\mu = \mathbf{1}/\sqrt{d}$, and we again consider three estimators for $\mathbb{E}[\|X\|_2^2]$. As before, the empirical estimator is simply $n^{-1} \sum_{i=1}^n \|X_i\|_2^2$, and the plug-in estimator uses the knowledge $\mathbb{E}_{X \sim \mathcal{N}(\widehat{\mu}, I_d)}[\|X\|_2^2] = \|\widehat{\mu}\|_2^2 + d$ and estimates $\widehat{\mu} = \bar{X}$. As for the amplified estimator, it first amplifies the sample X^n into Y^{n+m} via sufficiency (cf. Example 4.1), and then uses the empirical estimator based on Y^{n+m}. The plots of the mean absolute errors are displayed in Figure 1b. Here the empirical estimator outperforms the plug-in estimator due to a smaller bias, while the sample amplification further reduces the MAE as long as the size of amplification m is not too large. This could be explained by the bias-variance tradeoff, where the amplified estimator interpolates between the empirical estimator (with no bias) and the plug-in estimator (with the smallest asymptotic variance).
- Binary classification: here we observe two clusters of covariates $X_1, \cdots, X_{n/2} \sim \mathcal{N}(e_1, I_d)$ (with label 1) and $X_{n/2+1}, \cdots, X_n \sim \mathcal{N}(-e_1, I_d)$ (with label -1), with $n = 50, d = 10$ and e_1 being the first basis vector. The target is to train a classifier with a high classification accuracy on the test data with the same distribution. The standard classifier is via logistic regression, which does not use the knowledge of Gaussianity. To apply sample amplification, we first amplify the sample in each class via either sufficiency (cf. Example 4.1) or learning (cf. Example A.8), and then run logistic regression on the amplified datasets. Figure 1c displays the classification errors of all three approaches, and shows that both amplification procedures help reduce the classification error even for small values of m.

The above experiments demonstrate the potential for sample amplification to leverage knowledge of the data distribution to produce a larger dataset that is then fed into downstream distribution-agnostic algorithms. Some experiments (e.g. Figure 1b) also suggest a limit beyond which the amplification procedure alters the data distribution too much. We believe that rigorously examining amplification through the lens of the performance of downstream estimators and algorithms, including those illustrated in our numerical simulations, would be a fruitful direction for future work.

2. Connections, limitations and future work. As discussed above, it is commonplace in machine learning to increase the size of datasets using various heuristics, often resulting in large gains in downstream learning performance. However, a clear statistical understanding of when this is possible and what techniques are useful for this is missing. A natural starting point to get a better understanding is the formulation we consider that asks the extent to which datasets can be amplified in a perfect sense—where any verifier who knows the true distribution is not able to distinguish the amplified dataset from a set of i.i.d. draws.

A limitation of the sample amplification formulation described above is that the additive amplification factor m is rather small (e.g., $O(n\epsilon/\sqrt{d})$ for d dimensional exponential families). Moreover, we show matching lower bounds demonstrating that this factor can not be improved even when n is large enough to learn the distribution to non-trivial accuracy. However, it might be possible to achieve larger amplification factors with restricted verifiers, for instance, the class of verifiers corresponding to learning algorithms used for downstream tasks (see [2] for other possible classes of verifiers). Investigating the sample amplification problem with such restricted verifiers may be a practically fruitful future direction.

Despite this limitation, the sample amplification formulation does yield high-level insights that can inform the way datasets are amplified in practice. For instance, from the results in this paper, we know that sample amplification is possible for a broad class of distributions even when learning is not possible. Moreover, both our sufficiency or learning based approaches modify the original data points in general, conforming to the lower bound in [2] that optimal

amplification may be impossible if the amplifier returns a superset of the input dataset. These observations show that the folklore way of enlarging datasets by learning the data distribution and adding more samples from the learned distribution can be far from optimal.

Connections with other statistical notions. An equivalent view of Definition 1.1 is through Le Cam's distance [31], a classical concept in statistics. The formal definition of Le Cam's distance $\Delta(\mathcal{M}, \mathcal{N})$ is summarized in Definition 3.1; roughly speaking, it measures the fundamental difference in power in the statistical models $\mathcal{M}$ and $\mathcal{N}$, without resorting to specific estimation procedures. The sample amplification problem is equivalent to the study of Le Cam's distance $\Delta(\mathcal{P}^{\otimes n}, \mathcal{P}^{\otimes (n+m)})$ between product models, where (1.1) is precisely equivalent to $\Delta(\mathcal{P}^{\otimes n}, \mathcal{P}^{\otimes (n+m)}) \le \epsilon$. However, in the statistics literature, Le Cam's distance was mainly used to study the *asymptotic* equivalence, where a typical target is to show that $\lim_{n\to\infty} \Delta(\mathcal{M}_n, \mathcal{N}_n) = 0$ for certain sequences of statistical models. For example, showing that localized regular statistical models converge to Gaussian location models is the fundamental idea behind the Hájek–Le Cam asymptotic statistics; see [31, 32, 33] and [16, Chapter 9]. In nonparametric statistics, there is also a rich line of research [15, 16, 17, 42] establishing asymptotic (non-)equivalences, based on Le Cam's distance, between density models, regression models, and Gaussian white noise models. In the above lines of work, only asymptotic results were typically obtained with a fixed dimension and possibly slow convergence rate. In contrast, we aim to obtain a non-asymptotic characterization of $\Delta(\mathcal{P}^{\otimes n}, \mathcal{P}^{\otimes (n+m)})$ in (n, m) and the dimension of the problem, a task which is largely underexplored in the literature.

Another related angle is from reductions between statistical models. Over the past decade there has been a growing interest in constructing polynomial-time reductions between various statistical models (typically from the planted clique) to prove statistical-computational gaps, see, e.g. [10, 13, 14, 37]. The sample amplification falls into the reduction framework, and aims to perform reductions from a product model $\mathcal{P}^{\otimes n}$ to another product model $\mathcal{P}^{\otimes (n+m)}$. While previous reduction techniques were mainly constructive and employed to prove computational lower bounds, in this paper we also develop general tools to prove limitations of all possible reductions purely from the statistical perspective.

Organization. The rest of this paper is organized as follows. Section 3 lists some notations and preliminaries for this paper, and in particular introduces the concept of Le Cam's distance. Section 4 introduces a sufficiency-based procedure for sample amplification, with asymptotic properties for general exponential families and non-asymptotic performances in several specific examples. Section 5 is devoted to a learning-based procedure for sample amplification, with a general relationship between sample amplification and the χ^2 estimation error, as well as its applications in several examples. Section 6 presents the general idea of establishing lower bounds for sample amplification, with a universal result specializing to product models. Section 7 discusses more examples in sample amplification and learning, and shows that these tasks are in general non-comparable. More concrete examples of both the upper and lower bounds, auxiliary lemmas, and proofs are relegated to the appendices in the supplementary material.

3. Preliminaries. We use the following notations throughout this paper. For a random variable X, let $\mathcal{L}(X)$ be the law (i.e. probability distribution) of X. For a probability distribution P on a probability space Ω and a measurable map $T: \Omega \to \Omega'$, let $P \circ T^{-1}$ denotes the pushforward probability measure, i.e. $\mathcal{L}(T(X))$ with $\mathcal{L}(X) = P$. For a probability measure P, let $P^{\otimes n}$ be the n-fold product measure. For a positive integer n, let $[n] \triangleq \{1, \cdots, n\}$, and $x^n \triangleq (x_1, \cdots, x_n)$. We adopt the following asymptotic notations: for two non-negative sequences (a_n) and (b_n), we use $a_n = O(b_n)$ to denote that $\limsup_{n\to\infty} a_n/b_n < \infty$, and $a_n = \Omega(b_n)$ to denote $b_n = O(a_n)$, and $a_n = \Theta(b_n)$ to denote both $a_n = O(b_n)$ and $b_n = O(a_n)$. We also use the notations O_c, Ω_c, Θ_c to denote the respective meanings with

hidden constants depending on c. For probability measures P, Q defined on the same probability space, the total variation (TV) distance, Hellinger distance, Kullback–Leibler (KL) divergence, and the chi-squared divergence are defined as follows:

$$\|P - Q\|_{\text{TV}} = \frac{1}{2} \int |\mathrm{d}P - \mathrm{d}Q|, \qquad H(P,Q) = \left(\frac{1}{2} \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2 \right)^{\frac{1}{2}},$$

$$D_{\text{KL}}(P\|Q) = \int \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}, \qquad \chi^2(P\|Q) = \int \frac{(\mathrm{d}P - \mathrm{d}Q)^2}{\mathrm{d}Q}.$$

We will frequently use the following inequalities between the above quantities [15, Chapter 2]:

$$H^2(P,Q) \leq \|P - Q\|_{\text{TV}} \leq H(P,Q)\sqrt{2 - H^2(P,Q)}, \tag{3.1}$$

$$\|P - Q\|_{\text{TV}} \leq \sqrt{\frac{1}{2}D_{\text{KL}}(P\|Q)} \leq \sqrt{\frac{1}{2}\log(1 + \chi^2(P\|Q))}. \tag{3.2}$$

Next we define several quantities related to Definition 1.1. For a given distribution class $\mathcal{P}$ and sample sizes n and m, the *minimax error of sample amplification* is defined as

$$\epsilon^\star(\mathcal{P}, n, m) \triangleq \inf_T \sup_{P \in \mathcal{P}} \|P^{\otimes(n+m)} - P^{\otimes n} \circ T^{-1}\|_{\text{TV}}, \tag{3.3}$$

where the infimum is over all (possibly randomized) measurable mappings $T : \mathcal{X}^n \to \mathcal{X}^{n+m}$. For a given error level ϵ, the *maximum size of sample amplification* is the largest m such that there exists an $(n, n + m, \epsilon)$ sample amplification, i.e.

$$m^\star(\mathcal{P}, n, \epsilon) \triangleq \max\{m \in \mathbb{N} : \epsilon^\star(\mathcal{P}, n, m) \leq \epsilon\}. \tag{3.4}$$

For the ease of presentation, we often choose ϵ to be a small constant (say 0.1) and abbreviate the above quantity as $m^\star(\mathcal{P}, n)$; we remark that all our results work for a generic $\epsilon \in (0, 1)$. Finally, we also define the *sample amplification complexity* as the smallest n such that an amplification from n to $n + 1$ samples is possible:

$$n^\star(\mathcal{P}) \triangleq \min\{n \in \mathbb{N} : m^\star(\mathcal{P}, n) \geq 1\}. \tag{3.5}$$

Note that all the above notions are instance-wise in the distribution class $\mathcal{P}$.

The minimax error of sample amplification (3.3) is precisely known as the *Le Cam's distance* in the statistics literature. We adopt the standard framework of statistical decision theory [51]. A statistical model (or experiment) $\mathcal{M}$ is a tuple $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ where an observation $X \sim P_\theta$ is drawn for some $\theta \in \Theta$. A *decision rule* δ is a regular conditional probability kernel from $\mathcal{X}$ to the family of probability distributions on a general action space $\mathcal{A}$, and there is a (measurable) loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$. The *risk function* of a given decision rule δ is defined as

$$R_{\mathcal{M}}(\theta, \delta, L) \triangleq \mathbb{E}_\theta[L(\theta, \delta(X))] = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a)\delta(da \mid x)P_\theta(dx). \tag{3.6}$$

Based on the definition of risk functions, we are ready to define a metric, known as Le Cam's distance, between statistical models.

DEFINITION 3.1 (Le Cam's distance; see [31, 32, 33]). *For two statistical models* $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ *and* $\mathcal{N} = (\mathcal{Y}, (Q_\theta)_{\theta \in \Theta})$, *Le Cam's distance* $\Delta(\mathcal{M}, \mathcal{N})$ *is defined as*

$$\Delta(\mathcal{M}, \mathcal{N}) = \max \left\{ \sup_L \sup_{\delta_{\mathcal{N}}} \inf_{\delta_{\mathcal{M}}} \sup_{\theta \in \Theta} |R_{\mathcal{M}}(\theta, \delta_{\mathcal{M}}, L) - R_{\mathcal{N}}(\theta, \delta_{\mathcal{N}}, L)|, \right.$$

Algorithm 1 SAMPLE AMPLIFICATION VIA SUFFICIENCY

1: **Input:** samples $X_1, \cdots, X_n$, a given transformation f between sufficient statistics
2: Compute the sufficient statistic $T_n = T_n(X_1, \cdots, X_n)$.
3: Apply f to the sufficient statistic and compute $\widehat{T}_{n+m} = f(T_n)$.
4: Generate $(\widehat{X}_1, \cdots, \widehat{X}_{n+m}) \sim P_{X^{n+m}|T_{n+m}}(\cdot \mid \widehat{T}_{n+m})$.
5: **Output:** amplified samples $(\widehat{X}_1, \cdots, \widehat{X}_{n+m})$.

$$\left. \sup_L \sup_{\delta_{\mathcal{M}}} \inf_{\delta_{\mathcal{N}}} \sup_{\theta \in \Theta} |R_{\mathcal{M}}(\theta, \delta_{\mathcal{M}}, L) - R_{\mathcal{N}}(\theta, \delta_{\mathcal{N}}, L)| \right\}$$

$$= \max\left\{ \inf_{T_1} \sup_{\theta} \|P_\theta \circ T_1^{-1} - Q_\theta\|_{\mathrm{TV}}, \ \inf_{T_2} \sup_{\theta} \|Q_\theta \circ T_2^{-1} - P_\theta\|_{\mathrm{TV}} \right\},$$

where the loss function is taken over all measurable functions $L : \Theta \times \mathcal{A} \to [0, 1]$.

In the language of model deficiency introduced in [30], Le Cam's distance is the smallest $\epsilon > 0$ such that the model $\mathcal{M}$ is ϵ-deficient to the model $\mathcal{N}$, and $\mathcal{N}$ is also ϵ-deficient to $\mathcal{M}$. In the sample amplification problem, $(P_\theta)_{\theta \in \Theta} = \{P^{\otimes n} : P \in \mathcal{P}\}, (Q_\theta)_{\theta \in \Theta} = \{P^{\otimes(n+m)} : P \in \mathcal{P}\}$. Here, choosing $T_2(x^{n+m}) = x^n$ in Definition 3.1 shows that $\mathcal{N}$ is 0-deficient to $\mathcal{M}$, and the remaining quantity involving T_1 exactly reduces to the minimax error of sample amplification in (3.3). Therefore, studying the complexity of sample amplification is equivalent to the characterization of the quantity $\Delta(\mathcal{P}^{\otimes n}, \mathcal{P}^{\otimes(n+m)})$.

4. Sample amplification via sufficient statistics. The first idea we present for sample amplification is the classical idea of reduction by sufficiency. Albeit simple, the sufficiency-based idea reduces the problem of generating multiple random vectors to a simpler problem of generating only a few vectors, achieves the optimal complexity of sample amplification in many examples, and is easy to implement.

4.1. *The general idea.* We first recall the definition of sufficient statistics: in a statistical model $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ and $X \sim P_\theta$, a statistic $T = T(X) \in \mathcal{T}$ is *sufficient* if and only if both $\theta - X - T$ and $\theta - T - X$ are Markov chains. A classical result in statistical decision theory is *reduction by sufficiency*, i.e. only the sufficient statistic needs to be maintained to perform statistical tasks as $P_{X|T,\theta}$ does not depend on the unknown parameter θ. In terms of Le Cam's distance, let $\mathcal{M} \circ T^{-1} = (\mathcal{T}, (P_\theta \circ T^{-1})_{\theta \in \Theta})$ be the statistical experiment associated with T, then sufficiency of T implies that $\Delta(\mathcal{M}, \mathcal{M} \circ T^{-1}) = 0$. Hereafter, we will call $\mathcal{M} \circ T^{-1}$ the T-*reduced model*, or simply *reduced model* in short.

Reduction by sufficiency could be applied to sample amplification in a simple way, with a general algorithm displayed in Algorithm 1. Suppose that both models $\mathcal{P}^{\otimes n}$ and $\mathcal{P}^{\otimes(n+m)}$ admit sufficient statistics $T_n = T_n(X^n)$ and $T_{n+m} = T_{n+m}(X^{n+m})$, respectively. Algorithm 1 claims that it suffices to perform sample amplification on the reduced models $\mathcal{P}^{\otimes n} \circ T_n^{-1}$ and $\mathcal{P}^{\otimes(n+m)} \circ T_{n+m}^{-1}$, i.e. construct a randomization map f from T_n to T_{n+m}. Concretely, the algorithm decomposes into three steps:

1. Step I: map X^n to T_n. This step is straightforward: we simply compute $T_n = T_n(X_1, \cdots, X_n)$.
2. Step II: apply a randomization map in the reduced model. Upon choosing the map f, we simply compute $\widehat{T}_{n+m} = f(T_n)$ with the target that the TV distance $\|\mathcal{L}(\widehat{T}_{n+m}) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}}$ is uniformly small. The concrete choice of f depends on specific models.
3. Step III: map T_{n+m} to X^{n+m}. By sufficiency of T_{n+m}, the conditional distribution $P_{X^{n+m}|T_{n+m}}$ does not depend on the unknown model. Therefore, after replacing the true

statistic T_{n+m} by $\widehat{T}_{n+m}$, it is feasible to generate $\widehat{X}^{n+m} \sim P_{X^{n+m}|T_{n+m}}(\cdot \mid \widehat{T}_{n+m})$. To simulate this random vector, it suffices to choose any distribution $P_0 \in \mathcal{P}$ and generate $\widehat{X}^{n+m} \sim (P_0^{\otimes(n+m)} \mid T_{n+m}(\widehat{X}^{n+m}) = \widehat{T}_{n+m})$. This step may suffer from computational issues which will be discussed in Section 4.2.

The validity of this idea simply follows from

$$\Delta(\mathcal{P}^{\otimes n}, \mathcal{P}^{\otimes(n+m)}) = \Delta(\mathcal{P}^{\otimes n} \circ T_n^{-1}, \mathcal{P}^{\otimes(n+m)} \circ T_{n+m}^{-1}),$$

or equivalently, under each $P \in \mathcal{P}$,

$$\|\mathcal{L}(\widehat{X}^{n+m}) - \mathcal{L}(X^{n+m})\|_{\mathrm{TV}} = \|\mathcal{L}(\widehat{T}_{n+m}) \times P_{X^{n+m}|T_{n+m}} - \mathcal{L}(T_{n+m}) \times P_{X^{n+m}|T_{n+m}}\|_{\mathrm{TV}}$$

$$\overset{(a)}{=} \|\mathcal{L}(\widehat{T}_{n+m}) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}} = \|\mathcal{L}(f(T_n)) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}},$$

where (a) is due to the identity $\|P_X P_{Y|X} - Q_X P_{Y|X}\|_{\mathrm{TV}} = \|P_X - Q_X\|_{\mathrm{TV}}$. In other words, it suffices to work on reduced models and find the map f between sufficient statistics.

This idea of reduction by sufficiency simplifies the design of sample amplification procedures. Unlike in original models where X^n and X^{n+m} typically take values in spaces of different dimensions, in reduced models the sufficient statistics T_n and T_{n+m} are usually drawn from the same space. A simple example is as follows.

EXAMPLE 4.1 (Gaussian location model with known covariance). *Consider the observations $X_1, \cdots, X_n$ from the Gaussian location model $P_\theta = \mathcal{N}(\theta, \Sigma)$ with an unknown mean $\theta \in \mathbb{R}^d$ and a known covariance $\Sigma \in \mathbb{R}^{d \times d}$. To amplify to $n + m$ samples, note that the sample mean vector is a sufficient statistic here, with*

$$T_n(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(\theta, \Sigma/n).$$

Now consider the identity map between sufficient statistics $\widehat{T}_{n+m} = T_n$ used with algorithm 1. The amplified samples $(\widehat{X}_1, \cdots, \widehat{X}_{n+m})$ are drawn from $\mathcal{N}(0, \Sigma)$ conditioned on the event that $T_{n+m}(\widehat{X}^{n+m}) = \widehat{T}_{n+m} = T_n(X^n)$. For every mean vector $\theta \in \mathbb{R}^d$ we can upper bound the amplification error of this approach:

$$\|\mathcal{L}(\widehat{T}_{n+m}) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}} = \|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}}$$

$$= \|\mathcal{N}(\theta, \Sigma/n) - \mathcal{N}(\theta, \Sigma/(n+m))\|_{\mathrm{TV}}$$

$$\overset{(a)}{\leq} \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathcal{N}(\theta, \Sigma/n) \| \mathcal{N}(\theta, \Sigma/(n+m)))}$$

$$= \sqrt{\frac{d}{4}\left(\frac{m}{n} - \log\left(1 + \frac{m}{n}\right)\right)} = O\left(\frac{m\sqrt{d}}{n}\right),$$

where (a) is due to (3.2), and the last step holds whenever $m = O(n)$. Therefore, we could amplify $\Omega(n/\sqrt{d})$ additional samples based on n observations, and the complexity of sample amplification in (3.5) is $n^\star = O(\sqrt{d})$. In contrast, learning this distribution within a small TV distance requires $n = \Omega(d)$ samples, which is strictly harder than sample amplification. This example recovers the upper bound of [2] with a much simpler analysis, and in later sections we will show that this approach is exactly minimax optimal.

We make two remarks for the above example. First, the amplified samples $\widehat{X}^{n+m}$ are no longer independent, either marginally or conditioned on X^n. Therefore, the above approach

is fundamentally different from first estimating the distribution and then generating independent samples from the estimated distribution. Second, the amplified samples do not contain the original samples as a subset. In contrast, a tempting approach for sample amplification is to add m fake samples to the original n observations. However, [2] showed that any sample amplification approach containing the original samples cannot succeed if $n = o(d/\log d)$ in the above model, and our approach conforms to this result. More examples will be presented in Appendix A.1.

4.2. *Computational issues.*   A natural computational question in Algorithm 1 is how to sample $\widehat{X}^{n+m} \sim P_{X^{n+m}|T_{n+m}}(\cdot \mid \widehat{T}_{n+m})$ in a computationally efficient way. With an additional mild assumption that the sufficient statistic T is also complete (which is easy to find in exponential families), the conditional distribution $P_{X|T}$ could be efficiently sampled if we could find a statistic $S = S(X)$ with the following two properties:

1. S is *ancillary*, i.e. $\mathcal{L}(S)$ is independent of the model parameter θ;
2. There is a (measurable) bijection g between (T, S) and X, i.e. $X = g(T, S)$ almost surely.

In fact, if such an S exists, then under any $\theta \in \Theta$,

$$\mathcal{L}(X \mid T = t) \overset{\text{(a)}}{=} \mathcal{L}(g(T, S) \mid T = t) \overset{\text{(b)}}{=} \mathcal{L}(g(t, S)),$$

where (a) is due to the assumed bijection g between (T, S) and X, and (b) is due to a classical result of Basu [7, 8] that S and T are independent. Therefore, by the ancillarity of S, we could sample $X \sim P_{\theta_0}$ with any $\theta_0 \in \Theta$ and compute the statistic S from X, then $g(t, S)$ follows the desired conditional distribution $P_{X|T=t}$. An example of this procedure is illustrated below.

EXAMPLE 4.2 (Computation in Gaussian location model).   *Consider the setting of Example 4.1 where $P_\theta = \mathcal{N}(\theta, \Sigma)^{\otimes(n+m)}, T_{n+m} = (n + m)^{-1} \sum_{i=1}^{n+m} X_i$, and the target is to sample from the distribution $P_{X^{n+m}|T_{n+m}}$. In this model, T_{n+m} is complete and sufficient, and we choose $S = S(X^{n+m}) = (S_1, \cdots, S_{n+m-1})$ with $S_i = X_{i+1} - X_1$ for all i. Clearly S is ancillary, and X^{n+m} could be recovered from (T_{n+m}, S) via*

$$X_1 = T_{n+m} - \frac{\sum_{i=1}^{n+m-1} S_i}{n + m}, \qquad X_{i+1} = X_1 + S_i, \quad i \in [n + m - 1].$$

Therefore, the choice of S satisfies both conditions. Consequently, we can draw $Z^{n+m} \sim \mathcal{N}(0, \Sigma)^{\otimes(n+m)}$, compute $S = S(Z^{n+m})$ (where $S_i = Z_{i+1} - Z_1$), and recover X^{n+m} from (T_{n+m}, S).

The proper choice of S depends on specific models and may require some effort to find; we refer to Appendix A.1 for more examples. We remark that in general there is no golden rule to find S. One tempting approach is to find a *maximal* ancillary statistic S such that any other ancillary statistic S' must be a function of S. This idea is motivated by the existence of the minimal sufficient statistic under mild conditions and a known computationally efficient approach to compute it. However, for ancillary statistics there is typically no such a maximal one in the above sense, and there may exist uncountably many "maximal" ancillary statistics which are not equivalent to each other. From the measure theoretic perspective, this is due to the fact that the family of all ancillary sets is not closed under intersection and thus not a σ-algebra. In addition, even if a proper notion of "maximal" or "essentially maximal" could be defined, there is no guarantee that such an ancillary statistic satisfies the bijection condition, and it is hard to determine whether a given ancillary statistic is maximal or not. We refer to [9, 34] for detailed discussions on ancillarity from mathematical statistics.

There is also another sampling procedure of $P_{X^n|T_n}$ in the conditional inference literature [52]. Specifically, for each $i \in [n]$, this approach sequentially generates the observation X_i from the one-dimensional distribution $P_{X_i|X^{i-1},T_n}$, which is a simple task as long as we know its CDF. Although this works in simple models such as the Gaussian location model above, in more complicated models exact computation of the CDF is typically not feasible.

4.3. *General theory for exponential families.* In this section, we show that a general $(n, n + \Omega(n\epsilon/\sqrt{d}), \epsilon)$ sample amplification phenomenon holds for a rich class of exponential families, and is achieved by the sufficiency-based procedure in Algorithm 1. Specifically, we consider the following natural exponential family.

DEFINITION 4.3 (Exponential family). *The* exponential family $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ *of probability measures is determined by*

$$dP_\theta(x) = \exp(\theta^\top T(x) - A(\theta))d\mu(x),$$

where $\theta \in \Theta$ is the natural parameter with $\Theta = \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$, $T(x)$ is the sufficient statistic, $A(\theta)$ is the log-partition function, and μ is the base measure.

The exponential family includes many widely used probability distributions such as Gaussian, Gamma, Poisson, Exponential, Beta, etc. In the exponential family, the statistic $T(x)$ is sufficient and complete, and several well-known identities include $\mathbb{E}_\theta[T(X)] = \nabla A(\theta)$, and $\mathsf{Cov}_\theta[T(X)] = \nabla^2 A(\theta)$. We refer to [24] for a mathematical theory of the exponential family.

To establish a general theory of sample amplification for exponential families, we shall make the following assumptions on the exponential family.

ASSUMPTION 1 (Continuity). *The parameter set Θ has a non-empty interior. Under each $\theta \in \Theta$, the probability distribution $\mathcal{L}(T(X))$ is absolutely continuous with respect to the Lebesgue measure.*

ASSUMPTION 2 (Moment condition M_k). *For a given integer $k > 0$, it holds that*

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\left\| (\nabla^2 A(\theta))^{-1/2}(T(X) - \nabla A(\theta)) \right\|_2^k \right] < \infty.$$

We call it the moment condition M_k.

Assumption 1 requires an exponential family of continuous distributions. The motivation is that for continuous exponential family, the sufficient statistics $T_n(X)$ and $T_{n+m}(X)$ for different sample sizes take continuous values in the same space, and it is easier to construct a general transformation. We will propose a different sample amplification approach for discrete statistical models in Section 5. Assumption 2 is a moment condition on the normalized statistic $(\nabla^2 A(\theta))^{-1/2}(T(X) - \nabla A(\theta))$, whose moments always exist as the moment generating function of $T(X)$ exists around the origin. The moment condition M_k claims that the above normalized statistic has a uniformly bounded k-th moment for all $\theta \in \Theta$, which holds in several examples (such as Gaussian, exponential, Pareto) or by considering a slightly smaller $\Theta_0 \subseteq \Theta$ bounded away from the boundary. The following lemma presents a sufficient criterion for the moment condition M_k.

LEMMA 4.4. *If the log-partition function $A(\theta)$ satisfies*

$$\sup_{\theta \in \Theta} \sup_{u \in \mathbb{R}^d \setminus \{0\}} \frac{|\nabla^3 A(\theta)[u; u; u]|}{(\nabla^2 A(\theta)[u; u])^{3/2}} \le M < \infty,$$

then the exponential family satisfies the moment condition M_k for all $k \in \mathbb{N}$. Here for a k-tensor T and vectors $u_1, \cdots, u_k$, $T[u_1; \cdots ; u_k]$ denotes the value of $\langle T, u_1 \otimes \cdots \otimes u_k \rangle$.

The condition in Lemma 4.4 is called the self-concordant condition, which is a key concept in the interior point method for convex optimization [13]. For example, all quadratic functions and logarithmic functions are self-concordant (which correspond to Gaussian, exponential, and Pareto distributions), and the self-concordance is always fulfilled when Θ is compact.

Given any exponential family $\mathcal{P}$ satisfying Assumptions 1 and 2, we will show that a simple sample amplification procedure gives a size $\Omega(n/\sqrt{d})$ of sample amplification. Let $X_1, \cdots, X_n$ be i.i.d. samples drawn from P_θ taking a general form in Definition 4.3, then it is clear that the sample average

$$T_n(X^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} T(X_i)$$

is a sufficient statistic by the factorization theorem. We will apply the general Algorithm 1 with an identity map between sufficient statistics, i.e. $\widehat{T}_{n+m} = T_n$. The next theorem shows the performance of this approach.

THEOREM 4.5. *If the exponential family $\mathcal{P}$ satisfies Assumptions 1 and 2 with $k = 3$, then for $\theta \in \Theta$, it holds that*

$$\epsilon^\star(\mathcal{P}, n, m) \leq \|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}} \leq \frac{C}{\sqrt{n}} + \frac{m\sqrt{d}}{n},$$

where $C < \infty$ is an absolute constant depending only on d and the moment upper bound in Assumption 2. In particular, for sufficiently large n, a sample amplification of size $\Omega(n/\sqrt{d})$ is achievable.

Theorem 4.5 shows that the above simple procedure could achieve a sample amplification from n to $n + \Omega(n/\sqrt{d})$ samples in general continuous exponential families, provided that n is large enough. The main idea behind the proof of Theorem 4.5 is also simple. We show that the distribution of T_n is approximately $G_n \sim \mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/n)$ by CLT, apply the same CLT for T_{n+m}, and then compute the TV distance between two Gaussians as in Example 4.1. Theorem 4.5 is then a direct consequence of the triangle inequality:

$$\|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}}$$
$$\leq \|\mathcal{L}(T_n) - \mathcal{L}(G_n)\|_{\mathrm{TV}} + \|\mathcal{L}(G_n) - \mathcal{L}(G_{n+m})\|_{\mathrm{TV}} + \|\mathcal{L}(T_{n+m}) - \mathcal{L}(G_{n+m})\|_{\mathrm{TV}}.$$

Note that Assumption 1 ensures a vanishing TV distance for the Gaussian approximation, and Assumption 2 enables us to apply Berry–Esseen type arguments and obtain an $O(1/\sqrt{n})$ convergence rate for the Gaussian approximation.

The main drawback of Theorem 4.5 is that there is a hidden constant C depending on the dimension d, thus it does not mean that an $(n, n+1, \epsilon)$ sample amplification is possible as long as $n = \Omega(\sqrt{d}/\epsilon)$. To tackle this issue, we need to improve the first term in Theorem 4.5 and find the best possible dependence of the constant C on d. We remark that this is a challenging task in probability theory: although the convergence rates of both TV [3, 4, 41, 46] and KL [6, 12] in the CLT result were studied, almost all of them solely focused on the convergence rate on n, leaving the tight dependence on d still open. Moreover, direct computation of the quantity $\|\mathcal{L}(T_n) - \mathcal{L}(G_n)\|_{\mathrm{TV}}$ shows that even if the random vector T_n has independent components, this quantity is typically at least $\Omega(\sqrt{d/n})$. Therefore, $C = \Omega(\sqrt{d})$ under this proof technique, and a vanishing first term in Theorem 4.5 requires that $n = \Omega(d)$, which is already larger than the anticipated sample amplification complexity $n = O(\sqrt{d})$.

To overcome the above difficulties, we make the following changes to both the assumption and analysis. First, to avoid the unknown dependence on d, we additionally assume a *product*

exponential family, i.e. $P_\theta(dx) = \prod_{i=1}^d p_{\theta_i}(dx_i)$, where each $p_{\theta_i}(x_i)$ is a one-dimensional exponential family. Exploiting the product structure enables to find a constant C depending linearly on d. Second, we improve the $O(1/\sqrt{n})$ dependence on n by applying a higher-order CLT result to T_n and T_{n+m}, known as the *Edgeworth expansion* [11]. For any $k \geq 2$ and $n \in \mathbb{N}$, the signed measure of the Edgeworth expansion on $\mathbb{R}^d$ is

$$(4.1) \qquad \Gamma_{n,k}(dx) = \gamma(x)\left(1 + \sum_{\ell=1}^{\lfloor k/3 \rfloor} \frac{\mathcal{K}_\ell(x)}{n^{\ell/2}}\right)dx,$$

where $\gamma(x)$ is the density of a standard normal random variable on $\mathbb{R}^d$, and $\mathcal{K}_m(x)$ is a polynomial of degree $3m$ which depends only on the first $3m$ moments of the distribution. We note that unlike CLT, the general Edgeworth expansion is a signed measure with possibly negative densities; however, it is close to Gaussian with an $O(n^{-1/2})$ approximation error. Such a higher-order expansion enables us to have better Berry-Esseen type bounds, but upper bounding $\|\Gamma_{n,k} - \Gamma_{n+m,k}\|_{\mathrm{TV}}$ becomes more complicated and requires to handle the Gaussian part and the correction part separately; see Appendix B.2 for details. In particular, we could improve the error dependence on n from $O(1/\sqrt{n})$ to $O(1/n^2)$.

Formally, the next theorem shows a better sample amplification performance for product exponential families.

THEOREM 4.6. *Let $(\mathcal{X}, \mathcal{P} = (P_\theta)_{\theta \in \Theta})$ be a product exponential family, where each one-dimensional component satisfies Assumptions 1 and 2 with $k = 10$. Then for $\theta \in \Theta$, it holds that*

$$\epsilon^\star(\mathcal{P}, n, m) \leq \|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\|_{\mathrm{TV}} \leq C\left(\frac{d}{n^2} + \frac{m\sqrt{d}}{n}\right),$$

where $C < \infty$ is an absolute constant independent of (n, d). In particular, as long as $n = \Omega(\sqrt{d}/\epsilon)$, an $(n, n+m, \epsilon)$ sample amplification of size $m = \Omega(n\epsilon/\sqrt{d})$ is achievable.

Theorem 4.6 shows that for product exponential family, we not only achieve the amplification size $m = \Omega(n\epsilon/\sqrt{d})$, but also attain a sample complexity $n = O(\sqrt{d}/\epsilon)$ for sample amplification. This additional result on sample complexity is important in the sense that, even if distribution learning is impossible, it is possible to perform sample amplification. Although the independence or even the exponential family assumption could be strong practically, in Appendix A.1 we show that both phenomena $m = \Omega(n\epsilon/\sqrt{d})$ and $n = O(\sqrt{d}/\epsilon)$ hold in many natural models.

5. Sample amplification via learning. The sufficiency-based approach of sample amplification is not always desirable. First, models outside the exponential family typically do not admit non-trivial sufficient statistics, and therefore the reduction by sufficiency may not be very helpful. Second, the identity map applied to the sufficient statistics only works for continuous models, and incurs a too large TV distance when the underlying model is discrete. Third, previous approaches are not directly related to learning the model, so a general relationship between learning and sample amplification is largely missing. In this section, we propose another sample amplification approach, and show that how a good learner helps to obtain a good sample amplifier.

5.1. *The general result.* For a class of distribution $\mathcal{P}$ and n i.i.d. observations drawn from an unknown $P \in \mathcal{P}$, we define the following notion of the χ^2-estimation error.

DEFINITION 5.1 (χ^2-estimation error). *For a class of distributions $\mathcal{P}$ and sample size n, the χ^2-estimation error $r_{\chi^2}(\mathcal{P}, n)$ is defined to be the minimax estimation error under the expected χ^2-divergence:*

$$r_{\chi^2}(\mathcal{P}, n) \triangleq \inf_{\widehat{P}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\chi^2(\widehat{P}_n, P)],$$

where the infimum is taken over all possible distribution estimators $\widehat{P}_n$ based on n samples.

Roughly speaking, the χ^2-estimation error in the above definition characterizes the complexity of the distribution class $\mathcal{P}$ in terms of distribution learning under the χ^2-divergence. The main result of this section is to show that, the error of sample amplification in (3.3) could be upper bounded by using the χ^2-estimation error.

THEOREM 5.2. *For general $\mathcal{P}$ and $n, m \geq 0$, it holds that*

$$\epsilon^\star(\mathcal{P}, n, m) \leq \sqrt{\frac{m^2}{n} \cdot r_{\chi^2}(\mathcal{P}, n/2)}.$$

The following corollary is immediate from Theorem 5.2.

COROLLARY 5.3. *An $(n, n+m, \epsilon)$ sample amplification is possible if $m = O(\epsilon \sqrt{n/r_{\chi^2}(\mathcal{P}, n/2)})$. Moreover, the sample complexity of amplification in (3.5) satisfies*

$$n^\star(\mathcal{P}) = O\left(\min\left\{n \in \mathbb{N} : r_{\chi^2}(\mathcal{P}, n/2) \leq n\right\}\right).$$

REMARK 5.4. *Although the error of sample amplification in (3.3) is measured under the TV distance, the same result holds for the squared root of the KL divergence (which by (3.2) is no smaller than the TV distance).*

The above result provides a quantitative guarantee that the sample amplification is easier than learning (under the χ^2-divergence). Specifically, the sample complexity of learning is the smallest $n \in \mathbb{N}$ such that $r_{\chi^2}(\mathcal{P}, n) = O(1)$, while Corollary 5.3 shows that the complexity for amplification is at most the smallest $n \in \mathbb{N}$ such that $r_{\chi^2}(\mathcal{P}, n/2) = O(n)$. As $r_{\chi^2}(\mathcal{P}, n)$ is non-increasing in n, this means that the learning complexity is in general larger.

When the distribution class $\mathcal{P}$ has a product structure $\mathcal{P} = \prod_{j=1}^d \mathcal{P}_j$, the next theorem shows a better relationship between the amplification error and the learning error.

THEOREM 5.5. *For $\mathcal{P} = \prod_{j=1}^d \mathcal{P}_j$ and $n, m \geq 0$, it holds that*

$$\epsilon^\star(\mathcal{P}, n, m) \leq \sqrt{\frac{m^2}{n} \sum_{j=1}^d r_{\chi^2}(\mathcal{P}_j, n/2)}.$$

COROLLARY 5.6. *For product models, an $(n, n+m, \epsilon)$ sample amplification is achievable if*

$$m = O\left(\epsilon \sqrt{\frac{n}{\sum_{j=1}^d r_{\chi^2}(\mathcal{P}_j, n/2)}}\right).$$

Algorithm 2 SAMPLE AMPLIFICATION VIA SHUFFLING: GENERAL MODEL

1: **Input:** samples $X_1, \cdots, X_n$, a given class of distributions $\mathcal{P}$.
2: Based on samples $X_1, \cdots, X_{n/2}$, find an estimator $\widehat{P}_n$ such that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\chi^2(\widehat{P}_n, P)] \leq C \cdot r_{\chi^2}(\mathcal{P}, n/2).$$

3: Draw m additional samples $Y_1, \cdots, Y_m$ from $\widehat{P}_n$.
4: Uniformly at random, shuffle the pool of $X_{n/2+1}, \cdots, X_n, Y_1, \cdots, Y_m$ to obtain $(Z_1, \cdots, Z_{n/2+m})$.
5: **Output:** amplified samples $(X_1, \cdots, X_{n/2}, Z_1, \cdots, Z_{n/2+m})$.

Moreover, the sample complexity of amplification in (3.5) *satisfies*

$$n^\star(\mathcal{P}) = O\left(\min\left\{ n \in \mathbb{N} : \sum_{j=1}^d r_{\chi^2}(\mathcal{P}_j, n/2) \leq n \right\} \right).$$

We observe that the result of Theorem 5.5 typically improves over Theorem 5.2 for product models. In fact, since

$$\chi^2\left(\prod_{j=1}^d P_j, \prod_{j=1}^d Q_j \right) = \prod_{j=1}^d (1 + \chi^2(P_j, Q_j)) - 1 \geq \sum_{j=1}^d \chi^2(P_j, Q_j),$$

the inequality $\sum_{j=1}^d r_{\chi^2}(\mathcal{P}_j, n/2) \leq r_{\chi^2}(\mathcal{P}, n/2)$ typically holds. Moreover, there are scenarios where we have $\sum_{j=1}^d r_{\chi^2}(\mathcal{P}_j, n/2) \ll r_{\chi^2}(\mathcal{P}, n/2)$, thus Theorem 5.5 provides a substantial improvement over Theorem 5.2. For example, when $\mathcal{P} = (\mathcal{N}(\theta, I_d))_{\theta \in \mathbb{R}^d}$, it could be verified that $r_{\chi^2}(\mathcal{P}_j, n/2) = O(1/n)$ for each $j \in [d]$, while $r_{\chi^2}(\mathcal{P}, n/2) = \exp(\Omega(d/n)) - 1$. Hence, in the important regime $\sqrt{d} \ll n \ll d$ where learning is impossible but the sample amplification is possible, Theorem 5.5 is strictly stronger than Theorem 5.2.

REMARK 5.7. *In the above Gaussian location model, there is an alternative way to conclude that Theorem 5.5 is strictly stronger than Theorem 5.2. We will see that the shuffling approach achieving the bound in Theorem 5.2 keeps all the observed samples, whereas [2] shows that all such approaches must incur a sample complexity $n = \Omega(d/\log d)$ for the Gaussian model. In contrast, Theorem 5.5 and Corollary 5.6 give a sample complexity $n = O(\sqrt{d})$ of amplification in the Gaussian location model.*

5.2. *The shuffling approach.* This section presents the sample amplification approaches to achieve Theorems 5.2 and 5.5. The idea is simple: we find a good distribution learner $\widehat{P}_n$ which attains the rate-optimal χ^2-estimation error, draw additional m samples $Y_1, \cdots, Y_m$ from $\widehat{P}_n$, and shuffle them with the original samples $X_1, \cdots, X_n$ uniformly at random. This approach suffices to achieve the sample amplification error in Theorem 5.2, while for Theorem 5.5 an additional trick is applied: instead of shuffling the whole vectors, we independently shuffle each coordinate instead. For technical reasons, in both approaches we apply the sample splitting: the first $n/2$ samples are used for the estimation of $\widehat{P}_n$, while the second $n/2$ samples are used for shuffling. The algorithms are summarized in Algorithms 2 and 3.

The following lemma is the key to analyze the performance of the shuffling approach.

LEMMA 5.8. *Let $X_1, \cdots, X_n$ be i.i.d. drawn from P, and $Y_1, \cdots, Y_m$ be i.i.d. drawn from Q independent of $(X_1, \cdots, X_n)$. Let $(Z_1, \cdots, Z_{n+m})$ be a uniformly random permutation of $(X_1, \cdots, X_n, Y_1, \cdots, Y_m)$, and P_{mix} be the distribution of the random mixture*

Algorithm 3 SAMPLE AMPLIFICATION VIA SHUFFLING: PRODUCT MODEL

1: **Input:** samples $X_1, \cdots, X_n$, a given class of product distributions $\mathcal{P} = \prod_{j=1}^d \mathcal{P}_j$
2: **for** $j = 1, 2, \cdots, d$ **do**
3:    Based on samples $X_{1,j}, \cdots, X_{n/2,j}$, find an estimator $\widehat{P}_{n,j}$ such that

$$\sup_{P_j \in \mathcal{P}_j} \mathbb{E}_{P_j}[\chi^2(\widehat{P}_{n,j}, P_j)] \leq C \cdot r_{\chi^2}(\mathcal{P}_j, n/2).$$

4:    Draw m additional samples $Y_{1,j}, \cdots, Y_{m,j}$ from $\widehat{P}_{n,j}$.
5:    Uniformly at random, shuffle $X_{n/2+1,j}, \cdots, X_{n,j}, Y_{1,j}, \cdots, Y_{m,j}$ to obtain $(Z_{1,j}, \cdots, Z_{n/2+m,j})$.
6: **end for**
7: For each $i \in [n/2 + m]$, form the vector $Z_i = (Z_{i,1}, \cdots, Z_{i,d})$.
8: **Output:** amplified samples $(X_1, \cdots, X_{n/2}, Z_1, \cdots, Z_{n/2+m})$.

$(Z_1, \cdots, Z_{n+m})$. *Then*

$$\chi^2\left(P_{\mathrm{mix}}, P^{\otimes(n+m)}\right) \leq \left(1 + \frac{m}{n+m}\chi^2(Q, P)\right)^m - 1.$$

Based on Lemma 5.8, the advantage of random shuffling is clear: if we simply append $Y_1, \cdots, Y_m$ to the end of the original sequence $X_1, \cdots, X_n$, then the χ^2-divergence is exactly $(1 + \chi^2(Q, P))^m - 1$. By comparing with the upper bound in Lemma 5.8, we observe that a smaller coefficient $m/(n + m)$ is applied to the individual χ^2-divergence after a random shuffle. The proofs of Theorems 5.2 and 5.5 are then clear, where we simply take $Q = \widehat{P}_n$ and apply the above lemma. Note that the statement of Lemma 5.8 requires that $Y_1, \cdots, Y_m$ be independent of $X_1, \cdots, X_n$, which is exactly the reason why we apply sample splitting in Algorithms 2 and 3. The proof of Lemma 5.8 is presented in Appendix C, and the complete proofs of Theorems 5.2 and 5.5 are relegated to Appendix B. We also include concrete examples of the shuffling approach in Appendix A.2.

6. Minimax lower bounds. In this section we establish minimax lower bounds for sample amplification in different statistical models. Section 6.1 presents a general and tight approach for establishing the lower bound, which leads to an exact sample amplification result for the Gaussian location model. Based on this result, we show that for d-dimensional continuous exponential families, the sample amplification size cannot exceed $\omega(n\epsilon/\sqrt{d})$ for sufficiently large sample size n. Section 6.2 provides a specialized criterion for product models, where we show that $n = \Omega(\sqrt{d}/\epsilon)$ and $m = O(n\epsilon/\sqrt{d})$ are always valid lower bounds, with hidden constants independent of all parameters. Section A.3 lists several concrete examples where our general idea could be properly applied to provide tight and non-asymptotic results.

6.1. *General idea.* The main tool to establish the lower bound is the first equality in the Definition 3.1 of Le Cam's distance. Specifically, for a class of distributions $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$, let μ be a given prior distribution on Θ, and $L : \Theta \times \mathcal{A} \to [0, 1]$ be a given non-negative loss function upper bounded by one. Given n i.i.d. samples from an unknown distribution in $\mathcal{P}$, define the following *Bayes risk* and *minimax risk*:

$$r_{\mathrm{B}}(\mathcal{P}, n, L, \mu) = \inf_{\widehat{\theta}} \int_\Theta \mathbb{E}_\theta[L(\theta, \widehat{\theta}(X^n))]\mu(d\theta),$$

$$r(\mathcal{P}, n, L) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \widehat{\theta}(X^n))],$$

where the infimum is over all possible estimators $\widehat{\theta}(\cdot)$ taking value in $\mathcal{A}$. The following result is a direct consequence of Definition 3.1.

LEMMA 6.1. *For any integer $n, m > 0$, any class of distributions $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$, any prior μ on Θ, and any loss function $L : \Theta \times \mathcal{A} \to [0, 1]$, the minimax error of sample amplification $\epsilon^\star(\mathcal{P}, n, m)$ in (3.3) satisfies that*

$$\epsilon^\star(\mathcal{P}, n, m) \geq r_B(\mathcal{P}, n, L, \mu) - r_B(\mathcal{P}, n + m, L, \mu),$$
$$\epsilon^\star(\mathcal{P}, n, m) \geq r(\mathcal{P}, n, L) - r(\mathcal{P}, n + m, L).$$

Based on Lemma 6.1, it suffices to find an appropriate prior distribution μ and a loss function L, and then compute (or lower bound) the difference between the Bayes or minimax risks with different sample sizes. We note that the lower bound technique in [2], albeit seemingly different, is a special case of Lemma 6.1. Specifically, the authors of [2] designed a set-valued mapping $A_n : \theta \to \mathcal{P}(\mathcal{X}^n)$ for each $n \in \mathbb{N}$ such that $\mathbb{P}_\theta(X^{n+m} \in A_{n+m}(\theta)) \geq 0.99$ for all $\theta \in \Theta$, while there is a prior distribution μ on Θ such that

$$(6.1) \qquad \mathbb{E}_{X^n} \left[\sup_{x^n \in \mathcal{X}^n} \mathbb{P}_{\theta|X^n}(x^n \in A_n(\theta)) \right] \leq 0.5.$$

If the above conditions hold, then an $(n, n+m)$ sample amplification is impossible. Note that the probability term in (6.1) is the maximum coverage probability of the sets $A_n(\theta)$ where θ follows the posterior distribution $\mathbb{P}_{\theta|X^n}$, which is a well-defined geometric object when both $A_n(\theta)$ and the posterior are known. To see that the above approach falls into our framework, consider the loss function $L : \Theta \times \cup_{n \geq 1} \mathcal{X}^n \to [0, 1]$ with $L(\theta, X^n) = \mathbb{1}(X^n \notin A_n(\theta))$. Then the first condition ensures that $r_B(\mathcal{P}, n + m, L, \nu) \leq 0.01$ for each prior ν, and the second condition (6.1) exactly states that $r_B(\mathcal{P}, n, L, \mu) \geq 0.5$ for the chosen prior μ.

A first application of Lemma 6.1 is an *exact* lower bound in Gaussian location models.

THEOREM 6.2. *For the Gaussian location model $\mathcal{P} = \{\mathcal{N}(\theta, \Sigma)\}_{\theta \in \mathbb{R}^d}$ with a fixed co-variance $\Sigma \in \mathbb{R}^{d \times d}$, the minimax error of sample amplification in (3.3) is exactly*

$$\epsilon^\star(\mathcal{P}, n, m) = \left\| \mathcal{N}\left(0, \frac{I_d}{n}\right) - \mathcal{N}\left(0, \frac{I_d}{n + m}\right) \right\|_{\mathrm{TV}}.$$

In particular, the sufficiency-based approach in Example 4.1 is exactly minimax optimal.

Theorem 6.2 shows that an exact error characterization for the Gaussian location model is possible through the general lower bound approach in Lemma 6.1. This result is also asymptotically useful to a rich family of models: note that by CLT, the sufficient statistic in a continuous exponential family follows a Gaussian distribution asymptotically, with a vanishing TV distance. This idea was used in Section 4.3 to establish the $O(n\epsilon/\sqrt{d})$ upper bound, and the same observation could lead to an $\Omega(n\epsilon/\sqrt{d})$ lower bound as well, under slightly different assumptions. Specifically, we drop Assumption 2 while introducing an additional assumption.

ASSUMPTION 3 (Linear independence). *The components of sufficient statistic $T(x)$ are linearly independent, i.e. $a^\top T(x) = 0$ for μ-almost all $x \in \mathcal{X}$ implies $a = 0$.*

Assumption 3 ensures that the true dimension of the exponential family is indeed d. Whenever Assumption 3 does not hold, we could transform it into a minimal exponential family with a lower dimension fulfilling this assumption. Note that when Assumptions 1 and 3 hold, the mean mapping $\theta \mapsto \nabla A(\theta)$ is a diffeomorphism between Θ and $\nabla A(\theta)$; see, e.g. [12, Theorem 1.22]. Therefore, $\nabla A(\cdot)$ is an open map, and the set $\{\nabla A(\theta) : \theta \in \Theta\}$ contains a

d-dimensional ball. This fact enables us to obtain a d-dimensional Gaussian location model after we apply the CLT.

The following theorem characterizes an asymptotic lower bound for every exponential family satisfying Assumptions 1 and 3.

THEOREM 6.3. *Given a d-dimensional exponential family $\mathcal{P}$ satisfying Assumptions 1 and 3, for every $n, m \in \mathbb{N}$, the minimax error of sample amplification satisfies*

$$\epsilon^\star(\mathcal{P}, n, m) \geq c \cdot \left(\frac{m\sqrt{d}}{n} \wedge 1\right) - C \cdot \left(\frac{\log n}{n}\right)^{\frac{1}{3}},$$

where $c > 0$ is an absolute constant independent of $(n, m, d, \mathcal{P})$, and constant $C > 0$ depends only on the exponential family (and thus on d).

Theorem 6.3 shows that there exists some $n_0 > 0$ depending only on the exponential family, such that sample amplification from n to $n + \omega(n\epsilon/\sqrt{d})$ samples is impossible for all $n > n_0$. However, similar to the nature of the upper bound in Theorem 4.5, this asymptotic result does not imply that the sample amplification is impossible if $n = o(\sqrt{d}/\epsilon)$. Nevertheless, in the following sections we show that the sample complexity lower bound $n = \Omega(\sqrt{d}/\epsilon)$ indeed holds in product families, as well as several other concrete examples.

6.2. *Product models.* Although Lemma 6.1 presents a lower bound argument in general, the computation of exact Bayes or minimax risks could be very challenging, and the usual rate-optimal analysis (i.e. bounding the risks within a multiplicative constant) will not lead to meaningful results. In addition, choosing the right prior and loss is a difficult task which may change from instance to instance. Therefore, it is helpful to propose specialized versions of Lemma 6.1 which are easier to work with. Surprisingly, such a simple version exists for product models, which is summarized in the following theorem.

THEOREM 6.4. *Let $\epsilon \in (0, 1)$ and $P_\theta = \prod_{j=1}^d p_{\theta_j}$ be a product model with $(\theta_1, \cdots, \theta_d) \in \prod_{j=1}^d \Theta_j$. Suppose for each $j \in [d]$, there exist two points $\theta_{j,+}, \theta_{j,-} \in \Theta_j$ such that*

(6.2) $$\|p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}} \leq \alpha_j - \frac{\epsilon}{\sqrt{d}},$$

(6.3) $$\|p_{\theta_{j,+}}^{\otimes(n+m)} - p_{\theta_{j,-}}^{\otimes(n+m)}\|_{\mathrm{TV}} \geq \alpha_j + \frac{\epsilon}{\sqrt{d}},$$

with $\alpha_j \in (\underline{\alpha}, \overline{\alpha})$, where $\underline{\alpha}, \overline{\alpha} \in (0, 1)$ are absolute constants. Then there exists an absolute constant $c = c(\underline{\alpha}, \overline{\alpha}) > 0$ such that

$$\epsilon^\star(\mathcal{P}, n, m) \geq c\epsilon.$$

Theorem 6.4 leaves the choices of the prior and loss function in Lemma 6.1 implicit, and provides a simple criterion for product models. The usual routine of applying Theorem 6.4 is as follows: fix any constant α and a target error ϵ, find for each $j \in [d]$ two points $\theta_{j,+}, \theta_{j,-} \in \Theta_j$ such that the condition (6.2) holds for a given sample size n. Then the condition (6.3) becomes an inequality solely for m, from which we could solve the smallest $m_j \in \mathbb{N}$ such that (6.3) holds along the j-th coordinate. Finally, the sample amplification from n to $n + m$ samples is impossible by the above theorem, where $m = \max_{j \in [d]} m_j$. Although Theorem 6.5 is only for product models, similar ideas could also be applied to non-product models; we refer to Section A.3 for concrete examples.

Theorem 6.4 also provides some intuition on why the sample complexity lower bound for amplification is typically smaller than that of learning. Specifically, for learning under the TV distance, a small TV distance $\|\prod_{j=1}^{d} p_{\theta_{j,+}}^{\otimes n} - \prod_{j=1}^{d} p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}}$ between product distributions is required. This requirement typically leads to a much smaller individual TV distance $\|p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}}$, e.g. $O(1/\sqrt{d})$ for many regular models. In contrast, the conditions (6.2) and (6.3) only require a constant individual TV distance, which leads to a smaller sample complexity n in the sample amplification lower bound. To understand why a larger individual TV distance works for sample amplification, in the proof of Theorem 6.4 we consider the uniform prior on 2^d points $\prod_{j=1}^{d}\{\theta_{j,+}, \theta_{j,-}\}$. Under this prior, the test accuracy for each dimension is precisely $(1 + \mathrm{TV}_j)/2$, which is slightly smaller than $(1 + \alpha)/2$ with n samples, and slightly larger than $(1 + \alpha)/2$ with $n + m$ samples (assuming $\alpha_j \equiv \alpha$). Therefore, if a unit loss is incurred when the fraction of correct tests does not exceed $(1+\alpha)/2$, the current scaling in (6.2), (6.3) shows that there is an $\Omega(\epsilon)$ difference in the expected loss under different sample sizes. In other words, such an aggregate voting test helps to have a larger individual TV distance. The details of the proof are deferred to Appendix B.

Theorem 6.4 has a far-reaching consequence: with almost no assumption on the product model $\mathcal{P}$, for any $c > 0$ it always holds that $\epsilon^\star(\mathcal{P}, n, \lceil c\epsilon n/\sqrt{d}\rceil) \geq c'\epsilon$ for some absolute constant $c' > 0$ *independent* of the product model $\mathcal{P}$. The only assumption (besides the product structure) we make on $\mathcal{P}$ is as follows (here $n \in \mathbb{N}$ is a given sample size):

ASSUMPTION 4. *Let $\mathcal{P}$ possess the product structure as in Theorem 6.4. For each $j \in [d]$, there exists two points $\theta_{j,+}, \theta_{j,-} \in \Theta_j$ such that $1/(10n) \leq H^2(p_{\theta_{j,+}}, p_{\theta_{j,-}}) \leq 1/(5n)$.*

Assumption 4 is a mild assumption that requires that for each coordinate, the map $\theta_j \mapsto p_{\theta_j}$ is continuous under the Hellinger distance. This assumption is satisfied for almost all practical models, either discrete or continuous, and is invariant with model reparametrizations or bijective transformation of observations. We note that the coefficients $1/10$ and $1/5$ are not essential, and could be replaced by any smaller constants. The next theorem states that if Assumption 4 holds, we always have a lower bound $n = \Omega(\sqrt{d})$ for the sample complexity and an upper bound $m = O(n/\sqrt{d})$ for the size of sample amplification.

THEOREM 6.5. *Let $\mathcal{P}$ be a product model satisfying Assumption 4. Then for any $c > 0$, there is some $c' > 0$ depending only on c (thus independent of $n, d, \epsilon, \mathcal{P}$) such that*

$$\epsilon^\star\left(\mathcal{P}, n, \left\lceil\frac{c\epsilon n}{\sqrt{d}}\right\rceil\right) \geq c'\epsilon.$$

Theorem 6.5 is a general lower bound for sample amplification in product models, with intriguing properties that it is instance-wise in the model $\mathcal{P}$, while the constants c and c' are *independent* of $\mathcal{P}$. As a result, the sample complexity is uniformly $\Omega(\sqrt{d}/\epsilon)$, and the maximum size of sample amplification is uniformly $O(n\epsilon/\sqrt{d})$ for all product models. In comparison, the matching upper bound in Theorem 4.6 for product models has a hidden constant depending on the statistical model. We note that it is indeed natural to have sample amplification results independent of the underlying statistical model. For example, it is clear by definition that sample amplifications are invariant with bijective transformation of observations. However, Assumption 2 depends on such transformations, so it possibly contains some redundancy. In contrast, Assumption 4 remains invariant, which is therefore more natural.

The proof idea of Theorem 6.5 is best illustrated for the case $d = 1$. Using the two points θ_+, θ_- in Assumption 4, one could show that the TV distance between n copies of p_{θ_+} and

p_{θ_-} is bounded from above by a small constant. Similarly, for a large $C > 0$, the TV distance between Cn copies of them is lower bounded by a large constant. Consequently, if $m = (C - 1)n$, Theorem 6.4 applied with $d = 1$ gives an $\Omega(1)$ lower bound on $\epsilon^\star(\mathcal{P}, n, m)$. What happens if $m = c\epsilon n$ with a small c? The idea is to consider the TV distances between $n, n + c\epsilon n, n + 2c\epsilon n, \cdots, Cn$ copies of p_{θ_+} and p_{θ_-}, which is an increasing sequence. Now by the pigeonhole principle, there must be two adjacent TV distances differing by at least $\Omega(c\epsilon/C) = \Omega(\epsilon)$, and Theorem 6.4 could be applied to this pair of sample sizes. This idea is easily generalized to any dimensions, with the full proof in Appendix B.

We note that the lower bounds in Theorem 6.3 (as well as 6.2) and Theorem 6.5 are on two different ends of the spectrum. In Theorems 6.2 and 6.3, an asymptotic setting (i.e. d fixed and $n \to \infty$) is essentially considered, and a Gaussian limit is crucially used as long as there is local asymptotic normality. In comparison, Theorem 6.5 deals with a high-dimensional scenario (n, d can grow together) but restricts to a product model. However, looking at product submodels and/or exploiting its proof techniques could still lead to tight lower bounds for several non-product models, as shown in the examples in Appendix A.3.

7. Discussions on sample amplification versus learning. In all the examples we have seen in the previous sections, there is always a squared root relationship between the statistical complexities of sample amplification and learning. Specifically, when the dimensionality of the problem is d, the complexity of learning the distribution (under a small TV distance) is typically $n = \Theta(d)$, whereas that of the sample amplification is typically $n = \Theta(\sqrt{d})$. In this section, we give examples where this relationship could break down in either direction, thus show that there is no universal scaling between the sample complexities of amplification and learning.

7.1. *An example where the complexity of sample amplification is $o(\sqrt{d})$.* We first provide an example where the distribution learning is hard, but an $(n, n + 1, 0.1)$ sample amplification is easy. Consider the following class $\mathcal{P}_{d,t}$ of discrete distributions:

$$\mathcal{P}_{d,t} = \left\{ (p_0, \cdots, p_d) : p_i \geq 0, \sum_{i=0}^{d} p_i = 1, p_0 = t \right\},$$

where it is the same as the class of all discrete distributions over $d + 1$ points, except that the learner has the perfect knowledge of $p_0 = t$ for some known $t \in [1/(2\sqrt{d}), 1/2]$. It is a classical result (see, e.g. [27]) that the sample complexity of learning the distribution over $\mathcal{P}_{d,t}$ with a small TV distance is still $n = \Theta(d)$, regardless of t. However, the next theorem shows that the complexity of sample amplification is much smaller.

THEOREM 7.1. *For the class $\mathcal{P}_{d,t}$ with $t \in [1/(2\sqrt{d}), 1/2]$, an $(n, n + 1, 0.1)$ sample amplification is possible if and only if*

$$n = \Omega\left(\frac{1}{t}\right).$$

Note that for the choice of $t = \Theta(d^{-\alpha})$ with $\alpha \in [0, 1/2]$, the complexity of sample amplification could possibly be $n = \Theta(d^\alpha)$ for every $\alpha \in [0, 1/2]$, showing that it could be $o(\sqrt{d})$ with an arbitrary polynomial scale in d. Moreover, if $t = o(1/\sqrt{d})$, the complexity of sample amplification reduces to $n = \Theta(\sqrt{d})$, the case without the knowledge of t. The main reason why sample amplification is easier here is that the additional fake sample could be chosen as the first symbol, which has a large probability. In contrast, learning the distribution requires the estimation of all other probability masses, so the existence of a probable symbol does not help much in learning.

7.2. *An example where the complexity of sample amplification is $\omega(\sqrt{d})$.*   Next we provide an example where the complexity of sample amplification is the same as that of learning. Consider a low-rank covariance estimation model: $X_1, \cdots, X_n \sim \mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ could be written as $\Sigma = UU^\top$ with $U \in \mathbb{R}^{p \times d}$ and $U^\top U = I_d$. In other words, the covariance matrix Σ is isotropic on some d-dimensional subspace. Here $n \geq d$ samples suffice to estimate Σ and thus the whole distribution perfectly, for the d-dimensional subspace could be recovered using d i.i.d. samples with probability one. Therefore, the complexity of learning the distribution is $n = d$. The following theorem states that this is also the complexity of sample amplification.

THEOREM 7.2.   *For the above low-rank covariance estimation model with $p \geq d + 1$, an $(n, n + 1, 0.1)$ sample amplification is possible if and only if $n \geq d$.*

Theorem 7.2 shows that as opposed to learning, sample amplification fails to exploit the low-rank structure in the covariance estimation problem. As a result, the complexity of sample amplification coincides with that of learning in this example. Note that sample amplification is always no harder than learning: the learner could always estimate the distribution, generate one observation from the distribution and append it to the original samples. Therefore, Theorem 7.2 provides an example where the relationship between sample amplification and learning is the worst possible.

7.3. *An example where the TV distance is not the right metric.*   Finally we provide an example showing that the TV distance is not the right metric for the learning-based approach in Section 5, and thereby partially illustrate the necessity of using the χ^2 divergence. This example also goes beyond parametric families for sample amplification. Let $\mathcal{P}$ be the class of all L-Lipschitz densities supported on $[0, 1]$, i.e. the density f satisfies $|f(x) - f(y)| \leq L|x - y|$ for all $x, y \in [0, 1]$. For $c \in (0, 1)$, also let $\mathcal{P}_c \subseteq \mathcal{P}$ be the subclass of densities lower bounded by c everywhere, i.e. $f(x) \geq c$ for all $x \in [0, 1]$. It is a classical result (see, e.g. [15]) that the minimax density estimation error under TV distance is $\Theta(n^{-1/3})$ for both $\mathcal{P}$ and $\mathcal{P}_c$. The next theorem shows that the sample complexities for amplification are actually different.

THEOREM 7.3.   *Let $L \geq 8$ and $c \in (0, 1)$ be fixed. It holds that*
$$m^\star(\mathcal{P}_c, n) \asymp n^{5/6}, \quad \text{while} \quad m^\star(\mathcal{P}, n) \lesssim n^{3/4}.$$

Theorem 7.3 shows that, although assuming a density lower bound does not alter the TV estimation error, it boosts the size of amplified samples from $O(n^{3/4})$ to $\Theta(n^{5/6})$. In fact, the χ^2-estimation error is also reduced from $\mathcal{P}$ to $\mathcal{P}_c$: in the proof of Theorem 7.3 we show that $r_{\chi^2}(\mathcal{P}_c, n) \lesssim n^{-2/3}$, but $m^\star(\mathcal{P}, n) \lesssim n^{3/4}$ together with Theorem 5.2 imply that $r_{\chi^2}(\mathcal{P}, n) \gtrsim n^{-1/2}$. Therefore, this is an example suggesting that measuring the estimation error under the χ^2 divergence might be a better indicator for the complexity of sample amplification than the TV distance.

SUPPLEMENTARY MATERIAL

Supplement to "On the Statistical Complexity of Sample Amplification"

We provide proofs of main theorems (Theorem 4.5, 4.6, 5.2, 5.5, 6.2, 6.3, 6.4, 6.5, 7.1, 7.2, 7.3) and lemmas (Lemma 4.4, 5.8).

REFERENCES

[1] ANTONIOU, A., STORKEY, A. and EDWARDS, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

[2] AXELROD, B., GARG, S., SHARAN, V. and VALIANT, G. (2020). Sample amplification: Increasing dataset size even when learning is impossible. In *International Conference on Machine Learning* 442–451. PMLR.

[3] BAI, Y., KADAVATH, S., KUNDU, S., ASKELL, A., KERNION, J., JONES, A., CHEN, A., GOLDIE, A., MIRHOSEINI, A., MCKINNON, C. et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

[4] BALLY, V. and CARAMELLINO, L. (2014). On the distances between probability density functions. *Electronic Journal of Probability* **19**.

[5] BALLY, V. and CARAMELLINO, L. (2016). Asymptotic development for the CLT in total variation distance. *Bernoulli* **22** 2442–2485.

[6] BARRON, A. R. (1986). Entropy and the central limit theorem. *The Annals of probability* 336–342.

[7] BASU, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **15** 377–380.

[8] BASU, D. (1958). On statistics independent of sufficient statistics. *Sankhyā: The Indian Journal of Statistics* 223–226.

[9] BASU, D. (1959). The family of ancillary statistics. *Sankhyā: The Indian Journal of Statistics* 247–256.

[10] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics* **41** 1780–1815.

[11] BHATTACHARYA, R. N. and RAO, R. R. (2010). *Normal approximation and asymptotic expansions*. SIAM.

[12] BOBKOV, S. G., CHISTYAKOV, G. P. and GÖTZE, F. (2014). Berry–Esseen bounds in the entropic central limit theorem. *Probability Theory and Related Fields* **159** 435–478.

[13] BRENNAN, M. and BRESLER, G. (2019). Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. In *Conference on Learning Theory* 469–470. PMLR.

[14] BRENNAN, M. and BRESLER, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory* 648–847. PMLR.

[15] BROWN, L. D., CAI, T. T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of statistics* **30** 688–707.

[16] BROWN, L. D., CARTER, A. V., LOW, M. G. and ZHANG, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *The Annals of Statistics* **32** 2074–2097.

[17] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics* **24** 2384–2398.

[18] CALIMERI, F., MARZULLO, A., STAMILE, C. and TERRACINA, G. (2017). Biomedical data augmentation using generative adversarial neural networks. In *International conference on artificial neural networks* 626–634. Springer.

[19] CHATZIAGAPI, A., PARASKEVOPOULOS, G., SGOUROPOULOS, D., PANTAZOPOULOS, G., NIKANDROU, M., GIANNAKOPOULOS, T., KATSAMANIS, A., POTAMIANOS, A. and NARAYANAN, S. (2019). Data Augmentation Using GANs for Speech Emotion Recognition. In *Interspeech* 171–175.

[20] CHEN, D., QI, X., ZHENG, Y., LU, Y. and LI, Z. (2022). Deep Data Augmentation for Weed Recognition Enhancement: A Diffusion Probabilistic Model and Transfer Learning Based Approach. *arXiv preprint arXiv:2210.09509*.

[21] CHLAP, P., MIN, H., VANDENBERG, N., DOWLING, J., HOLLOWAY, L. and HAWORTH, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology* **65** 545–563.

[22] CUBUK, E. D., ZOPH, B., MANE, D., VASUDEVAN, V. and LE, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 113–123.

[23] CUBUK, E. D., ZOPH, B., SHLENS, J. and LE, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* 702–703.

[24] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *The Annals of statistics* 269–281.

[25] FRID-ADAR, M., DIAMANT, I., KLANG, E., AMITAI, M., GOLDBERGER, J. and GREENSPAN, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321** 321–331.

[26] HAN, C., RUNDO, L., ARAKI, R., NAGANO, Y., FURUKAWA, Y., MAURI, G., NAKAYAMA, H. and HAYASHI, H. (2019). Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. *IEEE Access* **7** 156966–156977.

[27] HAN, Y., JIAO, J. and WEISSMAN, T. (2015). Minimax Estimation of Discrete Distributions Under ℓ_1 Loss. *IEEE Transactions on Information Theory* **61** 6343–6354.

[28] HENDRYCKS, D., MU, N., CUBUK, E. D., ZOPH, B., GILMER, J. and LAKSHMINARAYANAN, B. (2020). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

[29] KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**.

[30] LE CAM, L. (1964). Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics* 1419–1455.

[31] LE CAM, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California.

[32] LE CAM, L. (1986). *Asymptotic methods in statistical decision theory*. Springer.

[33] LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*. Springer, New York.

[34] LEHMANN, E. L. and SCHOLZ, F. (1992). Ancillarity. *Lecture Notes-Monograph Series* **17** 32–51.

[35] LU, Y., SHEN, M., WANG, H., WANG, X., VAN RECHEM, C. and WEI, W. (2023). Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*.

[36] LUZI, L., SIAHKOOHI, A., MAYER, P. M., CASCO-RODRIGUEZ, J. and BARANIUK, R. (2022). Boomerang: Local sampling on image manifolds using diffusion models. *arXiv preprint arXiv:2210.12100*.

[37] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics* **43** 1089–1116.

[38] MADANI, A., MORADI, M., KARARGYRIS, A. and SYEDA-MAHMOOD, T. (2018). Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical imaging 2018: Image processing* **10574** 415–420. SPIE.

[39] MIESCKE, K.-J. and LIESE, F. (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer.

[40] NESTEROV, Y. (2003). *Introductory lectures on convex optimization: A basic course* **87**. Springer Science & Business Media.

[41] PROHOROV, Y. V. (1952). A local theorem for densities. In *Doklady Akad. Nauk SSSR (NS)* **83** 797–800.

[42] RAY, K. and SCHMIDT-HIEBER, J. (2019). Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **55** 2195–2208. Institut Henri Poincaré.

[43] SANDFORT, V., YAN, K., PICKHARDT, P. J. and SUMMERS, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Nature Scientific reports* **9** 1–9.

[44] SHORTEN, C. and KHOSHGOFTAAR, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data* **6** 1–48.

[45] SIMARD, P. Y., STEINKRAUS, D., PLATT, J. C. et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of International Conference on Document Analysis and Recognition* **3**. Edinburgh.

[46] SIRAZHDINOV, S. K. and MAMATOV, M. (1962). On convergence in the mean for densities. *Theory of Probability & Its Applications* **7** 424–428.

[47] TOKOZUME, Y., USHIKU, Y. and HARADA, T. (2018). Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 5486–5494.

[48] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

[49] VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

[50] VERMA, V., LAMB, A., BECKHAM, C., NAJAFI, A., MITLIAGKAS, I., LOPEZ-PAZ, D. and BENGIO, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning* 6438–6447. PMLR.

[51] WALD, A. (1950). *Statistical decision functions.* Wiley.

[52] WILLIAMS, E. (1982). Some classes of conditional inference procedures. *Journal of Applied Probability* **19** 293–303.

[53] YI, W., SUN, Y. and HE, S. (2018). Data augmentation using conditional GANs for facial emotion recognition. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)* 710–714. IEEE.

[54] YI, X., WALIA, E. and BABYN, P. (2019). Generative adversarial network in medical imaging: A review. *Medical image analysis* **58** 101552.

[55] YUN, S., HAN, D., OH, S. J., CHUN, S., CHOE, J. and YOO, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision* 6023–6032.

[56] ZHANG, H., CISSE, M., DAUPHIN, Y. N. and LOPEZ-PAZ, D. (2018). mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

SUPPLEMENT TO "ON THE STATISTICAL COMPLEXITY OF SAMPLE AMPLIFICATION"

BY BRIAN AXELROD, SHIVAM GARG, YANJUN HAN, VATSAL SHARAN, AND GREGORY VALIANT

APPENDIX A: CONCRETE EXAMPLES OF SAMPLE AMPLIFICATION

In this section we include concrete examples of sample amplification omitted in the main text due to space limitations, including the non-asymptotic upper bounds in Section 4 and 5, and the lower bounds for non-product models in Section 6.

A.1. Concrete examples of amplification via sufficiency.   In this section, in contrast to the mostly asymptotic results in Section 4.3, we investigate several non-asymptotic examples of sample amplification in concrete models. We show that for many natural models, including exponential family with dependent coordinates and non-exponential family which are not covered in the general theory, the sufficiency-based sample amplification approach could still amplify $\Omega(n/\sqrt{d})$ additional samples. We also illustrate the computational idea in Section 4.2 via more involved examples.

Our first example concerns the Gaussian model with a known mean but an unknown covariance. It is a folklore result that estimating the unknown covariance in a vanishing Frobenius norm requires $n = \Omega(d^2)$ samples [6, Corollary 1.2], which is also the sample complexity for learning the distribution within a small TV distance. The following example shows that $n = O(d)$ samples suffice for sample amplification.

EXAMPLE A.1 (Gaussian covariance model with known mean).   *Consider the i.i.d. observations $X_1, \cdots, X_n$ drawn from $\mathcal{N}(0, \Sigma)$ with zero mean and an unknown covariance $\Sigma \in \mathbb{R}^{d \times d}$. Here a minimal sufficient statistic is the sample covariance matrix*

$$\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

Lemma A.2 shows that $\|\mathcal{L}(\widehat{\Sigma}_n) - \mathcal{L}(\widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \epsilon$ as long as $m = O(\epsilon n/d)$, therefore drawing samples from $P_{X^{n+m}|\widehat{\Sigma}_{n+m}}$ achieves sample amplification of size $m = \Omega(n/d)$. This coincides with the general $O(n/\sqrt{D})$ result where $D \asymp d^2$ is the parameter dimension.

In order to sample from this conditional distribution, consider the following statistic

$$S_{n+m} = [(n+m)\widehat{\Sigma}_{n+m}]^{-1/2}[X_1, X_2, \cdots, X_{n+m}] \in \mathbb{R}^{d \times (n+m)},$$

which always exists even if $\widehat{\Sigma}_{n+m}$ is not invertible. Clearly there is a bijection between X^{n+m} and $(\widehat{\Sigma}_{m+n}, S_{n+m})$, and Lemma A.2 shows that S_{n+m} is an ancillary statistic. In particular, S_{n+m} always follows the uniform distribution on the following set:

$$A = \{U \in \mathbb{R}^{d \times (n+m)} : UU^\top = I_d\}.$$

Consequently an $(n, n + \Omega(n\epsilon/d), \epsilon)$ sample amplification is efficiently achievable in the Gaussian covariance model using the following algorithm:

1. *Given samples $[X_1, X_2, \cdots, X_n]$ compute $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$.*
2. *Sample $[Z_1, Z_2, \cdots, Z_{n+m}]$ where $Z_i \sim N(0, I)$ for all i. Compute $\widehat{\Sigma}_{n+m} = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top$. Based on these compute $S_{n+m} = [(n+m)\widehat{\Sigma}_{n+m}]^{-1/2}[Z_1, Z_2, \cdots, Z_{n+m}]$.*
3. *Output $(n+m)$ samples $X^{n+m} = [(n+m)\widehat{\Sigma}_n]^{1/2}S_{n+m}$.*

LEMMA A.2. *Under the notations of Example A.1, for $n \geq 4 \max\{m, d\}$ it holds that*

$$\|\mathcal{L}(\widehat{\Sigma}_n) - \mathcal{L}(\widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \frac{2md}{n}.$$

In addition, S_{n+m} is uniformly distributed on $A = \{U \in \mathbb{R}^{d \times (n+m)} : UU^\top = I_d\}$.

The next example shows that the sample amplification result does not change much if both the mean and covariance are unknown, though the sampling procedure becomes slightly more involved.

EXAMPLE A.3 (Gaussian model with unknown mean and covariance). *Next we consider the most general Gaussian model where $X_1, \cdots, X_n \sim \mathcal{N}(\theta, \Sigma)$ with unknown mean vector $\theta \in \mathbb{R}^d$ and unknown covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. In this case, a minimal sufficient statistic is the pair $(\overline{X}_n, \widehat{\Sigma}_n)$, with*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \qquad \widehat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^\top.$$

Lemma A.4 shows that $\|\mathcal{L}(\overline{X}_n, \widehat{\Sigma}_n) - \mathcal{L}(\overline{X}_{n+m}, \widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \epsilon$ as long as $m = O(n\epsilon/d)$, therefore drawing amplified samples from $P_{X^{n+m}|(\overline{X}_{n+m}, \widehat{\Sigma}_{n+m})}$ achieves sample amplification of size $m = \Omega(n\epsilon/d)$. For the computation, consider the following statistic

$$S_{n+m} = [(n+m-1)\widehat{\Sigma}_{n+m}]^{-1/2}[X_1 - \overline{X}_{n+m}, \cdots, X_{n+m} - \overline{X}_{n+m}] \in \mathbb{R}^{d \times (n+m)},$$

and it is clear that the whole samples X^{n+m} could be recovered from $(\overline{X}_{n+m}, \widehat{\Sigma}_{n+m}, S_{n+m})$. Again, Lemma A.4 shows that S_{n+m} is ancillary and uniformly distributed on the following set (assuming $n + m - 1 \geq d$):

$$A = \{U \in \mathbb{R}^{d \times (n+m)} : UU^\top = I_d, U\mathbf{1} = \mathbf{0}\}.$$

LEMMA A.4. *Under the notations of Example A.3, for $n \geq 4 \max\{m, d\}$ it holds that*

$$\|\mathcal{L}(\overline{X}_n, \widehat{\Sigma}_n) - \mathcal{L}(\overline{X}_{n+m}, \widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \frac{3md}{n-1}.$$

In addition, S_{n+m} is uniformly distributed on $A = \{U \in \mathbb{R}^{d \times (n+m)} : UU^\top = I_d, U\mathbf{1} = \mathbf{0}\}$.

The following example concerns the product exponential distributions, or general Gamma distributions with a fixed shape parameter.

EXAMPLE A.5 (Product exponential distribution). *In this example, we consider the product exponential model where $X_1, \cdots, X_n \sim \prod_{i=1}^d \mathrm{Exp}(\lambda_i)$ with unknown rate vector $\lambda = (\lambda_1, \cdots, \lambda_d)$. Again, in this model the sample mean $\overline{X}_n$ is sufficient, and follows a product Gamma distribution $\prod_{i=1}^d \mathrm{Gamma}(n, n\lambda_i)$. Consequently,*

$$D_{\mathrm{KL}}(\mathcal{L}(\overline{X}_n)\|\mathcal{L}(\overline{X}_{n+m})) = d \cdot \left(m - (n+m)\log\left(1 + \frac{m}{n}\right) + \log\frac{\Gamma(n+m)}{\Gamma(n)} - m\psi(n)\right),$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ and $\psi(x) = \frac{d}{dx}[\log\Gamma(x)]$ are the gamma and digamma functions, respectively. Using the definition of $f(n, m, d)$ in (C.3), we note that the above KL divergence is precisely $d \cdot f(2n, 2m, 1)$, thus by the proof of Lemma A.2 is further at most $O(dm^2/n^2)$. Consequently, sample amplification is possible as long as $n = \Omega(\sqrt{d}/\epsilon)$ and

$m = O(n\epsilon/\sqrt{d})$. *Alternatively, the same result could also follow from Theorem 4.6, for both assumptions 1 and 2 hold for the exponential distribution.*

To draw amplified samples $X_1, \cdots, X_{n+m}$ conditioned on $\overline{X}_{n+m}$, note that the following statistic $S \in \mathbb{R}^{d \times (n+m)}$ with i-th row being

$$S_i = \left(\frac{X_{1,i}}{\overline{X}_{n+m,i}}, \frac{X_{2,i}}{\overline{X}_{n+m,i}}, \cdots, \frac{X_{n+m,i}}{\overline{X}_{n+m,i}} \right),$$

is ancillary. In fact, each S_i follows the Dirichlet distribution $\mathrm{Dir}(1, 1, \cdots, 1)$. Then computational efficiency follows from the obvious fact that X^{n+m} is determined by $(\overline{X}_{n+m}, S)$.

Our final example is a non-exponential family which is not even differentiable in quadratic mean, but the $n = O(\sqrt{d})$ sample complexity still holds.

EXAMPLE A.6 (Uniform distribution over a rectangle). *In this example, let $X_1, \cdots, X_n$ be i.i.d. samples from the uniform distribution on an unknown rectangle $\prod_{j=1}^{d}[a_j, b_j]$ in $\mathbb{R}^d$. Note that this is not an exponential family. However, the sufficiency-based sample amplification could still be applied in this case. Specifically, here the sufficient statistics are $X_{\min}^n$ and $X_{\max}^n$, where for $j \in [d]$,*

$$X_{\min,j}^n = \min_{i \in [n]} X_{i,j}, \qquad X_{\max,j}^n = \max_{i \in [n]} X_{i,j}.$$

It is not hard to find that, the joint density of $(X_{\min,j}^n, X_{\max,j}^n)$ is

$$f_n(a,b) = \frac{n(n-1)(b-a)^{n-2}}{(b_j - a_j)^n} \cdot \mathbb{1}(a_j \le a \le b \le b_j).$$

Then after some algebra, the KL divergence between the sufficient statistics is

$$D_{\mathrm{KL}}(\mathcal{L}(X_{\min}^n, X_{\max}^n) \| \mathcal{L}(X_{\min}^{n+m}, X_{\max}^{n+m})) = d \left(\frac{m}{n} - \log\left(1 + \frac{m}{n}\right) + \frac{m}{n-1} - \log\left(1 + \frac{m}{n-1}\right) \right),$$

which is $O(dm^2/n^2)$ when $m = O(n)$. Therefore, sample amplification for uniform distributions is still possible whenever $m = O(n\epsilon/\sqrt{d})$ and $n = \Omega(\sqrt{d}/\epsilon)$, same as exponential families.

To draw amplified samples $X_1, \cdots, X_{m+n}$ conditioned on $(X_{\min}^{n+m}, X_{\max}^{m+n})$, note that the following statistic $S \in \mathbb{R}^{d \times (n+m)}$ with i-th row being

$$S_i = \left(\frac{X_{1,i} - X_{\min}^{n+m}}{X_{\max}^{m+n} - X_{\min}^{m+n}}, \cdots, \frac{X_{m+n,i} - X_{\min}^{n+m}}{X_{\max}^{m+n} - X_{\min}^{m+n}} \right),$$

is ancillary. This is due to the invariance property of the uniform distribution: if $X \sim \mathsf{U}(0,1)$, then $aX + b \sim \mathsf{U}(b, a+b)$. Since $(X_1, \cdots, X_{n+m})$ is determined by $(X_{\min}^{n+m}, X_{\max}^{n+m}, S)$, the computational efficiency follows.

Although all of the above examples work exclusively for continuous models and apply an identity map between sufficient statistics, we remark that more sophisticated maps based on learning could be useful and work for discrete models. The idea of sample amplification via learning is presented in Section 5, and an example which combines both the sufficiency and learning ideas could be found in Example A.14.

A.2. Concrete examples of amplification via learning. In this section we show how learning-based approaches achieve optimal performances of sample amplification in several examples, including both continuous and discrete models. In some scenarios the following strengthened lemma may be useful to deal with too large χ^2-divergence.

LEMMA A.7. *The same results of Theorem 5.2 and 5.5 hold for the following modification of the χ^2-estimation error:*

$$r_{\bar{\chi}^2}(\mathcal{P}, n) = \inf_{\widehat{P}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\chi^2(\widehat{P}_n, P) \wedge n \right].$$

The idea behind Lemma A.7 is that for some models, it might happen that $\chi^2(\widehat{P}_n, P) = \infty$ with a small probability. However, for the TV distance we always have $\|\widehat{P}_n - P\|_{\mathrm{TV}} \leq 1$, so a large χ^2-divergence could still lead to a meaningful TV distance. The proof of Lemma A.7 could be found in Appendix C.

The first example is again the Gaussian location model in Example 4.1, where we show that the shuffling-based approach also achieves the complexity $n = O(\sqrt{d}/\epsilon)$ and the size $m = \Omega(n\epsilon/\sqrt{d})$.

EXAMPLE A.8 (Gaussian location model with known covariance, continued). *Consider the setting of Example 4.1, where the family of distributions is $\mathcal{P} = \{\mathcal{N}(\theta, I_d)\}_{\theta \in \mathbb{R}^d}$. Here $\mathcal{P}$ has a product structure, with $\mathcal{P}_j = \{\mathcal{N}(\theta_j, 1)\}_{\theta_j \in \mathbb{R}}$ for each $j \in [d]$. To find an upper bound on the χ^2-estimation error, consider the distribution estimator $\widehat{P}_{n,j} = \mathcal{N}(\widehat{\theta}_j, 1)$, where $\widehat{\theta}_j = n^{-1} \sum_{i=1}^n X_{i,j}$. Consequently,*

$$\chi^2(\widehat{P}_{n,j}, P_j) = \chi^2(\mathcal{N}(\widehat{\theta}_j, 1), \mathcal{N}(\theta_j, 1)) = \exp((\widehat{\theta}_j - \theta_j)^2) - 1,$$

and using $\widehat{\theta}_j - \theta_j \sim \mathcal{N}(0, 1/n)$, we have

$$\mathbb{E}[\chi^2(\widehat{P}_{n,j}, P_j)] = \sqrt{\frac{n}{n-2}} - 1 = O\left(\frac{1}{n}\right)$$

whenever $n \geq 3$. Consequently, $r_{\chi^2}(\mathcal{P}_j, n) = O(1/n)$ for all $j \in [d]$, and Theorem 5.5 implies that an $(n, n+m, \epsilon)$ sample amplification is possible if $n = \Omega(\sqrt{d}/\epsilon)$ and $m = O(n\epsilon/\sqrt{d})$.

The next example is the discrete distribution model considered in [2].

EXAMPLE A.9 (Discrete distribution model). *Let $\mathcal{P}$ be the class of all discrete distributions supported on k elements. In this case, a natural learner is the empirical distribution $\widehat{P}_n = (\widehat{p}_1, \cdots, \widehat{p}_k)$, with*

$$\widehat{p}_j = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i = j), \qquad j \in [k].$$

Consequently,

$$\mathbb{E}[\chi^2(\widehat{P}_n, P)] = \mathbb{E}\left[\sum_{j=1}^k \frac{(\widehat{p}_j - p_j)^2}{p_j} \right] = \sum_{j=1}^k \frac{1 - p_j}{n} = \frac{k-1}{n},$$

meaning that $r_{\chi^2}(\mathcal{P}, n) \leq (k-1)/n$. Hence, Theorem 5.2 implies that sample amplification is possible whenever $n = \Omega(\sqrt{k}/\epsilon)$ and $m = O(n\epsilon/\sqrt{k})$. In this case, Algorithm 2 essentially subsamples from the original data and add them back with random shuffling, which is the same algorithm in [2]. However, the analysis here is much simplified.

The following example revisits the uniform distribution in Example A.6, and again shows that the same amplification performance could be achieved by random shuffling.

EXAMPLE A.10 (Uniform distribution, continued). *Consider the setting in Example A.6, where the distribution class $\mathcal{P}$ is the family of all uniform distributions on a rectangle $\prod_{j=1}^{d}[a_j, b_j]$. For each j, a natural distribution estimator $\widehat{P}_{n,j}$ is simply the uniform distribution on $[X_{\min,j}^n, X_{\max,j}^n]$, where these quantities are defined in Example A.6. For this learner, it holds that*

$$\chi^2\left(\widehat{P}_{n,j}, P_j\right) = \frac{(a_j - b_j)^2}{(X_{\max,j}^n - X_{\min,j}^n)^2} - 1.$$

Note that Example A.6 shows that the joint density of $(X_{\min,j}^n, X_{\max,j}^n)$ is given by

$$f_n(a, b) = \frac{n(n-1)(b-a)^{n-2}}{(b_j - a_j)^n} \cdot \mathbb{1}(a_j \le a \le b \le b_j),$$

the expected χ^2-divergence could then be computed as

$$\mathbb{E}\left[\chi^2\left(\widehat{P}_{n,j}, P_j\right)\right] = -1 + \iint_{a_j \le a \le b \le b_j} \frac{n(n-1)(b-a)^{n-4}}{(b_j - a_j)^{n-2}} da\, db = \frac{4n-6}{(n-2)(n-3)},$$

which is $O(n^{-1})$ if $n \ge 4$. Hence, we have $r_{\chi^2}(\mathcal{P}_j, n) = O(1/n)$ for each $j \in [d]$, and Theorem 5.5 shows an $(n, n+m, \epsilon)$ sample amplification if $n = \Omega(\sqrt{d}/\epsilon)$ and $m = O(n\epsilon/\sqrt{d})$.

The next example is the exponential distribution model studied in Example A.5, where the modified χ^2-learning error $r_{\bar{\chi}^2}(\mathcal{P}, n)$ in Lemma A.7 will be useful.

EXAMPLE A.11 (Exponential distribution, continued). *Consider the setting in Example A.5, where the distribution class $\mathcal{P}$ is a product of exponential distributions with unknown rate parameters. In this case, a natural distribution learner is to estimate each rate parameter as $\widehat{\lambda}_{n,j} = n/\sum_{i=1}^{n} X_{i,j}$, and use $\mathrm{Exp}(\widehat{\lambda}_{n,j})$ to estimate the truth $\mathrm{Exp}(\lambda_j)$. Note that*

$$\chi^2\left(\mathrm{Exp}(\widehat{\lambda}_{n,j}), \mathrm{Exp}(\lambda_j)\right) = \frac{(\widehat{\lambda}_{n,j} - \lambda_j)^2}{\lambda_j(2\widehat{\lambda}_{n,j} - \lambda_j)}$$

whenever $2\widehat{\lambda}_{n,j} > \lambda_j$. However, if $2\widehat{\lambda}_{n,j} \le \lambda_j$ the χ^2-divergence will be unbounded, which occurs with a small but positive probability.

To address this issue, we note that when $\lambda_j = 1$, the sub-exponential concentration claims that $|\sum_{i=1}^{n} X_{i,j} - n| \le n/3$ with probability at least $1 - \exp(-\Omega(n))$. By a simple scaling, the above event implies that $\widehat{\lambda}_{n,j}/\lambda_j \in [3/4, 3/2]$. Hence,

$$\mathbb{E}\left[\chi^2\left(\mathrm{Exp}(\widehat{\lambda}_{n,j}), \mathrm{Exp}(\lambda_j)\right) \wedge n\right] \le 2 \cdot \mathbb{E}\left[\left(\frac{\widehat{\lambda}_{n,j}}{\lambda_j} - 1\right)^2\right] + n \cdot \exp(-\Omega(n)) = O\left(\frac{1}{n}\right),$$

which means that $r_{\bar{\chi}^2}(\mathcal{P}_j, n) = O(1/n)$. Therefore, by Lemma A.7, the sample amplification is possible whenever $n = \Omega(\sqrt{d}/\epsilon)$ and $m = O(n\epsilon/\sqrt{d})$, the same as Example A.5.

The following example considers an interesting non-product model, i.e. the Gaussian distribution with a sparse mean vector.

EXAMPLE A.12 (Sparse Gaussian model). *Consider the Gaussian location model* $\mathcal{P} = \{\mathcal{N}(\theta, I_d)\}_{\theta \in \Theta}$*, with an additional constraint that the mean vector* θ *is* s*-sparse, i.e.* $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s\}$*. For the learning problem, it is well-known (cf. [7, Theorem 1]) that the soft-thresholding estimator* $\widehat{\theta}_n$ *with*

$$\widehat{\theta}_{n,j} = \text{sign}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right) \cdot \left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right| - \sqrt{\frac{C \log n}{n}}\right)_{+}$$

and any constant $C > 2$ *achieves that* $\sup_{\theta \in \Theta} \mathbb{E}[\|\widehat{\theta}_n - \theta\|_2^2] = O(s \log d / n)$*. Therefore, the sample complexity of learning sparse Gaussian distributions is* $n = O(s \log d)$*.*

For the complexity of sample amplification, for each $j \in [d]$ *we apply* $\mathcal{N}(\widehat{\theta}_{n,j}, 1)$ *as the distribution estimator. The* χ^2*-estimation performance of this estimator is summarized in Lemma A.13. Therefore, by a simple adaptation of Lemma A.7, an* $(n, n+m)$ *sample amplification is possible as long as* $n = \Omega(\sqrt{s \log d}/\epsilon)$ *and* $m = O(n\epsilon/\sqrt{s \log d})$*.*

LEMMA A.13. *Under the setting of Example A.12, it holds that*

$$\sup_{\theta \in \Theta} \sum_{j=1}^{d} \mathbb{E}\left[\chi^2\left(\mathcal{N}(\widehat{\theta}_{n,j}, 1), \mathcal{N}(\theta_j, 1)\right) \wedge n\right] = O\left(\frac{s \log d}{n}\right).$$

The final example is a special example where the sole application of either sufficient statistics or learning will fail to achieve the optimal sample amplification. The solution is to use a combination of both ideas.

EXAMPLE A.14 (Poisson distribution). *Consider the product Poisson model* $\prod_{j=1}^{d} \text{Poi}(\lambda_j)$ *with* $\lambda \in \mathbb{R}_+^d$*. We first show that a naïve application of either sufficiency-based or shuffling-based idea will not lead to a desired sample amplification. In fact, the sufficient statistic here is* $T_n = \sum_{i=1}^{n} X_i$ *which follows a product Poisson distribution* $\prod_{j=1}^{d} \text{Poi}(n\lambda_j)$*; as* T_n *takes discrete values, applying any linear map between* T_n *to* T_{n+m} *will not result in a small TV distance between sufficient statistics.*

The argument for the shuffling-based approach is subtler. A natural distribution estimator for $\text{Poi}(\lambda_j)$ *is* $\text{Poi}(\widehat{\lambda}_{n,j})$*, with* $\widehat{\lambda}_{n,j} = n^{-1}\sum_{i=1}^{n} X_{i,j}$*. Lemma A.15 shows that this distribution estimator could suffer from an expected* χ^2*-estimation error far greater than* $\Omega(1/n)$*, so we cannot conclude an* $(n, n + \Omega(n\epsilon/\sqrt{d}), \epsilon)$ *sample amplification from Theorem 5.5.*

Now we show that a combination of the sufficient statistic and learning leads to the rate-optimal sample amplification in this model. Specifically, we split samples and compute the empirical rate parameter $\widehat{\lambda}_{n/2}$ *based on the first* $n/2$ *samples. Next, conditioned on the first half of samples, the sufficient statistic for the remaining half is* $T_{n/2} = \sum_{i=n/2+1}^{n} X_i$*. Define*

$$\widehat{T}_{n/2+m} = T_{n/2} + Z,$$

where $Z \sim \prod_{j=1}^{d} \text{Poi}(m\widehat{\lambda}_{n/2,j})$ *is independent of* $T_{n/2}$ *conditioning on the first half samples. Finally, we generate the* $(n/2 + m)$ *amplified samples from the conditional distribution and append them to* $(X_1, \cdots, X_{n/2})$*. By the second statement of Lemma A.15, this achieves an* $(n, n + m, \epsilon)$ *sample amplification whenever* $n = \Omega(\sqrt{d}/\epsilon)$ *and* $m = O(n\epsilon/\sqrt{d})$*.*

LEMMA A.15. *Under the settings of Example A.14, there exists some* $\lambda_j > 0$ *such that*

$$\mathbb{E}\left[\chi^2\left(\text{Poi}(\widehat{\lambda}_{n,j}), \text{Poi}(\lambda_j)\right) \wedge n\right] = \Omega\left(\frac{1}{\log n}\right) \gg \frac{1}{n}.$$

In addition, the proposed sufficiency+learning approach satisfies

$$\|\mathcal{L}(\widehat{X}^{n+m}) - \mathcal{L}(X^{n+m})\|_{\mathrm{TV}} \leq \frac{m\sqrt{2d}}{n}.$$

A.3. Non-asymptotic examples of lower bounds. In this section, we apply the lower bound ideas to several concrete examples in Section 4 and 5, and show that the previously established upper bounds for sample amplification are indeed tight. Since Theorem 6.5 already handles all product models (including the Gaussian location model in Example 4.1, 4.2, and A.8, exponential model in Example A.5 and A.11, uniform model in Example A.6 and A.10, and Poisson model in Example A.14), we are only left with the remaining non-product models. In the sequel, we will make use of the general Lemma 6.1 to prove non-asymptotic lower bounds in these examples.

The lower bound of the discrete distribution model was obtained in [2], where the learning approach in Example A.9 is rate optimal. Our first example concerns the "Poissonized" version of the discrete distribution model, and we show that the results become slightly different.

EXAMPLE A.16 ("Poissonized" discrete distribution model). *We consider the following "Poissonized" discrete distribution model, where we have n i.i.d. samples drawn from $\prod_{j=1}^{k} \mathsf{Poi}(p_j)$, with $(p_1, \cdots, p_k)$ being an unknown probability vector. Although the Poissonization does not affect the optimal rate of estimation in many problems, Lemma A.17 shows the following distinction when it comes to sample amplification: the optimal amplification size is $m = \Theta(n\epsilon/\sqrt{k} + \sqrt{n}\epsilon)$ under the Poissonized model, while it is $m = \Theta(n\epsilon/\sqrt{k})$ under the non-Poissonized model [2].*

The complete proof of Lemma A.17 is relegated to the appendix, but we briefly comment on why Theorem 6.5 is not directly applicable when $k \gg n$. The reason here is that to apply Theorem 6.5, we will construct a parametric submodel which is a product model:

$$P_\theta = \prod_{j=1}^{k_0} \left[\mathsf{Poi}\left(\frac{1}{k} + \theta_j\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta_j\right) \right], \qquad \theta \in \Theta \triangleq \left[-\frac{1}{k}, \frac{1}{k} \right]^{k_0},$$

where $k = 2k_0$. However, for $\theta, \theta' \in [0, 1/k]$, the range of the squared Hellinger distance

$$H^2\left(\mathsf{Poi}\left(\frac{1}{k} + \theta\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta\right), \mathsf{Poi}\left(\frac{1}{k} + \theta'\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta'\right) \right)$$

is only $[0, \Theta(1/k)]$, so Assumption 4 does not hold when $k \gg n$. This is precisely the subtlety in the Poissonized model.

LEMMA A.17. *Under the Poissonized discrete distribution model, an $(n, n+m, \epsilon)$ sample amplification is possible if and only if $n = \Omega(1)$ and $m = O(n\epsilon/\sqrt{k} + \sqrt{n}\epsilon)$.*

Our next example is the sparse Gaussian model in Example A.12, where we establish the tightness of $n = \Omega(\sqrt{s \log d}/\epsilon)$ and $m = O(n\epsilon/\sqrt{s \log d})$ directly using Lemma 6.1.

EXAMPLE A.18 (Sparse Gaussian location model). *Consider the setting of the sparse Gaussian location model in Example A.12, and we aim to prove a matching lower bound for sample amplification. In the sequel, we first handle the case $s = 1$ to reflect the main idea, and postpone the case of general s to Lemma A.19.*

For $s = 1$, we apply Lemma 6.1 to a proper choice of the prior μ and loss L. Fixing a parameter $t > 0$ to be chosen later, let μ be the uniform distribution on the finite set of vectors $\{te_1, \cdots, te_d\}$, where $e_1, \cdots, e_d$ are the canonical vectors in $\mathbb{R}^d$. Moreover, for an estimator

$\widehat{\theta} \in \mathbb{R}^d$ *of the unknown mean vector, the loss function is chosen to be* $L(\theta, \widehat{\theta}) = \mathbb{1}(\widehat{\theta} \neq \theta)$. *For the above prior and loss, it is clear that the maximum likelihood estimator* $\widehat{\theta} = t e_{\widehat{j}}$ *with*

$$\widehat{j} = \arg\max_{j \in [d]} \sum_{i=1}^{n} X_{i,j}$$

is the Bayes estimator, and the Bayes risk admits the following expression:

$$r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) = 1 - \mathbb{E}\left[\frac{\exp(\sqrt{n}t(\sqrt{n}t + Z_1))}{\exp(\sqrt{n}t(\sqrt{n}t + Z_1)) + \sum_{j=2}^{d}\exp(\sqrt{n}t Z_j)}\right] \triangleq 1 - p_d(\sqrt{n} \cdot t),$$

where $Z_1, \cdots, Z_d \sim \mathcal{N}(0, 1)$ *are i.i.d. standard normal random variables. Similarly, the Bayes risk under* $n + m$ *samples is* $r_{\mathrm{B}}(\mathcal{P}, n+m, \mu, L) = 1 - p_d(\sqrt{n+m} \cdot t)$, *and it remains to investigate the property of the function* $p_d(\cdot)$. *Lemma A.19 summarizes a useful property for the lower bound, i.e. the function* $p_d(z)$ *enjoys a phase transition around* $z \sim \sqrt{2 \log d}$.

Based on Lemma A.19, the pigeonhole principle implies that for every given $c > 0$, *there exists some* $z \in [\sqrt{2 \log d} - C, \sqrt{2 \log d} + C]$ *such that* $p_d(z + c\epsilon) - p_d(z) = \Omega_c(\epsilon)$. *Therefore, if* $m = \lceil cn\epsilon/\sqrt{\log d} \rceil$, *choosing* $t = z/\sqrt{n}$ *for the above* z *yields that* $p_d(\sqrt{n+m} \cdot t) - p_d(\sqrt{n} \cdot t) = \Omega_c(\epsilon)$, *and consequently*

$$\epsilon^\star(\mathcal{P}, n, m) \geq r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) - r_{\mathrm{B}}(\mathcal{P}, n+m, \mu, L) = p_d(\sqrt{n+m} \cdot t) - p_d(\sqrt{n} \cdot t) = \Omega_c(\epsilon).$$

Therefore, we must have $n = \Omega(\sqrt{\log d}/\epsilon)$ *and* $m = O(n\epsilon/\sqrt{\log d})$ *for sample amplification.*

The following lemma summarizes the phase-transition property of p_d used in the above example.

LEMMA A.19. *For the function* $p_d(z)$ *in Example A.18, there exists an absolute constant* C *independent of* d *such that*

$$p_d(\sqrt{2 \log d} - C) \leq 0.1,$$
$$p_d(\sqrt{2 \log d} + C) \geq 0.9.$$

Moreover, for general $s < d/2$, *an* $(n, n+m)$ *sample amplification is possible under the sparse Gaussian model only if* $n = \Omega(\sqrt{s \log(d/s)}/\epsilon)$ *and* $m = O(n\epsilon/\sqrt{s \log(d/s)})$.

Our final example proves the tightness of the sufficiency-based approaches in Examples A.1 and A.3 for Gaussian models with unknown covariance. The proof relies on the computation of the minimax risks in Lemma 6.1, as well as the statistical theory of invariance.

EXAMPLE A.20 (Gaussian model with unknown covariance). *We establish the matching lower bounds of sample amplification in Example A.1 with a known mean, which imply the lower bounds in Example A.3. We make use of the minimax risk formulation of Lemma 6.1, and consider the following loss:*

$$L(\Sigma, \widehat{\Sigma}) = \mathbb{1}\left(\ell(\Sigma, \widehat{\Sigma}) \geq g(n + m - 1 - d, d) + C \cdot \frac{d}{n}\right),$$

where function g *is given by* (A.2) *in the Lemma C.3 below,* $C > 0$ *is a large absolute constant to be determined later, and* $\ell(\Sigma, \widehat{\Sigma})$ *is the Stein's loss*

$$\ell(\Sigma, \widehat{\Sigma}) = \mathrm{tr}(\Sigma^{-1}\widehat{\Sigma}) - \log\det(\Sigma^{-1}\widehat{\Sigma}) - d.$$

To search for the minimax estimator under the above loss, similar arguments in [11] based on the theory of invariance show that it suffices to consider estimators taking the form

$$\widehat{\Sigma}_n = L_n D_n L_n^\top, \tag{A.1}$$

where $D_n \in \mathbb{R}^{d \times d}$ is a diagonal matrix independent of the observations, and $L_n \in \mathbb{R}^{d \times d}$ is the lower triangular matrix satisfying that $L_n L_n^\top = \sum_{i=1}^n X_i X_i^\top$. Moreover, the risk of the above estimator is invariant with the unknown Σ, so in the sequel we assume that $\Sigma = I_d$. Note that when $D_n = I_d/n$, the estimator $\widehat{\Sigma}_n$ is the sample covariance; however, other choices of the diagonal matrix D_n could give a uniform improvement over the sample covariance, see [11].

The proof idea is to show that with $n + m$ samples, there exists an estimator $\widehat{\Sigma}_{n+m}$ with a specific choice of D_{n+m} in (A.1) such that (cf. Lemma A.21)

$$|\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_{n+m})] - g(n + m + 1 - d, d)| \le \frac{5d}{n+m}, \qquad \sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_{n+m}))} \le \frac{4d}{n+m}.$$

Consequently, by Chebyshev's inequality, for $C > 0$ large enough $r(\mathcal{P}, n + m, L) \le 0.1$.

To lower bound the minimax risk $r(\mathcal{P}, n, L)$ with n samples, we need to enumerate over all possible estimators taking the form of (A.1). It turns out that for any choice of D_n, the first two moments of $\ell(\Sigma, \widehat{\Sigma}_n)$ admit explicit expressions, and it always holds that (cf. Lemma A.21)

$$\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] \ge g(n + 1 - d, d) + C_1 \left(\sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n))} - \frac{4d}{n} \right) - \frac{6d}{n}$$

for any given constant $C_1 > 0$, provided that $n, d \ge C_2$ with C_2 depending only on C_1. Choosing $C_1 > 0$ large enough, Chebyshev's inequality leads to $\ell(\Sigma, \widehat{\Sigma}_n) \ge g(n+1-d, d) - 5C_1 d/n$ with probability at least 0.9 for every $\widehat{\Sigma}_n$ taking the form of (A.1). By the last statement of Lemma A.21, for $m = C_3 n/d$ with a large enough $C_3 > 0$, the above event implies that $r(\mathcal{P}, n, L) \ge 0.9$.

Combining the above scenarios and applying the pigeonhole principle, an application of Lemma 6.1 claims that $m = O(n\epsilon/d)$ is necessary for an $(n, n + m, \epsilon)$ sample amplification.

The following lemma summarizes the necessary technical results for Example A.20.

LEMMA A.21. *Let $n \ge d$. For $D_n = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ with $\lambda_j = 1/(n + d + 1 - 2j)$ for all $j \in [d]$, the corresponding estimator $\widehat{\Sigma}_n$ in (A.1) satisfies*

$$|\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] - g(n + 1 - d, d)| \le \frac{5d}{n}, \qquad \sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n))} \le \frac{4d}{n},$$

where

$$g(u, v) \triangleq \frac{(u + 2v) \log(u + 2v) + u \log u}{2} - (u + v) \log(u + v). \tag{A.2}$$

Meanwhile, for any choice of D_n and any absolute constant $C_1 > 0$, the estimator $\widehat{\Sigma}_n$ in (A.1) satisfies

$$\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] \ge g(n + 1 - d, d) + C_1 \left(\sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n))} - \frac{4d}{n} \right) - \frac{6d}{n},$$

as long as $n, d \ge C_2$ for some large enough constant C_2 depending only on C_1.

Finally, the function g defined in (A.2) satisfies the following inequality: for $n \ge 2d$ and $0 \le m \le n$, it holds that

$$g(n + 1 - d, d) - g(n + m + 1 - d, d) \ge \frac{md^2}{13n^2}.$$

APPENDIX B: PROOF OF MAIN THEOREMS

B.1. Proof of Theorem 4.5. We recall the following result in [3, Theorem 2.6]: given Assumptions 1 and 2 with $k = 3$, there exists a constant $C > 0$ depending only on d and the moment upper bound such that

$$\sup_{\theta \in \Theta} \|\mathcal{L}(\sqrt{n}[\nabla^2 A(\theta)]^{-1/2}(T_n - \nabla A(\theta))) - \mathcal{N}(0, I_d)\|_{\text{TV}} \leq \frac{C}{\sqrt{n}}.$$

By an affine transformation, it is then clear that

$$\sup_{\theta \in \Theta} \|\mathcal{L}(T_n) - \mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/n)\|_{\text{TV}} \leq \frac{C}{\sqrt{n}}.$$

Moreover, the computation in Example 4.1 shows that

$$\sup_{\theta \in \Theta} \|\mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/n) - \mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/(n+m))\|_{\text{TV}} \leq \frac{m\sqrt{d}}{n}.$$

Now the desired result follows from the above inequalities and a triangle inequality.

B.2. Proof of Theorem 4.6. As the density in the Edgeworth expansion (4.1) may be negative at some point, throughout the proof we extend the formal definition of the TV distance to any signed measures P and Q, just as half of the L_1 distance. Under this new definition, it is clear that the triangle inequality $\|P - R\|_{\text{TV}} \leq \|P - Q\|_{\text{TV}} + \|Q - R\|_{\text{TV}}$ still holds. The following tensorization property also holds: for general signed measures $P_1, \cdots, P_d$ and $Q_1, \cdots, Q_d$ with $\max_{i \in [d]} \max\{|P_i|(\Omega), |Q_i|(\Omega)\} \leq r$, we have

$$\|P_1 \times \cdots \times P_d - Q_1 \times \cdots \times Q_d\|_{\text{TV}}$$

$$\leq \sum_{i=1}^{d} \|P_1 \times \cdots \times P_{i-1} \times Q_i \times \cdots \times Q_d - P_1 \times \cdots \times P_i \times Q_{i+1} \times \cdots \times Q_d\|_{\text{TV}}$$

$$\leq \sum_{i=1}^{d} \|P_i - Q_i\|_{\text{TV}} \cdot \prod_{j<i} |P_j|(\Omega) \cdot \prod_{k>i} |Q_k|(\Omega)$$

(B.1) $$\leq r^d \sum_{i=1}^{d} \|P_i - Q_i\|_{\text{TV}}.$$

Fix any $\theta \in \Theta$, let $P_{n,i}$ (resp. $P_{n+m,i}$) be the probability distribution of the i-th coordinate of T_n (resp. T_{n+m}), and $Q_{n,i}$ (resp. $Q_{n+m,i}$) be the signed measure of the corresponding Edgeworth expansion taking the form (4.1) with $k = 9$. We note that the polynomials $\mathcal{K}_\ell(x)$ in (4.1) are the same for $Q_{n,i}$ and $Q_{n+m,i}$, and their coefficients are uniformly bounded over $\theta \in \Theta$ thanks to Assumption 2. Then based on Assumptions 1 and 2 with $k = 10$, the result of [3, Theorem 2.7] claims that

$$\|P_{n,i} - Q_{n,i}\|_{\text{TV}} \leq \frac{C}{n^2},$$

with $C > 0$ independent of (n, d, θ). Moreover, the signed measure of the Edgeworth expansion in (4.1) could be negative only if $|x| = \Omega(\sqrt{n})$, and therefore the total variation of each $Q_{n,i}$ satisfies

$$|Q_{n,i}|(\mathbb{R}) = |\Gamma_{n,9}|(\mathbb{R}) \leq \Gamma_{n,9}([-c\sqrt{n}, c\sqrt{n}]) + |\Gamma_{n,9}|(\mathbb{R} \setminus [-c\sqrt{n}, c\sqrt{n}]) \leq 1 + \exp(-\Omega(n)),$$

where the last inequality follows from integrating the Gaussian tails. Finally, let $Q_{n,i}^+$ be the positive part of $Q_{n,i}$ in the Jordan decomposition, the tensorization property (B.1) leads to

$$\text{(B.2)} \qquad \left\| \mathcal{L}(T_n) - \prod_{i=1}^d Q_{n,i}^+ \right\|_{\text{TV}} \le \left\| \mathcal{L}(T_n) - \prod_{i=1}^d Q_{n,i} \right\|_{\text{TV}} \le \frac{Cd}{n^2} \cdot (1 + \exp(-\Omega(n)))^d,$$

and similar result holds for $\mathcal{L}(T_{n+m})$.

Next it remains to upper bound the TV distance between $\prod_{i=1}^d Q_{n,i}^+$ and $\prod_{i=1}^d Q_{n+m,i}^+$. To this end, we also generalize the formal definition of the Hellinger distance between general measures which are not necessarily probabilities. Then the following inequality between generalized TV and Hellinger distance holds: for measures P and Q on Ω,

$$\|P - Q\|_{\text{TV}} = \frac{1}{2} \int |dP - dQ| = \frac{1}{2} \int |\sqrt{dP} - \sqrt{dQ}|(\sqrt{dP} + \sqrt{dQ})$$

$$\le \frac{1}{2} \sqrt{\int (\sqrt{dP} - \sqrt{dQ})^2 \cdot \int (\sqrt{dP} + \sqrt{dQ})^2}$$

$$\text{(B.3)} \qquad \le H(P, Q) \cdot \sqrt{P(\Omega) + Q(\Omega)}.$$

Also, the following tensorization property holds for the Hellinger distance between general measures: for (not necessarily probability) measures P_i, Q_i on Ω_i,

(B.4)
$$H^2 \left(\prod_{i=1}^d P_i, \prod_{i=1}^d Q_i \right) = \frac{\prod_{i=1}^d P_i(\Omega_i) + \prod_{i=1}^d Q_i(\Omega_i)}{2} - \prod_{i=1}^d \left(\frac{P_i(\Omega_i) + Q_i(\Omega_i)}{2} - H^2(P_i, Q_i) \right).$$

Consequently, it suffices to prove an upper bound on the Hellinger distance $H(P_i, Q_i)$, and the TV distance on the product measure is a direct consequence of (B.3) and (B.4).

To upper bound the individual Hellinger distance, note that after a proper affine transformation, the densities of $Q_{n,i}$ and $Q_{n+m,i}$ are as follows:

$$\text{(B.5)} \qquad Q_{n,i}(dx) = \gamma_n(x) \left(1 + \sum_{\ell=1}^3 \frac{\mathcal{K}_{\ell,i}(\sqrt{n} \cdot x)}{n^{\ell/2}} \right) dx,$$

$$\text{(B.6)} \qquad Q_{n+m,i}(dx) = \gamma_{n+m}(x) \left(1 + \sum_{\ell=1}^3 \frac{\mathcal{K}_{\ell,i}(\sqrt{n+m} \cdot x)}{(n+m)^{\ell/2}} \right) dx,$$

where γ_n is the density of $\mathcal{N}(0, 1/n)$, and $\mathcal{K}_{\ell,i}$ is a polynomial of degree 3ℓ with uniformly bounded coefficients. By (B.5) and (B.6), we observe that there exists an absolute constant $c > 0$ independent of (n, m) such that $Q_{n,i}(x)/\gamma_n(x), Q_{n+m,i}(x)/\gamma_{n+m}(x) \in [1/2, 3/2]$ whenever $|x| \le c$. Consequently, the squared Hellinger distance could then be expressed as

$$H^2(Q_{n,i}^+, Q_{n+m,i}^+) = \frac{1}{2} \int_{|x| \le c} \left(\sqrt{Q_{n,i}(x)} - \sqrt{Q_{n+m,i}(x)} \right)^2 dx + \frac{1}{2} \int_{|x| > c} \left(\sqrt{Q_{n,i}^+(x)} - \sqrt{Q_{n+m,i}^+(x)} \right)^2 dx$$

$$\le \int_{|x| \le c} (\sqrt{\gamma_n(x)} - \sqrt{\gamma_{n+m}(x)})^2 \left(1 + \sum_{\ell=1}^3 \frac{\mathcal{K}_{\ell,i}(\sqrt{n+m} \cdot x)}{(n+m)^{\ell/2}} \right) dx$$

$$+ \int_{|x| \le c} \gamma_n(x) \left(\sqrt{1 + \sum_{\ell=1}^3 \frac{\mathcal{K}_{\ell,i}(\sqrt{n} \cdot x)}{n^{\ell/2}}} - \sqrt{1 + \sum_{\ell=1}^3 \frac{\mathcal{K}_{\ell,i}(\sqrt{n+m} \cdot x)}{(n+m)^{\ell/2}}} \right)^2 dx$$

$$+ \int_{|x|>c} \left(Q^+_{n,i}(x) + Q^+_{n+m,i}(x) \right) \mathrm{d}x$$

$$\equiv A_1 + A_2 + A_3.$$

Next we upper bound the terms A_1, A_2, and A_3 separately.

1. Upper bounding A_1: note that by the definition of c, the multiplication factor in A_1 is at most $3/2$. Therefore,

$$A_1 \leq \frac{3}{2} \int_{|x| \leq c} (\sqrt{\gamma_n(x)} - \sqrt{\gamma_{n+m}(x)})^2 \mathrm{d}x$$

$$\leq \frac{3}{2} \int_{\mathbb{R}} (\sqrt{\gamma_n(x)} - \sqrt{\gamma_{n+m}(x)})^2 \mathrm{d}x$$

$$\overset{(a)}{=} 3 \left(1 - \left(\frac{n(n+m)}{(n+m/2)^2} \right)^{1/4} \right)$$

(B.7)
$$\overset{(b)}{\leq} \frac{3m^2}{4n^2},$$

where (a) is due to the direct computation of the squared Hellinger distance between $\mathcal{N}(0, 1/n)$ and $\mathcal{N}(0, 1/(n+m))$, and (b) makes use of the inequality $1 - x^{1/4} \leq 1 - x$ for $x \in [0, 1]$.

2. Upper bounding A_2: using $|\sqrt{a} - \sqrt{b}| \leq |a - b|$ for $a, b \geq 1/2$, we have

$$A_2 \leq \int_{|x| \leq c} \gamma_n(x) \left(\sum_{\ell=1}^{3} \frac{\mathcal{K}_{\ell,i}(\sqrt{n} \cdot x)}{n^{\ell/2}} - \sum_{\ell=1}^{3} \frac{\mathcal{K}_{\ell,i}(\sqrt{n+m} \cdot x)}{(n+m)^{\ell/2}} \right)^2 \mathrm{d}x$$

$$\leq 3 \sum_{\ell=1}^{3} \int_{\mathbb{R}} \gamma_n(x) \left(\frac{\mathcal{K}_{\ell,i}(\sqrt{n} \cdot x)}{n^{\ell/2}} - \frac{\mathcal{K}_{\ell,i}(\sqrt{n+m} \cdot x)}{(n+m)^{\ell/2}} \right)^2 \mathrm{d}x$$

$$= 3 \sum_{\ell=1}^{3} \int_{\mathbb{R}} \gamma(x) \left(\frac{\mathcal{K}_{\ell,i}(x)}{n^{\ell/2}} - \frac{\mathcal{K}_{\ell,i}(\sqrt{1+m/n} \cdot x)}{(n+m)^{\ell/2}} \right)^2 \mathrm{d}x,$$

where the last step is a change of measure, and γ is the density of $\mathcal{N}(0,1)$. Writing $\mathcal{K}_{\ell,i}(x) = \sum_{j=0}^{3\ell} a_{i,j} x^j$, then

$$\left| \frac{\mathcal{K}_{\ell,i}(x)}{n^{\ell/2}} - \frac{\mathcal{K}_{\ell,i}(\sqrt{1+m/n} \cdot x)}{(n+m)^{\ell/2}} \right| = \left| \sum_{j=0}^{3\ell} \frac{a_{i,j} x^j}{n^{\ell/2}} \left(1 - \left(1 + \frac{m}{n} \right)^{\frac{i-\ell}{2}} \right) \right| \lesssim \frac{m(1 + x^{3\ell})}{n^{\ell/2+1}}$$

whenever $m = O(n)$. Combining the above two inequalities yields

(B.8)
$$A_2 \lesssim \frac{m^2}{n^3}.$$

3. Upper bounding A_3: note that the tail of $\gamma_n(x)$ is at least $\exp(-\Omega(n))$ when $x \geq c$, integrating the tail leads to

(B.9)
$$A_3 = \exp(-\Omega(n)).$$

In summary, a combination of (B.7), (B.8), and (B.9) leads to

$$H^2(Q^+_{n,i}, Q^+_{n+m,i}) = O \left(\frac{m^2}{n^2} + \exp(-\Omega(n)) \right).$$

Now using (B.3), (B.4), and $|Q_{n,i}|(\mathbb{R}) \leq 1 + \exp(-\Omega(n))$, we conclude that

$$(B.10) \qquad \left\| \prod_{i=1}^{d} Q_{n,i}^{+} - \prod_{i=1}^{d} Q_{n+m,i}^{+} \right\|_{\mathrm{TV}} = O\left(\frac{m\sqrt{d}}{n} + (1 + \exp(-\Omega(n)))^{d} \right).$$

Hence, when $n = \Omega(\sqrt{d})$ and $m = O(n)$, the desired result follows from (B.2) and (B.10). For the other scenarios, we simply use that the TV distance is upper bounded by one, and the result still holds.

B.3. Proof of Theorem 5.2. Let $\widehat{P}_n$ be the distribution learned from the first $n/2$ samples which achieves the χ^2-learning error $r_{\chi^2}(\mathcal{P}, n/2)$, and P_{mix} be the distribution of the shuffled samples $(Z_1, \cdots, Z_{n/2+m})$ in Algorithm 2. Note that both distributions depend on the first $n/2$ samples and are therefore random. Then the final TV distance of sample amplification is

$$\|P_{X^{n/2}} \times P_{\mathrm{mix}}(X^{n/2}) - P^{\otimes(n+m)}\|_{\mathrm{TV}} = \mathbb{E}_{X^{n/2}}\|P_{\mathrm{mix}}(X^{n/2}) - P^{\otimes(n/2+m)}\|_{\mathrm{TV}},$$

which is the expected TV distance between the mixture distribution and the product distribution.

By Lemma 5.8, for any realization of $X^{n/2}$ it holds that

$$\chi^2\left(P_{\mathrm{mix}}, P^{\otimes(n/2+m)} \right) \leq \left(1 + \frac{m}{n/2+m}\chi^2(\widehat{P}_n, P) \right)^m - 1.$$

Since (3.2) shows that $D_{\mathrm{KL}}(P\|Q) \leq \log(1 + \chi^2(P,Q))$, we have

$$D_{\mathrm{KL}}(P_{\mathrm{mix}}\|P^{\otimes(n/2+m)}) \leq m \log\left(1 + \frac{m}{n/2+m}\chi^2(\widehat{P}_n, P) \right) \leq \frac{m^2}{n/2+m}\chi^2(\widehat{P}_n, P).$$

Consequently, by (3.2) again and the concavity of $x \mapsto \sqrt{x}$, we have

$$\mathbb{E}_{X^{n/2}}\|P_{\mathrm{mix}}(X^{n/2}) - P^{\otimes(n/2+m)}\|_{\mathrm{TV}} \leq \mathbb{E}_{X^{n/2}}\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_{\mathrm{mix}}(X^{n/2})\|P^{\otimes(n/2+m)})}$$

$$\leq \mathbb{E}_{X^{n/2}}\sqrt{\frac{m^2}{n+2m}\chi^2(\widehat{P}_n(X^{n/2}), P)}$$

$$\leq \sqrt{\frac{m^2}{n+2m}\mathbb{E}_{X^{n/2}}[\chi^2(\widehat{P}_n(X^{n/2}), P)]}$$

$$\leq \sqrt{\frac{m^2}{n+2m} \cdot r_{\chi^2}(\mathcal{P}, n/2)}.$$

B.4. Proof of Theorem 5.5. The main arguments are essentially the same as Theorem 5.2. Note that by the same argument, for each $j \in [d]$ we have

$$D_{\mathrm{KL}}(P_{\mathrm{mix},j}\|P_j^{\otimes(n/2+m)}) \leq \frac{m^2}{n/2+m}\chi^2(\widehat{P}_{n,j}, P_j).$$

Note that both P_{mix} and P have product structures, the chain rule of KL divergence implies that

$$D_{\mathrm{KL}}(P_{\mathrm{mix}}\|P^{\otimes(n/2+m)}) \leq \frac{m^2}{n/2+m}\sum_{j=1}^{d}\chi^2(\widehat{P}_{n,j}, P_j).$$

Now the rest of the proof follows from the same last few lines of that of Theorem 5.2.

14

B.5. Proof of Theorem 6.2. The proof relies on Lemma 6.1 and a classical statistical result known as Anderson's lemma. Without loss of generality we assume $\Sigma = I_d$. Choose $L(\theta, \widehat{\theta}) = \ell(\theta - \widehat{\theta})$, with a bowl-shaped (i.e. symmetric and quasi-convex) loss function $\ell(\cdot) \in [0,1]$. Then Anderson's lemma (see, e.g. [16, Lemma 8.5]) implies that the minimax estimator under L and n samples is $\widehat{\theta}_n = n^{-1} \sum_{i=1}^n X_i$, thus

$$r(\mathcal{P}, n, L) = \mathbb{E}\left[\ell\left(\frac{Z}{\sqrt{n}}\right)\right],$$

with $Z \sim \mathcal{N}(0, I_d)$. Choosing $\ell(x) = \mathbb{1}(\|x\|_2 \geq r)$ with the parameter $r > 0$ determined by

$$\mathbb{P}\left(\left|\frac{Z}{\sqrt{n}}\right| \geq r\right) - \mathbb{P}\left(\left|\frac{Z}{\sqrt{n+m}}\right| \geq r\right) = \left\|\mathcal{N}\left(0, \frac{I_d}{n}\right) - \mathcal{N}\left(0, \frac{I_d}{n+m}\right)\right\|_{\mathrm{TV}},$$

then clearly $\ell(\cdot) \in \{0,1\}$ is bowl-shaped. Hence, for this choice of L, Lemma 6.1 gives that

$$\epsilon^\star(\mathcal{P}, n, m) \geq r(\mathcal{P}, n, L) - r(\mathcal{P}, n+m, L) = \left\|\mathcal{N}\left(0, \frac{I_d}{n}\right) - \mathcal{N}\left(0, \frac{I_d}{n+m}\right)\right\|_{\mathrm{TV}},$$

and a matching upper bound is presented in Example 4.1.

B.6. Proof of Theorem 6.3. Based on the discussions above Theorem 6.3, we choose an arbitrary open ball $\Theta_0 \subseteq \Theta$, and pick any d-dimensional ball $B_d(\mu_0; r)$ contained in $\nabla A(\Theta_0)$. Let θ_0 be the center of Θ_0, and after a proper affine transformation we assume that $\nabla A^2(\theta_0) = I_d$. Consider a truncated Gaussian prior ν with

$$\mu \sim \mathcal{N}(\mu_0, c_n r_n^2) \mid \mu \in B_d(\mu_0; r_n),$$

where $c_n > 0$ and $r_n \in (0, r)$ are parameters to be determined later. Using the diffeomorphism ∇A, there is a prior ν_0 on Θ_0 which induces the above prior on $\nabla A(\Theta_0)$. Moreover, as $\overline{\Theta}_0$, the closure of Θ_0, is a compact set, a weaker Assumption 2 with the supremum restricted to Θ_0 holds for $k = 3$.

Now we analyze the Bayes risks under the above prior ν_0 and the loss

$$L(\theta, \widehat{\mu}) = \ell(\nabla A(\theta) - \widehat{\mu}) \in [0,1],$$

where ℓ is a bowl-shaped function. By sufficiency, it remains to consider the class of estimators depending only on the sufficient statistic $T_n(X^n) = n^{-1} \sum_{i=1}^n T(X_i)$. Let $\widehat{\mu}(T_n)$ be such an estimator, then under each $\theta \in \Theta_0$, we have

$$\mathbb{E}_\theta[L(\theta, \widehat{\mu}(T_n))] \geq \mathbb{E}_\theta[L(\theta, \widehat{\mu}(Z_n))] - \|\mathcal{L}(T_n) - \mathcal{L}(Z_n)\|_{\mathrm{TV}},$$

where $Z_n \mid \theta \sim \mathcal{N}(\nabla A(\theta), I_d/n)$. To upper bound this TV distance, the proof of Theorem 4.5 together with Assumption 2 applied to Θ_0 gives that

$$\|\mathcal{L}(T_n) - \mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/n)\|_{\mathrm{TV}} \leq \frac{C_1}{\sqrt{n}},$$

where $C_1 > 0$ only depends on the exponential family. Moreover,

$$\|\mathcal{N}(\nabla A(\theta), I_d/n) - \mathcal{N}(\nabla A(\theta), \nabla^2 A(\theta)/n)\|_{\mathrm{TV}}$$

$$= \|\mathcal{N}(0, I_d) - \mathcal{N}(0, \nabla^2 A(\theta))\|_{\mathrm{TV}}$$

$$\overset{(a)}{\leq} \frac{3}{2}\|\nabla^2 A(\theta) - I_d\|_{\mathrm{F}} \overset{(b)}{\leq} C_2 r_n,$$

where (a) follows from [5, Theorem 1.1], and (b) makes use of the analytical property of $A(\theta)$ (see, e.g. [12, Theorem 1.17]), the assumption that $\|\nabla A(\theta) - \nabla A(\theta_0)\|_2 \leq r_n$, and

$\nabla^2 A(\theta_0) = I_d$. Again, here the constant $C_2 > 0$ is independent of n. Combining the above inequalities yields that

$$(B.11) \qquad \mathbb{E}_\theta[L(\theta, \widehat{\mu}(T_n))] \geq \mathbb{E}_\theta[L(\theta, \widehat{\mu}(Z_n))] - \frac{C_1}{\sqrt{n}} - C_2 r_n.$$

Next we lower bound the Bayes risk $\mathbb{E}_{\nu_0} \mathbb{E}_\theta[L(\theta, \widehat{\mu}(Z_n))]$ when T_n has been replaced by Z_n. Let ν' be the non-truncated Gaussian distribution $\mathcal{N}(\mu_0, c_n r_n^2)$, and $G_n \sim \mathcal{N}(0, I_d/n)$, then

$$\mathbb{E}_{\nu_0} \mathbb{E}_\theta[L(\theta, \widehat{\mu}(Z_n))] = \mathbb{E}_{\mu \sim \nu}[\ell(\mu - \widehat{\mu}(G_n + \mu))]$$

$$\geq \mathbb{E}_{\mu \sim \nu'}[\ell(\mu - \widehat{\mu}(G_n + \mu))] - \nu'(\{\mu \notin B_d(\mu_0; r_n)\})$$

$$\overset{(c)}{\geq} \mathbb{E}_{\mu \sim \nu'}[\ell(\mu - \widehat{\mu}(G_n + \mu))] - e^{-C_3/c_n}$$

$$\overset{(d)}{\geq} \mathbb{E}\left[\ell\left(\sqrt{\frac{nc_n r_n^2}{1 + nc_n r_n^2}} \cdot G_n\right)\right] - e^{-C_3/c_n}$$

$$(B.12) \qquad \overset{(e)}{\geq} \mathbb{E}[\ell(G_n)] - e^{-C_3/c_n} - \frac{C_4}{1 + nc_n r_n^2}.$$

Here (c) follows from the Gaussian tail probability, (d) makes use of Anderson's lemma and the fact that under ν', the posterior distribution of μ given Z_n is Gaussian with covariance $nc_n r_n^2 I_d/(1 + nc_n r_n^2)$. The final inequality (e) is due to the following upper bound on the TV distance:

$$\|\mathcal{N}(0, \Sigma) - \mathcal{N}(0, c\Sigma)\|_{\text{TV}} \leq \frac{3}{2}\|(c-1)I_d\|_{\text{F}} = \frac{3\sqrt{d}|c-1|}{2}.$$

Now combining (B.11) and (B.12), we obtain a lower bound on the Bayes risk:

$$r_{\text{B}}(\mathcal{P}, n, \nu_0, L) \geq \mathbb{E}[\ell(G_n)] - \frac{C_1}{\sqrt{n}} - C_2 r_n - e^{-C_3/c_n} - \frac{C_4}{1 + nc_n r_n^2}.$$

Similarly, by reversing all the above inequalities, an upper bound of the Bayes risk with $n + m$ samples is also available:

$$r_{\text{B}}(\mathcal{P}, n+m, \nu_0, L) \leq \mathbb{E}[\ell(G_{n+m})] + \frac{C_1}{\sqrt{n}} + C_2 r_n + e^{-C_3/c_n} + \frac{C_4}{1 + nc_n r_n^2}.$$

By the proof of Theorem 6.2, a proper choice of ℓ satisfies that

$$\mathbb{E}[\ell(G_n)] - \mathbb{E}[\ell(G_{n+m})] = \left\|\mathcal{N}\left(0, \frac{I_d}{n}\right) - \mathcal{N}\left(0, \frac{I_d}{n+m}\right)\right\|_{\text{TV}} = \Omega\left(\frac{m\sqrt{d}}{n} \wedge 1\right),$$

where the last step is again due to [5, Theorem 1.1]. Consequently, by choosing $c_n = \Theta(1/\log n)$ and $r_n = \Theta((\log n/n)^{1/3})$, the desired result follows from Lemma 6.1.

B.7. Proof of Theorem 6.4. We will apply Lemma 6.1 to the uniform prior μ over 2^d points $\prod_{j=1}^d \{\theta_{j,+}, \theta_{j,-}\}$, and the loss function $L : \Theta \times \Theta \to [0,1]$ with

$$L((\theta_1, \cdots, \theta_d), (\widehat{\theta}_1, \cdots, \widehat{\theta}_d)) = \mathbb{1}\left(\sum_{j=1}^d \mathbb{1}(\theta_j = \widehat{\theta}_j) \leq \frac{d}{2} + \frac{1}{2}\sum_{j=1}^d \alpha_j\right).$$

We compute the Bayes risks $r_{\text{B}}(\mathcal{P}, n, \mu, L)$ and $r_{\text{B}}(\mathcal{P}, n+m, \mu, L)$ in this scenario.

Under the uniform prior μ, it is straightforward to see that the posterior distribution of θ given X^n is a product distribution $\prod_{j=1}^{d} p_{\theta_j | X_j^n}$, with $p_{\theta_j | X_j^n}$ supported on two elements $\{\theta_{j,+}, \theta_{j,-}\}$, and

$$p_{\theta_j | X_j^n}(\theta_{j,+}) = \frac{\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j})}{\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j}) + \prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j})},$$

$$p_{\theta_j | X_j^n}(\theta_{j,-}) = \frac{\prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j})}{\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j}) + \prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j})}.$$

Then given X^n, the Bayes estimator $\widehat{\theta}(X^n) \in \Theta$ which minimizes the expected loss L is a minimizer $\xi \in \Theta$ of the above expression

$$\mathbb{P}_{\theta | X^n} \left(\sum_{j=1}^{d} \mathbb{1}(\theta_j = \xi_j) \leq \frac{d}{2} + \frac{1}{2} \sum_{j=1}^{d} \alpha_j \right),$$

which is easily seen to be

$$\widehat{\theta}_j(X^n) = \widehat{\theta}_j(X_j^n) = \theta_{j,-} + (\theta_{j,+} - \theta_{j,-}) \cdot \mathbb{1}\left(\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j}) \geq \prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j}) \right).$$

For the above Bayes estimator, the random variables $\mathbb{1}(\theta_j = \widehat{\theta}_j(X^n))$ are mutually independent, with the mean value

$$p_{n,j} = \mathbb{P}(\theta_j = \widehat{\theta}_j(X^n))$$

$$= \frac{1}{2} p_{\theta_{j,+}}^{\otimes n} \left(\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j}) < \prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j}) \right) + \frac{1}{2} p_{\theta_{j,-}}^{\otimes n} \left(\prod_{i=1}^{n} p_{\theta_{j,+}}(X_{i,j}) \geq \prod_{i=1}^{n} p_{\theta_{j,-}}(X_{i,j}) \right)$$

$$= \frac{1}{2} \left(1 + \|p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}} \right).$$

Consequently, we have $p_{n,j} \leq (1 + \alpha_j)/2 - \epsilon/(2\sqrt{d})$ for each $j \in [d]$ by (6.2), and thus

$$r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) = \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}(p_{n,j}) \leq \frac{d}{2} + \frac{1}{2} \sum_{j=1}^{d} \alpha_j \right)$$

(B.13)
$$\geq \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} - \frac{\epsilon}{2\sqrt{d}} \right) \leq \frac{d}{2} + \frac{1}{2} \sum_{j=1}^{d} \alpha_j \right).$$

An entirely symmetric argument leads to

(B.14)
$$r_{\mathrm{B}}(\mathcal{P}, n+m, \mu, L) \leq \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} + \frac{\epsilon}{2\sqrt{d}} \right) \leq \frac{d}{2} + \frac{1}{2} \sum_{j=1}^{d} \alpha_j \right).$$

To further lower bound (B.13) and upper bound (B.14), the following lemma shows that we may assume that the above Bernoulli distributions have the same parameter.

LEMMA B.1 (Theorems 4 and 5 of [8]). *Let $X_i \sim \mathsf{Bern}(p_i)$ be independent Bernoulli random variables for $i = 1, \cdots, n$, with $\bar{p} = n^{-1} \sum_{i=1}^{n} p_i$. Then for $0 \leq k \leq n\bar{p} - 1$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq k \right) \leq \mathbb{P}\left(\mathsf{B}(n, \bar{p}) \leq k \right),$$

and for $k \geq n\bar{p}$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq k\right) \geq \mathbb{P}\left(\mathsf{B}(n,\bar{p}) \leq k\right).$$

In particular, for integers (b,c) with $b \leq n\bar{p} \leq c$, we have

$$\mathbb{P}\left(b \leq \sum_{i=1}^{n} X_i \leq c\right) \geq \mathbb{P}\left(b \leq \mathsf{B}(n,\bar{p}) \leq c\right).$$

For $d \geq 4/\epsilon^2$, based on the second statement of Lemma B.1, the quantity in (B.13) satisfies

$$\text{(B.15)} \qquad r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) \geq \mathbb{P}\left(\mathsf{B}\left(d, \frac{1+\alpha}{2} - \frac{\epsilon}{2\sqrt{d}}\right) \leq \frac{1+\alpha}{2} \cdot d\right),$$

with $\alpha \triangleq d^{-1}\sum_{j=1}^{d} \alpha_j$. Similarly, based on the first statement of Lemma B.1, for (B.14) we have

$$\text{(B.16)} \qquad r_{\mathrm{B}}(\mathcal{P}, n+m, \mu, L) \leq \mathbb{P}\left(\mathsf{B}\left(d, \frac{1+\alpha}{2} + \frac{\epsilon}{2\sqrt{d}}\right) \leq \frac{1+\alpha}{2} \cdot d\right).$$

Since

$$\frac{\mathrm{d}\mathbb{P}(\mathsf{B}(n,x)=k)}{\mathrm{d}x} = n\left(\mathbb{P}(\mathsf{B}(n-1,x)=k-1) - \mathbb{P}(\mathsf{B}(n-1,x)=k)\right),$$

we invoke Lemma 6.1 and lower bound the Bayes risk difference as

$$\epsilon^\star(\mathcal{P}, n, m) \geq r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) - r_{\mathrm{B}}(\mathcal{P}, n+m, \mu, L)$$

$$\geq \mathbb{P}\left(\mathsf{B}\left(d, \frac{1+\alpha}{2} - \frac{\epsilon}{2\sqrt{d}}\right) \leq \frac{1+\alpha}{2} \cdot d\right) - \mathbb{P}\left(\mathsf{B}\left(d, \frac{1+\alpha}{2} + \frac{\epsilon}{2\sqrt{d}}\right) \leq \frac{1+\alpha}{2} \cdot d\right)$$

$$= - \sum_{0 \leq k \leq (1+\alpha)d/2} \int_{\frac{1+\alpha}{2} - \frac{\epsilon}{2\sqrt{d}}}^{\frac{1+\alpha}{2} + \frac{\epsilon}{2\sqrt{d}}} \frac{\mathrm{d}\mathbb{P}(\mathsf{B}(d,x)=k)}{\mathrm{d}x} \mathrm{d}x$$

$$= d \int_{\frac{1+\alpha}{2} - \frac{\epsilon}{2\sqrt{d}}}^{\frac{1+\alpha}{2} + \frac{\epsilon}{2\sqrt{d}}} \mathbb{P}\left(\mathsf{B}(d-1,x) = \left\lfloor \frac{1+\alpha}{2} \cdot d \right\rfloor\right) \mathrm{d}x$$

$$\overset{(a)}{\geq} d \int_{\frac{1+\alpha}{2} - \frac{\epsilon}{2\sqrt{d}}}^{\frac{1+\alpha}{2} + \frac{\epsilon}{2\sqrt{d}}} \frac{c(\underline{\alpha}, \overline{\alpha})}{\sqrt{d}} \mathrm{d}x = c(\underline{\alpha}, \overline{\alpha})\epsilon,$$

where (a) is due to

$$\text{(B.17)} \qquad \min_{p_0 \leq p \leq p_1, |k-np| \leq C\sqrt{n}} \mathbb{P}(\mathsf{B}(n,p)=k) = \Omega_{p_0,p_1,C}\left(\frac{1}{\sqrt{n}}\right)$$

for any $p_0, p_1 \in (0,1)$ and $C > 0$, by Stirling's approximation.

For $d < 4/\epsilon^2$, we first note the following identity:

$$\frac{\mathrm{d}}{\mathrm{d}x}\bigg|_{x=0} \mathbb{P}\left(\sum_{i=1}^{n} \mathsf{Bern}(p_i+x) = k\right)$$

$$= \frac{\mathrm{d}}{\mathrm{d}x}\bigg|_{x=0} \sum_{w \in \{0,1\}^n : \sum_{i=1}^{n} w_i = k} \prod_{i=1}^{n} (p_i+x)^{w_i}(1-p_i-x)^{1-w_i}$$

$$= \sum_{i=1}^{n} \left(\sum_{w_{\backslash i} \in \{0,1\}^{n-1}: \sum_{j \neq i} w_j = k-1} \prod_{j \neq i} p_j^{w_j}(1-p_j)^{1-w_j} - \sum_{w_{\backslash i} \in \{0,1\}^{n-1}: \sum_{j \neq i} w_j = k} \prod_{j \neq i} p_j^{w_j}(1-p_j)^{1-w_j} \right)$$

$$= \sum_{i=1}^{n} \left[\mathbb{P}\left(\sum_{j \neq i} \mathsf{Bern}(p_j) = k-1 \right) - \mathbb{P}\left(\sum_{j \neq i} \mathsf{Bern}(p_j) = k \right) \right].$$

Based on (B.13) and (B.14), we then have

$$\epsilon^\star(\mathcal{P}, n, m) \geq r_\mathrm{B}(\mathcal{P}, n, \mu, L) - r_\mathrm{B}(\mathcal{P}, n+m, \mu, L)$$

$$\geq \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} - \frac{\epsilon}{2\sqrt{d}} \right) \leq \frac{(1+\alpha)d}{2} \right) - \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} + \frac{\epsilon}{2\sqrt{d}} \right) \leq \frac{(1+\alpha)d}{2} \right)$$

$$= - \sum_{0 \leq k \leq (1+\alpha)d/2} \int_{-\frac{\epsilon}{2\sqrt{d}}}^{\frac{\epsilon}{2\sqrt{d}}} \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{P}\left(\sum_{j=1}^{d} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} + x \right) = k \right) \mathrm{d}x$$

$$= \sum_{i=1}^{d} \int_{-\frac{\epsilon}{2\sqrt{d}}}^{\frac{\epsilon}{2\sqrt{d}}} \mathbb{P}\left(\sum_{j \neq i} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} + x \right) = \left\lfloor \frac{(1+\alpha)d}{2} \right\rfloor \right) \mathrm{d}x.$$

We will show that the integrand is uniformly of the order $\Omega(1/\sqrt{d})$. To this end, note that for $|x| \leq \epsilon/(2\sqrt{d}) < 1/d$ as $d < 4/\epsilon^2$, it holds that

$$\sum_{j \neq i} \left(\frac{1+\alpha_j}{2} + x \right) < \sum_{j=1}^{d} \frac{1+\alpha_j}{2} + 1 \leq \left\lfloor \frac{(1+\alpha)d}{2} \right\rfloor + 2,$$

$$\sum_{j \neq i} \left(\frac{1+\alpha_j}{2} + x \right) > \sum_{j=1}^{d} \frac{1+\alpha_j}{2} - 1 - 1 \geq \left\lfloor \frac{(1+\alpha)d}{2} \right\rfloor - 2.$$

Consequently, by the last statement of Lemma B.1, one has

$$\mathbb{P}\left(\left| \sum_{j \neq i} \mathsf{Bern}\left(\frac{1+\alpha_j}{2} + x \right) - \left\lfloor \frac{(1+\alpha)d}{2} \right\rfloor \right| \leq 2 \right)$$

$$\geq \mathbb{P}\left(\left| \mathsf{B}\left(d-1, \frac{1}{d-1} \sum_{j \neq i} \left(\frac{1+\alpha_j}{2} + x \right) \right) - \left\lfloor \frac{(1+\alpha)d}{2} \right\rfloor \right| \leq 2 \right) = \Omega_{\underline{\alpha}, \overline{\alpha}}\left(\frac{1}{\sqrt{d}} \right),$$

where the last step is again due to (B.17). The above display is a lower bound for the probability of a size-5 set, and in view of the following lemma, the same $\Omega_{\underline{\alpha}, \overline{\alpha}}(1/\sqrt{d})$ lower bound also holds for the probability of any singleton. Plugging this lower bound back into the integral then yields to $\epsilon^\star(\mathcal{P}, n, m) = \Omega_{\underline{\alpha}, \overline{\alpha}}(\epsilon)$, as desired.

LEMMA B.2.   *Let $p_1, \cdots, p_n \in [a, b]$ with $0 < a \leq b < 1$, and $cn \leq k < k+1 \leq (1-c)n$ for some $c > 0$. Define*

$$f(k) = \mathbb{P}\left(\sum_{i=1}^{n} \mathsf{Bern}(p_i) = k \right).$$

Then there exists an absolute constant $C = C(a, b, c) < \infty$ such that

$$C^{-1} \leq \frac{f(k+1)}{f(k)} \leq C.$$

PROOF. Let $W_k = \{w \in \{0, 1\}^n : \sum_{i=1}^n w_i = k\}$, and $g(w) = \prod_{i=1}^n p_i^{w_i}(1 - p_i)^{1-w_i}$. Then it is clear that

$$f(k) = \sum_{w \in W_k} g(w).$$

Call two binary vectors w and w' as *neighbors* (denoted by $w \sim w'$) if w and w' only differ in one coordinate. It is clear that every $w \in W_k$ has $n - k$ neighbors in W_{k+1}, and every $w \in W_{k+1}$ has $k + 1$ neighbors in W_k. Moreover, for $w \sim w'$,

$$\frac{g(w)}{g(w')} \leq \max\left\{\frac{b}{a}, \frac{1-a}{1-b}\right\} =: \rho.$$

Consequently, by double counting,

$$f(k+1) = \sum_{w \in W_{k+1}} g(w) = \frac{1}{k+1} \sum_{w' \in W_k} \sum_{w \in W_{k+1}: w \sim w'} g(w)$$

$$\leq \frac{\rho}{k+1} \sum_{w' \in W_k} \sum_{w \in W_{k+1}: w \sim w'} g(w') = \frac{(n-k)\rho}{k+1} \sum_{w' \in W_k} g(w')$$

$$= \frac{(n-k)\rho}{k+1} f(k) \leq \frac{(1-c)\rho}{c} f(k).$$

The other inequality can be established analogously. $\square$

B.8. Proof of Theorem 6.5. For each $j \in [d]$, we pick two points $\theta_{j,+}, \theta_{j,-} \in \Theta_j$ in Assumption 4. We first aim to show that

(B.18) $$0.09 \triangleq \epsilon_1' \leq \|p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}} \leq \epsilon_1 \triangleq 0.6,$$

(B.19) $$0.99995 \triangleq \epsilon_2' \geq \|p_{\theta_{j,+}}^{\otimes 20n} - p_{\theta_{j,-}}^{\otimes 20n}\|_{\mathrm{TV}} \geq \epsilon_2 \triangleq 0.86.$$

The proofs of (B.18) and (B.19) rely on the tensorization property of the Hellinger distance

$$H^2\left(\prod_{i=1}^d P_i, \prod_{i=1}^d Q_i\right) = 1 - \prod_{i=1}^d \left(1 - H^2(P_i, Q_i)\right),$$

and the relationship between TV and Hellinger distance in (3.1). For example, for (B.18), we have

$$\|p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n}\|_{\mathrm{TV}} \leq \sqrt{1 - (1 - H^2(p_{\theta_{j,+}}^{\otimes n}, p_{\theta_{j,-}}^{\otimes n}))^2}$$

$$= \sqrt{1 - \left(1 - H^2(p_{\theta_{j,+}}, p_{\theta_{j,-}})\right)^{2n}}$$

$$\leq \sqrt{1 - \left(1 - \frac{1}{5n}\right)^{2n}}$$

$$\leq \sqrt{\frac{9}{25}} = 0.6.$$

Applying the other inequality, for (B.19) we have

$$\|p_{\theta_{j,+}}^{\otimes 20n} - p_{\theta_{j,-}}^{\otimes 20n}\|_{\mathrm{TV}} \geq H^2(p_{\theta_{j,+}}^{\otimes 20n}, p_{\theta_{j,-}}^{\otimes 20n})$$

$$\geq 1 - \left(1 - \frac{1}{10n}\right)^{20n}$$

$$\geq 1 - \exp(-2) > 0.86.$$

The other inequalities involving ϵ_1' and ϵ_2' could be established analogously.

Next, for the choice of $m = \lceil c\epsilon n/\sqrt{d} \rceil$ in the statement of Theorem 6.5, we show that there exists $n_j \in [n, 20n]$ such that

$$(B.20) \qquad \|p_{\theta_{j,+}}^{\otimes(n_j+m)} - p_{\theta_{j,-}}^{\otimes(n_j+m)}\|_{\mathrm{TV}} - \|p_{\theta_{j,+}}^{\otimes n_j} - p_{\theta_{j,-}}^{\otimes n_j}\|_{\mathrm{TV}} \geq \frac{\epsilon_2 - \epsilon_1}{\lceil 19\sqrt{d}/(c\epsilon) \rceil}.$$

To prove (B.20), first note that $t \mapsto f_j(t) \triangleq \|p_{\theta_{j,+}}^{\otimes t} - p_{\theta_{j,-}}^{\otimes t}\|_{\mathrm{TV}}$ is non-decreasing by the data-processing property of the TV distance. Moreover, by (B.18) and (B.19), we have $f_j(n) \leq \epsilon_1$ and $f_j(20n) \geq \epsilon_2$. Consequently,

$$\epsilon_2 - \epsilon_1 \leq f_j(20n) - f_j(n)$$

$$\leq \sum_{k=1}^{\lceil 19n/m \rceil - 1} [f_j(n+km) - f_j(n+(k-1)m)] + [f_j(20n) - f_j(20n - m)]$$

$$\leq \left\lceil \frac{19n}{m} \right\rceil \cdot \max_{n \leq n_j \leq 20n-m} [f_j(n_j + m) - f_j(n_j)],$$

which gives (B.20). In addition, we also have $\epsilon_1' \leq f_j(n_j) \leq f_j(n_j + m) \leq \epsilon_2'$.

Next we are about to apply Theorem 6.4. By (B.20), there exist $\alpha_j \in (\epsilon_1', \epsilon_2')$ such that

$$\|p_{\theta_{j,+}}^{\otimes n_j} - p_{\theta_{j,-}}^{\otimes n_j}\|_{\mathrm{TV}} \leq \alpha_j - \Omega_c\left(\frac{\epsilon}{\sqrt{d}}\right),$$

$$\|p_{\theta_{j,+}}^{\otimes(n_j+m)} - p_{\theta_{j,-}}^{\otimes(n_j+m)}\|_{\mathrm{TV}} \geq \alpha_j + \Omega_c\left(\frac{\epsilon}{\sqrt{d}}\right).$$

Therefore, Theorem 6.4 shows that there is an absolute constant $c' > 0$ depending only on $(c, \epsilon_1', \epsilon_2')$ such that

$$\epsilon^\star(\mathcal{P}, (n_1, \cdots, n_d), m) \geq c'\epsilon,$$

where the above quantity denotes the minimax error in a new sample amplification problem: suppose we draw n_j independent samples from $\mathcal{P}_j$, also independently for each $j \in [d]$, and we aim to amplify into $n_j + m$ independent samples from $\mathcal{P}_j$. In other words, the sample sizes for different dimensions may not be equal in the new problem, but the target is still to generate m more samples. We claim that $\epsilon^\star(\mathcal{P}, (n_1, \cdots, n_d), m) \leq \epsilon^\star(\mathcal{P}, n, m)$, and thereby complete the proof. To show the claim, note that $n_j \geq n$ for all $j \in [d]$, hence in the new problem we could keep $n_j - n$ samples unused for each j, use the remaining samples to amplify into $n + m$ vectors, and add the above unused samples back to form the final amplification.

B.9. Proof of Theorem 7.1. We first prove the upper bound. Consider the distribution estimator $\widehat{P}_n = (1, 0, \cdots, 0)$, which has a χ^2-divergence

$$\chi^2(\widehat{P}_n \| P) = \frac{1}{t} - 1, \qquad \forall P \in \mathcal{P}_{d,t}.$$

Consequently, the χ^2-estimation error $r_{\chi^2}(\mathcal{P}_{d,t}, n)$ is at most $1/t$, and Theorem 5.2 states that the random shuffling approach achieves an $(n, n+1, 0.1)$ sample amplification if $n = \Theta(1/t)$.

Next we prove the lower bound. Let $n = 1/(100t)$ and d be a multiple of 100. Consider the following prior μ over $\mathcal{P}_{d,t}$: μ is the uniform prior over $(p_0, \cdots, p_d) \in \mathcal{P}_{d,t}$ with $p_0 = t$ and the remaining $1 - t$ mass evenly distributed on a uniformly random subset of $[d]$ of size $d/100$. The action space $\mathcal{A}$ is chosen to be $\mathcal{X}^{n+1}$, and the loss function is

$$L(P, x^{n+1}) = 1 - \mathbb{1}(x^{n+1} \text{ belongs to the support of } P,$$

$$\text{does not contain symbol 0 or repeated symbols}).$$

We first show that $r_{\mathrm{B}}(\mathcal{P}_{d,t}, n+1, \mu) \leq 0.1$. In fact, after observing $n+1$ samples X^{n+1}, we simply use X^{n+1} as the estimator under the above loss. Clearly X^{n+1} belongs to the support of P. For the remaining conditions,

$$\mathbb{P}(X^{n+1} \text{ contains symbol 0}) = 1 - (1-t)^{n+1} = 1 - (1-t)^{1/(100t)+1} \leq 0.05,$$

$$\mathbb{P}(X^{n+1} \text{ contains repeated symbols}) \leq \binom{n+1}{2}\left(t^2 + \sum_{j=1}^{d/100} \frac{1}{(d/100)^2}\right)$$

$$\leq n^2\left(t^2 + \frac{100}{d}\right) \leq \frac{1}{10000} + \frac{1}{100} < 0.05.$$

By the union bound, this estimator achieves a Bayes risk at most 0.1, and thus $r_{\mathrm{B}}(\mathcal{P}_{d,t}, n+1, \mu) \leq 0.1$.

Next we show that $r_{\mathrm{B}}(\mathcal{P}_{d,t}, n, \mu) \geq 0.9$, which combined with Lemma 6.1 gives the desired lower bound. To show this, consider the new symbol in the estimator x^{n+1} not in X^n when the learner observes X^n. Since x^{n+1} has length $n+1 > n$, there is at least one such symbol. In order to have $L(P, x^{n+1}) = 0$, this new symbol cannot be 0 or appear in X^n. Moreover, the posterior distribution of the support of $P \sim \mu$ given X^n is uniformly distributed over

$$\{0, X_1, \cdots, X_n\} \cup \left\{ S \subseteq [d] \backslash \{X_1, \cdots, X_n\} : |S| = \frac{d}{100} - n \right\}.$$

Therefore, the posterior probability of the new symbol being outside the support of P (recall that it could be neither one of $\{X_1, \cdots, X_n\}$ nor 0) is at least

$$1 - \frac{d/100 - n}{d - n} = \frac{0.99d}{d - n} \geq 0.99 > 0.9,$$

giving the desired inequality $r_{\mathrm{B}}(\mathcal{P}_{d,t}, n, \mu) \geq 0.9$.

B.10. Proof of Theorem 7.2. The upper bound of sample amplification directly follows from that of learning, and it remains to show the lower bound $n \geq d$. If $n \leq d - 1$, with probability one the observations X^n spans an n-dimensional subspace of the row space of Σ. An $(n, n+1, 0.1)$ sample amplification calls for at least one additional observation not in X^n, which with probability 1 should not belong to the n-dimensional subspace spanned by X^n for $n \leq d - 1$. However, since $p \geq d + 1$, under a uniformly chosen d-dimensional row space of Σ, the posterior probability of the additional observation belonging to the row space of Σ is zero. Consequently, an $(n, n+1, 0.1)$ sample amplification is impossible if $n < d$.

B.11. Proof of Theorem 7.3. We prove the three claims separately.

The proof of $m^\star(\mathcal{P}_c, n) \gtrsim_c n^{5/6}$**.** Classical theory of nonparametric estimation (see, e.g. [15]) tells that there exists a density estimator $\widehat{f}$ such that $\mathbb{E}_f \|\widehat{f} - f\|_2^2 \lesssim n^{-2/3}$. Since f is lower bounded by c, this implies that

$$\chi^2(\widehat{f}, f) \leq \frac{\|\widehat{f} - f\|_2^2}{c} \lesssim_c n^{-2/3}.$$

Consequently, $r_{\chi^2}(\mathcal{P}_c, n) \lesssim n^{-2/3}$, and Theorem 5.2 implies that $m^\star(\mathcal{P}_c, n) \gtrsim_c n^{5/6}$.

The proof of $m^\star(\mathcal{P}_c, n) \lesssim_c n^{5/6}$**.** We construct a parametric subfamily of $\mathcal{P}_c$ and invoke Theorem 6.5. Let g be a 1-Lipschitz function supported on $[0,1]$ with $\int_0^1 g(x)dx = 0$ and $\|g\|_2 > 0$; in particular $\|g\|_\infty \leq 1$. Let $h = n^{-1/3}$ and assume that $M := h^{-1}$ is an integer. For $u = (u_1, \cdots, u_M) \in \{\pm 1\}^M$, define

$$f_u(x) = 1 + c_0 \sum_{i=1}^M u_i h g\left(\frac{x - (i-1)h}{h}\right),$$

where $c_0 \in (0,1)$ is a small constant satisfying $c_0 h \leq 1 - c$. Consequently, $f_u \in \mathcal{P}_c$ for every $u \in \{\pm 1\}^M$. However, the density estimation model $X_1, \cdots, X_n \sim f_u$ is not a product model, so Theorem 6.5 cannot be directly applied.

To overcome the above issue, we consider a *Poissonized* model as follows: first we draw $N \sim \mathrm{Poi}(n)$, and then draw N i.i.d. samples $X_1, \cdots, X_N \sim f_u$. The Poissonized model satisfies the following two properties:

1. For any measurable set $A \subseteq [0,1]$, we have

$$M(A) := |\{i \in [N] : X_i \in A\}| \sim \mathrm{Poi}\left(n \int_A f(x)dx\right).$$

2. For any collection of disjoint subsets $\{A_i, i \geq 1\}$, the random variables $\{M(A_i), i \geq 1\}$ are mutually independent.

For $i \in [M]$, let $A_i = [(i-1)/M, i/M)$, so that $\int_{A_i} f_u(x)du = 1/M$ for every $u \in \{\pm 1\}^M$. Clearly, there is a one-to-one correspondence between $(X_1, \cdots, X_N)$ and $(Y_1, \cdots, Y_M)$, where Y_i is the collection of observations in $X_1, \cdots, X_N$ that falls into the set A_i. By the above two properties, $(Y_1, \cdots, Y_M)$ are mutually independent, and $Y_i \sim f_{i,u_i}^{\otimes n}$ under f_u. Here f_{i,u_i} is the probability distribution of the following process: sample $N_i \sim \mathrm{Poi}(1/M)$, and draw N_i i.i.d. samples from the density

$$M\left(1 + c_0 u_i h g\left(\frac{x - (i-1)h}{h}\right)\right)$$

supported on A_i. In addition,

$$H^2(f_{i,+1}, f_{i,-1}) \asymp \mathbb{E}[N_i] \cdot M c_0^2 h^3 \|g\|_2^2 \asymp \frac{1}{n},$$

so the Poissonized model is a product model which satisfies the prerequisite of Theorem 6.5.

Next we denote the i.i.d. sampling model by $\mathcal{P}_c^{\otimes n}$, and the Poissonized model by $\mathcal{P}_c^{\mathrm{Poi}(n)}$. Let $n_1 = n + C\sqrt{n}, n_2 = n + m - C\sqrt{n+m}$, where $C > 0$ is a large universal constant to be chosen later. By Theorem 6.5 applied to $d = M \asymp n^{1/3}$, Le Cam's distance in Definition 3.1 between Poissonized models satisfies

$$\Delta(\mathcal{P}_c^{\mathrm{Poi}(n_1)}, \mathcal{P}_c^{\mathrm{Poi}(n_2)}) = \Omega(1),$$

as long as $m = \Omega(n^{5/6})$. In addition, the model $\mathcal{P}_c^{\text{Poi}(n_1)}$ is more informative than $\mathcal{P}_c^{\otimes n}$ with probability at least $1 - \epsilon_1$, where $\epsilon_1 = \mathbb{P}(\text{Poi}(n_1) < n)$. Similarly, $\mathcal{P}_c^{\otimes(n+m)}$ is more informative than $\mathcal{P}_c^{\text{Poi}(n_2)}$ with probability at least $1 - \epsilon_2$, where $\epsilon_2 = \mathbb{P}(\text{Poi}(n_2) > n+m)$. Therefore,

$$\Delta(\mathcal{P}_c^{\otimes n}, \mathcal{P}_c^{\otimes(m+n)}) \geq \Delta(\mathcal{P}_c^{\text{Poi}(n_1)}, \mathcal{P}_c^{\text{Poi}(n_2)}) - \epsilon_1 - \epsilon_2 = \Omega(1),$$

where in the last step we have $\epsilon_1, \epsilon_2 \to 0$ by choosing $C > 0$ large enough.

The proof of $m^\star(\mathcal{P}, n) \lesssim n^{3/4}$. Suppose $n \geq m \geq Cn^{3/4}$ for a large constant $C > 0$. Consider the following prior μ on the density f: let $M = \sqrt{n}$, and $u = (u_1, \cdots, u_M)$ be uniformly distributed over all vectors in $\{0, 1\}^M$ such that the number of 1's is $M/100$. Given u, we construct

$$f_u(x) = \begin{cases} u_i(2/M - 4|x - (2i-1)/2M|) & \text{if } x \in A_i \triangleq [(i-1)/M, i/M), i \in [M], \\ (8 - 2/(25M))(x - 1/2) & \text{if } 1/2 \leq x \leq 1, \end{cases}$$

and let $f = f_u$. It is not hard to verify that each f_u is a density and 8-Lipschitz, so $f_u \in \mathcal{P}$.

Next under the context of Lemma 6.1, we choose the action space to be $[0, 1]^{n+m}$, as well as the following loss:

$$L(f, x^{n+m}) = 1 - \mathbb{1}(x^{n+m} \text{ belongs to the support of } f,$$

$$\text{and fall into at least } [1 - (1 - n^{-1})^n]\sqrt{n}/100 + C_0 n^{1/4} \text{ sets in } \{A_i\}_{i=1}^M),$$

where $C_0 > 0$ is a large enough constant. We aim to show that under the loss L and prior μ, the Bayes risk satisfies $r_B(\mathcal{P}^{\otimes(n+m)}, \mu) \leq 0.1$ and $r_B(\mathcal{P}^{\otimes n}, \mu) \geq 0.9$. Then an application of Lemma 6.1 completes the proof of $m^\star(\mathcal{P}, n) \lesssim n^{3/4}$.

We first show that $r_B(\mathcal{P}^{\otimes(n+m)}, \mu) \leq 0.1$. We simply use the observed sample X^{n+m} as the estimator x^{n+m} under the above loss, then clearly x^{n+m} belongs to the support of f. For the second event in the loss function, let U_{n+m} be the number of sets in $\{A_i\}_{i=1}^M$ which receive any observations in X^{n+m}. By the linearity of expectation, as well as the negative dependence across different set counts, we compute for every f_u that

$$\mathbb{E}[U_{n+m}] = \sum_{i=1}^{M/100} \left[1 - \left(1 - \frac{1}{M^2}\right)^{n+m} \right] = \frac{\sqrt{n}}{100}\left[1 - \left(1 - \frac{1}{n}\right)^n\right] + \Omega\left(\frac{m}{\sqrt{n}}\right),$$

$$\text{Var}(U_{n+m}) \leq \mathbb{E}[U_{n+m}] = O(\sqrt{n}).$$

By Chebyshev's inequality, for $m \geq Cn^{3/4}$ with a large enough $C > 0$, we have $r_B(\mathcal{P}^{\otimes(n+m)}, \mu) \leq 0.1$.

Next we show that $r_B(\mathcal{P}^{\otimes n}, \mu) \geq 0.9$. Let $(i_1, \cdots, i_k) \in [M]^k$ be the indices such that the set A_{i_j} is hit by the observations X^n. Clearly the posterior distribution of the support of the vector u is uniformly distributed over

$$\{i_1, \cdots, i_k\} \cup \left\{ S \subseteq [M] \setminus \{i_1, \cdots, i_k\} : |S| = \frac{M}{100} - k \right\}.$$

On one hand, if the estimator x^{m+n} hits a set A_j with $j \notin \{i_1, \cdots, i_k\}$, the probability that x^{n+m} does not belong to the support of f is at least

$$1 - \frac{M/100 - k}{M - k} = \frac{0.99M}{M - k} \geq 0.99.$$

On the other hand, if the estimator x^{m+n} never hits a set A_j with $j \notin \{i_1, \cdots, i_k\}$, then x^{m+n} fall into at most U_n sets in $\{A_i\}_{i=1}^M$, where U_n is defined in a similar way as U_{n+m}. Similar

to the previous computation, we have

$$\mathbb{E}[U_n] = \frac{\sqrt{n}}{100}\left[1 - \left(1 - \frac{1}{n}\right)^n\right], \qquad \mathsf{Var}(U_n) = O(\sqrt{n}).$$

By Chebyshev's inequality, for $C_0 > 0$ large enough, the probability that x^{n+m} violates the second event in the loss function is at least 0.99. Now a combination of the above two cases implies that $r_{\mathrm{B}}(\mathcal{P}^{\otimes n}, \mu) \geq 0.9$, as desired.

APPENDIX C: PROOF OF MAIN LEMMAS

C.1. Proof of Lemma 4.4. Recall that the log-partition function $A(\theta)$ is defined as

$$A(\theta) = \log \int_{\mathcal{X}} \exp(\theta^\top T(x)) \mathrm{d}\mu(x),$$

for any vector $\lambda \in \mathbb{R}^d$ with $\theta + (\nabla^2 A(\theta))^{-1/2}\lambda \in \Theta$, we have

$$\mathbb{E}_\theta\left[\exp\left(\lambda^\top(\nabla^2 A(\theta))^{-1/2}(T(X) - \nabla A(\theta))\right)\right]$$

$$= \int_{\mathcal{X}} \exp\left((\theta + (\nabla^2 A(\theta))^{-1/2}\lambda)^\top T(x) - A(\theta) - [(\nabla^2 A(\theta))^{-1/2}\lambda]^\top \nabla A(\theta)\right)\mathrm{d}\mu(x)$$

(C.1)

$$= \exp\left(A(\theta + (\nabla^2 A(\theta))^{-1/2}\lambda) - A(\theta) - [(\nabla^2 A(\theta))^{-1/2}\lambda]^\top \nabla A(\theta)\right).$$

It remains to show that when $\|\lambda\|_2$ is sufficiently small, we always have $\theta + (\nabla^2 A(\theta))^{-1/2}\lambda \in \Theta$, and the exponent of (C.1) is uniformly bounded from above over $\theta \in \Theta$ and $\lambda \in \mathbb{R}^d$. Then the existence of uniformly bounded MGF around zero implies a uniformly bounded moment of any order.

The result of [13, Theorem 4.1.6] shows that for a self-concordant and convex function f, we have $\nabla^2 f(y) \preceq 4\nabla^2 f(x)$ whenever $(y - x)^\top \nabla^2 f(x)(y - x) \leq M^2/16$. Consequently, for $\|\lambda\|_2 \leq M/4$, a Taylor expansion with a Lagrange remainder gives

$$A(\theta + (\nabla^2 A(\theta))^{-1/2}\lambda) - A(\theta) - [(\nabla^2 A(\theta))^{-1/2}\lambda]^\top \nabla A(\theta)$$

(C.2)

$$= \frac{1}{2}\lambda^\top(\nabla^2 A(\theta))^{-1/2} \cdot \nabla^2 A(\xi) \cdot (\nabla^2 A(\theta))^{-1/2}\lambda,$$

where ξ lies on the line segment between θ and $\theta + (\nabla^2 A(\theta))^{-1/2}\lambda$. Consequently, we have

$$(\xi - \theta)^\top \cdot \nabla^2 A(\theta) \cdot (\xi - \theta) \leq \lambda^\top(\nabla^2 A(\theta))^{-1/2} \cdot \nabla^2 A(\theta) \cdot (\nabla^2 A(\theta))^{-1/2}\lambda = \|\lambda\|_2^2 \leq \frac{M^2}{16},$$

and therefore $\nabla^2 A(\xi) \preceq 4\nabla^2 A(\theta)$. Plugging it back into (C.2) establishes the boundedness of (C.1), as well as the finiteness of $A(\theta + (\nabla^2 A(\theta))^{-1/2}\lambda)$, or equivalently, $\theta + (\nabla^2 A(\theta))^{-1/2}\lambda \in \Theta$.

C.2. Proof of Lemma A.2. For $X_1, \cdots, X_n \sim \mathcal{N}(0, \Sigma)$ and $n \geq d$, it is well known that the empirical covariance $\widehat{\Sigma}_n$ follows a Wishart distribution $\mathcal{W}_d(\Sigma/n, n)$, where the density of $\mathcal{W}_d(V, n)$ is given by

$$f_{V,n}(X) = \frac{\det(X)^{(n-d-1)/2}\exp(-\mathsf{Tr}(V^{-1}X)/2)}{2^{nd/2}\det(V)^{n/2}\Gamma_d(n/2)},$$

where $\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma(x - (i-1)/2)$ is the multivariate Gamma function, and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the usual Gamma function. After some algebra, we have

$$D_{\mathrm{KL}}(\mathcal{W}_d(\Sigma/n, n) \| \mathcal{W}_d(\Sigma/(n+m), n+m))$$

(C.3)
$$= \frac{d}{2}\left(m - (n+m)\log\left(1 + \frac{m}{n}\right)\right) + \log\frac{\Gamma_d((n+m)/2)}{\Gamma_d(n/2)} - \frac{m}{2}\psi_d\left(\frac{n}{2}\right),$$

where $\psi_d(x) = \frac{d}{dx}[\log\Gamma_d(x)]$ is the multivariate digamma function. Note that the above KL divergence (C.3) does not depend on Σ; we denote it by $f(n, m, d)$.

By (3.2), it suffices to establish an upper bound of $f(n, m, d)$. Applying infinite Taylor series to $\log\Gamma_d(x)$ at $x = n/2$ yields

$$\log\Gamma_d\left(\frac{n+m}{2}\right) = \log\Gamma_d\left(\frac{n}{2}\right) + \frac{m}{2}\psi_d\left(\frac{n}{2}\right) + \sum_{t=2}^\infty \frac{1}{t!}\left(\frac{m}{2}\right)^t \psi_d^{(t-1)}\left(\frac{n}{2}\right)$$

$$= \log\Gamma_d\left(\frac{n}{2}\right) + \frac{m}{2}\psi_d\left(\frac{n}{2}\right) + \sum_{t=2}^\infty \frac{1}{t!}\left(\frac{m}{2}\right)^t \sum_{k=1}^d \psi^{(t-1)}\left(\frac{n-k+1}{2}\right),$$

where $\psi^{(t-1)}(x) = \frac{d^t}{dx^t}[\log\Gamma(x)]$ is the polygamma function. For any $t \geq 2$ and $x \geq 1$, the following inequality holds for the polygamma function [1, Equation 6.4.10]:

$$\left|\psi^{(t-1)}(x) - (-1)^t\frac{(t-2)!}{x^{t-1}}\right| \leq \frac{(t-1)!}{x^t}.$$

As a result, for the following modification

(C.4)
$$g(n, m, d) \triangleq \frac{d}{2}\left(m - (n+m)\log\left(1 + \frac{m}{n}\right)\right) + m\sum_{k=1}^d\sum_{t=2}^\infty \frac{(-1)^t}{2t(t-1)}\left(\frac{m}{n-k+1}\right)^{t-1},$$

it holds that

$$|f(n, m, d) - g(n, m, d)| \leq \sum_{k=1}^d\sum_{t=2}^\infty \frac{1}{t!}\left(\frac{m}{2}\right)^t \cdot \frac{(t-1)!}{[(n-k+1)/2]^t}$$

$$= \sum_{k=1}^d\sum_{t=2}^\infty \frac{1}{t}\left(\frac{m}{n-k+1}\right)^t$$

(C.5)
$$\leq d \cdot \sum_{t=2}^\infty \frac{1}{2}\left(\frac{2m}{n}\right)^t \leq \frac{4dm^2}{n^2},$$

where we have used the assumption $n \geq 4\max\{m, d\}$.

Next we establish an upper bound of $g(n, m, d)$. Using the identity

$$h(x) \triangleq \frac{(1+x)\log(1+x) - x}{x} = \sum_{t=2}^\infty \frac{(-1)^t x^{t-1}}{t(t-1)}$$

and some algebra, we have

$$g(n, m, d) = -\frac{dm}{2}h\left(\frac{m}{n}\right) + \frac{m}{2}\sum_{k=1}^d h\left(\frac{m}{n-k+1}\right) = \frac{m}{2}\sum_{k=1}^d\left[h\left(\frac{m}{n-k+1}\right) - h\left(\frac{m}{n}\right)\right].$$

Note that for $x \in [0,1]$, we have

$$h'(x) = \frac{x - \log(1+x)}{x^2} \in \left[0, \frac{1}{2}\right],$$

we conclude that

(C.6) $$g(n,m,d) \leq \frac{m}{2} \sum_{k=1}^{d} \frac{1}{2} \left[\frac{m}{n-k+1} - \frac{m}{n} \right] \leq \frac{md}{2} \cdot \frac{2md}{n^2} = \frac{m^2 d^2}{n^2}.$$

Finally, by (C.5), (C.6) and (3.2), we conclude that

$$\|\mathcal{L}(\widehat{\Sigma}_n) - \mathcal{L}(\widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \sqrt{\frac{f(n,m,d)}{2}} \leq \frac{2md}{n}.$$

For the claim that S_{n+m} follows the uniform distribution on the set $A = \{U \in \mathbb{R}^{d \times (n+m)} : UU^\top = I_d\}$, we need the following auxiliary definitions and results. A topological group is a group $(G,+)$ with a topology such that the operation $+ : G \times G \to G$ is continuous. A (right) group action of G on X is a function $\phi : X \times G \to X$ such that $\phi(\phi(x,g),g') = \phi(x,gg')$ and $\phi(x,e) = x$, where e is the identity element of G. A group action is called *transitive* if for every $x, x' \in A$, there exists some $g \in G$ such that $\phi(x,g) = x'$. A group action is called *proper* if for any compact $K \subseteq X$ and $x \in X$, the map $\phi_x : G \to X$ with $g \mapsto \phi(x,g)$ satisfies that $\phi_x^{-1}(K) \subseteq G$ is compact. The following lemma is useful.

LEMMA C.1 (Chapter 14, Theorem 25 of [14]).   *Let G be a locally compact group acting transitively and properly on a locally compact Hausdorff space X. Then there is a unique (up to multiplicative factors) Baire measure on X which is invariant under the action of G.*

Now for the claimed result, it is easy to verify that (a proper version of) S_{n+m} always takes value in A, assuming $n + m \geq d$. To show that S_{n+m} is uniform on A, note that for any orthogonal matrix $V \in \mathbb{R}^{(n+m) \times (n+m)}$, it is easy to verify

$$[X_1, \cdots, X_{n+m}] \stackrel{d}{=} [X_1, \cdots, X_{n+m}]V.$$

Denote the RHS by $[Y_1, \cdots, Y_{n+m}]$, we also have $\widehat{\Sigma}_{n+m}(Y^{n+m}) = \widehat{\Sigma}_{n+m}(X^{n+m})$. Consequently,

$$\begin{aligned}
S_{n+m}(X^{m+n}) \cdot V &= [(n+m)\widehat{\Sigma}_{n+m}(X^{n+m})]^{-1/2}[Y_1, Y_2, \cdots, Y_{n+m}] \\
&= [(n+m)\widehat{\Sigma}_{n+m}(Y^{n+m})]^{-1/2}[Y_1, Y_2, \cdots, Y_{n+m}] \\
&\stackrel{d}{=} [(n+m)\widehat{\Sigma}_{n+m}(X^{n+m})]^{-1/2}[X_1, X_2, \cdots, X_{n+m}] = S_{n+m}(X^{m+n}),
\end{aligned}$$

meaning that the distribution of S_{n+m} is invariant with right multiplication of an orthogonal matrix. Let G be the orthogonal group $\mathsf{O}(n+m)$, the map $\phi : A \times G \to A$ with $\phi(U,V) = UV$ is a group action of G on A. This action is transitive as for any $U, U' \in A$, we could add more rows to U, U' to obtain $\widetilde{U}, \widetilde{U}' \in G$, and then $V = \widetilde{U}^\top \widetilde{U}' \in G$ maps U to U'. This action is also proper as G itself is compact. Hence, Lemma C.1 below shows that S_{n+m} is uniformly distributed on A.

C.3. Proof of Lemma A.4. The proof of Lemma A.4 is a simple consequence of several known results. First, Basu's theorem claims that $\overline{X}_n$ and $\widehat{\Sigma}_n$ are independent. Second, the computation in Example 4.1 shows

$$\|\mathcal{L}(\overline{X}_n) - \mathcal{L}(\overline{X}_{n+m})\|_{\mathrm{TV}} \leq \frac{m\sqrt{d}}{n}.$$

Finally, as $\widehat{\Sigma}_n \sim \mathcal{W}_d(\Sigma/(n-1), n-1)$ again follows a Wishart distribution, the proof of Lemma A.2 shows that

$$\|\mathcal{L}(\widehat{\Sigma}_n) - \mathcal{L}(\widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \frac{2md}{n-1}.$$

In conclusion, we have

$$\|\mathcal{L}(\overline{X}_n, \widehat{\Sigma}_n) - \mathcal{L}(\overline{X}_{n+m}, \widehat{\Sigma}_{n+m})\|_{\mathrm{TV}} \leq \|\mathcal{L}(\overline{X}_n) - \mathcal{L}(\overline{X}_{n+m})\|_{\mathrm{TV}} + \|\mathcal{L}(\widehat{\Sigma}_n) - \mathcal{L}(\widehat{\Sigma}_{n+m})\|_{\mathrm{TV}}$$

$$\leq \frac{3md}{n-1}.$$

For the distribution of S_{n+m}, it is clear that (a proper version of) S_{n+m} always takes value in A, and we show that the distribution of S_{n+m} is invariant with proper group actions in A. Consider the set

$$G = \{V \in \mathbb{R}^{(n+m)\times(n+m)} : VV^\top = I_{n+m}, V\mathbf{1} = \mathbf{1}\}$$

with the usual matrix multiplication. We show that G is a group: clearly $I_{n+m} \in G$; for $V, V' \in G$, it is clear that $VV'\mathbf{1} = V\mathbf{1} = \mathbf{1}$ and therefore $VV' \in G$; for $V \in G$, it holds that $V^{-1}\mathbf{1} = V^{-1}V\mathbf{1} = \mathbf{1}$ and thus $V^{-1} \in G$. Next we show that the action $\phi: A \times G \to A$ with $\phi(U, V) = UV$ is a group action on A, and it suffices to show that $UV \in A$. This is true as $(UV)(UV)^\top = UVV^\top U^\top = UU^\top = I_d$, and $UV\mathbf{1} = U\mathbf{1} = \mathbf{0}$. This group action is also transitive: for any $U, U' \in A$, we may properly add rows to them and obtain $\widetilde{U}, \widetilde{U}' \in \mathsf{O}(n+m)$, where one of the added rows is a scalar multiple of $\mathbf{1}^\top$, which is feasible as $U\mathbf{1} = \mathbf{0}$. Consequently, the matrix $V = \widetilde{U}^{-1}\widetilde{U}' \in \mathsf{O}(n+m)$ maps U to U', and also $\mathbf{1}^\top$ to $\mathbf{1}^\top$; hence $V \in G$. Finally, we show that any group action of G on S_{m+n} does not change the distribution of S_{m+n}. To see this, for any $V \in G$ we have

$$[X_1, \cdots, X_{n+m}] \stackrel{d}{=} [X_1, \cdots, X_{n+m}]V,$$

$$\overline{X}_{n+m}([X_1, \cdots, X_{n+m}]V) = \overline{X}_{n+m}([X_1, \cdots, X_{n+m}]),$$

$$\widehat{\Sigma}_{n+m}([X_1, \cdots, X_{n+m}]V) = \widehat{\Sigma}_{n+m}([X_1, \cdots, X_{n+m}]).$$

Therefore, following the same arguments as in Example A.1 we arrive at the desired invariance, and the uniform distribution of S_{m+n} is a direct consequence of Lemma C.1.

C.4. Proof of Lemma 5.8. For any subset $S \subseteq [n+m]$ with $|S| = m$, let P_S be the distribution of $(Z_1, \cdots, Z_{n+m})$ when the samples $(Y_1, \cdots, Y_m)$ are placed in the index set S of the pool $(Z_1, \cdots, Z_{n+m})$. Then it is clear that $P_{\mathrm{mix}} = \mathbb{E}[P_S]$, with S uniformly distributed on all size-m subsets of $[n+m]$. To compute the χ^2-divergence where the first distribution is a mixture, the following identity holds [10]:

$$\chi^2\left(P_{\mathrm{mix}}, P^{\otimes(n+m)}\right) = \mathbb{E}_{S,S'}\left[\int \frac{\mathrm{d}P_S \mathrm{d}P_{S'}}{\mathrm{d}P^{\otimes(n+m)}}\right] - 1,$$

where S' is an independent copy of S. By the independence assumption, we have

$$\frac{\mathrm{d}P_S}{\mathrm{d}P^{\otimes(n+m)}}(z_1, \cdots, z_{n+m}) = \prod_{i \in S} \frac{\mathrm{d}Q}{\mathrm{d}P}(z_i),$$

and consequently

$$\int \frac{\mathrm{d}P_S \mathrm{d}P_{S'}}{\mathrm{d}P^{\otimes(n+m)}} = \mathbb{E}_P\left[\prod_{i\in S}\frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)\prod_{i\in S'}\frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)\right]$$

$$= \prod_{i\in S\cap S'}\mathbb{E}_P\left[\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)\right)^2\right]\cdot\prod_{i\in S\Delta S'}\mathbb{E}_P\left[\frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)\right]$$

$$= (1+\chi^2(Q,P))^{|S\cap S'|}.$$

It remains to upper bound the expectation with respect to the random variable $|S\cap S'|$. Note that $|S\cap S'|$ follows the hypergeometric distribution with parameter $(n+m,m,m)$, which corresponds to sampling *without replacement*. The counterpart for sampling *with replacement* corresponds to a Binomial distribution $\mathsf{B}(m,\frac{m}{n+m})$, and the following lemma shows that the latter dominates the former in terms of the convex order:

LEMMA C.2. *[9, Theorem 4] Let the population be $\mathcal{C} = \{c_1,\cdots,c_N\}$. Let $X_1,\cdots,X_n$ denote random samples without replacement from $\mathcal{C}$ and $Y_1,\cdots,Y_n$ denote random samples with replacement. If $f(\cdot)$ is convex, then*

$$\mathbb{E}\left[f\left(\sum_{i=1}^n X_i\right)\right] \le \mathbb{E}\left[f\left(\sum_{i=1}^n Y_i\right)\right].$$

Applying Lemma C.2 to the convex function $x \mapsto (1+\chi^2(Q,P))^x$ yields

$$\mathbb{E}_{S,S'}[(1+\chi^2(Q,P))^{|S\cap S'|}] \le \mathbb{E}[(1+\chi^2(Q,P))^{\mathsf{B}(m,m/(n+m))}]$$

$$= \left(\mathbb{E}[(1+\chi^2(Q,P))^{\mathsf{Bern}(m/(n+m))}]\right)^m$$

$$= \left(1+\frac{m}{n+m}\chi^2(Q,P)\right)^m,$$

as desired.

C.5. Proof of Lemma A.7. Note that in the proof of Theorem 5.2, we have

$$\|P_{\mathrm{mix}}(X^{n/2}) - P^{\otimes(n/2+m)}\|_{\mathrm{TV}} \le \sqrt{\frac{m^2}{n}\chi^2(\widehat{P}_n(X^{n/2}),P)}.$$

Since the TV distance is always upper bounded by one, the following upper bound is also true:

$$\|P_{\mathrm{mix}}(X^{n/2}) - P^{\otimes(n/2+m)}\|_{\mathrm{TV}} \le \sqrt{\frac{m^2}{n}\cdot\left(\chi^2(\widehat{P}_n(X^{n/2}),P)\wedge n\right)}.$$

Consequently, taking the expectation over $X^{n/2}$ leads to the claim.

C.6. Proof of Lemma A.13. First we note that

$$\chi^2\left(\mathcal{N}(\widehat{\theta}_{n,j},1),\mathcal{N}(\theta_j,1)\right) = \exp\left((\widehat{\theta}_{n,j}-\theta_j)^2\right) - 1,$$

By the triangle inequality,

$$\mathbb{E}|\widehat{\theta}_{n,j} - \theta_j| \leq \mathbb{E}\left|\widehat{\theta}_{n,j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right| + \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \theta_j\right|$$

$$(C.7) \qquad\qquad \leq \sqrt{\frac{C\log n}{n}} + \sqrt{\frac{2}{n\pi}} = O\left(\sqrt{\frac{\log n}{n}}\right).$$

Moreover, it is straightforward to verify that $|\widehat{\theta}_{n,j} - \theta_j|$ is $(1/n)$-Lipschitz with respect to $(X_{1,j}, \cdots, X_{n,j})$. Therefore, the Gaussian Lipschitz concentration (see, e.g. [4, Theorem 10.17]) gives that

$$(C.8) \qquad \mathbb{P}\left(|\widehat{\theta}_{n,j} - \theta_j| \geq \mathbb{E}|\widehat{\theta}_{n,j} - \theta_j| + t\right) \leq \exp\left(-\frac{nt^2}{2}\right)$$

for any $t \geq 0$. Hence, combining (C.7) and (C.8), we conclude that $|\widehat{\theta}_{n,j} - \theta_j| = O(\sqrt{\log(nd)/n})$ holds with probability at least $1 - (nd)^{-2}$, and therefore

$$\mathbb{E}\left[\chi^2\left(\mathcal{N}(\widehat{\theta}_{n,j}, 1), \mathcal{N}(\theta_j, 1)\right) \wedge n\right] \leq 2 \cdot \mathbb{E}[(\widehat{\theta}_{n,j} - \theta_j)^2] + \frac{1}{nd},$$

where we have used that $e^x \leq 1 + 2x$ whenever $x \in [0, 1]$. Summing over $j \in [d]$ and using the property of the soft-thresholding estimator $\sup_{\theta \in \Theta} \mathbb{E}[\|\widehat{\theta}_n - \theta\|_2^2] = O(s\log d/n)$ gives the claimed result.

C.7. Proof of Lemma A.15. For the first claim, note that for Poisson models, we have

$$\chi^2\left(\mathsf{Poi}(\widehat{\lambda}_{n,j}), \mathsf{Poi}(\lambda_j)\right) = \exp\left(\frac{(\widehat{\lambda}_{n,j} - \lambda_j)^2}{\lambda_j}\right) - 1.$$

Consequently, for $\lambda_j = 1/(n^2\log n)$, we have

$$\mathbb{E}\left[\chi^2\left(\mathsf{Poi}(\widehat{\lambda}_{n,j}), \mathsf{Poi}(\lambda_j)\right) \wedge n\right] \geq \left[\left(\exp\left(\frac{(1/n - \lambda_j)^2}{\lambda_j}\right) - 1\right) \wedge n\right] \cdot \mathbb{P}(\widehat{\lambda}_{n,j} = 1/n)$$

$$= \Omega(n) \cdot e^{-n\lambda_j} n\lambda_j = \Omega\left(\frac{1}{\log n}\right) \gg \frac{1}{n},$$

establishing the first claim.

For the second claim, note that conditioning on $X^{n/2}$,

$$D_{\mathrm{KL}}\left(\mathcal{L}(\widehat{T}_{n/2+m}) \| \mathcal{L}(T_{n/2+m})\right) = \sum_{j=1}^{d} D_{\mathrm{KL}}\left(\mathsf{Poi}(n\lambda_j/2 + m\widehat{\lambda}_{n,j}) \| \mathsf{Poi}((n/2+m)\lambda_j)\right)$$

$$\leq \sum_{j=1}^{d} \frac{m^2(\widehat{\lambda}_{n,j} - \lambda_j)^2}{(n/2+m)\lambda_j}.$$

where we have used $D_{\mathrm{KL}}(\mathsf{Poi}(\lambda_1) \| \mathsf{Poi}(\lambda_2)) = \lambda_2 - \lambda_1 + \lambda_1\log(\lambda_1/\lambda_2) \leq (\lambda_1 - \lambda_2)^2/\lambda_2$. Hence,

$$\mathbb{E}_{X^{n/2}}\left[D_{\mathrm{KL}}\left(\mathcal{L}(\widehat{T}_{n/2+m}) \| \mathcal{L}(T_{n/2+m})\right)\right] \leq \frac{dm^2}{(n/2+m)n/2} \leq \frac{4dm^2}{n^2},$$

and the TV distance satisfies

$$\|\mathcal{L}(\widehat{X}^{n+m}) - \mathcal{L}(X^{n+m})\|_{\mathrm{TV}} = \mathbb{E}_{X^{n/2}}\left[\|\mathcal{L}(\widehat{T}_{n/2+m}) - \mathcal{L}(T_{n/2+m})\|_{\mathrm{TV}}\right]$$

$$\leq \mathbb{E}_{X^{n/2}}\left[\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathcal{L}(\widehat{T}_{n/2+m})\|\mathcal{L}(T_{n/2+m})\right)}\right]$$

$$\leq \sqrt{\frac{1}{2}\mathbb{E}_{X^{n/2}}\left[D_{\mathrm{KL}}\left(\mathcal{L}(\widehat{T}_{n/2+m})\|\mathcal{L}(T_{n/2+m})\right)\right]}$$

$$\leq \frac{m\sqrt{2d}}{n}.$$

C.8. Proof of Lemma A.17. The upper bound result is easy. The $m = O(n\epsilon/\sqrt{k})$ upper bound is a consequence of the general product Poisson model considered in Example A.14. For the $m = O(\sqrt{n}\epsilon)$ upper bound, we consider the sufficient statistic $T_n = \sum_{i=1}^{n} X_i \sim \prod_{j=1}^{k} \mathsf{Poi}(np_j)$, and simply apply the sufficiency-based algorithm to $\widehat{T}_{n+m} = T_n$. Since

$$D_{\mathrm{KL}}(\mathcal{L}(T_{n+m})\|\mathcal{L}(T_n)) = \sum_{j=1}^{k} D_{\mathrm{KL}}(\mathsf{Poi}((n+m)p_j)\|\mathsf{Poi}(np_j)) \leq \sum_{j=1}^{k} \frac{(mp_j)^2}{np_j} = \frac{m^2}{n},$$

where the last identity crucially makes use of the identity $\sum_{j=1}^{k} p_j = 1$, this procedure works.

Next we show that sample amplification is impossible when $m = \omega(n\epsilon/\sqrt{k})$ and $k = O(n)$. Note that this implies that $m = \omega(\max\{n\epsilon/\sqrt{k}, \sqrt{n}\epsilon\})$ is impossible in general, for the $k > n$ case is always not easier than the $k = n$ case. To prove the above claim, w.l.o.g. we assume that $k = 2k_0$ is even, and consider the following parametric submodel:

$$P_\theta = \prod_{j=1}^{k_0}\left[\mathsf{Poi}\left(\frac{1}{k} + \theta_j\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta_j\right)\right], \qquad \theta \in \Theta \triangleq \left[-\frac{1}{k}, \frac{1}{k}\right]^{k_0}.$$

Clearly P_θ is a parametric submodel, by setting $p_{2j-1} = 1/k + \theta_j$ and $p_{2j} = 1/k - \theta_j$ in the original model. This submodel is a product model, thus we could apply the result of Theorem 6.5 after we have verified Assumption 4. Note that when θ, θ' arbitrarily vary in $[-1/k, 1/k]$, the range of

$$H^2\left(\mathsf{Poi}\left(\frac{1}{k} + \theta\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta\right), \mathsf{Poi}\left(\frac{1}{k} + \theta'\right) \times \mathsf{Poi}\left(\frac{1}{k} - \theta'\right)\right)$$

is $[0, \Theta(1/k)]$, so Assumption 4 is fulfilled when $k \leq cn$ for a small constant $c > 0$. Consequently, Theorem 6.5 establishes the desired bound $m = O(n\epsilon/\sqrt{k})$.

C.9. Proof of Lemma A.19. For the first inequality, note that for a large $C_0 > 0$, both $Z_1 \leq C_0$ and $\max_{2\leq j\leq d} Z_j \geq \sqrt{2\log d} - C_0$ hold with probability at least 0.99. When both events hold, we have

$$\exp(t(t + Z_1)) \leq \exp(t(t + C_0)),$$

$$\sum_{j=2}^{d} \exp(tZ_j) \geq \exp(t(\sqrt{2\log d} - C_0)).$$

Consequently, for $C = 3C_0$ with $C_0 > 0$ large enough, we have

$$p_d(\sqrt{2\log d} - C) \leq 0.01 + \frac{\exp((\sqrt{2\log d} - 3C_0)(\sqrt{2\log d} - 2C_0))}{\exp((\sqrt{2\log d} - 3C_0)(\sqrt{2\log d} - C_0))}$$

$$= 0.01 + \exp\left(-C_0(\sqrt{2\log d} - 3C_0)\right) < 0.1.$$

For the second inequality, note that Jensen's inequality yields that

$$p_d(t) \geq \mathbb{E}_{Z_1}\left[\frac{\exp(t(t + Z_1))}{\exp(t(t + Z_1)) + \sum_{j=2}^{d} \mathbb{E}[\exp(tZ_j)]}\right]$$

$$\geq \mathbb{E}_{Z_1}\left[\frac{\exp(t(t + Z_1))}{\exp(t(t + Z_1)) + d\exp(t^2/2)}\right].$$

Again, with probability at least 0.99 we have $Z_1 \geq -C_0$, and therefore for $t = \sqrt{2\log d} + 2C_0$,

$$p_d(t) \geq 0.99 \times \frac{\exp(t(t - C_0))}{\exp(t(t - C_0)) + d\exp(t^2/2)}$$

$$= 0.99 \times \frac{1}{1 + \exp(t^2/2 + (t - 2C_0)^2/2 - t(t - C_0))}$$

$$= 0.99 \times \frac{1}{1 + \exp(-C_0 \cdot \sqrt{2\log d})} > 0.9,$$

for $C_0 > 0$ large enough.

For the last claim, consider any $s < d/2$. Let $d_0 \triangleq \lfloor d/s \rfloor \geq 2$, and consider the product prior $\mu^{\otimes s}$ to s blocks each of dimension d_0. In other words, we set exactly one of the first d_0 coordinates of the mean vector to t uniformly at random, and do the same for the next d_0 coordinates, and so on. Clearly the resulting mean vector is always s-sparse. Writing $\theta = (\theta_1, \cdots, \theta_s)$ with each $\theta_i \in \mathbb{R}^{d_0}$, let the loss function be

$$L(\theta, \widehat{\theta}) = \mathbb{1}\left(\sum_{j=1}^{s} \mathbb{1}(\theta_j \neq \widehat{\theta}_j) \geq N\right),$$

with an integer N to be specified later. Then in each block, we reduce to the case $s = 1$, and the error probability for this block is $1 - p_{d_0}(\sqrt{n} \cdot t)$ for sample size n. Moreover, the errors in different blocks are independent. Consequently,

$$r_{\mathrm{B}}(\mathcal{P}, n, \mu, L) - r_{\mathrm{B}}(\mathcal{P}, n + m, \mu, L)$$

(C.9) $$= \mathbb{P}\left(\mathsf{B}(s, 1 - p_{d_0}(\sqrt{n} \cdot t)) \geq N\right) - \mathbb{P}\left(\mathsf{B}(s, 1 - p_{d_0}(\sqrt{n + m} \cdot t)) \geq N\right).$$

Finally, again by the properties of $p_d(\cdot)$ summarized in Lemma A.19 and the pigeonhole principle, for $m = \lceil cn\epsilon/\sqrt{s\log d_0} \rceil$, we could always find some $t > 0$ such that $p_{d_0}(\sqrt{n + m} \cdot t) - p_{d_0}(\sqrt{n} \cdot t) = \Omega_c(\epsilon/\sqrt{s})$, with both quantities in $[0.1, 0.9]$. Now choosing

$$N = \left\lceil \frac{s(1 - p_{d_0}(\sqrt{n} \cdot t))}{2} + \frac{s(1 - p_{d_0}(\sqrt{n + m} \cdot t))}{2} \right\rceil$$

with the above choice of t, the Bayes risk difference will be lower bounded by $\Omega_c(\epsilon)$ by a similar argument to the proof of Theorem 6.4. Now by (C.9) and Lemma 6.1, we see that $n = \Omega(\sqrt{s\log(d/s)}/\epsilon)$ and $m = O(n\epsilon/\sqrt{s\log(d/s)})$ are necessary for sample amplification.

C.10. Proof of Lemma A.21. The following results are the key to the proof of Lemma A.21, which we will assume for now and prove in the subsequent sections.

LEMMA C.3. *For $D_n = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$, the following identities hold for the estimator $\widehat{\Sigma}_n$:*

$$(\text{C.10}) \qquad \mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] = \sum_{j=1}^{d} \left[(n+d+1-2j)\lambda_j - \log\lambda_j - \mathbb{E}\left[\log\chi^2_{n-j+1}\right] \right] - d,$$

$$(\text{C.11}) \qquad \mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n)) = \sum_{j=1}^{d} \left[2(n+d+1-2j)\lambda_j^2 - 4\lambda_j + \psi'\left(\frac{n+1-j}{2}\right) \right],$$

where χ^2_m denotes the chi-squared distribution with m degrees of freedom, and $\psi'(x)$ is the polygamma function of order 1. In particular, when $n \geq 2d$, the following inequalities hold:

$$(\text{C.12}) \qquad \left| \mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] - \left(g(n+1-d,d) + \sum_{j=1}^{d} h((n+d+1-2j)\lambda_j) \right) \right| \leq \frac{5d}{n},$$

$$(\text{C.13}) \qquad \mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n)) \leq \frac{16d^2}{n^2} + \sum_{j=1}^{d} \frac{4((n+d+1-2j)\lambda_j - 1)^2}{n},$$

where the functions g and h are given by (A.2) and

$$(\text{C.14}) \qquad h(u) \triangleq u - \log u - 1.$$

LEMMA C.4. *For the function h in (C.14) and $u_1, \cdots, u_d \in \mathbb{R}_+$, it holds that*

$$(\text{C.15}) \qquad \sum_{j=1}^{d} h(u_j) \geq \frac{1}{8} \min\left\{ \sum_{j=1}^{d} (u_j - 1)^2, \sqrt{\sum_{j=1}^{d} (u_j - 1)^2} \right\}.$$

Returning to the proof of Lemma A.21, the first claim is a direct application of (C.12) and (C.13). For the second claim, let $D_n = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$, and

$$V \triangleq \sum_{j=1}^{d} ((n+d+1-2j)\lambda_j - 1)^2.$$

Then by (C.12), (C.13) and Lemma C.4, we have

$$\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] \geq g(n+1-d,d) + \frac{V \wedge \sqrt{V}}{8} - \frac{5d}{n}.$$

Meanwhile, the variance satisfies

$$\sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n))} \leq \frac{4d}{n} + 2\sqrt{\frac{V}{n}}.$$

Note that for n, d larger than a constant depending only on C_1, we always have

$$\frac{d}{n} + \frac{V \wedge \sqrt{V}}{8} \geq 2C_1\sqrt{\frac{V}{n}}.$$

Therefore, for $n, d = \Omega(1)$ large enough, the above inequalities implies

$$\ell(\Sigma, \widehat{\Sigma}) \geq g(n + 1 - d, d) + C_1 \left(\sqrt{\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n))} - \frac{4d}{n} \right) - \frac{6d}{n},$$

establishing the second claim.

For the last statement, note that

$$\frac{\partial g(u, v)}{\partial u} = \frac{\log(u + 2v) + \log(v)}{2} - \log(u + v) = \frac{1}{2} \log \left(\frac{u(u + 2v)}{(u + v)^2} \right) \leq -\frac{v^2}{2(u + v)^2},$$

the intermediate value theorem then implies that

$$g(n + 1 - d, d) - g(n + m + 1 - d, d) \geq m \cdot \min_{u \in [n+1-d, n+m+1-d]} \frac{d^2}{2(u + d)^2} \geq \frac{md^2}{13n^2}.$$

C.11. Proof of Lemma C.3. We first recall the well-known Bartlett decomposition: for the lower triangular matrix $L_n = (L_{ij})_{i \geq j}$, the random variables $\{L_{ij}\}_{i \geq j}$ are mutually independent, with

$$L_{ij} \sim \mathcal{N}(0, 1), \quad i > j; \qquad L_{jj}^2 \sim \chi_{n+1-j}^2, \quad j \in [d].$$

For $\Sigma = I_d$ and $\widehat{\Sigma}_n = L_n D_n L_n^\top$, simple algebra gives

$$\ell(\Sigma, \widehat{\Sigma}_n) = \sum_{j=1}^d \left(\sum_{i \geq j} \lambda_j L_{ij}^2 - \log \lambda_j - \log L_{jj}^2 - 1 \right).$$

Consequently, the identity (C.10) follows from the above Bartlett decomposition. This identity was also obtained in [11].

For the identity (C.11) on the variance, by the mutual independence we have

(C.16)
$$\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n)) = \sum_{j=1}^d \sum_{i \geq j} \lambda_j^2 \cdot \mathsf{Var}(L_{ij}^2) + \sum_{j=1}^d \mathsf{Var}(\log L_{jj}^2) - 2 \sum_{j=1}^d \lambda_j \cdot \mathsf{Cov}(L_{jj}^2, \log L_{jj}^2).$$

Next we evaluate each term of (C.16). Clearly $\mathsf{Var}(L_{ij}^2) = \mathbb{E}[Z^4] - \mathbb{E}[Z^2]^2 = 2$ for $i > j$ and $Z \sim \mathcal{N}(0, 1)$. For the other random variables, we need to recall the following identity for χ_m^2:

(C.17)
$$\Lambda(t) \triangleq \log \mathbb{E}[(\chi_m^2)^t] = t \log 2 + \log \Gamma \left(\frac{m}{2} + t \right) - \log \Gamma \left(\frac{m}{2} \right).$$

Based on (C.17), we have

$$\mathsf{Var}(L_{jj}^2) = (n + 1 - j)(n + 3 - j) - (n + 1 - j)^2 = 2(n + 1 - j).$$

Moreover, differentiating $\Lambda(t)$ at $t = 0$ gives

(C.18)
$$\mathbb{E}[\log \chi_m^2] = \Lambda'(0) = \log 2 + \psi \left(\frac{m}{2} \right),$$

(C.19)
$$\mathbb{E}[(\log \chi_m^2)^2] - \left(\mathbb{E}[\log \chi_m^2] \right)^2 = \Lambda''(0) = \psi' \left(\frac{m}{2} \right).$$

Note that (C.19) leads to

$$\mathsf{Var}(\log L_{jj}^2) = \psi' \left(\frac{n + 1 - j}{2} \right).$$

Finally, differentiating $\Lambda(t)$ at $t = 1$ gives that

$$\mathbb{E}[L_{jj}^2 \log L_{jj}^2] = \mathbb{E}[L_{jj}^2] \cdot \left(\log 2 + \psi\left(\frac{n+1-j}{2} + 1 \right) \right)$$

$$= (n+1-j)\left(\log 2 + \psi\left(\frac{n+1-j}{2} + 1 \right) \right),$$

and hence the identity $\psi(x+1) = \psi(x) + x^{-1}$ for the digamma function together with (C.18) leads to

$$\mathsf{Cov}(L_{jj}^2, \log L_{jj}^2) = 2.$$

Therefore, plugging the above quantities in (C.16) gives the identity (C.11).

Next we prove the remaining inequalities when $n \geq 2d$. For (C.12), note that (C.10) gives an identity (together with (C.18))

$$\mathbb{E}[\ell(\Sigma, \widehat{\Sigma}_n)] - \left(g(n+1-d, d) + \sum_{j=1}^d h((n+d+1-2j)\lambda_j) \right)$$

$$= \sum_{j=1}^d \left(\log(n+d+1-2j) - \log 2 - \psi\left(\frac{n+1-j}{2} \right) \right) - g(n+1-d, d).$$

Since $|\psi(x) - \log(x)| \leq 1/x$ for all $x \geq 1$, replacing $\psi(\cdot)$ by $\log(\cdot)$ in the above expression only incurs an absolute difference at most

$$\sum_{j=1}^d \frac{2}{n+1-j} \leq 2 \int_0^d \frac{\mathrm{d}x}{n-x} = 2 \log\left(1 + \frac{d}{n-d} \right) \leq \frac{4d}{n}.$$

For the remaining terms, it is not hard to verify that

$$g(n+1-d, d) = \int_0^d \log \frac{n+1-d+2x}{n+1-d+x} dx.$$

As $x \mapsto \log(n+1-d+2x) - \log(n+1-d+x)$ is increasing on $[0, \infty)$, we have

$$0 \leq g(n+1-d, d) - \sum_{x=0}^{d-1} \log \frac{n+1-d+2x}{n+1-d+x} \leq \log \frac{n+1-d+2d}{n+1-d+d} \leq \frac{d}{n}.$$

Now (C.12) follows from a combination of the above inequalities.

For the inequality (C.13), we complete the square of (C.11) to obtain

$$\mathsf{Var}(\ell(\Sigma, \widehat{\Sigma}_n)) = \sum_{j=1}^d \frac{2((n+d+1-2j)\lambda_j - 1)^2}{n+d+1-2j} + \sum_{j=1}^d \left(\psi'\left(\frac{n+1-j}{2} \right) - \frac{2}{n+d+1-2j} \right)$$

$$\leq \frac{4}{n} \sum_{j=1}^d ((n+d+1-2j)\lambda_j - 1)^2 + \sum_{j=1}^d \left(\psi'\left(\frac{n+1-j}{2} \right) - \frac{2}{n+d+1-2j} \right).$$

To handle the second sum, note that [1, Equation 6.4.10] gives $|\psi'(x) - x^{-1}| \leq x^{-2}$ for $x \geq 1$. Therefore, the second term has an absolute value at most

$$\sum_{j=1}^d \left[\frac{2(d-j)}{(n+1-j)(n+d+1-2j)} + \left(\frac{2}{n+1-j} \right)^2 \right] \leq d \cdot \left[\frac{2d}{(n/2)^2} + \left(\frac{2}{n/2} \right)^2 \right] \leq \frac{16d^2}{n^2}.$$

C.12. Proof of Lemma C.4. Note that when $0 < u \leq 2$, Taylor expansion of $h(\cdot)$ at $u = 1$ gives

$$h(u) \geq \frac{(u-1)^2}{2} \min_{\theta \in [0,2]} h''(\theta) = \frac{(u-1)^2}{8}.$$

For $u > 2$, we have $u - 1 \geq \log_2 u > 10/7 \cdot \log u$, and therefore

$$h(u) = u - \log u - 1 \geq \frac{3(u-1)}{10}.$$

Therefore, in both cases we have

$$h(u) \geq \frac{1}{8} \min \left\{ (u-1)^2, |u-1| \right\}.$$

To prove (C.15), let $J \triangleq \{j \in [d] : |u_j - 1| \leq 1\}$, $S \triangleq \sum_{j=1}^{d} (u_j - 1)^2$. Using the above inequality, we have

$$
\begin{aligned}
\sum_{j=1}^{d} h(u_j) &\geq \frac{1}{8} \sum_{j \in J} (u_j - 1)^2 + \frac{1}{8} \sum_{j \notin J} |u_j - 1| \\
&\geq \frac{1}{8} \sum_{j \in J} (u_j - 1)^2 + \frac{1}{8} \sqrt{\sum_{j \notin J} (u_j - 1)^2} \\
&\geq \frac{1}{8} \min_{x \in [0,S]} \left(x + \sqrt{S - x} \right) \\
&= \frac{1}{8} \min\{S, \sqrt{S}\},
\end{aligned}
$$

which is precisely (C.15).

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* **55**. US Government printing office.

[2] AXELROD, B., GARG, S., SHARAN, V. and VALIANT, G. (2020). Sample amplification: Increasing dataset size even when learning is impossible. In *International Conference on Machine Learning* 442–451. PMLR.

[3] BALLY, V. and CARAMELLINO, L. (2016). Asymptotic development for the CLT in total variation distance. *Bernoulli* **22** 2442–2485.

[4] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

[5] DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2018). The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*.

[6] DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2020). The minimax learning rates of normal and Ising undirected graphical models. *Electronic Journal of Statistics* **14** 2338 – 2361. https://doi.org/10.1214/20-EJS1721

[7] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

[8] HOEFFDING, W. (1956). On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics* 713–721.

[9] HOEFFDING, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58** 13–30.

[10] INGSTER, Y. and SUSLINA, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models* **169**. Springer Science & Business Media.

[11] JAMES, W. and STEIN, C. (1961). Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

[12] MIESCKE, K.-J. and LIESE, F. (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer.

[13] NESTEROV, Y. (2003). *Introductory lectures on convex optimization: A basic course* **87**. Springer Science & Business Media.

[14] ROYDEN, H. L. (1988). *Real analysis (3rd edition)*. Prentice-Hall.

[15] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

[16] VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.