The Annals of Statistics
2024, Vol. 52, No. 6, 2767–2790
https://doi.org/10.1214/24-AOS2444

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]*Courant 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 [3], 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

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

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

Received June 2023; revised August 2024.
MSC2020 subject classifications. Primary 62B15; secondary 62C20.
Key words and phrases. Sample amplification, Le Cam's distance, minimax rate.

known invariances such as rotation or translation invariance to augment the dataset by including transformed versions of each original datapoint [23, 24, 30, 45, 46]. 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, 4, 19–22, 26, 27, 37, 38, 40, 44, 54, 55]. More opaque approaches such as MixUp [57] and related techniques [29, 48, 51, 56] 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 [3], 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, that is, 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 procedure than [3] 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 nontrivial 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

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

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: [3], 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

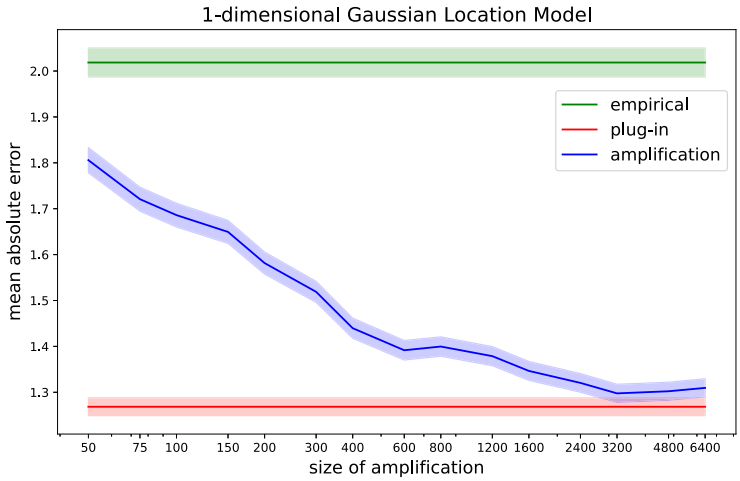
one idea over the other, and there exists an example where the above two ideas must be combined.

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.

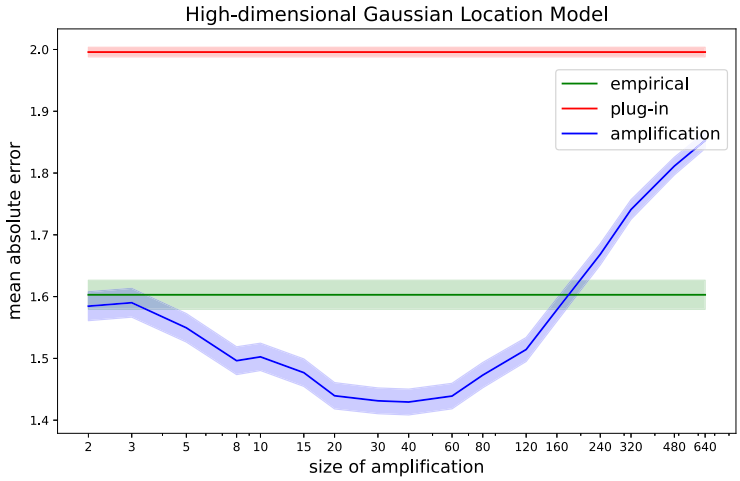
For non-product models, the above powerful result does not directly apply, but proper applications of the general recipe could still lead to tight lower bounds for sample amplification. Specifically, we obtain matching lower bounds for all examples listed in Table 1, including the non-product examples.

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:

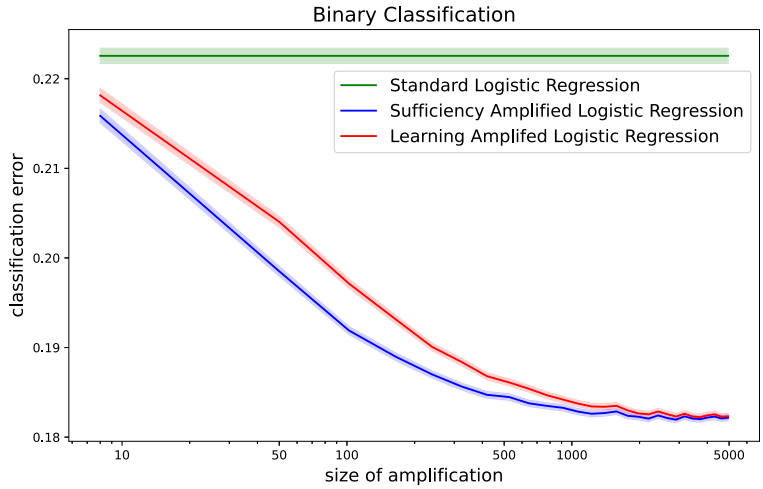
- Fourth moment estimation for one-dimensional Gaussian: here we observe $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ with $n = 100$ and $\mu = 1$, and we consider three estimators. The empirical estimator operates in a distribution-agnostic fashion and is simply the empirical fourth moment $n^{-1}\sum_{i=1}^n X_i^4$. 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 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}.



(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* 5000 *simulations.*

- Squared L_2 norm estimation for high-dimensional Gaussian: here we observe $X_1, \dots, X_n \sim \mathcal{N}(\mu, I_d)$ 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 $X_1, \dots, X_{n/2} \sim \mathcal{N}(e_1, I_d)$ (with label 1) and $X_{n/2+1}, \dots, X_n \sim \mathcal{N}(-e_1, I_d)$ (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 cannot be improved even when n is large enough to learn the distribution to nontrivial 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 [3] 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 [3] 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 [32], 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 [32–34] and [50], Chapter 9. In nonparametric statistics, there is also a rich line of research [16–18, 43] 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 nonasymptotic 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, for example, [11, 14, 15, 39]. 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 nonasymptotic 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 noncomparable. More concrete examples of both the upper and lower bounds, auxiliary lemmas and proofs are relegated to the appendices in the Supplementary Material [2].

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, that is, $\mathcal{L}(T(X))$ 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 $[n] \triangleq \{1, \dots, n\}$, and $x^n \triangleq (x_1, \dots, x_n)$. We adopt the following asymptotic notations: for two nonnegative 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:

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

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

We will frequently use the following inequalities between the above quantities [49], Chapter 2:

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

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

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

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

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, that is,

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

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:

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

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 [52]. 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

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

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 [32–34]). 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

$$\begin{aligned}\Delta(\mathcal{M}, \mathcal{N}) &= \max\Bigl\{\sup_L \sup_{\delta_{\mathcal{N}}} \inf_{\delta_{\mathcal{M}}} \sup_{\theta\in\Theta} \bigl|R_{\mathcal{M}}(\theta, \delta_{\mathcal{M}}, L) - R_{\mathcal{N}}(\theta, \delta_{\mathcal{N}}, L)\bigr|,\\ &\qquad \sup_L \sup_{\delta_{\mathcal{M}}} \inf_{\delta_{\mathcal{N}}} \sup_{\theta\in\Theta} \bigl|R_{\mathcal{M}}(\theta, \delta_{\mathcal{M}}, L) - R_{\mathcal{N}}(\theta, \delta_{\mathcal{N}}, L)\bigr|\Bigr\}\\ &= \max\Bigl\{\inf_{T_1} \sup_\theta \bigl\| P_\theta \circ T_1^{-1} - Q_\theta \bigr\|_{\mathrm{TV}}, \inf_{T_2} \sup_\theta \bigl\| Q_\theta \circ T_2^{-1} - P_\theta \bigr\|_{\mathrm{TV}}\Bigr\},\end{aligned}$$

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 [31], 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*, that is, 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 $\mathcal{P}^{\otimes(n+m)} \circ T_{n+m}^{-1}$, that is, 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 $T_n = T_n(X_1, \dots, X_n)$.
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

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

Algorithm 1 SAMPLE AMPLIFICATION VIA SUFFICIENCY

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

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

$$\begin{aligned}\big\|\mathcal{L}(\widehat{X}^{n+m}) - \mathcal{L}(X^{n+m})\big\|_{\mathrm{TV}} &= \big\|\mathcal{L}(\widehat{T}_{n+m}) \times P_{X^{n+m}|T_{n+m}} - \mathcal{L}(T_{n+m}) \times P_{X^{n+m}|T_{n+m}}\big\|_{\mathrm{TV}} \\ &\overset{\text{(a)}}{=} \big\|\mathcal{L}(\widehat{T}_{n+m}) - \mathcal{L}(T_{n+m})\big\|_{\mathrm{TV}} = \big\|\mathcal{L}\big(f(T_n)\big) - \mathcal{L}(T_{n+m})\big\|_{\mathrm{TV}},\end{aligned}$$

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 $X_1, \ldots, X_n$ 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

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

Now consider the identity map between sufficient statistics $\widehat{T}_{n+m} = T_n$ used with Algorithm 1. The amplified samples $(\widehat{X}_1, \ldots, \widehat{X}_{n+m})$ 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:

$$\begin{aligned}\big\|\mathcal{L}(\widehat{T}_{n+m}) - \mathcal{L}(T_{n+m})\big\|_{\mathrm{TV}} &= \big\|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\big\|_{\mathrm{TV}} \\ &= \big\|\mathcal{N}(\theta, \Sigma/n) - \mathcal{N}\big(\theta, \Sigma/(n+m)\big)\big\|_{\mathrm{TV}} \\ &\overset{\text{(a)}}{\le} \sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(\mathcal{N}(\theta, \Sigma/n)\|\mathcal{N}\big(\theta, \Sigma/(n+m)\big)\big)} \\ &= \sqrt{\frac{d}{4}\left(\frac{m}{n} - \log\left(1 + \frac{m}{n}\right)\right)} = O\left(\frac{m\sqrt{d}}{n}\right),\end{aligned}$$

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 [3] 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, [3] 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*, that is, $\mathcal{L}(S)$ is independent of the model parameter θ;
2. There is a (measurable) bijection g between (T, S) and X, that is, $X = g(T, S)$ almost surely.

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

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

where (a) is due to the assumed bijection g between (T, S) and X, and (b) is due to a classical result of Basu [8, 9] 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 $S = S(X^{n+m}) = (S_1, \ldots, S_{n+m-1})$ 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 [10, 35] 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 [53]. 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

$$\mathrm{d}P_\theta(x) = \exp\big(\theta^\top T(x) - A(\theta)\big)\,\mathrm{d}\mu(x),$$

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 [25] 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 nonempty 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

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

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 kth 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 $u_1, \ldots, u_k$, $T[u_1; \ldots; u_k]$ 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 [41]. 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 $X_1, \ldots, X_n$ 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, that is, $\widehat{T}_{n+m} = T_n$. 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*

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

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:

$$\begin{aligned}&\big\|\mathcal{L}(T_n) - \mathcal{L}(T_{n+m})\big\|_{\mathrm{TV}}\\ &\quad\le \big\|\mathcal{L}(T_n) - \mathcal{L}(G_n)\big\|_{\mathrm{TV}} + \big\|\mathcal{L}(G_n) - \mathcal{L}(G_{n+m})\big\|_{\mathrm{TV}} + \big\|\mathcal{L}(T_{n+m}) - \mathcal{L}(G_{n+m})\big\|_{\mathrm{TV}}.\end{aligned}$$

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 [5, 6, 42, 47] and KL [7, 13] 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, that is, $P_\theta(dx) = \prod_{i=1}^d p_{\theta_i}(dx_i)$, 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* [12]. For any $k \geq 2$ and $n \in \mathbb{N}$, the signed measure of the Edgeworth expansion on $\mathbb{R}^d$ is

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

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*

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

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 nontrivial 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:

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

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*

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

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 nonincreasing 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*

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

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 [3] 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 $Y_1, \ldots, Y_m$ from $\widehat{P}_n$, and shuffle them with the original samples $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ be i.i.d. drawn from P, and $Y_1, \ldots, Y_m$ be i.i.d. drawn from Q independent of $(X_1, \ldots, X_n)$. Let $(Z_1, \ldots, Z_{n+m})$ be a uniformly random permutation of $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$, and P_{mix} be the distribution of the random mixture $(Z_1, \ldots, Z_{n+m})$. Then*

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

Algorithm 2 SAMPLE AMPLIFICATION VIA SHUFFLING: GENERAL MODEL

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

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

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

Algorithm 3 SAMPLE AMPLIFICATION VIA SHUFFLING: PRODUCT MODEL

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

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

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

Based on Lemma 5.8, the advantage of random shuffling is clear: if we simply append $Y_1, \dots, Y_m$ to the end of the original sequence $X_1, \dots, X_n$, 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 $Y_1, \dots, Y_m$ be independent of $X_1, \dots, X_n$, 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. Appendix A.3 lists several concrete examples where our general idea could be properly applied to provide tight and nonasymptotic 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 nonnegative 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*:

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

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

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*

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

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 [3], albeit seemingly different, is a special case of Lemma 6.1. Specifically, the authors of [3] 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

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

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 $L : \Theta \times \bigcup_{n\geq 1} \mathcal{X}^n \to [0, 1]$ with $L(\theta, X^n) = \mathbb{1}(X^n \notin A_n(\theta))$. Then the first condition ensures that $r_{\mathrm{B}}(\mathcal{P}, n+m, L, \nu) \leq 0.01$ for each prior ν, and the second condition (6.1) exactly states that $r_{\mathrm{B}}(\mathcal{P}, n, L, \mu) \geq 0.5$ 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 covariance $\Sigma \in \mathbb{R}^{d\times d}$, 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, that is, $a^\top T(x) = 0$ 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, for example, [36], 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 $(\theta_1, \ldots, \theta_d) \in \prod_{j=1}^d \Theta_j$. Suppose for each $j \in [d]$, there exist two points $\theta_{j,+}, \theta_{j,-} \in \Theta_j$ such that*

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

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

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 jth 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 Appendix 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 $\| p_{\theta_{j,+}}^{\otimes n} - p_{\theta_{j,-}}^{\otimes n} \|_{\mathrm{TV}}$, for example, $O(1/\sqrt{d})$ 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 $n, n+c\epsilon n, n+2c\epsilon n, \ldots, Cn$ 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:

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

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., [28]) 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: $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$, 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 $[0, 1]$, that is, 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, that is, $f(x) \geq c$ for all $x \in [0, 1]$. It is a classical result (see, e.g., [49]) 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*

$$m^\star(\mathcal{P}_c, n) \asymp n^{5/6} \quad \text{while } m^\star(\mathcal{P}, n) \lesssim n^{3/4}.$$

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.

Acknowledgments. Thank you to anonymous reviewers for helpful feedback on earlier drafts of this paper.

Funding. Shivam Garg conducted this research while affiliated with Stanford University and was supported by a Stanford Interdisciplinary Graduate Fellowship.

Yanjun Han was supported by a Simons-Berkeley research fellowship and the Norbert Wiener postdoctoral fellowship in statistics at MIT IDSS.

Vatsal Sharan was supported by NSF CAREER Award CCF-2239265 and an Amazon Research Award.

Gregory Valiant was supported by NSF Awards AF-2341890, CCF-1704417, CCF-1813049, UT Austin's Foundation of ML NSF AI Institute, and a Simons Foundation Investigator Award.

SUPPLEMENTARY MATERIAL

Supplement to "On the statistical complexity of sample amplification" (DOI: 10.1214/24-AOS2444SUPP; .pdf). We provide proofs of main theorems (Theorems 4.5, 4.6, 5.2, 5.5, 6.2, 6.3, 6.4, 6.5, 7.1, 7.2, 7.3) and lemmas (Lemmas 4.4, 5.8).

REFERENCES

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

[2] AXELROD, B.,GARG, S.,HAN, Y.,SHARAN, V.,VALIANT, G. (2024). Supplement to "On the statistical complexity of sample amplification." https://doi.org/10.1214/24-AOS2444SUPP

[3] 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.

[4] BAI, Y., KADAVATH, S., KUNDU, S., ASKELL, A., KERNION, J., JONES, A., CHEN, A., GOLDIE, A., MIRHOSEINI, A. et al. (2022). Constitutional AI: Harmlessness from AI feedback. Preprint. Available at arXiv:2212.08073.

[5] BALLY, V. and CARAMELLINO, L. (2014). On the distances between probability density functions. *Electron. J. Probab.* **19** no. 110, 33 pp. MR3296526 https://doi.org/10.1214/EJP.v19-3175

[6] BALLY, V. and CARAMELLINO, L. (2016). Asymptotic development for the CLT in total variation distance. *Bernoulli* **22** 2442–2485. MR3498034 https://doi.org/10.3150/15-BEJ734

[7] BARRON, A. R. (1986). Entropy and the central limit theorem. *Ann. Probab.* **14** 336–342. MR0815975

[8] BASU, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā* **15** 377–380. MR0074745 https://doi.org/10.1007/978-1-4419-5825-9_14

[9] BASU, D. (1958). On statistics independent of sufficient statistics. *Sankhyā* **20** 223–226. MR0105758 https://doi.org/10.1007/978-1-4419-5825-9_16

[10] BASU, D. (1959). The family of ancillary statistics. *Sankhyā* **21** 247–256. MR0110115 https://doi.org/10.1007/978-1-4419-5825-9_18

[11] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849 https://doi.org/10.1214/13-AOS1127

[12] BHATTACHARYA, R. N. and RAO, R. R. (2010). *Normal Approximation and Asymptotic Expansions*. *Classics in Applied Mathematics* **64**. SIAM, Philadelphia, PA. MR3396213 https://doi.org/10.1137/1.9780898719895.ch1

[13] BOBKOV, S. G., CHISTYAKOV, G. P. and GÖTZE, F. (2014). Berry–Esseen bounds in the entropic central limit theorem. *Probab. Theory Related Fields* **159** 435–478. MR3230000 https://doi.org/10.1007/s00440-013-0510-3

[14] 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.

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

[16] BROWN, L. D., CAI, T. T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538 https://doi.org/10.1214/aos/1028674838

[17] 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. *Ann. Statist.* **32** 2074–2097. MR2102503 https://doi.org/10.1214/009053604000000012

[18] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

[19] 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, Berlin.

[20] 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.

[21] 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. Preprint. Available at arXiv:2210.09509.

[22] 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. *J. Med. Imag. Radiat. Oncol.* **65** 545–563. https://doi.org/10.1111/1754-9485.13261

[23] 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.

[24] 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.

[25] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[26] 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.

[27] 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.

[28] HAN, Y., JIAO, J. and WEISSMAN, T. (2015). Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Trans. Inf. Theory* **61** 6343–6354. MR3418968 https://doi.org/10.1109/TIT.2015.2478816

[29] 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*.

[30] KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**.

[31] LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Stat.* **35** 1419–1455. MR0207093 https://doi.org/10.1214/aoms/1177700372

[32] 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.

[33] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7

[34] LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*: *Some Basic Concepts. Springer Series in Statistics*. Springer, New York. MR1066869 https://doi.org/10.1007/978-1-4684-0377-0

[35] LEHMANN, E. L. and SCHOLZ, F.-W. (1992). Ancillarity. In *Current Issues in Statistical Inference*: *Essays in Honor of D. Basu. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **17** 32–51. IMS, Hayward, CA. MR1194408 https://doi.org/10.1214/lnms/1215458837

[36] LIESE, F. and MIESCKE, K.-J. (2008). *Statistical Decision Theory*: *Estimation, Testing, and Selection. Springer Series in Statistics*. Springer, New York. MR2421720

[37] LU, Y., SHEN, M., WANG, H., WANG, X., VAN RECHEM, C. and WEI, W. (2023). Machine learning for synthetic data generation: A review. Preprint. Available at arXiv:2302.04062.

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

[39] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698 https://doi.org/10.1214/14-AOS1300

[40] 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.

[41] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

[42] PROKHOROV, Y. V. (1952). A local theorem for densities. *Dokl. Akad. Nauk SSSR* **83** 797–800. MR0049501

[43] RAY, K. and SCHMIDT-HIEBER, J. (2019). Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. *Ann. Inst. Henri Poincaré Probab. Stat.* **55** 2195–2208. MR4029152 https://doi.org/10.1214/18-AIHP946

[44] 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. *Nat. Sci. Rep.* **9** 1–9.

[45] SHORTEN, C. and KHOSHGOFTAAR, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* **6** 1–48.

[46] 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.

[47] SIRAZHDINOV, S. K. and MAMATOV, M. (1962). On convergence in the mean for densities. *Theory Probab. Appl*. **7** 424–428.

[48] 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.

[49] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[50] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[51] 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.

[52] WALD, A. (1950). *Statistical Decision Functions*. Wiley, New York. MR0036976

[53] WILLIAMS, E. J. (1982). Some classes of conditional inference procedures. *J. Appl. Probab.* **19** 293–303. MR0633198

[54] 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.

[55] YI, X., WALIA, E. and BABYN, P. (2019). Generative adversarial network in medical imaging: A review. *Med. Image Anal*. **58** 101552. https://doi.org/10.1016/j.media.2019.101552

[56] 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.

[57] ZHANG, H., CISSE, M., DAUPHIN, Y. N. and LOPEZ-PAZ, D. (2018). Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.