Robustness Implies Generalization via Data-Dependent Generalization Bounds

Kenji Kawaguchi [1] **Zhun Deng** [2] **Kyle Luh** [3] **Jiaoyang Huang** [4]

Abstract

This paper proves that robustness implies generalization via data-dependent generalization bounds. As a result, robustness and generalization are shown to be connected closely in a data-dependent manner. Our bounds improve previous bounds in two directions, to solve an open problem that has seen little development since 2010. The first is to reduce the dependence on the covering number. The second is to remove the dependence on the hypothesis space. We present several examples, including ones for lasso and deep learning, in which our bounds are provably preferable. The experiments on real-world data and theoretical models demonstrate near-exponential improvements in various situations. To achieve these improvements, we do not require additional assumptions on the unknown distribution; instead, we only incorporate an observable and computable property of the training samples. A key technical innovation is an improved concentration bound for multinomial random variables that is of independent interest beyond robustness and generalization.

1. Introduction

Robust optimization (Ben-Tal & Nemirovski, 1998; Bertsimas et al., 2011; Gabrel et al., 2014) is an influential paradigm for dealing with noisy or uncertain data. Many optimization problems are sensitive to perturbations in their parameters. Using powerful concepts derived from convexity and duality, robust optimization aims to find a solution to these optimization problems that is feasible with respect to all possible realizations of noisy or unknown parameters. Essentially, this criterion solves the optimization problem for the worst-case choice of the possible parameters. Robust optimization has been successfully applied in a variety of fields, e.g., machine learning, to deal with inaccurately observed training samples and strengthen the robustness of deep learning (Bhattacharyya et al., 2004; Globerson & Roweis, 2006; Deng et al., 2021b; Rice et al., 2021; Robey et al., 2021; Pedraza et al., 2022; Chen et al., 2022).

Inspired by robust optimization, Xu & Mannor (2010; 2012) showed that robust algorithms generalize to unseen data well for various models including deep neural networks. Thus, the notion of robustness provides an alternative view in the topic of generalization (Vapnik, 1998; Bartlett & Mendelson, 2002; Bousquet & Elisseeff, 2002; Kawaguchi et al., 2017; Arora et al., 2019; Kawaguchi & Huang, 2019; Deng et al., 2021a; Hu et al., 2021; Pham et al., 2021; Zhang et al., 2021a;b).

A learning algorithm $\mathcal{A}$ is said to be robust if the loss ℓ of the hypothesis $\mathcal{A}_S$ (returned by the learning algorithm $\mathcal{A}$ under the training dataset S) behaves similarly on two samples that are near each other:

Definition 1. A learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-*robust*, for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z}^n \to \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{\mathcal{C}_k\}_{k=1}^K$, such that the following holds for all $S \in \mathcal{Z}^n$: $\forall s \in S, \forall z \in \mathcal{Z}, \forall k \in [K]$, if $s, z \in \mathcal{C}_k$, then $|\ell(\mathcal{A}_S, s) - \ell(\mathcal{A}_S, z)| \leq \epsilon(S)$.

Here, a training dataset $S = (z_i)_{i=1}^n$ consists of n samples and $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ is the per-sample loss, where $\mathcal{H}$ is a hypothesis space and $z_i \in \mathcal{Z}$ is the i-th training data point. That is, a learning algorithm $\mathcal{A}$ is a mapping from $S \in \mathcal{Z}^n$ to $\mathcal{A}_S \in \mathcal{H}$.

Using Definition 1, Xu & Mannor (2010; 2012) proved that the generalization error of hypothesis $\mathcal{A}_S$ has an upper bound that scales proportionally to $\epsilon(S) + \sqrt{K/n}$. This bound is consequential in the theory of invariant classifiers (Sokolic et al., 2017a), adversarial examples (Cisse et al., 2017), majority voting (Xu & Mannor, 2012; Devroye et al., 2013), support vector machines (Xu & Mannor, 2012; Xu et al., 2009; Qi et al., 2013), lasso (Xu & Mannor, 2012; Hastie et al., 2019), principle component analysis (Xu & Mannor, 2012; Jolliffe & Cadima, 2016), deep neural networks (Xu & Mannor, 2012; Sokolic et al., 2017b; Cisse et al., 2017; Sener & Savarese, 2017; Gouk et al., 2021; Jia et al., 2019), metric learning (Bellet & Habrard, 2015; Shi et al., 2014), facial recognition (Ding et al., 2015; Tao et al., 2016), matrix completion (Luo et al., 2015), spectral clustering (Liu et al., 2017), domain adaption (Redko et al.,

[1] National University of Singapore [2] Harvard University [3] University of Colorado Boulder [4] New York University. Correspondence to: Kenji Kawaguchi <kenji@nus.edu.sg>.

Proceedings of the 39^{th} *International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

2020), numerical analysis (Shen et al., 2020) and stochastic algorithms (Zahavy et al., 2016).

The bound based on algorithmic robustness (Definition 1) has gained considerable interest in the community and has been discussed in much literature as listed above, partially because the dependence on the robustness term $\epsilon(S)$ is natural and intuitive. However, the square root dependence on the partition size (or covering number) K is problematic, because K can be prohibitively large in many applications, especially in high-dimensional data where the covering number can be exponential in the dimension (Van Der Vaart et al., 1996; Vershynin, 2018).

Indeed, the K dependence is one of the chief disadvantages of the robust algorithm framework and Xu & Mannor (2010; 2012) conjectured that it would be possible to reduce the dependency on K via adaptive partitioning but remarked that extending their proof to achieve this is complex, leaving this issue as an open problem.

The proof of the algorithmic robustness bound relies on the concentration results for multinomial random variables, in particular the ℓ_1 deviations (Xu & Mannor, 2012; Wellner et al., 2013). Spurred by applications in learning theory, the concentration of multinomial random variables has been an active area of research by itself beyond algorithmic robustness (Weissman et al., 2003; Devroye, 1983; Agrawal & Jia, 2017; Qian et al., 2020), where a particular attention has been paid to the dependence of the bound on K — the number of possible outcomes in the definition of the multinomial random variable. In the robust algorithmic context, K corresponds exactly to K in Definition 1. In a paper previously published in NeurIPS (Agrawal & Jia, 2017), a significant improvement in the $\sqrt{K}$ dependence was claimed which was later refuted by Qian et al. (2020) with the refutation being acknowledged by the authors (Agrawal & Jia, 2020). Thus, to date, there has been no success in reducing the $\sqrt{K}$ term reported in the literature despite its importance and several previous attempts.

Importantly, Qian et al. (2020) established a lower bound that already scales as $\sqrt{K}$; that is, we have matching upper and lower bounds in terms of K. Thus, it may seem that the open question posed in (Xu & Mannor, 2012) has been settled negatively and any attempts to reduce the $\sqrt{K}$ dependence is futile. However, similar to other lower bounds in machine learning, the lower bound given in (Qian et al., 2020) only means that there exists a worst-case distribution for which the (lower) bound cannot be further improved.

It is plausible that this worst-case distribution is neither representative nor commonplace. Thus, by incorporating information from the training data, it may be possible to extract the properties of the underlying distribution, which may allow one to reduce the $\sqrt{K}$ dependence. In fact, by probing beyond the worst-case analysis, we show that *non-uniform* and *purely data-dependent* bounds can greatly outperform these previous bounds (that are implicitly derived for the worst-case distributions). Here, a bound is said to be *non-uniform* if the bound differs for different data-distributions. Unlike the standard data-dependence through the outcome of the learning algorithm A_S (e.g., in the robustness term $\epsilon(S)$), a bound is said to be *purely data-dependent* if the bound contains a term that is independent of the algorithm $\mathcal{A}$ and differs for different training data S. We summarize our main contributions below:

1. In Section 3, we address the open problem of reducing the $\sqrt{K}$ dependence without making any additional assumptions about the data distribution. The key insight (and challenge) here is to prove an *purely data-dependent* bound where the $\sqrt{K}$ dependence is replaced by an easily computable quantity that is a function of the training samples. This allows us to reduce the $\sqrt{K}$ dependence without presuming strong prior knowledge of the probability space and the learning algorithm.
2. A crucial technical innovation is a series of *non-uniform* and *purely data-dependent* bounds for multinomial random variables that greatly improve the classical bounds under certain conditions. A representative of our new bounds is stated in Section 3 (and others are presented in Appendices A and B). These bounds are likely of independent interest in the broader literature beyond robustness and generalization.
3. In addition to our main theorems, we provide abundant numerical simulations and several theoretical examples in which our bounds are *provably* superior in Sections 4 and 5. As a consequence of our improvements to algorithmic robustness, we can deduce immediate improvements to the many applications of algorithmic robustness listed above, ranging from invariant classifiers to numerical analysis and stochastic learning algorithms.

2. Preliminaries

This section introduces notation and previous results.

2.1. Notation

For an integer n, we use $[n]$ to denote the set of integers $1, \dots, n$. For a finite set $\mathcal{B}$, we let $|\mathcal{B}|$ represent the number of elements in $\mathcal{B}$. For a set $\mathcal{S}$ equipped with metric ρ, we define $\hat{\mathcal{S}}$ as an ε-cover of $\mathcal{S}$ if for all $s \in \mathcal{S}$, there exists $\hat{s} \in \hat{\mathcal{S}}$ such that $\rho(s, \hat{s}) \leq \varepsilon$. We then define the ε-covering number as

$$\mathcal{N}(\varepsilon, \mathcal{S}, \rho) = \min\{|\hat{\mathcal{S}}| : \hat{\mathcal{S}} \text{ is an } \varepsilon\text{-cover of } \mathcal{S}\}.$$

We use $\mathbb{1}(\cdot)$ as an indicator function, and $\|\cdot\|_p$ is the standard p-norm for a vector.

2.2. Problem Setting and Background

In this study, we are interested in bounding the expected loss $\mathbb{E}_z[\ell(\mathcal{A}_S, z)]$, where $\mathbb{E}_z$ denotes the expectation with respect to the sampling distribution. This is a quantity that cannot be computed or accessed. Accordingly, we obtain an upper bound by using the training loss $\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i)$, which is a computable quantity, and by invoking other computable terms. A previous study (Xu & Mannor, 2012) used algorithmic robustness (Definition 1) to achieve the following result:

Proposition 1. *(Xu & Mannor, 2012) Assume that for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$, the loss is upper bounded by B i.e., $\ell(h, z) \leq B$. If the learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{\mathcal{C}_k\}_{k=1}^{K}$), then for any $\delta > 0$, with probability at least $1 - \delta$ over an iid draw of n samples $S = (z_i)_{i=1}^{n}$, the following holds:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \qquad (1)$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i) + \epsilon(S) + B\sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}}.$$

For example, in the special case of supervised learning, the sample space can be decomposed as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. However, note that $\mathcal{Z}$ can differ from the original space of the data points. For example, if the original data point is $\tilde{z}$, we can use $z = g(\tilde{z})$ for any fixed-function g.

The previous paper (Xu & Mannor, 2012) also proves the same upper bound on $\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i) - \mathbb{E}_z[\ell(\mathcal{A}_S, z)]$, instead of $\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i)$. However, the empirical loss $\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i)$ can be minimized during training; hence, we are typically interested in the upper bound on $\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i)$. The focus on this quantity.

The relationship between algorithmic robustness and the multinomial distribution is apparent when we consider independent samples from the sample space of $\{\mathcal{C}_k\}_{k=1}^{K}$. Then, the number of samples from each class, $\mathcal{C}_k$, is multinomially distributed with $p_k = \mathbb{P}(z \in \mathcal{C}_k)$. The actual values of p_k are not available to us. Therefore, it is natural that the concentration of the multinomial values around these expectations is required in the argument.

The concentration of a multinomial random variable is of interest in theoretical probability and practical use in applied fields such as statistics and computer science (Van Der Vaart et al., 1996). Consequently, several concentration bounds have been proposed in the literature (Weissman et al., 2003; Devroye, 1983; Agrawal & Jia, 2017; Qian et al., 2020; Van Der Vaart et al., 1996), for example:

Proposition 2 (Bretagnolle-Huber-Carol inequality)**.** *(Van Der Vaart et al., 1996, Proposition A.6.6) If $X_1, \ldots, X_K$ are multinomially distributed with parameters n and $p_1, \ldots, p_K$, then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sum_{k=1}^{K}\left|p_k - \frac{X_k}{n}\right| \leq \sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}} \qquad (2)$$

Crucially, the bounds in the literature are uniform in the parameters p_k, meaning that the right-hand side of the inequality is true for any set of parameters. A key step in our reduction of the $\sqrt{K}$ dependence in algorithmic robustness is the non-uniform (and purely data-dependent) enhancement of the above concentration bound, which may be of independent interest beyond algorithmic robustness.

3. Main Theorems

In this section, we record our improvements to Proposition 1 along with our upgraded bounds for the multinomial distribution. We discuss our main contributions and relegate the complete proofs of theoretical results to the appendices.

3.1. Algorithmic Robustness

One of the main contributions of this study is the following refinement of the algorithmic robustness bound:

Theorem 1. *If the learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{\mathcal{C}_k\}_{k=1}^{K}$), then for any $\delta > 0$, with probability at least $1 - \delta$ over an iid draw of n samples $S = (z_i)_{i=1}^{n}$, the following holds:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S, z_i) + \epsilon(S) \qquad (3)$$
$$+ \zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right),$$

where $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$,

$$\zeta(\mathcal{A}_S) := \max_{z \in \mathcal{Z}}\{\ell(\mathcal{A}_S, z)\}, \text{ and}$$
$$\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}.$$

Theorem 1 is a significant improvement over the previous bound (1) of Proposition 1 as (3) has a far better dependence on K. In terms of K, we have reduced $\sqrt{K}$ to $\sqrt{\ln K}$. Overall, if we ignore the log factor, K in the previous bound is replaced by $|\mathcal{T}_S|$ in our bound. Here, $|\mathcal{T}_S|$ is the number of distinct classes, $\mathcal{C}_k$, *that actually appear in the single specific training dataset* S; thus, $|\mathcal{T}_S| \leq K$ by the definition and it is shown to have $|\mathcal{T}_S| \ll K$ in many general cases later in Proposition 3 and Sections 4.2 and 5. For example, our experimental results in Section 5 indicate that $|\mathcal{T}_S| \ll K$ in many natural settings, where we see an exponential discrepancy in a variety of real-world data sets and theoretical models.

Intuitively, $|\mathcal{T}_S|$ is likely to be significantly less than K when there are many sparsely populated classes $\mathcal{C}_k$. Therefore, it is improbable that many of these classes are represented in the sample data. Our theoretical and experimental and results demonstrate that such scenarios are prevalent in the field.

The following proposition shows that $|\mathcal{T}_S|$ is indeed independent of K and only scales as $\ln n$ under a general mild condition on p_k, proving that we have $|\mathcal{T}_S| \ll K$ and $|\mathcal{T}_S| \ll n$ in a general case:

Proposition 3. *Under the assumptions of Theorem 1, we denote $p_k = \mathbb{P}(z \in \mathcal{C}_k)$ where $p_1 \geq p_2 \geq \cdots \geq p_K$. If there are some constants $\alpha, \beta, C > 0$ such that p_k decays as $p_k \leq Ce^{-(k/\beta)^\alpha}$, then with probability at least $1-\delta$,*

$$|\mathcal{T}_S| \leq \beta(\ln n)^{1/\alpha} + C(e-1)\frac{\beta}{\alpha} + \log(1/\delta) \tag{4}$$

In Proposition 3, α controls the rate of decay for p_k. For real-world data sets, we expect the data distribution concentrates on a lower dimension manifold or around small number of modes. In such settings, we expect the probability $p_k = \mathbb{P}(z \in \mathcal{C}_k)$ (arranging them in decaying order) exhibits fast decays. If $\alpha = \infty$, p_k concentrates on *unknown* β bins and we have $|\mathcal{T}_S| \leq \beta$. If $\alpha < \infty$, we have $p_k \neq 0$ for all $k \in [K]$, but $|\mathcal{T}_S|$ is still upper bound by β times the constant up to a logarithmic factor and is independent of K.

Proposition 3 also demonstrates the fact that even with perfect prior knowledge of the data distribution, $|\mathcal{T}_S|$ can be much smaller than K because $|\mathcal{T}_S|$ is more adaptive according to the training data while K cells need to cover all the possible parts that the distribution has positive mass on. Without the perfect knowledge, $|\mathcal{T}_S|$ can be more significantly smaller than K.

A crucial aspect of Theorem 1 is that $\mathcal{T}_S$ depends exclusively on the training sample data and not the actual background distribution. Accordingly, our result is of practical value in statistical learning settings, where information about the actual distribution can only be obtained through the training sample data.

Although the main breakthrough in Theorem 1 is the reduced K dependence, there is also a substantially refined dependence on the upper bound of the loss value — the replacement of B with $\zeta(\mathcal{A}_S)$, where $\zeta(\mathcal{A}_S) \leq B$; i.e., we replace a maximum over the entire hypothesis space with the single hypothesis returned by the algorithm. This can be a significant advantage for common loss functions, such as square loss and cross-entropy loss. Note that B in the previous bound is defined to be larger than $\ell(h, z)$ for all $h \in \mathcal{H}$, meaning that B is dependent on *the entire hypothesis space $\mathcal{H}$*. In contrast, $\zeta(\mathcal{A}_S)$ in our bound depends only on the single actual hypothesis, $\mathcal{A}_S$, returned by the specific algorithm for each data set S.

With a more refined analysis, we also prove a stronger (yet more complicated) version of Theorem 1:

Theorem 2. *If the learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{\mathcal{C}_k\}_{k=1}^K$), then for any $\delta > 0$, with probability at least $1-\delta$ over an iid draw of n examples $S = (z_i)_{i=1}^n$, the following holds:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \leq \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{A}_S, z_i) + \epsilon(S) + \mathcal{Q}_1\sqrt{\frac{\ln(2K/\delta)}{n}} + \frac{2\mathcal{Q}_2\ln(2K/\delta)}{n} \tag{5}$$

where

$$\mathcal{Q}_1 := \sum_{k \in \mathcal{T}_S}\left(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) + \sqrt{2}\alpha_k(\mathcal{A}_S)\right)\sqrt{\frac{|\mathcal{I}_k^S|}{n}},$$

$$\mathcal{Q}_2 := \alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)|\mathcal{T}_S| + \sum_{k \in \mathcal{T}_S}\alpha_k(\mathcal{A}_S),$$

with $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$, $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$, $\alpha_k(h) := \mathbb{E}_z[\ell(h, z)|z \in \mathcal{C}_k]$, $\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) := \max_{k \in \mathcal{T}_S^c}\alpha_k(\mathcal{A}_S)$, and $\mathcal{T}_S^c := [K] \setminus \mathcal{T}_S$.

Note that $\sum_{k \in \mathcal{T}_S}\alpha_k(\mathcal{A}_S) \leq |\mathcal{T}_S|\zeta(\mathcal{A}_S)$ and $\sum_{k \in \mathcal{T}_S}\sqrt{|\mathcal{I}_k^S|/n} \leq \sqrt{|\mathcal{T}_S|}$ by the Cauchy-Schwarz inequality. Thus, Theorem 2 is always as strong as Theorem 1. Furthermore, Theorem 2 significantly upgrades Theorem 1 approximately when $\sum_{k \in \mathcal{T}_S}\alpha_k(\mathcal{A}_S) \ll |\mathcal{T}_S|\zeta(\mathcal{A}_S)$ or $\sum_{k \in \mathcal{T}_S}\sqrt{|\mathcal{I}_k^S|/n} \ll \sqrt{|\mathcal{T}_S|}$. Otherwise put, if the maximum expected loss of the classes is much larger than the typical expected loss or the distribution of samples in the classes is lopsided, Theorem 2 will be an even tighter bound.

The complete proofs of Theorems 1 and 2 are provided in Sections C and E of the Appendix. We remark that our proof of Theorem 1 proves a stronger theoretical statement, where $\zeta(\mathcal{A}_S)$ is replaced by $\max_{k \in [K]}\mathbb{E}_z[\ell(\mathcal{A}_S, z)|z \in \mathcal{C}_k]$ ($\leq \zeta(\mathcal{A}_S)$). This formulation may have advantages over that in Theorem 1 if the problem context reveals more information about the conditional expectation.

3.2. Concentration Bounds for the Multinomial Distribution

Let the vector $X = (X_1, \ldots, X_K)$ follow a multinomial distribution with parameters n and $p = (p_1, \ldots, p_K)$. As shown in the proof of Theorem 1, the key technical hurdle is to avoid an explicit $\sqrt{K}$ dependence as we upper bound a quantity of the following form:

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right), \tag{6}$$

where a_i is an arbitrary function with $a_i(X) \geq 0$ for all $i \in \{1, \dots, K\}$.

Importantly, $a_i(X)$ are functions of $X_1, \dots, X_K$, which makes this problem particularly challenging and further complicates the non-trivial correlations already present in X_i. This difficulty is avoided in (Qian et al., 2020; Bellet & Habrard, 2015; Xu & Mannor, 2012) by using the global maximum of the loss function with the $\sqrt{K}$ dependence. Allowing $a_i(X)$ to depend on X is critical in our analysis and underpins the improvement from B in Proposition 1 to $\zeta(\mathcal{A}_S)$ in Theorem 1, in addition to the improvement from $\sqrt{K}$ to $\sqrt{\ln K}$.

One example of our new multinomial bounds that are non-uniform is the following lemma.

Lemma 1. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \leq \left(\sum_{i=1}^{K} a_i(X)\sqrt{p_i}\right)\sqrt{\frac{2\ln(K/\delta)}{n}}.$$

Since a_i is arbitrary, Lemma 1 holds, for example, with $a_i(X) = \operatorname{sign}(p_i - \frac{X_i}{n})$, where $\operatorname{sign}(q)$ is the sign of q. In this case, the left-hand side in Lemma 1 becomes that in Proposition 2. By comparing such a special case of Lemma 1 to Proposition 2, in some range of parameters, we have essentially replaced $\sqrt{K}$ with $\sum_i \sqrt{p_i}$. If $p_i = 1/K$ for all i, then we recover (2). Conversely, in the other extreme case, if $p_1 \approx 1$ and the remaining p_i are near zero, then $\sum \sqrt{p_i} \approx 1$. Thus, our result interpolates between these cases, and there is a wide range of possible distributions in which our bound is significantly better than Proposition 2.

While Lemma 1 is of independent interest, it is likely difficult to be used directly in machine learning because it depends on the probability distribution p, which is typically unknown. To overcome this issue, we further refine Lemma 1 and remove the dependence on p. To keep the notation consistent, we write $p_k = \mathbb{P}(z \in \mathcal{C}_k)$ and $X_k = \sum_{i=1}^{n} \mathbb{1}\{z_i \in \mathcal{C}_k\}$ with $\{\mathcal{C}_k\}_{k=1}^{K}$ being arbitrary in this subsection. Since $\{\mathcal{C}_k\}_{k=1}^{K}$ is arbitrary here, this still represents general multinomial distributions.

One of the main contributions of this study is the following refinement of the concentration bound on general multinomial distributions:

Theorem 3. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \tag{7}$$
$$\leq \tilde{\mathcal{Q}}\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)\frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n},$$

where $\tilde{\mathcal{Q}} = a_{\mathcal{T}_S}(X)\sqrt{2} + a_{\mathcal{T}_S^c}(X)$, $a_{\mathcal{T}_S}(X) = \max_{i \in \mathcal{T}_S} a_i(X)$ and $a_{\mathcal{T}_S^c}(X) = \max_{i \in \mathcal{T}_S^c} a_i(X)$ with $\mathcal{T}_S^c = [K] \setminus \mathcal{T}_S$.

Further results of this nature and their technical proofs can be found in Sections A and B of the Appendix.

3.3. Discussion and Extensions

3.3.1. Proof Ideas and Challenges

The proof of Theorem 1 can be divided into three phases. In the first, we prove (a stronger version of) Lemma 1. Next, we invoke Lemma 1 to prove Theorem 3. Finally, we deduce Theorem 1 from Theorem 3. Thus, the first obstacle is to establish Lemma 1, which supplants Proposition 2. After this key step, the next challenge lies in going from Lemma 1 to Theorem 3, which requires us to remove the $\sum_i \sqrt{p_i}$ term in Lemma 1 without incurring the $\sqrt{K}$ dependence. For example, if we naively bound $\sum_i \sqrt{p_i}$ with the Cauchy–Schwarz inequality, we have that

$$\sum_{i=1}^{K}\sqrt{p_i} \leq \sqrt{\sum_{i=1}^{K} p_i}\sqrt{\sum_{i=1}^{K} 1} = \sqrt{K}.$$

Similarly, if we apply Jensen's inequality, which relies on the concavity of the square root function in this domain, we find that

$$\frac{1}{K}\sum_{i=1}^{K}\sqrt{p_i} \leq \sqrt{\frac{1}{K}\sum_{i=1}^{K} p_i} = \sqrt{\frac{1}{K}},$$

which again implies that $\sum_{i=1}^{K}\sqrt{p_i} \leq \sqrt{K}$. Thus, both approaches reproduce the $\sqrt{K}$ dependence, illustrating the challenges of removing the $\sum_i \sqrt{p_i}$ term without the $\sqrt{K}$ dependence. Our novel observation is to first decompose the sum as

$$\sum_{i=1}^{K}\left(p_i - \frac{X_i}{n}\right) = \sum_{i \in \mathcal{T}_S}\left(p_i - \frac{X_i}{n}\right) + \sum_{i \notin \mathcal{T}_S} p_i \tag{8}$$
$$= \sum_{i \in \mathcal{T}_S}\left(p_i - \frac{X_i}{n}\right) + \left(1 - \sum_{i \in \mathcal{T}_S} p_i\right),$$

and find an upper bound on the second term $1 - \sum_{i \in \mathcal{T}_S} p_i$ by using a *lower* bound on $\sum_{i \in \mathcal{T}_S}\left(p_i - \frac{X_i}{n}\right)$. That is, if $\sum_{i \in \mathcal{T}_S}\left(p_i - \frac{X_i}{n}\right) \geq -c$ for some $c > 0$, then $1 -$

$\sum_{i\in\mathcal{T}_S} p_i \le 1 + c - \sum_{i\in\mathcal{T}_S}\frac{X_i}{n} = c$. The second line of (8) is a conceptually crucial step where we convert the question of *upper* bounding $\sum_{i\notin\mathcal{T}_S} p_i$ to that of *lower* bounding $\sum_{i\in\mathcal{T}_S} p_i$. If we directly upper bound $\sum_{i\notin\mathcal{T}_S} p_i$, it will sustain a cost of $\sqrt{K - |\mathcal{T}_S|}$, resulting in a $\sqrt{K}$ dependence. This step of decomposing the sum and of converting the upper bound to a lower bound is designed to avoid the $\sqrt{K}$ dependence. Now, the problem has been reduced to finding tight lower and upper bounds on these quantities, which is still nontrivial and is described in Appendices A and B.

3.3.2. Pseudo-Robustness

Xu & Mannor (2012) also introduced a more general notion of *pseudo-robustness* which relaxes algorithmic robustness by only requiring the nearness of the loss functions to hold for a subset of the training samples:

Definition 2. A learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ *pseudo-robust* , for $K \in \mathbb{N}$, $\epsilon(\cdot) : \mathcal{Z}^n \to \mathbb{R}$, and $\hat{n}(\cdot) : \mathcal{Z}^n \to \{1, \dots, n\}$, where $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{\mathcal{C}_k\}_{k=1}^K$, such that for all $S \in \mathcal{Z}^n$, there exists a subset of training samples $\hat{S}$ with $|\hat{S}| = \hat{n}(S)$ and the following holds: $\forall s \in \hat{S}, \forall z \in \mathcal{Z}, \forall k = 1, \dots, K$: if $s, z \in \mathcal{C}_k$, then $|\ell(\mathcal{A}_S, s) - \ell(\mathcal{A}_S, z)| \le \epsilon(S)$.

For pseudo-robustness, we have proved the following analog of Theorem 1:

Theorem 4. *If the learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ is pseudo-robust (with $\{\mathcal{C}_k\}_{k=1}^K$), then for any $\delta > 0$, with probability at least $1 - \delta$ over an iid draw of n examples $S = (z_i)_{i=1}^n$, the following holds:*

$$\begin{aligned}\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \le \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{A}_S, z_i) + \frac{\hat{n}(S)}{n}\epsilon(S) \\ + \frac{n - \hat{n}(S)}{n}\hat{\zeta}(\mathcal{A}, S) \\ + \zeta(\mathcal{A}_S)\Bigg((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} \\ + \frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\Bigg),\end{aligned}$$

where $\hat{\zeta}(\mathcal{A}, S) := \max_{(k,i)\in[K]\times[n]} |\alpha_k(\mathcal{A}_S) - \ell(\mathcal{A}_S, z_i)|$,

$$\zeta(\mathcal{A}_S) := \max_{k\in[K]} \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k]$$

and

$$\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \ge 1\}$$

with $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$.

Theorem 2 exhibits an analogous generalization. A precise statement and proof can be found in Appendix F. These theorems offer concomitant improvements to the pseudo-robustness bounds attained in (Xu & Mannor, 2012).

4. Examples

Our contributions augment the abstract framework of algorithmic robustness. We do not use application-dependent information nor do we append any restrictive assumptions, and, therefore, can deduce improvements to the many applications that employ the structure of algorithmic robustness. After presenting known examples in Section 4.1, we provide new theoretical comparisons via examples in Section 4.2.

4.1. Robust Algorithms

Although our Theorems 1 and 2 are applicable to a wide range of applications, this section provides only a few simple examples from (Xu & Mannor, 2012) to which our Theorems 1 and 2 can be applied. When we have the decomposition $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$ as the input space and $\mathcal{Y}$ as the output space, we let $z^{(x)} \in \mathcal{X}$ and $z^{(y)} \in \mathcal{Y}$ denote the $\mathcal{X}$ and $\mathcal{Y}$ components of a sample point z, respectively, with $z = (z^{(x)}, z^{(y)})$. We also write $S = (s_1, \dots, s_n)$.

Example 1. (Xu & Mannor, 2012) (Lipschitz continuous functions) Broad classes of learning problems are set in spaces with natural metrics. If the loss function is Lipschitz, which is a simple and natural condition, then the algorithm is robust. More precisely, if $\mathcal{Z}$ is compact with regarding a metric ρ and $\ell(\mathcal{A}_S, \cdot)$ is Lipschitz continuous with Lipschitz constant $c(S)$, that is,

$$|\ell(\mathcal{A}_S, z) - \ell(\mathcal{A}_S, z')| \le c(S)\rho(z, z'), \quad \forall z, z' \in \mathcal{Z},$$

then $\mathcal{A}$ is $(\mathcal{N}(\gamma/2, \mathcal{Z}, \rho), c(S)\gamma)$-robust for all $\gamma > 0$.

Example 2. (Xu & Mannor, 2012) (Lasso) Lasso is a workhorse of modern machine learning (Hastie et al., 2019). We assume that $\mathcal{Z}$ is compact, and we use the loss function $\ell(\mathcal{A}_S, z) = |z^{(y)} - \mathcal{A}_S(z^{(x)})|$. Lasso can be formulated as an optimization problem.

$$\underset{w}{\text{minimize}} : \frac{1}{n}\sum_{i=1}^n (s_i^{(y)} - w^\top s_i^{(x)})^2 + c\|w\|_1.$$

This algorithm is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^n (s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.

Example 3. (Xu & Mannor, 2012) (Principal Component Analysis) For $\mathcal{Z} \subset \mathbb{R}^m$, a set with the maximum ℓ_2 norm bounded by B. If we use the loss function

$$\ell((w_1, \dots, w_d), z) = \sum_{j=1}^d (w_j^\top z)^2,$$

then finding the first d principal components via the optimization problem:

$$\text{Maximize: } \sum_{i=1}^n \sum_{j=1}^d (w_j^\top s_i)^2$$

with the constraint that $\|w_j\|_2 = 1$ and $w_i^\top w_j = 0$ for $i \ne j$ is $(\mathcal{N}(\gamma/2, \mathcal{Z}, \|\cdot\|_2), 2d\gamma B)$-robust, for all $\gamma > 0$.

Theorems 1 and 2 can be used as a black box mathematical tool for many more of the existing applications cataloged in the introduction.

4.2. Theoretical Comparisons

Here, we further demonstrate the advantage of using our bounds in Section 3 over the bounds provided by Xu & Mannor (2012) and the bounds obtained via uniform stability (Bousquet & Elisseeff, 2002), using Lasso, least square regression, neural networks, and discrete-valued neural communication as examples.

The first example demonstrates that when the data are embedded with high probability on a low-dimensional manifold in the data space, and our bound is much stronger than that of (Xu & Mannor, 2012):

Example 4 (Lasso)**.** Recall that in Example 2, Lasso is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^n (s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$. Consider $z^{(y)} \in \mathbb{R}$ and $z^{(x)} \in \mathbb{R}^d$. Given any $\nu > 0$, let z follow a distribution $\mathcal{D}_z$, such that $z^{(x)} = ({x^{(1)}}^\top, {x^{(2)}}^\top)^\top$, where $x^{(1)} \sim N(0, I_p)|_{[-1,1]^p}$ (truncated Gaussian on $[-1,1]^p$), $x^{(2)} \sim N(\mu, \sigma^2 I_r)|_{[-1,1]^r}$, and $r = d - p$, $z^{(y)} = {w^*}^\top z^{(x)}$, where $\|w^*\|_1 \leq 1$. For sufficiently small σ, we can check that the ν-covering of the data space $[-1,1]^d$ satisfies Proposition 3, with $\beta = (2/\nu)^p$ and $\alpha = 2$. As a consequence, we have that $|\mathcal{T}_S| = \Theta((2/\nu)^{p+1})$ with a probability of at least $1-\delta$. Since $\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty) = \Theta((2/\nu)^{d+1})$, there exists D, N such that for any $d > D$ and $n > N$, when $d \gg p$, there exists (μ, σ) such that the bound in Theorem 1 is much tighter than that in Proposition 1 as $|\mathcal{T}_S| = \Theta((2/\nu)^{p+1}) \ll \Theta((2/\nu)^{d+1}) = \mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty)$. See Appendix G for more details.

In the next example, we can see that our bound is much tighter than the bound obtained via uniform stability when there are outliers in the data:

Example 5 (Regularized least square regression)**.** We refer to the example in (Bousquet & Elisseeff, 2002) for regularized least squares regression. Specifically, $z^{(y)} \in [0, B]$ and $z^{(x)} \in [0,1]$. The regularized least squares regression is defined as $\mathcal{A}_S = \text{argmin}_w \frac{1}{n}\sum_{i=1}^n \ell(w, z_i) + \lambda|w|^2$, where $\ell(w,z) = (w \cdot z^{(x)} - z^{(y)})^2$ and $w \in \mathbb{R}$. For this example, Bousquet & Elisseeff (2002) observe that $0 \leq \ell(w,z) \leq \sqrt{B/\lambda}$, and establish the stability bound $\beta \leq \frac{2B^2}{\lambda n}$. Now, consider the following distribution on z: $z^{(y)} = w^* \cdot z^{(x)} + \epsilon \mathbf{1}(|\epsilon| < B)$. In addition, $z^{(x)}$ follows a continuous distribution on $[0,1]$, for instance, a uniform distribution on $[0,1]$, and $\epsilon \sim N(0, \sigma^2)$. Without loss of generality, let $w^* = 1$. In this example, we can possibly observe some large outlier with small probability. One can check that by suitably chosen σ, Proposition 3 holds with $\beta = 2/\nu$ and $\alpha = 2$. Thus, there exists a threshold N such that for any $n > N$, there exists $\nu > 0, \sigma > 0$ and $\delta > 0$, with a probability of at least $1-\delta$, $|\mathcal{T}_S| = \Theta(2\nu)$. Thus, if $B^2/\lambda \gg 2/\nu$, the bound in Theorem 1 is a far more precise bound than that obtained via uniform stability. For details of the proof, we refer the reader to Appendix G.

Although Example 5 compares our robustness bound and the uniform stability bound, we emphasize that robustness and stability are very different properties and have distinct strengths and weaknesses: i.e., one setting prefers the robustness framework and another setting favors the stability approach. For example, a learning algorithm may be robust but not stable, e.g., lasso regression (Xu et al., 2009), and vice versa. Accordingly, this paper focuses on the fundamental advancements of the robustness framework and of the statistical bound for general multinomial distributions.

Furthermore, the following examples illustrate some of the immediate improvements for robust margin deep neural networks and discrete-valued neural communication:

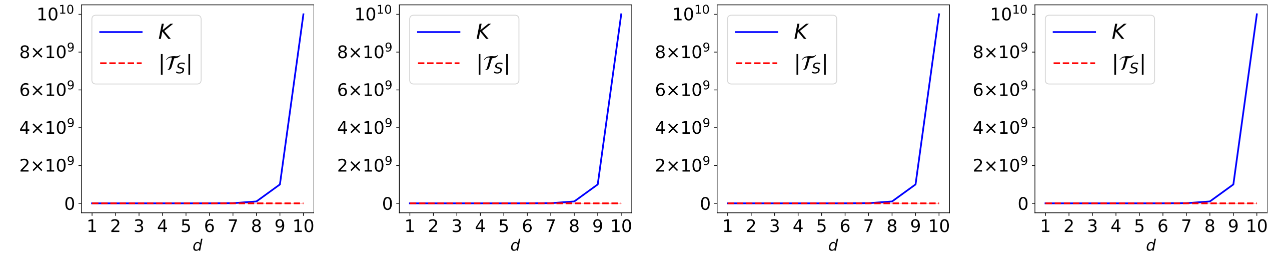
Example 6 (Robust margin deep neural networks)**.** The previous paper (Sokolic et al., 2017b) uses Proposition 1 (in our paper) with the ϵ-covering of $\mathcal{X}$ for robust margin neural networks. Our new theorem (Theorem 1 or Theorem 2) immediately improve their bounds by replacing $K = \frac{2^{k+1}(C_M)^k}{\gamma_b^k}$ in their bounds with $|\mathcal{T}_S|$. In our paper, the comparison of K v.s. $|\mathcal{T}_S|$ for the ϵ-covering of $\mathcal{X}$ is shown in Figure 3 for the real-life datasets. By plugging these values of K and $|\mathcal{T}_S|$ into the previous bounds and our versions, we yield exponential improvements over the previous bounds for robust margin deep neural networks.

Example 7 (Discrete-valued neural communication)**.** The bound in (previous) theorem 3 of the recent paper on *discrete-valued neural communication* (Liu et al., 2021) scales at the rate of L^G (which is the size of the discrete bottleneck). Since their proof uses Proposition 2 (in our paper) to bound the left-hand-side of (7) (in our Theorem 3), by applying our Theorem 3, we yield an improvement by replacing L^G with $|\mathcal{T}_S|$ for discrete-valued neural communication.

5. Experiments

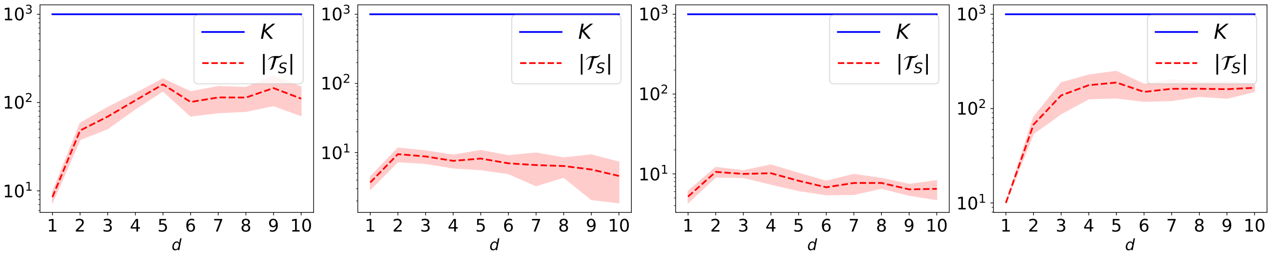
This section establishes the advantage of our new bounds via experiments using both synthetic data and real-world data. We generated synthetic data by sampling from beta distributions and Gaussian mixture distributions with a variety of hyperparameters. For real-world data, we adopted the standard benchmark datasets: MNIST (LeCun et al., 1998), CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), Fashion-MNIST (FM-NIST) (Xiao et al., 2017), Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2019), and Semeion (Srl & Brescia, 1994).

The value of $\epsilon(S)$ is exactly the same for the previous bound



(a) Beta(0.1, 0.1) (b) Beta(0.1, 10) (c) Gauss mix (0.01) (d) Gauss mix (1.0)

Figure 1. The values of K versus $|\mathcal{T}_S|$ with synthetic data and the ϵ-covering of the original space. The plot shows the mean of 10 random trials. The value of K increases exponentially as d increases linearly, whereas $|\mathcal{T}_S|$ does not.



(a) Beta(0.1, 0.1) (b) Beta(0.1, 10) (c) Gauss mix (0.01) (d) Gauss mix (1.0)

Figure 2. The values of K versus $|\mathcal{T}_S|$ with synthetic data and the inverse image of the ϵ-covering in randomly projected spaces. The plot shows the mean and one standard deviation of 10 random trials. We still have $|\mathcal{T}_S| < K$ with random projections to reduce K.

and our bound in all the experiments. To choose the partition $\{\mathcal{C}_k\}_{k=1}^K$, in addition to other examples, we used the ϵ-covering of the original input space $\mathcal{X}$ as our primary example, as this is the default option in (Xu & Mannor, 2012). The data space is normalized such that $\mathcal{X} \subseteq [0,1]^d$ for the dimensionality d of each input data. Accordingly, we used the infinity norm and a diameter of 0.1 for the ϵ-covering in all experiments. See Appendix H for more details on the experimental setup.

5.1. Synthetic data

Figure 1 shows the values of K and $|\mathcal{T}_S|$ for the synthetic data with the partition $\{\mathcal{C}_k\}_{k=1}^K$ being the ϵ-covering of $\mathcal{X}$. Here, Beta(α, β) indicates the Beta distribution with hyperparameters α and β, and Gauss mix (σ) means the mixture of five Gaussian distributions with a standard deviation σ. Appendix H presents more results with different distributions, showing the same qualitative behavior in all cases.

While the ϵ-covering of the original input space $\mathcal{X}$ is the default example from the previous paper (Xu & Mannor, 2012), in Figure 1 we see that K grows rapidly as d increases. Therefore, to reduce K significantly, we also propose utilizing the inverse image of the ϵ-covering in a randomly projected space. That is, given a random matrix A, we use the ϵ-covering of the space of $u = Ax$ to define the pre-partition $\{\tilde{\mathcal{C}}_k\}_{k=1}^K$. Then, the partition $\{\mathcal{C}_k\}_{k=1}^K$ is defined by $\mathcal{C}_k = \{x \in \mathcal{X} : Ax \in \tilde{\mathcal{C}}_k\}$. We randomly generated matrix $A \in \mathbb{R}^{3\times d}$ in each trial.

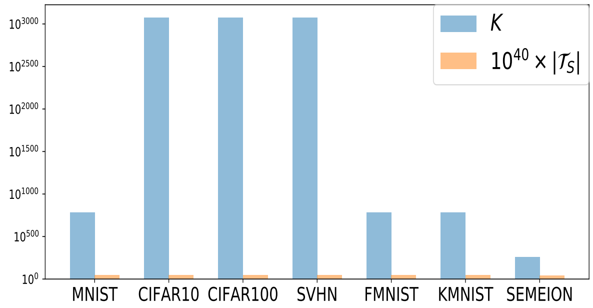


Figure 3. The values of K versus $|\mathcal{T}_S|$ with real-world data and the ϵ-covering. The values of $|\mathcal{T}_S|$ are extremely small compared to those of K in all datasets.

Figure 2 shows the values of K versus $|\mathcal{T}_S|$ for the synthetic data with the partition $\{\mathcal{C}_k\}_{k=1}^K$ being the inverse image of the ϵ-covering in randomly projected spaces. As can be seen, even in the case where K is reduced via random projection, we have $|\mathcal{T}_S| \ll K$. Thus, in both cases, our bounds are significantly tighter than the previous bound for these synthetics data.

5.2. Real-world data

Figure 3 shows the values of K versus $|\mathcal{T}_S|$ for the real-world data with the partition $\{\mathcal{C}_k\}_{k=1}^K$ being the ϵ-covering of $\mathcal{X}$. All the training data points of each dataset were used. As can be seen, we have $|\mathcal{T}_S| \ll K$ for the real-world data.

To reduce the value of K, we additionally propose the fol-

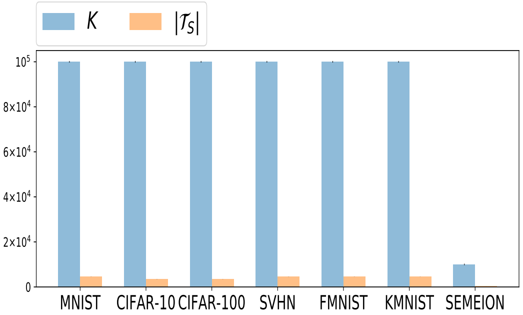


Figure 4. The values of K versus $|\mathcal{T}_S|$ with real-world data and the clustering using unlabeled data. With clustering to reduce K, we still have $|\mathcal{T}_S| < K$. Here, $|\mathcal{T}_S|$ was close to zero for Semeion.

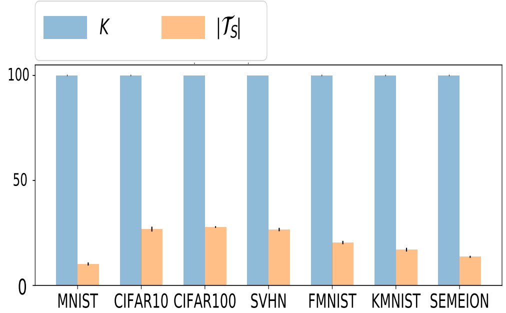


Figure 5. The values of K versus $|\mathcal{T}_S|$ with real-world data and random projection. With random projection to reduce K, we still have $|\mathcal{T}_S| < 30 < K = 100 < n \approx 60,000$ for the real-life datasets. Here, n is the full train data size of each dataset: e.g., $n = 60,000$ for MNIST.

lowing new method; i.e., as we have unlabeled data in many applications, we propose to use them to help define the partition $\{\mathcal{C}_k\}_{k=1}^K$. The key idea here was that the choice of partition $\{\mathcal{C}_k\}_{k=1}^K$ had to be independent of the labeled data used in the training loss in Theorems 1 and 2, but it could depend on the unlabeled data. Otherwise expressed, given a set of unlabeled data points $\{\bar{x}_k\}_{k=1}^K$, the partition $\{\mathcal{C}_k\}_{k=1}^K$ is defined by the clustering with the unlabeled data as $\mathcal{C}_k = \{x \in \mathcal{X} : k = \operatorname{argmin}_{k' \in [K]} \|x - \bar{x}_{k'}\|_2\}$. Following the literature on semi-supervised learning, we split the training data points into labeled data points (500 for Semeion and 5000 for all other datasets) and unlabeled data points (the remainder of the training data).

Figure 4 shows the values of K versus $|\mathcal{T}_S|$ for the real-world data with the partition $\{\mathcal{C}_k\}_{k=1}^K$ being the clustering with the unlabeled data. As can be seen, even in this case with the significantly reduced K, we still have $|\mathcal{T}_S| \ll K$.

Figure 5 shows the values of K v.s. $|\mathcal{T}_S|$ for real-life datasets with the partition being the inverse image of the ϵ-covering in randomly projected spaces. The random projection was conducted in the same manner without unlabeled data as in Figure 2. The projection reduced the value of K significantly, and yet we still have $|\mathcal{T}_S| \ll K$. Thus, in all three cases, our bounds are significantly tighter than the previous bound for these real-world data.

6. Conclusion

Since its introduction in 2010, algorithmic robustness has been a popular approach for analyzing learning algorithms (Xu & Mannor, 2010; 2012; Bellet & Habrard, 2015; Jolliffe & Cadima, 2016). In the original manuscript, which initiated the study of algorithmic robustness, Xu & Mannor (2010; 2012) pointed out that one disadvantage of their method is the dependence of the bound on the covering number of the sample space. To the community, they posed the open problem of finding a mechanism to improve this dependence. Despite the popularity and several unsuccessful attempts, no significant progress has been made in this regard (Qian et al., 2020; Agrawal & Jia, 2017; 2020).

In this study, we provide tighter bounds for algorithmic robustness and general multinomial distributions. Our results establish natural and easily verified conditions in which the dependence of K can be greatly reduced. Additionally, we demonstrate that the expected loss can be controlled by examining the single hypothesis returned by an algorithm, whereas in (Xu & Mannor, 2012), the entire hypothesis space has to be analyzed. This is a considerable gain against several common loss functions.

Our bound is both practical and effective in the machine learning setting, as it depends only on the training samples. Furthermore, we provided theoretical and numerical examples in which our bounds proved superior to those of Xu & Mannor (2012) and those that follow from uniform stability (Bousquet & Elisseeff, 2002). Our experimental simulations show that on common datasets and popular theoretical models, this bound is exponentially better than the algorithmic robustness bound (Xu & Mannor, 2012). These improvements to the foundations of algorithmic robustness have immediate impacts on applications ranging from metric learning to invariant classifiers.

The main limitation of our approach is that we cannot know the values of the bounds until specifying training data; i.e., $|\mathcal{T}_S|$ in our bound is data-dependent, whereas K in the previous bound is data-independent. This data-dependence might not be preferable in some applications, where we may want to compute a bound before seeing training data.

Acknowledgements

The research of J.H. is supported by the Simons Foundation as a Junior Fellow at the Simons Society of Fellows, and NSF grant DMS-2054835. The research of Z.D. is supported by the Sloan Foundation grants, the NSF grant 1763665, and the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

References

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Agrawal, S. and Jia, R. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv update as a correction of the NeurIPS 2017 paper of the same authors*, 2020.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bellet, A. and Habrard, A. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

Ben-Tal, A. and Nemirovski, A. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Bertsimas, D., Brown, D. B., and Caramanis, C. Theory and applications of robust optimization. *SIAM review*, 53(3): 464–501, 2011.

Bhattacharyya, C., Pannagadatta, K., and Smola, A. J. A second order cone programming formulation for classifying missing data. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pp. 153–160, 2004.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Chen, B., Feng, Y., Dai, T., Bai, J., Jiang, Y., Xia, S.-T., and Wang, X. Adversarial examples generation for deep product quantization networks on image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pp. 854–863. PMLR, 2017.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. In *NeurIPS Creativity Workshop 2019*, 2019.

Deng, Z., He, H., and Su, W. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning*, pp. 2590–2600. PMLR, 2021a.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Y. Adversarial training helps transfer learning via better representations. *Advances in Neural Information Processing Systems*, 34, 2021b.

Devroye, L. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pp. 896–904, 1983.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Ding, C., Xu, C., and Tao, D. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24 (3):980–993, 2015.

Gabrel, V., Murat, C., and Thiele, A. Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3):471–483, 2014.

Globerson, A. and Roweis, S. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pp. 353–360, 2006.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

Hu, Z., Jagtap, A. D., Karniadakis, G. E., and Kawaguchi, K. When do extended physics-informed neural networks (xpinns) improve generalization? *arXiv preprint arXiv:2109.09444*, 2021.

Jia, K., Li, S., Wen, Y., Liu, T., and Tao, D. Orthogonal deep neural networks. *arXiv preprint arXiv:1905.05929*, 2019.

Jolliffe, I. T. and Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Kawaguchi, K. and Huang, J. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 92–99. IEEE, 2019.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, D., Lamb, A. M., Kawaguchi, K., ALIAS PARTH GOYAL, A. G., Sun, C., Mozer, M. C., and Bengio, Y. Discrete-valued neural communication. *Advances in Neural Information Processing Systems*, 34, 2021.

Liu, H., Wu, J., Liu, T., Tao, D., and Fu, Y. Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. *IEEE transactions on knowledge and data engineering*, 29(5):1129–1143, 2017.

Luo, Y., Liu, T., Tao, D., and Xu, C. Multiview matrix completion for multilabel image classification. *IEEE Transactions on Image Processing*, 24(8):2355–2368, 2015.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

Pedraza, A., Deniz, O., and Bueno, G. Lyapunov stability for detecting adversarial image examples. *Chaos, Solitons & Fractals*, 155:111745, 2022.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., Tan, M., and Le, Q. V. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021. doi: 10.48550/arXiv.2111.10050. URL https://arxiv.org/abs/2111.10050.

Qi, Z., Tian, Y., and Shi, Y. Robust twin support vector machine for pattern classification. *Pattern Recognition*, 46(1):305–316, 2013.

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. Concentration inequalities for multinoulli random variables. *arXiv preprint arXiv:2001.11595*, 2020.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Rice, L., Bair, A., Zhang, H., and Kolter, J. Z. Robustness between the worst and average case. *Advances in Neural Information Processing Systems*, 34, 2021.

Robey, A., Chamon, L., Pappas, G. J., Hassani, H., and Ribeiro, A. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34:6198–6215, 2021.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Shen, X., Cheng, X., and Liang, K. Deep euler method: solving odes by approximating the local truncation error of the euler method. *arXiv preprint arXiv:2003.09573*, 2020.

Shi, Y., Bellet, A., and Sha, F. Sparse compositional metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pp. 1094–1103, 2017a.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 2017b.

Srl, B. T. and Brescia, I. Semeion handwritten digit data set. *Semeion Research Center of Sciences of Communication, Rome, Italy*, 1994.

Tao, D., Guo, Y., Song, M., Li, Y., Yu, Z., and Tang, Y. Y. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6): 2726–2738, 2016.

Van Der Vaart, A. W., van der Vaart, A. W., van der Vaart, A., and Wellner, J. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Vapnik, V. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wellner, J. et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, H. and Mannor, S. Robustness and generalization. In *Conference on Learning Theory (COLT)*, 2010.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Zahavy, T., Kang, B., Sivak, A., Feng, J., Xu, H., and Mannor, S. Ensemble robustness and generalization of stochastic deep learning algorithms. *arXiv preprint arXiv:1602.02389*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Zhang, L., Deng, Z., and Kawaguchi, K. How does mixup help with robustness and generalization? In *International Conference on Learning Representations (ICLR)*, 2021b.

A. Multinomial Concentration Bounds with Probability Distribution Dependence

Let the vector $X = (X_1, \dots, X_K)$ follow the multinomial distribution with parameters m and $p = (p_1, \dots, p_K)$. We wish to upper bound the following quantity:

$$\sum_{i=1}^{K} a_i(X) \left(p_i - \frac{X_i}{m} \right), \tag{9}$$

where $a_i(X) \geq 0$ for all $i \in \{1, \dots, K\}$. Recall that $a_i(X)$ depend on $X_1, \dots, X_K$, which is the heart of the problem.

In this section, we first establish concentration results for scaled multinomial random variables, in other words with $a_i(X)$ fixed independent of X. This is the first step towards dealing with the dependent coefficients in the above equation. After this first step, we will analyze the quantity with the dependent a_i.

A.1. Multinomial Distribution with fixed a

In this setting, we obtain the following sharp lower and upper bounds.

Lemma 2. *Let $\bar{a}_1, \dots, \bar{a}_K \geq 0$ be fixed such that $\sum_{i=1}^K \bar{a}_i p_i \neq 0$. Then, for any $M > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(p_i - \frac{X_i}{m} \right) < -M \right) \leq \exp\left(-\frac{mM}{2\bar{a}} \min\left\{ 1, \frac{\bar{a}M}{\beta} \right\} \right)$$

where $\bar{a} := \max_{i \in [K]} \bar{a}_i$ and $\beta := 2\sum_{i=1}^K \bar{a}_i^2 p_i$.

Proof. Note that $\mathbb{P}\left(\sum_{i=1}^K \bar{a}_i \left(p_i - \frac{X_i}{m}\right) < -M\right) = \mathbb{P}\left(\sum_{i=1}^K \bar{a}_i \left(\frac{X_i}{m} - p_i\right) > M\right)$. By Markov's inequality, the following holds for any $\nu > 0$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i \right) > M \right) &= \mathbb{P}\left(\nu \sum_{i=1}^{K} \bar{a}_i \frac{X_i}{m} > \nu \left(M + \sum_{i=1}^{K} \bar{a}_i p_i \right) \right) \\
&= \mathbb{P}\left(e^{\nu \sum_{i=1}^K \bar{a}_i \frac{X_i}{m}} > e^{\nu\left(M + \sum_{i=1}^K \bar{a}_i p_i\right)} \right) \\
&\leq \mathbb{P}\left(e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}} \geq e^{\nu\left(M + \sum_{i=1}^K \bar{a}_i p_i\right)} \right) \\
&\leq e^{-\nu\left(M + \sum_{i=1}^K \bar{a}_i p_i\right)} \mathbb{E}[e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}}].
\end{aligned} \tag{10}$$

To evaluate the moment generating function, we use the probability mass function of the multinomial distribution and the multinomial theorem to find that

$$\begin{aligned}
\mathbb{E}[e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}}] &= \sum \frac{m!}{X_1! \dots X_K!} p_1^{X_1} p_2^{X_2} \dots p_K^{X_K} e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}} \\
&= \sum \frac{m!}{X_1! \dots X_K!} (p_1 e^{\frac{\nu \bar{a}_1}{m}})^{X_1} (p_2 e^{\frac{\nu \bar{a}_2}{m}})^{X_2} \dots (p_K e^{\frac{\nu \bar{a}_K}{m}})^{X_K} \\
&= (\sum_{i=1}^{K} p_i e^{\frac{\nu \bar{a}_i}{m}})^m,
\end{aligned} \tag{11}$$

where the sum in the first two lines is taken over all possible values of the random variable $(X_1, \dots, X_K)$ (i.e., this is the sum in the definition of the expectation on the left-hand side of the equation). Pugging this into (10), we conclude that

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i \right) > M \right) \leq e^{-\nu\left(M + \sum_{i=1}^K \bar{a}_i p_i\right)} \left(\sum_{i=1}^{K} p_i e^{\frac{\nu \bar{a}_i}{m}} \right)^m. \tag{12}$$

We recall the following simple bounds for exponential functions:

$$e^x \geq 1 + x \qquad \text{for } x \geq 0 \tag{13}$$

$$e^x \leq 1 + x + x^2 \ \text{ for } 0 \leq x \leq 1. \tag{14}$$

Then, for $0 < \nu \leq \frac{m}{\bar{a}} \leq \frac{m}{\bar{a}_i}$, using (13)–(14), we have that

$$\sum_{i=1}^{K} p_i e^{\frac{\nu \bar{a}_i}{m}} \leq \sum_{i=1}^{K} p_i \left(1 + \frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{m^2}\right) \leq e^{\sum_{i=1}^{K} p_i \left(\frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{m^2}\right)}.$$

Pugging this into (12), we deduce that

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i\right) > M\right) &\leq e^{-\nu\left(M + \sum_{i=1}^{K} \bar{a}_i p_i\right)} \left(e^{\sum_{i=1}^{K} p_i \left(\frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{m^2}\right)}\right)^m \\
&= e^{-\nu\left(M + \sum_{i=1}^{K} \bar{a}_i p_i\right)} e^{\sum_{i=1}^{K} \nu p_i \bar{a}_i + \sum_{i=1}^{K} p_i \frac{\nu^2 \bar{a}_i^2}{m}} \\
&= e^{-\nu M + \sum_{i=1}^{K} p_i \frac{\nu^2 \bar{a}_i^2}{m}} \\
&= e^{-\nu M + \frac{\nu^2 \beta}{2m}}
\end{aligned} \tag{15}$$

Here, we have $\beta := 2 \sum_{i=1}^{K} \bar{a}_i^2 p_i > 0$ since $\sum_{i=1}^{K} \bar{a}_i p_i \neq 0$ and $\bar{a}_i, p_i \geq 0$ by assumption.

We consider two possible cases. If $M \leq \frac{\beta}{\bar{a}}$, we take $0 < \nu = \frac{mM}{\beta} \leq \frac{m}{\bar{a}}$, which yields in (15):

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i\right) > M\right) \leq e^{-\frac{mM^2}{2\beta}}.$$

If $M \geq \frac{\beta}{\bar{a}}$, we take $0 < \nu = \frac{m}{\bar{a}}$, which yields in (15):

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i\right) > M\right) \leq e^{-\frac{m}{\bar{a}}\left(M - \frac{\beta}{2\bar{a}}\right)} \leq e^{-\frac{mM}{2\bar{a}}}.$$

Combining these conclusions yields

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(\frac{X_i}{m} - p_i\right) > M\right) \leq \exp\left(-\frac{mM}{2\bar{a}} \min\left\{1, \frac{\bar{a}M}{\beta}\right\}\right).$$

□

For the other tail, we establish the following estimate.

Lemma 3. *Let $\bar{a}_1, \ldots, \bar{a}_K \geq 0$ be fixed such that $\sum_{i=1}^{K} \bar{a}_i p_i \neq 0$. Then, for any $M > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > M\right) \leq \exp\left(-\frac{mM^2}{\beta}\right)$$

where $\beta := 2 \sum_{i=1}^{K} \bar{a}_i^2 p_i$.

Proof. By Markov's inequality, the following holds *for any $\nu < 0$:*

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{K} \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > M\right) &= \mathbb{P}\left(\nu \sum_{i=1}^{K} \bar{a}_i \frac{X_i}{m} > \nu \left(\sum_{i=1}^{K} \bar{a}_i p_i - M\right)\right) \\
&= \mathbb{P}\left(e^{\nu \sum_{i=1}^{K} \bar{a}_i \frac{X_i}{m}} > e^{\nu\left(\sum_{i=1}^{K} \bar{a}_i p_i - M\right)}\right) \\
&\leq \mathbb{P}\left(e^{\nu \sum_{i=1}^{K} \frac{\bar{a}_i X_i}{m}} \geq e^{\nu\left(\sum_{i=1}^{K} \bar{a}_i p_i - M\right)}\right)
\end{aligned}$$

$$\leq e^{-\nu\left(\sum_{i=1}^K \bar{a}_i p_i - M\right)} \mathbb{E}[e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}}]$$

Using (11) from the proof of Lemma 2, we have $\mathbb{E}[e^{\nu \sum_{i=1}^K \frac{\bar{a}_i X_i}{m}}] = (\sum_{i=1}^K p_i e^{\frac{\nu \bar{a}_i}{m}})^m$, yielding that

$$\mathbb{P}\left(\sum_{i=1}^K \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > M\right) \leq e^{-\nu\left(\sum_{i=1}^K \bar{a}_i p_i - M\right)} (\sum_{i=1}^K p_i e^{\frac{\nu \bar{a}_i}{m}})^m.$$

We recall the following bounds for exponential functions:

$$\begin{aligned} &e^x \leq 1 + x + \frac{x^2}{2} \ \text{ for any } x \leq 0 \\ &e^x \geq 1 + x \ \text{ for any } x \in \mathbb{R} \end{aligned}$$

Using these bounds,

$$\sum_{i=1}^K p_i e^{\frac{\nu \bar{a}_i}{m}} \leq \sum_{i=1}^K p_i (1 + \frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{2m^2}) \leq e^{\sum_{i=1}^K p_i (\frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{2m^2})}.$$

Combining what we have so far,

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^K \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > M\right) &\leq e^{-\nu\left(\sum_{i=1}^K \bar{a}_i p_i - M\right)} (e^{\sum_{i=1}^K p_i (\frac{\nu \bar{a}_i}{m} + \frac{\nu^2 \bar{a}_i^2}{2m^2})})^m \\ &= e^{\nu M + \sum_{i=1}^K p_i \frac{\nu^2 \bar{a}_i^2}{2m}} \\ &= e^{\nu M + \frac{\nu^2 \beta}{4m}} \end{aligned}$$

Here, we have $\beta := 2\sum_{i=1}^K \bar{a}_i^2 p_i > 0$ since $\sum_{i=1}^K \bar{a}_i p_i \neq 0$ and $\bar{a}_i, p_i \geq 0$ by assumption. Finally, we set $\nu = -\frac{2m\nu}{\beta} (< 0)$ to conclude that

$$\mathbb{P}\left(\sum_{i=1}^K \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > M\right) \leq e^{-\frac{mM^2}{\beta}}.$$

□

A.2. Multinomial Distribution without a

In this section, we establish some bounds for $a_i = 1$ which is of special interest in applications.

Lemma 4. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $i \in \{1, \ldots, K\}$:*

$$p_i \geq \begin{cases} \frac{X_i}{m} - \sqrt{\frac{p_i \ln(K/\delta)}{m}} & \textit{if } p_i > \frac{\ln(K/\delta)}{4m} \\ \frac{X_i}{m} - \frac{2\ln(K/\delta)}{m} & \textit{if } p_i \leq \frac{\ln(K/\delta)}{4m} \end{cases}$$

Proof. For each $i \in [K]$, if $p_i = 0$, then the desired statement holds vacuously, because $0 = p_i \geq \frac{X_i}{m} - \frac{2\ln(K/\delta)}{m}$ where $\frac{X_i}{m} = 0$ (since $p_i = 0$) and $\frac{2\ln(K/\delta)}{m} \geq 0$. Thus, for the remainder of the proof, we consider the case where $p_i \neq 0$. For each $i \in [K]$, we use Lemma 2 with $\bar{a}_i = 1$ and $\bar{a}_j = 0$ for all $j \neq i$ (which satisfies $\sum_{i=1}^K \bar{a}_i p_i \neq 0$ since $p_i \neq 0$), yielding that for any $M > 0$,

$$\mathbb{P}\left(p_i - \frac{X_i}{m} < -M\right) \leq \exp\left(-\frac{mM}{2} \min\left\{1, \frac{M}{2p_i}\right\}\right).$$

In other words, for any $M > 0$,

$$\forall M \leq 2p_i, \ \mathbb{P}\left(\frac{X_i}{m} - p_i > M\right) \leq e^{-\frac{mM^2}{4p_i}}$$

and

$$\forall M \geq 2p_i, \ \mathbb{P}\left(\frac{X_i}{m} - p_i > M\right) \leq e^{-\frac{mM}{2}}.$$

We now consider the two cases on the value of p_i for an arbitrary $\delta > 0$.

Case $p_i \geq \frac{\ln(K/\delta)}{4m}$: in this case, we set $M = \sqrt{\frac{p_i \ln(K/\delta)}{m}}$. Then, we have that $M \leq 2p_i$ since the condition $p_i \geq \frac{\ln(K/\delta)}{4m}$ implies that $4p_i^2 \geq \frac{p_i \ln(K/\delta)}{m}$ and hence $2p_i \geq \sqrt{\frac{p_i \ln(K/\delta)}{m}} = M$. Therefore, if $p_i \geq \frac{\ln(K/\delta)}{4m}$,

$$\mathbb{P}\left(\frac{X_i}{m} - p_i > \sqrt{\frac{p_i \ln(K/\delta)}{m}}\right) \leq \frac{\delta}{K}.$$

Case $p_i \leq \frac{\ln(K/\delta)}{4m}$: here, we set $M = \frac{2\ln(K/\delta)}{m}$. Then, we have that $M \geq 2p_i$ since the condition $p_i \leq \frac{\ln(K/\delta)}{4m}$ implies that $2p_i \leq \frac{\ln(K/\delta)}{2m} \leq \frac{2\ln(K/\delta)}{m} = M$. Thus, we have that if $p_i \leq \frac{\ln(K/\delta)}{4m}$,

$$\mathbb{P}\left(\frac{X_i}{m} - p_i > \frac{2\ln(K/\delta)}{m}\right) \leq \frac{\delta}{K}.$$

Taking union bounds over all $i \in \{1, \dots, K\}$, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $i \in \{1, \dots, K\}$:

$$\frac{X_i}{m} - p_i \leq \begin{cases} \sqrt{\frac{p_i \ln(K/\delta)}{m}} & \text{if } p_i > \frac{\ln(K/\delta)}{4m} \\ \frac{2\ln(K/\delta)}{m} & \text{if } p_i \leq \frac{\ln(K/\delta)}{4m}. \end{cases}$$

In other words, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $i \in \{1, \dots, K\}$:

$$p_i \geq \begin{cases} \frac{X_i}{m} - \sqrt{\frac{p_i \ln(K/\delta)}{m}} & \text{if } p_i > \frac{\ln(K/\delta)}{4m} \\ \frac{X_i}{m} - \frac{2\ln(K/\delta)}{m} & \text{if } p_i \leq \frac{\ln(K/\delta)}{4m}. \end{cases}$$

□

We have a simpler statement for the other tail.

Lemma 5. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $i \in \{1, \dots, K\}$:*

$$p_i - \frac{X_i}{m} \leq \sqrt{\frac{2p_i \ln(K/\delta)}{m}}. \tag{16}$$

Proof. For each $i \in [K]$, if $p_i = 0$, then the desired statement holds trivially because $p_i - \frac{X_i}{m} = -\frac{X_i}{m} \leq \sqrt{\frac{2p_i \ln(K/\delta)}{m}}$ where $\frac{X_i}{m} = 0$ and $\sqrt{\frac{2p_i \ln(K/\delta)}{m}} = 0$. Thus, for the rest of the proof, we consider the case where $p_i \neq 0$. For each $i \in [K]$, we use Lemma 3 with $\bar{a}_i = 1$ and $\bar{a}_j = 0$ for all $j \neq i$ (which satisfies $\sum_{i=1}^K \bar{a}_i p_i \neq 0$ since $p_i \neq 0$), yielding that for any $M > 0$,

$$\mathbb{P}\left(p_i - \frac{X_i}{m} > M\right) \leq \exp\left(-\frac{mM^2}{2p_i}\right).$$

By setting $M = \sqrt{\frac{2p_i \ln(K/\delta)}{m}}$,

$$\mathbb{P}\left(p_i - \frac{X_i}{m} > \sqrt{\frac{2p_i \ln(K/\delta)}{m}}\right) \leq \frac{\delta}{K}.$$

We conclude the proof by taking a union bound over all $i \in [K]$. □

A.3. Proof of Lemma 1

The proof of Lemma 1 is obtained by combining the results thus far:

Proof of Lemma 1. Lemma 5 implies Lemma 1 by summing up both sides of (16) with the coefficients $a_i(X) \geq 0$. That is, since $a_i(X) \geq 0$, Lemma 5 implies that for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $i \in \{1, \dots, K\}$:

$$a_i(X)\left(p_i - \frac{X_i}{m}\right) \leq a_i(X)\sqrt{\frac{2p_i \ln(K/\delta)}{m}}.$$

This then implies that for any $\delta > 0$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K} a_i(X)\left(p_i - \frac{X_i}{m}\right) \leq \left(\sum_{k=1}^{K} a_i(X)\sqrt{p_i}\right)\sqrt{\frac{2\ln(K/\delta)}{m}}.$$

□

B. Multinomial Concentration Bounds with Data Dependence

Recall that $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$ with $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$. Additionally, we defined $a_{\mathcal{T}_S}(X) = \max_{i \in \mathcal{T}_S} a_i(X)$ and $a_{\mathcal{T}_S^c}(X) = \max_{i \in \mathcal{T}_S^c} a_i(X)$ where $\mathcal{T}_S^c = [K] \setminus \mathcal{T}_S$.

Let the vector $X = (X_1, \dots, X_K)$ follows the multinomial distribution with parameter n and $p = (p_1, \dots, p_K)$, where $p_k = \mathbb{P}(z \in \mathcal{C}_k)$ and $X_k = \sum_{i=1}^{n} \mathbb{1}\{z_i \in \mathcal{C}_k\}$. We want to upper bound the following quantity:

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right), \tag{17}$$

where $a_i(X) \geq 0$ for all $i \in \{1, \dots, K\}$. Here, the $a_i(X)$ depend on $X_1, \dots, X_K$.

B.1. Basic Version

In this subsection, using our new results from Appendix A, we prove Theorem 3, which is restated as Lemma 6 in the following and further refined in Appendix B.2:

Lemma 6. *For any $\delta > 0$, with probability at least $1-\delta$, the following holds:*

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \leq (a_{\mathcal{T}_S}(X)\sqrt{2} + a_{\mathcal{T}_S^c}(X))\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)\frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}.$$

Proof. By using Lemma 4 and Lemma 5 and takeing union bounds for the two, we have that for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $i \in [K]$,

$$p_i - \frac{X_i}{n} \leq \sqrt{\frac{2p_i \ln(2K/\delta)}{n}} \tag{18}$$

and

$$p_i \geq \begin{cases} \frac{X_i}{n} - \sqrt{\frac{p_i \ln(2K/\delta)}{n}} & \text{if } p_i > \frac{\ln(2K/\delta)}{4n} \\ \frac{X_i}{n} - \frac{2\ln(2K/\delta)}{n} & \text{if } p_i \leq \frac{\ln(2K/\delta)}{4n} \end{cases} \tag{19}$$

Recall that $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$ with $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$. Summing up both sides of (19) over $i \in \mathcal{T}_S$ and using $\sum_{i \in \mathcal{T}_S} \frac{X_i}{n} = 1$,

$$\sum_{i \in \mathcal{T}_S} p_i \geq 1 - \sum_{i \in I_1}\sqrt{\frac{p_i \ln(2K/\delta)}{n}} - \sum_{i \in I_2}\frac{2\ln(2K/\delta)}{n} = 1 - \sum_{i \in I_1}\sqrt{\frac{p_i \ln(2K/\delta)}{n}} - (|\mathcal{T}_S| - |I_1|)\frac{2\ln(2K/\delta)}{n}$$

where $I_1 = \{i \in \mathcal{T}_S : p_i > \frac{\ln(K/\delta)}{4n}\}$ and $I_2 = \{i \in \mathcal{T}_S : p_i \leq \frac{\ln(K/\delta)}{4n}\}$. This implies that

$$1 - \sum_{i \in \mathcal{T}_S} p_i \leq \sum_{i \in I_1}\sqrt{\frac{p_i \ln(2K/\delta)}{n}} + (|\mathcal{T}_S| - |I_1|)\frac{2\ln(2K/\delta)}{n}.$$

Since $\sum_{i=1}^{K} p_i = 1$ and hence $\sum_{i \notin \mathcal{T}_S} p_i = 1 - \sum_{i \in \mathcal{T}_S} p_i$, this implies that

$$\sum_{i \notin \mathcal{T}_S} p_i \leq \left(\sum_{i \in I_1} \sqrt{p_i} \right) \sqrt{\frac{\ln(2K/\delta)}{n}} + (|\mathcal{T}_S| - |I_1|) \frac{2\ln(2K/\delta)}{n}. \tag{20}$$

Using $a_{\mathcal{T}_S^c}(X) = \max_{i \in \mathcal{T}_S^c} a_i(X)$ with $a_i(X) \geq 0$,

$$\begin{aligned} \sum_{i=1}^{K} a_i(X) \left(p_i - \frac{X_i}{n} \right) &= \sum_{i \in \mathcal{T}_S} a_i(X) \left(p_i - \frac{X_i}{n} \right) + \sum_{i \notin \mathcal{T}_S} a_i(X) p_i \\ &\leq \sum_{i \in \mathcal{T}_S} a_i(X) \left(p_i - \frac{X_i}{n} \right) + a_{\mathcal{T}_S^c}(X) \sum_{i \notin \mathcal{T}_S} p_i \end{aligned}$$

Plugging (18) in the first term and (20) in the second term,

$$\begin{aligned} &\sum_{i=1}^{K} a_i(X) \left(p_i - \frac{X_i}{n} \right) \qquad (21) \\ &\leq \sum_{i \in \mathcal{T}_S} a_i(X) \sqrt{\frac{2p_i \ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X) \left(\sum_{i \in I_1} \sqrt{p_i} \right) \sqrt{\frac{\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)(|\mathcal{T}_S| - |I_1|) \frac{2\ln(2K/\delta)}{n} \\ &\leq \left(\sum_{i \in \mathcal{T}_S} a_i(X) \sqrt{p_i} \right) \sqrt{\frac{2\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X) \left(\sum_{i \in I_1} \sqrt{p_i} \right) \sqrt{\frac{\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)(|\mathcal{T}_S| - |I_1|) \frac{2\ln(2K/\delta)}{n} \end{aligned}$$

Using $a_{\mathcal{T}_S}(X) = \max_{i \in \mathcal{T}_S} a_i(X)$, since $\sum_{i \in \mathcal{T}_S} a_i(X) \sqrt{p_i} \leq a_{\mathcal{T}_S}(X) \sqrt{|\mathcal{T}_S|}$ and $\sum_{i \in I_1} \sqrt{p_i} \leq \sum_{i \in \mathcal{T}_S} \sqrt{p_i} \leq \sqrt{|\mathcal{T}_S|}$ by using the Cauchy-Schwarz inequality ($\sum_{i=1}^{r} b_i \sqrt{a_i} \leq \sqrt{\sum_{i=1}^{r} a_i} \sqrt{\sum_{i=1}^{r} b_i^2}$),

$$\begin{aligned} &\sum_{i=1}^{K} \alpha_i(X) \left(p_i - \frac{X_i}{n} \right) \\ &\leq a_{\mathcal{T}_S}(X) \sqrt{\frac{2|\mathcal{T}_S| \ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X) \sqrt{\frac{|\mathcal{T}_S| \ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X) \frac{2|\mathcal{T}_S| \ln(2K/\delta)}{n} \\ &\leq (a_{\mathcal{T}_S}(X)\sqrt{2} + a_{\mathcal{T}_S^c}(X)) \sqrt{\frac{|\mathcal{T}_S| \ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X) \frac{2|\mathcal{T}_S| \ln(2K/\delta)}{n}. \end{aligned}$$

□

B.2. Tighter version

Lemma 7. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds:*

$$\begin{aligned} &\sum_{i=1}^{K} \alpha_i(X) \left(p_i - \frac{X_i}{n} \right) \\ &\leq \sqrt{\frac{\ln(2K/\delta)}{n}} \left(\sum_{i \in \mathcal{T}_S} (a_{\mathcal{T}_S^c}(X) + \sqrt{2} a_i(X)) \sqrt{\frac{X_i}{n}} \right) + \frac{2\ln(2K/\delta)}{n} \left(a_{\mathcal{T}_S^c}(X) |\mathcal{T}_S| + \sum_{i \in \mathcal{T}_S} a_i(X) \right). \end{aligned}$$

Proof. We start with (21) from the proof of Lemma 6: i.e., for any $\delta > 0$, with probability at least $1 - \delta$,

$$p_i - \frac{X_i}{n} \leq \sqrt{\frac{2p_i \ln(2K/\delta)}{n}} \qquad \forall i \in [K], \tag{22}$$

and

$$\begin{aligned}
&\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \\
&\le \left(\sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{p_i}\right)\sqrt{\frac{2\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)\left(\sum_{i\in I_1}\sqrt{p_i}\right)\sqrt{\frac{\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)(|\mathcal{T}_S|-|I_1|)\frac{2\ln(2K/\delta)}{n}.
\end{aligned} \tag{23}$$

Instead of using the Cauchy-Schwarz inequality to bound the two terms $\sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{p_i}$ and $\sum_{i\in I_1}\sqrt{p_i}$ (which is done in the proof of Lemma 6), we now derive and use another probabilistic bound to bound these terms. For any $I \subseteq [K]$,

$$\begin{aligned}
\sum_{i\in I} a_i(X)\sqrt{p_i} &= \sum_{i\in I} a_i(X)\left(\sqrt{p_i} - \sqrt{\frac{X_i}{n}}\right) + \sum_{i\in I} a_i(X)\sqrt{\frac{X_i}{n}} \\
&= \sum_{i\in I} a_i(X)\frac{p_i - \frac{X_i}{n}}{\sqrt{p_i} + \sqrt{\frac{X_i}{n}}} + \sum_{i\in I} a_i(X)\sqrt{\frac{X_i}{n}}
\end{aligned}$$

Using (23) for $p_i - \frac{X_i}{n}$,

$$\begin{aligned}
\sum_{i\in I} a_i(X)\sqrt{p_i} &\le \sum_{i\in I} a_i(X)\frac{1}{\sqrt{p_i} + \sqrt{\frac{X_i}{n}}}\sqrt{\frac{2p_i\ln(2K/\delta)}{n}} + \sum_{i\in I} a_i(X)\sqrt{\frac{X_i}{n}} \\
&\le \sum_{i\in I} a_i(X)\sqrt{\frac{2p_i\ln(2K/\delta)}{np_i}} + \sum_{i\in I} a_i(X)\sqrt{\frac{X_i}{n}} \\
&= \sqrt{\frac{2\ln(2K/\delta)}{n}}\sum_{i\in I} a_i(X) + \sum_{i\in I} a_i(X)\sqrt{\frac{X_i}{n}}
\end{aligned}$$

Thus, by setting $I = \mathcal{T}_S$ as well as $I = I_1$ and $a_i(X) = 1$,

$$\sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{p_i} \le \sqrt{\frac{2\ln(2K/\delta)}{n}}\sum_{i\in\mathcal{T}_S} a_i(X) + \sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{\frac{X_i}{n}} \qquad \text{and,}$$

$$\sum_{i\in I_1}\sqrt{p_i} \le \sqrt{\frac{2\ln(2K/\delta)}{n}}|I_1| + \sum_{i\in I_1}\sqrt{\frac{X_i}{n}}$$

Plugging these into (23),

$$\begin{aligned}
&\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \\
&\le \left(\sqrt{\frac{2\ln(2K/\delta)}{n}}\sum_{i\in\mathcal{T}_S} a_i(X) + \sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{\frac{X_i}{n}}\right)\sqrt{\frac{2\ln(2K/\delta)}{n}} \\
&\qquad + a_{\mathcal{T}_S^c}(X)\left(\sqrt{\frac{2\ln(2K/\delta)}{n}}|I_1| + \sum_{i\in I_1}\sqrt{\frac{X_i}{n}}\right)\sqrt{\frac{\ln(2K/\delta)}{n}} + a_{\mathcal{T}_S^c}(X)(|\mathcal{T}_S|-|I_1|)\frac{2\ln(2K/\delta)}{n} \\
&= \frac{2\ln(2K/\delta)}{n}\sum_{i\in\mathcal{T}_S} a_i(X) + \sqrt{\frac{2\ln(2K/\delta)}{n}}\sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{\frac{X_i}{n}} \\
&\qquad + a_{\mathcal{T}_S^c}(X)\left(\frac{\ln(2K/\delta)}{n}\sqrt{2}|I_1| + \sqrt{\frac{\ln(2K/\delta)}{n}}\sum_{i\in I_1}\sqrt{\frac{X_i}{n}}\right) + a_{\mathcal{T}_S^c}(X)(|\mathcal{T}_S|-|I_1|)\frac{2\ln(2K/\delta)}{n} \\
&\le \frac{2\ln(2K/\delta)}{n}\sum_{i\in\mathcal{T}_S} a_i(X) + \sqrt{\frac{2\ln(2K/\delta)}{n}}\sum_{i\in\mathcal{T}_S} a_i(X)\sqrt{\frac{X_i}{n}}
\end{aligned}$$

$$+ a_{\mathcal{T}_S^c}(X)\sqrt{\frac{\ln(2K/\delta)}{n}}\sum_{i\in\mathcal{T}_S}\sqrt{\frac{X_i}{n}} + a_{\mathcal{T}_S^c}(X)|\mathcal{T}_S|\frac{2\ln(2K/\delta)}{n}$$
$$= \sqrt{\frac{\ln(2K/\delta)}{n}}\left(\sum_{i\in\mathcal{T}_S}(a_{\mathcal{T}_S^c}(X) + \sqrt{2}a_i(X))\sqrt{\frac{X_i}{n}}\right) + \frac{2\ln(2K/\delta)}{n}\left(a_{\mathcal{T}_S^c}(X)|\mathcal{T}_S| + \sum_{i\in\mathcal{T}_S}a_i(X)\right)$$

□

C. Proof of Theorem 1

This section culminates in the proof of Theorem 1 using our new results on multinomial distributions from Appendices A and B. We define

$$\mathcal{I}_k := \mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\},$$

and

$$\alpha_k(h) := \mathbb{E}_z[\ell(h,z)|z \in \mathcal{C}_k].$$

We start with the proof of the following lemma that relate the gap to the concentration of the multinomial distributions:

Lemma 8. *For any $h \in \mathcal{H}$ and $z_i \in \mathcal{Z}$ for all $i \in [n]$,*

$$\mathbb{E}_z[\ell(h,z)] - \frac{1}{n}\sum_{i=1}^n \ell(h,z_i) = \sum_{k=1}^K \alpha_k(h)\left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) + \frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k|\left(\alpha_k(h) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(h,z_i)\right).$$

Proof. We first write the expected error as the sum of the conditional expected error:

$$\mathbb{E}_z[\ell(h,z)] = \sum_{k=1}^K \mathbb{E}_z[\ell(h,z)|z\in\mathcal{C}_k]\Pr(z\in\mathcal{C}_k) = \sum_{k=1}^K \mathbb{E}_{z_k}[\ell(h,z_k)]\Pr(z\in\mathcal{C}_k),$$

where z_k is the random variable z conditioned on the event $z \in \mathcal{C}_k$. Using this, we decompose the generalization error into two terms:

$$\mathbb{E}_z[\ell(h,z)] - \frac{1}{n}\sum_{i=1}^n \ell(h,z_i) \tag{24}$$
$$= \sum_{k=1}^K \mathbb{E}_{z_k}[\ell(h,z_k)]\left(\Pr(z\in\mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) + \left(\sum_{k=1}^K \mathbb{E}_{z_k}[\ell(h,z_k)]\frac{|\mathcal{I}_k|}{n} - \frac{1}{n}\sum_{i=1}^n \ell(h,z_i)\right).$$

The second term in the right-hand side of (24) is further simplified by using

$$\frac{1}{n}\sum_{i=1}^n \ell(h,z_i) = \frac{1}{n}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\ell(h,z_i),$$

as

$$\sum_{k=1}^K \mathbb{E}_{z_k}[\ell(h,z_k)]\frac{|\mathcal{I}_k|}{n} - \frac{1}{n}\sum_{i=1}^n \ell(h,z_i) = \frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k|\left(\mathbb{E}_{z_k}[\ell(h,z_k)] - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(h,z_i)\right).$$

Substituting these into equation (24) yields

$$\mathbb{E}_z[\ell(h,z)] - \frac{1}{n}\sum_{i=1}^n \ell(h,z_i)$$
$$= \sum_{k=1}^K \mathbb{E}_{z_k}[\ell(h,z_k)]\left(\Pr(z\in\mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) + \frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k|\left(\mathbb{E}_{z_k}[\ell(h,z_k)] - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(h,z_i)\right).$$

□

The second term in the previous lemma is bounded by the following lemma using the robustness:

Lemma 9. *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust, then the following holds for any $S \in \mathcal{Z}^n$:*

$$\left| \frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left(\mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right) \right| \leq \epsilon(S).$$

Proof. By the triangle inequality,

$$\begin{aligned} &\left| \frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left(\mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right) \right| \\ &\leq \frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left| \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right|. \end{aligned}$$

Furthermore, again by the triangle inequality,

$$\begin{aligned} \left| \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{S_i \in \mathcal{I}_k} \ell(\mathcal{A}_S, S_i) \right| &= \left| \frac{1}{|\mathcal{I}_k|} \sum_{S_i \in \mathcal{I}_k} \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{S_i \in \mathcal{I}_k} \ell(\mathcal{A}_S, S_i) \right| \\ &\leq \frac{1}{|\mathcal{I}_k|} \sum_{S_i \in \mathcal{I}_k} \left| \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \ell(\mathcal{A}_S, S_i) \right| \\ &\leq \sup_{z \in \mathcal{Z} \cap \mathcal{C}_k, s \in S \cap \mathcal{C}_k} \frac{1}{|\mathcal{I}_k|} \sum_{S_i \in \mathcal{I}_k} |\ell(\mathcal{A}_S, z) - \ell(\mathcal{A}_S, s)| \\ &= \sup_{z \in \mathcal{Z} \cap \mathcal{C}_k, s \in S \cap \mathcal{C}_k} |\ell(\mathcal{A}_S, z) - \ell(\mathcal{A}_S, s)|. \end{aligned}$$

We now suppose that a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust. Then, $\sup_{z \in \mathcal{Z} \cap \mathcal{C}_k, s \in S \cap \mathcal{C}_k} |\ell(\mathcal{A}_S, z) - \ell(\mathcal{A}_S, s)| \leq \epsilon(S)$ for all $k = 1, \dots, K$ by the definition of a learning algorithm $\mathcal{A}$ being $(K, \epsilon(\cdot))$-robust. Thus, we have that

$$\frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left| \mathbb{E}_z[\ell(\mathcal{A}_S, z) | z \in \mathcal{C}_k] - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right| \leq \epsilon(S) \left(\frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \right) = \epsilon(S),$$

since $\sum_{k=1}^{K} |\mathcal{I}_k| = n$. □

Using these lemmas and our new concentration bounds on multinomial distributions from Appendices A and B, we can complete the proof of Theorem 1 as follows:

Proof of Theorem 1. From Lemma 8,

$$\begin{aligned} &\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n} \sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) \qquad (25) \\ &= \frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right) + \sum_{k=1}^{K} \alpha_k(\mathcal{A}_S) \left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n} \right). \end{aligned}$$

By using Lemma 6 with $a_k(X) = \alpha_k(\mathcal{A}_S)$ and noticing that $a_{\mathcal{T}_S}(X), a_{\mathcal{T}_S^c}(X) \leq \zeta(\mathcal{A}_S)$, we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sum_{k=1}^{K} \alpha_k(\mathcal{A}_S) \left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n} \right) \leq \zeta(\mathcal{A}_S) \left((\sqrt{2} + 1) \sqrt{\frac{|\mathcal{T}_S| \ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S| \ln(2K/\delta)}{n} \right). \qquad (26)$$

Invoking Lemma 9,

$$\frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right) \leq \left| \frac{1}{n} \sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(\mathcal{A}_S, z_i) \right) \right| \leq \epsilon(S). \qquad (27)$$

By combining (25)–(27), we obtain the desired statement. □

D. Proof of Proposition 3

Proof of Proposition 3. Take $m = \beta(\ln n)^{1/\alpha}$, then we have

$$\mathcal{T}_S \leq m + \sum_{k>m} \mathbf{1}(|\mathcal{I}_k^S| \geq 1). \tag{28}$$

Under our assumption that $p_k \leq Ce^{-(k/\beta)^\alpha}$, we have

$$\begin{aligned}\mathbb{P}(z \in \cup_{k>m}\mathcal{C}_k) &= \sum_{k>m} p_k \leq \sum_{k>m} Ce^{-(k/\beta)^\alpha} \\ &\leq \int_m^\infty Ce^{-(x/\beta)^\alpha}\mathrm{d}x \leq \frac{C\beta}{\alpha}e^{(-m/\beta)^\alpha} = \frac{C\beta}{n\alpha}\end{aligned}$$

Therefore, by the Chernoff bound, it follows

$$\begin{aligned}&\mathbb{P}\left(\sum_{k>m} \mathbf{1}(|\mathcal{I}_k^S| \geq 1) \geq \lambda\right) && (29)\\ &\leq \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}(z_i \in \cup_{k>m}\mathcal{C}_k) \geq \lambda\right) && (30)\\ &\leq e^{C(e-1)\beta/\alpha-\lambda} && (31)\end{aligned}$$

If we take $\lambda = C(e-1)\beta/\alpha + \log(1/\delta)$, we can conclude that

$$\mathbb{P}\left(\sum_{k>m} \mathbf{1}(|\mathcal{I}_k^S| \geq 1) \geq \lambda\right) \leq \delta, \tag{32}$$

and (4) follows. □

E. Proof of Theorem 2

In this section, we refine the proof of Theorem 1 to obtain a tighter bound by using a tighter version of our new concentration bounds on multinomial distributions from Appendices A and B.

Proof of Theorem 2. The proof begins in the same manner as in the previous section, but uses the sharper multinomial bound. From Lemma 8,

$$\begin{aligned}&\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{A}_S, z_i) && (33)\\ &= \frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(\mathcal{A}_S, z_i)\right) + \sum_{k=1}^K \alpha_k(\mathcal{A}_S)\left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right).\end{aligned}$$

By using Lemma 7 with $a_k(X) = \alpha_k(\mathcal{A}_S), a_{\mathcal{T}_S}(X) = \alpha_{\mathcal{T}_S}(\mathcal{A}_S)$, and $a_{\mathcal{T}_S^c}(X) = \alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)$, we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{k=1}^K \alpha_k(\mathcal{A}_S)\left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) && (34)\\ &\leq \sqrt{\frac{\ln(2K/\delta)}{n}}\left(\sum_{k\in\mathcal{T}_S}(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) + \sqrt{2}\alpha_k(\mathcal{A}_S))\sqrt{\frac{|\mathcal{I}_k|}{n}}\right) + \frac{2\ln(2K/\delta)}{n}\left(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)|\mathcal{T}_S| + \sum_{k\in\mathcal{T}_S}\alpha_k(\mathcal{A}_S)\right) && (35)\end{aligned}$$

Applying Lemma 9,

$$\frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(\mathcal{A}_S, z_i)\right) \leq \left|\frac{1}{n}\sum_{k=1}^K |\mathcal{I}_k| \left(\alpha_k(\mathcal{A}_S) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(\mathcal{A}_S, z_i)\right)\right| \leq \epsilon(S). \tag{36}$$

By combining these, we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) \leq \epsilon(S) + \mathcal{Q}_1\sqrt{\frac{\ln(2K/\delta)}{n}} + \frac{2\mathcal{Q}_2 \ln(2K/\delta)}{n},$$

where $\mathcal{Q}_1 = \sum_{k\in\mathcal{T}_S}(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) + \sqrt{2}\alpha_k(\mathcal{A}_S))\sqrt{\frac{|\mathcal{I}_k|}{n}}$ and $\mathcal{Q}_2 = \alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)|\mathcal{T}_S| + \sum_{k\in\mathcal{T}_S}\alpha_k(\mathcal{A}_S)$. □

F. Pseudo-robustness

This section is devoted to the generalizations of Theorems 1 and 2 in the pseudo-robust context.

Definition 3. A learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ *pseudo robust*, for $K \in \mathbb{N}$, $\epsilon(\cdot) : \mathcal{Z}^n \to \mathbb{R}$, and $\hat{n}(\cdot) : \mathcal{Z}^n \to \{1, \ldots, n\}$, if $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{\mathcal{C}_k\}_{k=1}^K$, such that for all $S \in \mathcal{Z}^n$, there exists a subset of training samples $\hat{S}$ with $|\hat{S}| = \hat{n}(S)$ and the following holds:

$$\forall s \in \hat{S}, \forall z \in \mathcal{Z}, \forall k = 1, \ldots, K: \text{ if } s, z \in \mathcal{C}_k, \text{ then } |\ell(\mathcal{A}_S, s) - \ell(\mathcal{A}_S, z)| \leq \epsilon(S).$$

Define $\hat{\zeta}(\mathcal{A}, S) = \max_{(k,i)\in[K]\times[n]} |\alpha_k(\mathcal{A}_S) - \ell(\mathcal{A}_S, z_i)|$ where $S = (z_1, \ldots, z_n)$ and $\alpha_k(h) = \mathbb{E}_z[\ell(h,z)|z \in \mathcal{C}_k]$.

F.1. Simple Version

Our first theorem is the analogue of Theorem 1.

Theorem 5. *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ pseudo robust (with $\{\mathcal{C}_k\}_{k=1}^K$), then for any $\delta > 0$, with probability at least $1-\delta$ over an iid draw of n examples $S = (z_i)_{i=1}^n$, the following holds:*

$$\begin{aligned}&\mathbb{E}_z[\ell(\mathcal{A}_S, z)]\\ &\leq \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) + \frac{\hat{n}(S)}{n}\epsilon(S) + \frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A}, S) + \zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right),\end{aligned}$$

where $\zeta(\mathcal{A}_S) := \max_{k\in[K]} \mathbb{E}_z[\ell(\mathcal{A}_S, z)|z \in \mathcal{C}_k]$, $\hat{\zeta}(\mathcal{A}, S) := \max_{(k,i)\in[K]\times[n]} |\alpha_k(\mathcal{A}_S) - \ell(\mathcal{A}_S, z_i)|$, and $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$ with $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$.

F.2. Stronger Version

The following statement is the analogue of Theorem 2 and is a strengthening of Theorem 5.

Theorem 6. *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ pseudo robust (with $\{\mathcal{C}_k\}_{k=1}^K$), then for any $\delta > 0$, with probability at least $1-\delta$ over an iid draw of n examples $S = (z_i)_{i=1}^n$, the following holds:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) + \frac{\hat{n}(S)}{n}\epsilon(S) + \frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A}, S) + \mathcal{Q}_1\sqrt{\frac{\ln(2K/\delta)}{n}} + \frac{2\mathcal{Q}_2\ln(2K/\delta)}{n}$$

where $\mathcal{Q}_1 := \sum_{k\in\mathcal{T}_S}(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) + \sqrt{2}\alpha_k(\mathcal{A}_S))\sqrt{\frac{|\mathcal{I}_k^S|}{n}}$, $\mathcal{Q}_2 := \alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)|\mathcal{T}_S| + \sum_{k\in\mathcal{T}_S}\alpha_k(\mathcal{A}_S)$, $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$ with $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$, $\alpha_k(h) := \mathbb{E}_z[\ell(h,z)|z \in \mathcal{C}_k]$, and $\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) := \max_{k\in\mathcal{T}_S^c}\alpha_k(\mathcal{A}_S)$ with $\mathcal{T}_S^c = [K] \setminus \mathcal{T}_S$.

F.3. Proof for pseudo robustness

We now prove Theorems 5 and 6.

Lemma 10. *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{n}(\cdot))$ pseudo robust, then the following holds for any $S \in \mathcal{Z}^n$:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) \leq \frac{\hat{n}(S)}{n}\epsilon(S) + \frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A}, S) + \sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right).$$

where $\hat{\zeta}(\mathcal{A}, S) := \max_{(k,i)\in[K]\times[n]} |\alpha_k(\mathcal{A}_S) - \ell(\mathcal{A}_S, z_i)|$.

Proof. Define $\hat{\mathcal{I}}_k := \{i \in [n] : z_i \in \hat{S}, z_i \in \mathcal{C}_k\}$. Then

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\left(\alpha_k(\mathcal{A}_S)-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(\mathcal{A}_S,z_i)\right)\right| \\
&=\left|\frac{1}{n}\sum_{k=1}^{K}\left(|\mathcal{I}_k|\alpha_k(\mathcal{A}_S)-\sum_{i\in\hat{\mathcal{I}}_k}\ell(\mathcal{A}_S,z_i)-\sum_{i\in\mathcal{I}_k\wedge i\notin\hat{\mathcal{I}}_k}\ell(\mathcal{A}_S,z_i)\right)\right| \\
&=\left|\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\hat{\mathcal{I}}_k}(\alpha_k(\mathcal{A}_S)-\ell(\mathcal{A}_S,z_i))+\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k\wedge i\notin\hat{\mathcal{I}}_k}(\alpha_k(\mathcal{A}_S)-\ell(\mathcal{A}_S,z_i))\right| \\
&\le\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\hat{\mathcal{I}}_k}|\alpha_k(\mathcal{A}_S)-\ell(\mathcal{A}_S,z_i)|+\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k\wedge i\notin\hat{\mathcal{I}}_k}|\alpha_k(\mathcal{A}_S)-\ell(\mathcal{A}_S,z_i)| \\
&\le\frac{\hat{n}(S)}{n}\epsilon(S)+\frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A},S),
\end{aligned}$$

where $\hat{\zeta}(\mathcal{A},S)=\max_{(k,i)\in[K]\times[n]}|\alpha_k(\mathcal{A}_S)-\ell(\mathcal{A}_S,z_i)|$. Combining this with Lemma 8 gives

$$\begin{aligned}
&\mathbb{E}_z[\ell(\mathcal{A}_S,z)]-\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S,z_i) \\
&=\frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\left(\alpha_k(\mathcal{A}_S)-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(\mathcal{A}_S,z_i)\right)+\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right) \\
&\le\frac{\hat{n}(S)}{n}\epsilon(S)+\frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A},S)+\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right).
\end{aligned}$$

□

We now have all the tools necessary to complete the proofs of the main theorems.

Proof of Theorem 5. By Lemma 10,

$$\mathbb{E}_z[\ell(\mathcal{A}_S,z)]-\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S,z_i)\le\frac{\hat{n}(S)}{n}\epsilon(S)+\frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A},S)+\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right).$$

By using Lemma 6 with $a_k(X)=\alpha_k(\mathcal{A}_S)$ and noticing that $a_{\mathcal{T}_S}(X), a_{\mathcal{T}_S^c}(X)\le\zeta(\mathcal{A}_S)$, we have that for any $\delta>0$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right)\le\zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}}+\frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right).$$

□

Proof of Theorem 6. By Lemma 10,

$$\mathbb{E}_z[\ell(\mathcal{A}_S,z)]-\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S,z_i)\le\frac{\hat{n}(S)}{n}\epsilon(S)+\frac{n-\hat{n}(S)}{n}\hat{\zeta}(\mathcal{A},S)+\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right).$$

By using Lemma 7 with $a_k(X)=\alpha_k(\mathcal{A}_S), a_{\mathcal{T}_S}(X)=\alpha_{\mathcal{T}_S}(\mathcal{A}_S)$, and $a_{\mathcal{T}_S^c}(X)=\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)$, we have that for any $\delta>0$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\alpha_k(\mathcal{A}_S)\left(\Pr(z\in\mathcal{C}_k)-\frac{|\mathcal{I}_k|}{n}\right)$$

$$\leq \sqrt{\frac{\ln(2K/\delta)}{n}} \left(\sum_{k\in\mathcal{T}_S} (\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S) + \sqrt{2}\alpha_k(\mathcal{A}_S))\sqrt{\frac{|\mathcal{I}_k|}{n}} \right) + \frac{2\ln(2K/\delta)}{n} \left(\alpha_{\mathcal{T}_S^c}(\mathcal{A}_S)|\mathcal{T}_S| + \sum_{k\in\mathcal{T}_S} \alpha_k(\mathcal{A}_S) \right).$$

□

G. Proof of Theoretical Comparisons

Proof of Example 4. First, in order to show the bound in Theorem 1 is tighter than the that of Proposition 1, we must show that

$$(\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n} \leq \sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}}. \tag{37}$$

It is not hard to see for any given δ, when $n > 2|\mathcal{T}_S|\ln(2K/\delta)$, $|\mathcal{T}_S| \ll K$, and $2K > 1/\delta$, the above inequality holds.

We can now divide each coordinate of $z^{(x)}$ into equal sized ν-length intervals (possibly excluding the last interval):

$$[-1,-1+\nu),[-1+\nu,-1+2\nu),\cdots[-1+i\nu,-1+(i+1)\nu),\cdots$$

Then, $\{\mathcal{C}_i\}$ is the Cartesian product of the intervals of each coordinate. Notice that by standard concentration for the maximum of a sequence of sub-gaussians, for any $\delta > 0$, there exists small enough $\sigma > 0$, such that $\|x^{(2)} - \mu\|_\infty < 1/\nu$ with probabilty at least $1-\delta$. Let us choose $\mu = c \cdot (1,1,\cdots,1)^\top$, where $c \in [-1,1]$ is the center of one of the intervals constructed above. Then, with probability at least $1-\delta$, $|\mathcal{T}_S| = \Theta((2/\nu)^{p+1})$. As a result, when $d \gg p$, we have the desired inequality.

Proof of Example 5. Recall the bound in Theorem 1 implies that

$$\mathbb{E}_z[\ell(\mathcal{A}_S,z)] \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S,z_i) + \epsilon(S) + \sqrt{\frac{B}{\lambda}}\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right) \tag{38}$$

On the other hand, the bound obtained via uniform stability is:

$$\mathbb{E}_z[\ell(\mathcal{A}_S,z)] \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}_S,z_i) + 2\beta + (4n\beta + \sqrt{\frac{B}{\lambda}})\sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{39}$$

When n and B are large, the dominating term is $\sqrt{\frac{B}{\lambda}}(\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}}$ and $4n\beta\sqrt{\frac{\ln(1/\delta)}{2n}}$, where we take $\beta = 2B^2/(\lambda n)$ as in Bousquet & Elisseeff (2002). We can divide $z^{(x)}$ into equal sized ν-length intervals (again with the possible exception of the last interval):

$$[0,\nu),[\nu,2\nu),\cdots,[i\nu,(i+1)\nu),\cdots$$

If there is no noise, i.e. $z^{(y)} = w^* z^{(x)}$, then all the points will fall on the line segment $(z^{(x)}, z^{(x)})$. When we have a Gaussian perturbation over $z^{(x)}$, by suitably choosing the variance parameter $\sigma > 0$, and concentration of Gaussian variables, we can let most of the data mass covered in the union of $\{\mathcal{C}_i\}$, where $\mathcal{C}_i = \{(x,y) : (x-\frac{i\nu}{2})^2 + (y-\frac{i\nu}{2})^2 \leq \frac{\nu^2}{2}\}$ is a circle with radius $\frac{\sqrt{2}\nu}{2}$ and its center is on the line segment $(z^{(x),z^{(x)}})$. We then have $|\mathcal{T}_S| \leq \Theta(2/\nu)$. Thus, when $B \gg 2/\nu$, our bound is strictly less than the uniform stability result (39).

H. Additional Experimental Results and Details

We report the additional experimental results in Figures 6, 7, and 8, where we can observe that our new bounds provide the significant improvements over the previous bounds.

For the real-world data, we adopted the standard benchmark datasets — MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009), CIFAR-100 (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), Fashion-MNIST (FMNIST) (Xiao et al., 2017), Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2019), and Semeion (Srl & Brescia, 1994). We used all the training samples exactly as provided by those datasets. For the synthetic data, we generated them by sampling the input $x \in \mathcal{X}$ from beta distributions and Gaussian mixture distributions with a variety of hyperparameters. Beta(α, β) indicates the Beta distribution with hyper-parameters α and β. Gauss mix (σ) means the mixture of five Gaussian distributions with a standard deviation σ. Beta mix (α, β)-(σ) represents the mixture of beta distributions generated by the following procedure:

$$x = 0.4 * v_0 + v_1 + v_2,$$

where v_0 is drawn from the uniform distribution on $[0, 1]$, $v_1 \sim$ Beta(α, β), and $v_2 \sim$ Beta(σ, σ). Similarly, beta-Gauss (α, β)-(σ) represents the mixture of distributions generated by the following procedure:

$$x = 0.4 * v_0 + v_1 + v_2,$$

where v_0 is drawn from the uniform distribution on $[0, 1]$, $v_1 \sim$ Beta(α, β), and v_2 is drawn from the Gaussian distribution with a standard deviation σ. For all the synthetic data, we generated and used 1000 training data points.

For the partition $\{\mathcal{C}_k\}_{k=1}^K$, we consider the division of the input space $\mathcal{X}$ because we can either (1) assume that there exists a function $\hat{y}$ such that $y = \hat{y}(x)$ or (2) notice that the partition $\{\mathcal{C}_k\}_{k=1}^K$ of the input space $\mathcal{X}$ can be dictated by the label $y \in \mathcal{Y}$; i.e., $K = |\mathcal{Y}| \times K'$ where K' is the size of the partition of the input space $\mathcal{X}$, which is used in the previous paper (Xu & Mannor, 2012). Thus, we can focus on partition of the input space $\mathcal{X}$ for the purpose of comparing K and $|\mathcal{T}_S|$.

The ϵ-covering of $\mathcal{X} \subseteq [0, 1]^d$ can be defined by the following. We first define

$$\mathcal{C}'_{k_1,\ldots,k_d} = \{x \in \mathcal{X} : 0.1(k_j - 1) \leq x_j < 0.1k_j + \mathbb{1}(k_j = 10), j = 1, \ldots, d\}, \qquad k_1, \ldots, k_d \in \{1, \ldots, 10\},$$

where $\mathbb{1}(k_j = 10)$ is one if $k_j = 10$ and is zero otherwise. Note that without this notation of $\mathbb{1}(k_j = 10)$, equivalently, we can define the condition by $0.1(k_j - 1) \leq x_j \leq 0.1k_j$ if $k_j = 10$ and $0.1(k_j - 1) \leq x_j < 0.1k_j$ if $k_j < 10$, since $\mathcal{X} \subseteq [0, 1]^d$. We then define $\mathcal{C}_k$ to be the flatten version of $\mathcal{C}'_{k_1,\ldots,k_d}$; i.e., $\{\mathcal{C}_k\}_{k=1}^K = \{\mathcal{C}'_{k_1,\ldots,k_d}\}_{k_1,\ldots,k_d \in [10]}$ with $C_1 = \mathcal{C}'_{1,1,\ldots,1}$, $C_2 = \mathcal{C}'_{2,1,\ldots,1}$, $C_{10} = \mathcal{C}'_{10,1,\ldots,1}$, $C_{10+1} = \mathcal{C}'_{1,2,1,\ldots,1}$, $C_{20} = \mathcal{C}'_{10,2,1,\ldots,1}$, and so on. While the ϵ-covering of the original input space $\mathcal{X}$ is the default example from the previous paper (Xu & Mannor, 2012), in Figure 1 we see that K grows rapidly as d increases. Therefore, to reduce K significantly, we also propose utilizing the inverse image of the ϵ-covering in a randomly projected space. That is, given a random matrix A, we use the ϵ-covering of the space of $u = Ax$ to define the pre-partition $\{\tilde{\mathcal{C}}_k\}_{k=1}^K$. More concretely, the random matrix A for the projection was generated by the following procedure:

1. Each entry of a random matrix $\tilde{A}$ is generated by the Uniform Distribution on $[0, 1]$ independently.
2. Each row of the random matrix $\tilde{A} \in \mathbb{R}^{3 \times d}$ is then normalized so that $Ax \in [0, 1]^3$; i.e.,

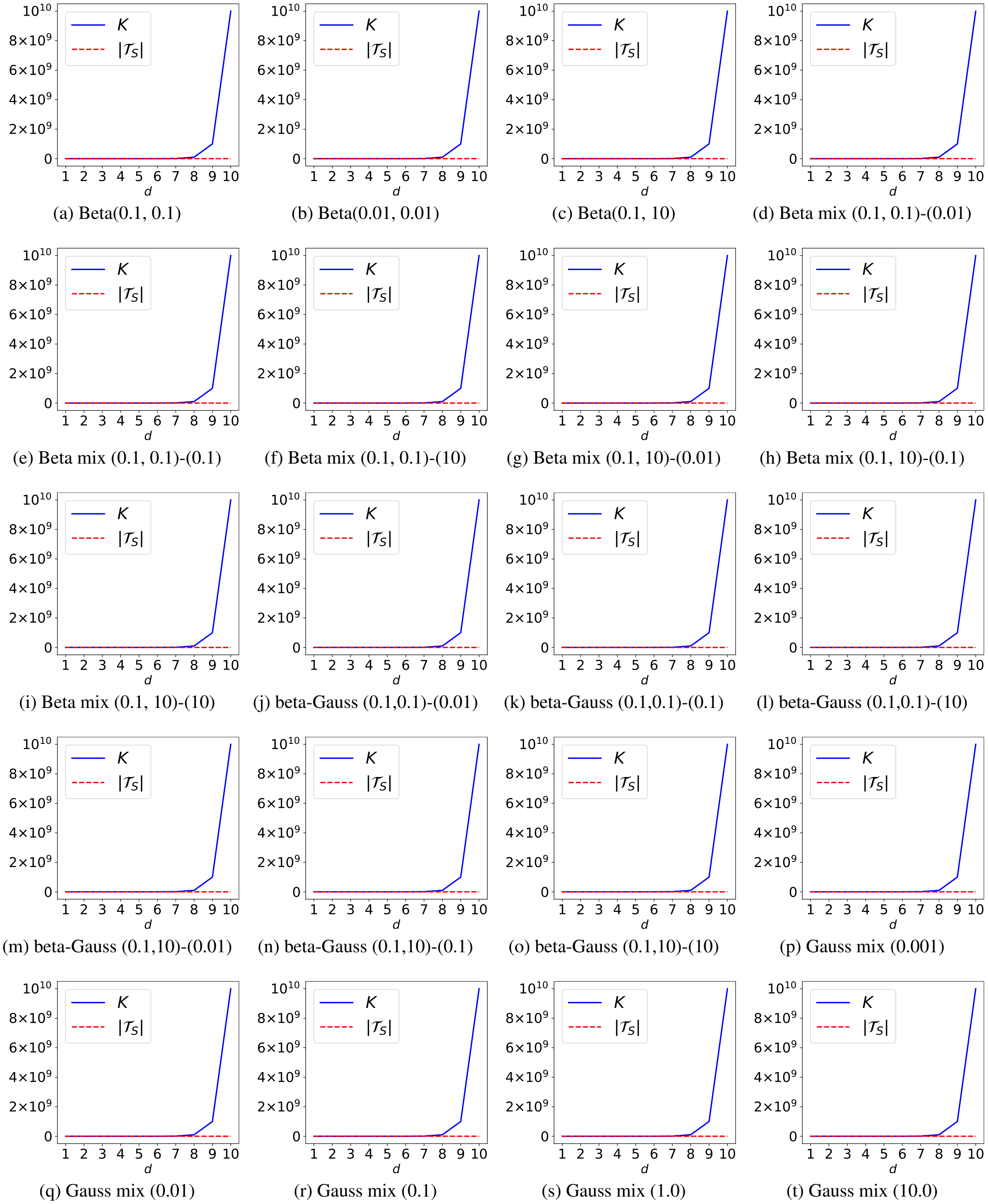
$$A_{ij} = \frac{\tilde{A}_{ij}}{\sum_{j=1}^d \tilde{A}_{ij}}.$$

Then, we can define

$$\tilde{\mathcal{C}}'_{k_1,k_2,k_3} = \{u \in [0, 1]^3 : 0.1(k_j - 1) \leq x_j < 0.1k_j + \mathbb{1}(k_j = 10), j = 1, 2, 3\}, \qquad k_1, k_2, k_3 \in \{1, \ldots, 10\}.$$

We then define $\tilde{\mathcal{C}}_k$ to be the flatten version of $\tilde{\mathcal{C}}'_{k_1,\ldots,k_d}$; i.e., $\{\tilde{\mathcal{C}}_k\}_{k=1}^K = \{\tilde{\mathcal{C}}'_{k_1,\ldots,k_d}\}_{k_1,\ldots,k_d \in [10]}$. Finally, the partition $\{\mathcal{C}_k\}_{k=1}^K$ is defined by $\mathcal{C}_k = \{x \in \mathcal{X} : Ax \in \tilde{\mathcal{C}}_k\}$. In this study, we randomly generated matrix $A \in \mathbb{R}^{3 \times d}$ in each trial.

For the clustering with unlabeled data, given a set of unlabeled data points $\{\bar{x}_k\}_{k=1}^K$, the partition $\{\mathcal{C}_k\}_{k=1}^K$ is defined by $\mathcal{C}_k = \{x \in \mathcal{X} : k = \operatorname{argmin}_{k' \in [K]} \|x - \bar{x}_{k'}\|_2\}$. Following the literature on semi-supervised learning, we randomly split the training data points into labeled data points (500 for Semeion and 5000 for all other datasets) and unlabeled data points (the remainder of all the training data).



(a) Beta(0.1, 0.1)

(b) Beta(0.01, 0.01)

(c) Beta(0.1, 10)

(d) Beta mix (0.1, 0.1)-(0.01)

(e) Beta mix (0.1, 0.1)-(0.1)

(f) Beta mix (0.1, 0.1)-(10)

(g) Beta mix (0.1, 10)-(0.01)

(h) Beta mix (0.1, 10)-(0.1)

(i) Beta mix (0.1, 10)-(10)

(j) beta-Gauss (0.1,0.1)-(0.01)

(k) beta-Gauss (0.1,0.1)-(0.1)

(l) beta-Gauss (0.1,0.1)-(10)

(m) beta-Gauss (0.1,10)-(0.01)

(n) beta-Gauss (0.1,10)-(0.1)

(o) beta-Gauss (0.1,10)-(10)

(p) Gauss mix (0.001)

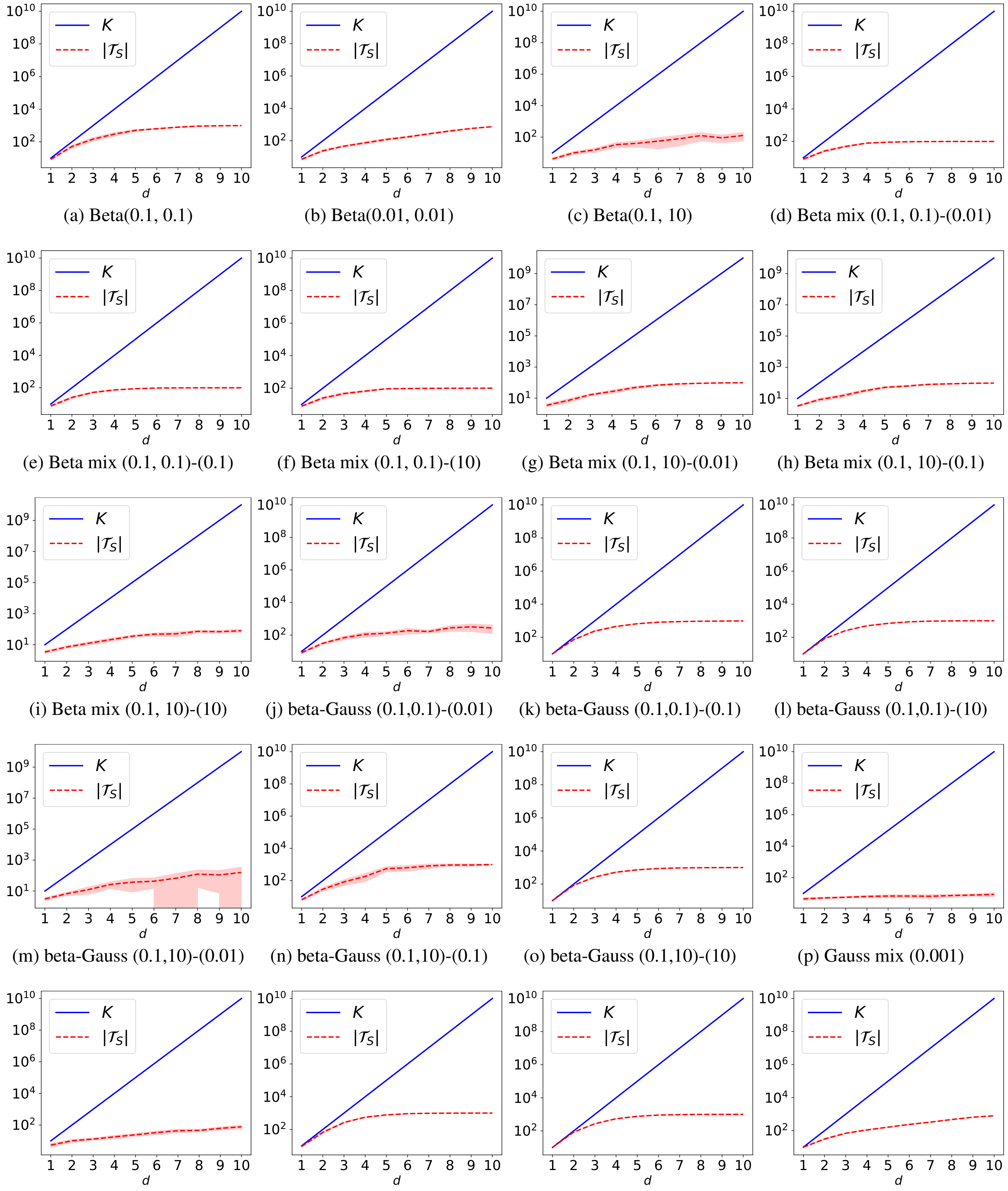
(q) Gauss mix (0.01)

(r) Gauss mix (0.1)

(s) Gauss mix (1.0)

(t) Gauss mix (10.0)

Figure 6. The values of K versus $|\mathcal{T}_S|$ with the ϵ-covering of the original space. The figures display the mean of 10 random trials along with one standard deviation.



(a) Beta(0.1, 0.1)

(b) Beta(0.01, 0.01)

(c) Beta(0.1, 10)

(d) Beta mix (0.1, 0.1)-(0.01)

(e) Beta mix (0.1, 0.1)-(0.1)

(f) Beta mix (0.1, 0.1)-(10)

(g) Beta mix (0.1, 10)-(0.01)

(h) Beta mix (0.1, 10)-(0.1)

(i) Beta mix (0.1, 10)-(10)

(j) beta-Gauss (0.1,0.1)-(0.01)

(k) beta-Gauss (0.1,0.1)-(0.1)

(l) beta-Gauss (0.1,0.1)-(10)

(m) beta-Gauss (0.1,10)-(0.01)

(n) beta-Gauss (0.1,10)-(0.1)

(o) beta-Gauss (0.1,10)-(10)

(p) Gauss mix (0.001)

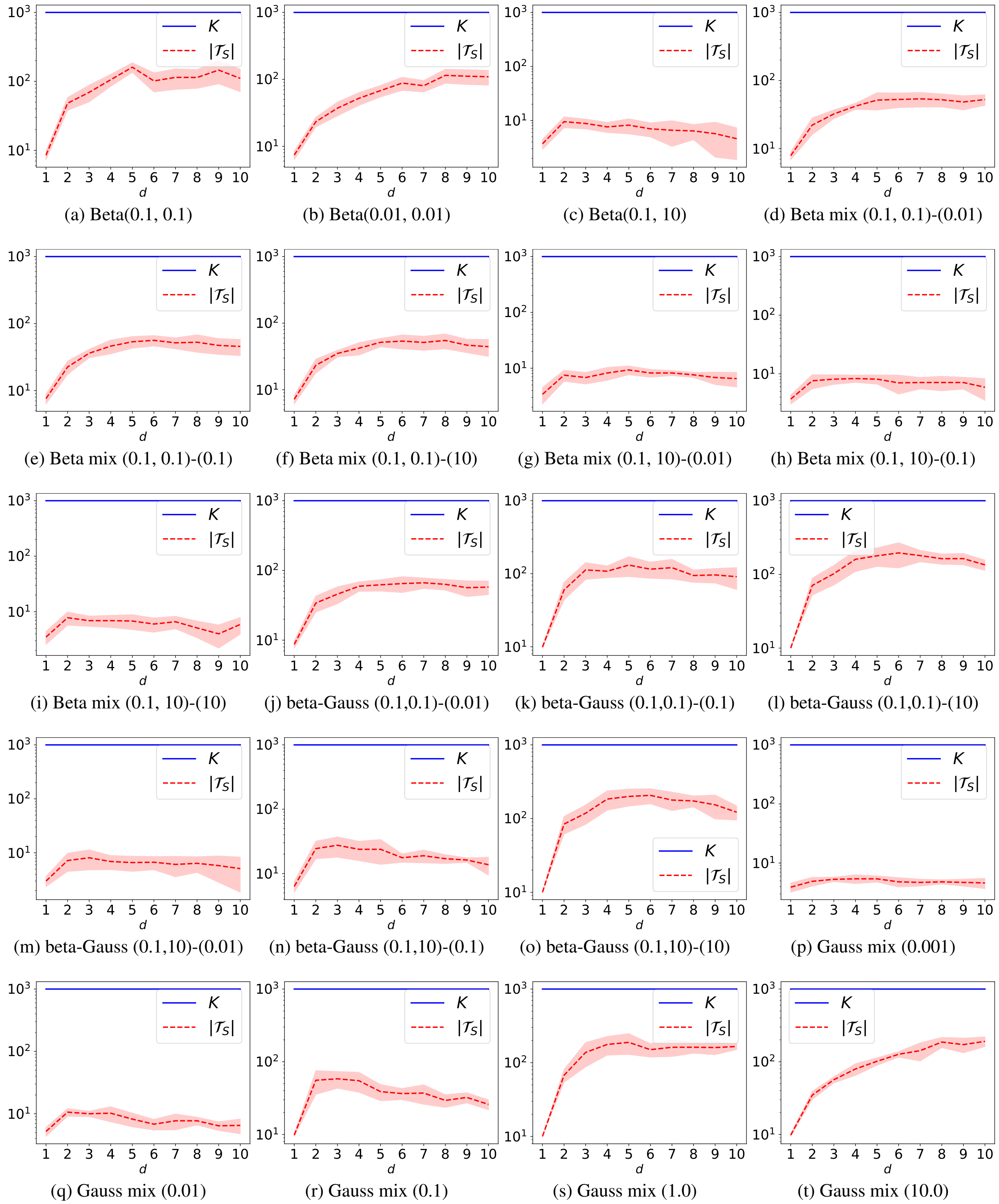
(q) Gauss mix (0.01)

(r) Gauss mix (0.1)

(s) Gauss mix (1.0)

(t) Gauss mix (10.0)

Figure 7. The values of K versus $|\mathcal{T}_S|$ with the ϵ-covering of the original space. These figures show the mean of 10 random trials and one standard deviation. They are plotted on a *logarithmic* scale to show the discrepancy in the rates of growth.



(a) Beta(0.1, 0.1)

(b) Beta(0.01, 0.01)

(c) Beta(0.1, 10)

(d) Beta mix (0.1, 0.1)-(0.01)

(e) Beta mix (0.1, 0.1)-(0.1)

(f) Beta mix (0.1, 0.1)-(10)

(g) Beta mix (0.1, 10)-(0.01)

(h) Beta mix (0.1, 10)-(0.1)

(i) Beta mix (0.1, 10)-(10)

(j) beta-Gauss (0.1,0.1)-(0.01)

(k) beta-Gauss (0.1,0.1)-(0.1)

(l) beta-Gauss (0.1,0.1)-(10)

(m) beta-Gauss (0.1,10)-(0.01)

(n) beta-Gauss (0.1,10)-(0.1)

(o) beta-Gauss (0.1,10)-(10)

(p) Gauss mix (0.001)

(q) Gauss mix (0.01)

(r) Gauss mix (0.1)

(s) Gauss mix (1.0)

(t) Gauss mix (10.0)

Figure 8. The values of K versus $|\mathcal{T}_S|$ with the inverse image of the ϵ-covering in randomly projected spaces. These figures show the mean of 10 random trials and one standard deviation. Here, the input space was first randomly projected onto a space of dimension 3. This data is plotted on a logarithmic scale.