Rate-Distortion Theoretic Bounds on Generalization Error for Distributed Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we use tools from rate-distortion theory to establish new upper bounds on the generalization error of statistical distributed learning algorithms. Specifically, there are K clients whose individually chosen models are aggregated by a central server. The bounds depend on the compressibility of each client's algorithm while keeping other clients' algorithms un-compressed, and leverage the fact that small changes in each local model change the aggregated model by a factor of only $1/K$. Adopting a recently proposed approach by Sefidgaran et al., and extending it suitably to the distributed setting, this enables smaller rate-distortion terms which are shown to translate into tighter generalization bounds. The bounds are then applied to the *distributed* support vector machines (SVM), *suggesting* that the generalization error of the distributed setting decays faster than that of the centralized one with a factor of $\mathcal{O}(\log(K)/\sqrt{K})$. A similar conclusion is obtained for a *multiple-round federated learning* setup where each client uses *stochastic gradient Langevin dynamics* (SGLD).

1 Intorduction

A key performance indicator of any stochastic learning algorithm that uses a given finite set of data points is how well it performs on points that are outside that set, i.e., unseen data. This is often captured through the so-called *generalization error*. The questions of what really controls the generalization error of a given stochastic algorithm, and how to make it sufficiently small, are still not yet well understood, however. For example, while classic approaches [SSBD14] suggest that algorithms with over-parametrized models are likely to *overfit*, it is now known that there exist a few such ones which do generalize well [ZBH$^+$17]. Common approaches to studying the generalization error of a statistical learning algorithm often consider the *effective hypothesis space* induced by the algorithm, rather than the entire hypothesis space, or the information leakage about the *training dataset*. Examples include information-theoretic (mutual information) approaches [RZ16, XR17, HRVSG21, HDMR21, NHD$^+$20, SZ20], compression-based approaches [AGNZ18, SAN20, HJTW21, BSE$^+$21, KLG$^+$21] and intrinsic-dimension or fractal based approaches [ŞSDE20, BLGŞ21, HŞKM21]. Recently, a novel approach [SGRS22] that generalizes the notion of algorithm compressibility by using lossy covering from source coding concepts was used to show that the compression error rate of an algorithm is strongly connected to its generalization error both in expectation and with high probability; and, consequently, establish new rate-distortion-based bounds on the generalization error. The bounds of [SGRS22] were shown to possibly improve strictly upon those of [XR17, BZV20] and [SZ20]. The approach also has the advantage to offer a unifying perspective on mutual-information, compressibility, and fractal-based frameworks.

Another major focus of machine learning research over the recent years has been the study of statistical learning algorithms when applied in distributed (network or graph) settings. In part, this is due to the emergence of new applications in which resources are constrained, data is distributed or the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

need to preserve privacy [VWK+20, MMR+17, KMY+16, KLN+11]. Example algorithms include the now popular *Federated Learning* [YAE+18, KMA+19, LSTS20, RCZ+20, KKM+20, YZR21], the *Split Learning* of [GR18], *Group Alternating Direction Method of Multipliers* [EPB+20], and the so-called *in-network learning* of [AZ19, MZ21]. Despite its importance, however, little is known about the generalization guarantees of distributed statistical learning algorithms. In fact, in the distributed learning setting the technical challenges are numerous, including the lack of a proper definition of the generalization error in this case [YMNS22, MSS19]. Notable exception works in this direction include [YDVP20] and [BDP22]. In [YDVP20], information-theoretic upper bounds on the generalization error of distributed statistical learning algorithms are obtained merely by viewing the entire distributed system, from the input data to each (local) algorithm to the output aggregated model, as a single (centralized) algorithm and applying to it the bounds of [BZV20, XR17]. While somewhat useful this, however, has left the difference between the bounds for the distributed learning setting and their counterparts for the centralized learning setting implicit in the involved mutual information terms. The problem of studying the generalization error of distributed statistical learning algorithms was further studied in the recent [BDP22]. Therein, using results from [BZV20], the authors establish bounds on the expectation of the generalization error for two special cases. For the first, linear or location models with Bergman divergence loss, the proposed bound on the generalization error for the distributed setup [BDP22, Theorem 4] is shown better (i.e., smaller) than its counter-part for the centralized learning setting by a factor of $\mathcal{O}(1/\sqrt{K})$, where K is the number of clients. This result, however, relies strongly on the assumed linearity of the loss with respect to the hypothesis. For the second, Lipschitz continuous loss [BDP22, Theorem 5], a similar behavior is shown by reducing the problem to the centralized case and using the triangle inequality.

In this work, we study the generalization error of distributed statistical learning algorithms. Essentially we extend suitably the approach of [SGRS22] to establish rate-distortion theoretic upper bounds on the generalization error. In doing so, we bring analysis of the distributed architecture of the learning problem into the bounds. The bounds, which hold with high probability and in expectation, allow to only consider the compressibility of each local algorithm – the latter having effect with a factor of only $1/K$ on the aggregated model order-wise. This is shown to result in more relaxed distortion criterion for the *local algorithm compressibility*, smaller rate-distortion terms; and, in turn, better generalization bounds. Furthermore, we apply our results to the *distributed support vector machines* (DSVM). The obtained bounds suggest that for the non-separable data, the generalization error of the distributed setting decays faster with that of the centralized one with a factor of $\log(K)/\sqrt{K}$. We also consider the related Federated learning setting; and derive bounds on the generalization error in two setups: when each client applies the *stochastic gradient Langevin dynamics* (SGLD) method and locally deterministic algorithms with Lipschitz loss. These latter results are deferred to the supplemental material for lack of space. In both cases our bounds suggest a decreasing behavior for the generalization performance as K grows.

Notation. Random variables, their realizations, and their domains are denoted respectively by upper-case, lower-case, and calligraphy fonts, *e.g.,* X, x, and $\mathcal{X}$. Their distributions and expectations are denoted by P_X and $\mathbb{E}[X]$. The random variable X is called *σ-subgaussian* if for all $t \in \mathbb{R}$, $\mathbb{E}[\exp(t(X - \mathbb{E}[X]))] \leqslant \exp(\sigma^2 t^2/2)$. A vector of $m \in \mathbb{N}$ numbers (or random variables) $(x_1, \ldots, x_m)$ are denoted by either $x_{1:m}$ or $\{x_i\}_{i=1}^m$, depending on the context, and the vector $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$ is denoted by $x_{1:m\setminus i}$. Similarly for $n, m \in \mathbb{N}$, a vector $(x_{1,1}, \ldots, x_{1,m}, x_{2,1}, \ldots, x_{2,m}, x_{n,1}, \ldots, x_{n,m})$ is denoted by $x_{1:n,1:m}$ or $\{x_{i,1:m}\}_{i\in[n]}$, or $\{x_{1:n,j}\}_{j\in[m]}$, where $[n] = \{1, \ldots, n\}$. Parts of our results are stated in terms of information-theoretic quantities: for random variables X and Y, we denote the differential *entropy* of X by $h(X)$, the *conditional differential entropy* of X given Y by $h(X|Y)$, and the mutual information between them by $I(X;Y)$. Moreover, the *Kullback–Leibler* (KL) divergence between distributions Q and P is denoted by $D_{KL}(Q\|P)$. For more details, we refer the reader to [CT06, PW14].

2 Preliminaries and problem setup

For convenience, we start with a brief review of the standard (centralized) statistical learning setup together with few definitions and recent results associated to it. Let the *input data* Z be distributed according to an unknown distribution μ over the *data space* $\mathcal{Z}$. A *training dataset* $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$ consists of n samples $\{Z_i\}$ generated independently each according to μ. A possibly stochastic learning algorithm $\mathcal{A}\colon \mathcal{Z}^n \mapsto \mathcal{W}$ (e.g., stochastic gradient descent) assigns an hypothesis $\mathcal{A}(S) = W$ chosen from the *hypothesis class* $\mathcal{W} \in \mathbb{R}^d$ to every $S \in \mathcal{Z}^n$. The map $\mathcal{A}$ induces

a conditional distribution $P_{W|S}$ which together with μ induce the joint dataset-hypothesis distribution $P_{S,W} = \mu^{\otimes n} P_{W|S}$. The quality of the prediction is measured using a loss function $\ell\colon \mathcal{Z} \times \mathcal{W} \mapsto \mathbb{R}^+$. The *generalization error* of an algorithm $\mathcal{A}$ is defined as $\mathrm{gen}(s,w) := \mathcal{L}(w) - \hat{\mathcal{L}}(s,w)$ where $\mathcal{L}(w) := \mathbb{E}_{Z\sim\mu}[\ell(z,w)]$ denotes the population risk and $\hat{\mathcal{L}}(s,w) := \frac{1}{n}\sum_{j\in[n]} \ell(z_j,w)$ denotes the empirical risk. In the binary classification context, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} = \{-1,+1\}$, we often consider the 0-1 loss function $\ell_0(z,w) := \mathbb{1}_{\{yf(x,w)<0\}}$, where the sign of $f(x,w)$, $f\colon \mathcal{X} \times \mathcal{W} \mapsto \mathbb{R}$, is the label prediction by hypothesis w and $\mathbb{1}$ is the indicator function. In this setup, it is common to assess the empirical risk with respect to 0-1 loss function with margin $\theta \in \mathbb{R}^+$ defined as $\ell_\theta(z,w) := \mathbb{1}_{\{yf(x,w)<\theta\}}$, while using 0-1 loss function for the population risk evaluation. We denote the corresponding empirical risk as $\hat{\mathcal{L}}_\theta(s,w) := \frac{1}{n}\sum_{j\in[n]} \ell_\theta(z_j,w)$ and the generalization error as $\mathrm{gen}_\theta(s,w) := \mathcal{L}(w) - \hat{\mathcal{L}}_\theta(s,w)$.

The exact analysis of the generalization error $\mathrm{gen}(S,W)$ seems out of reach; and, for this reason, as already mentioned upper bounds on it were developed using an information-theoretic approach. Essentially, such approach connects the generalization error of a statistical learning algorithm $\mathcal{A}$ with the mutual information between the input data sample S and the algorithm output $W = \mathcal{A}(S)$ (for details the reader may refer to a line of work that was initiated by Russo and Zou [RZ16] and Xu and Raginsky [RZ16] and since then improved by using various conditional versions of mutual information). Other approaches rely on the observation that the output W can be *compressible* in some suitable sense [AGNZ18] or the algorithm might generate a *fractal structure* [ŞSDE20].

Very recently, an approach [SGRS22] that relies on probabilistic ϵ-covering from source coding concepts was proposed and shown to possibly improve strictly over the aforementioned, seemingly un-related, approaches, while offering a unifying framework to them. The upper bounds of [SGRS22] are rate-distortion theoretic. Specifically, for every distribution Q defined over $\mathcal{S} \times \mathcal{W}$, let

$$\mathfrak{RD}(Q,\epsilon) := \inf_{P_{\hat{W}|S}} I(S;\hat{W}), \quad \text{s.t.} \quad \mathbb{E}[\mathrm{gen}(S,W) - \mathrm{gen}(S,\hat{W})] \leqslant \epsilon,$$

where $\hat{W} \in \hat{\mathcal{W}}$, $\mathrm{gen}(S,\hat{W})$ is defined with respect to a loss function $\hat{\ell}\colon \mathcal{Z} \times \hat{\mathcal{W}} \mapsto \mathbb{R}^+$ (possibly different from ℓ), the mutual information and expectation are with respect to $Q_S P_{\hat{W}|S}$ and $Q P_{\hat{W}|S}$, respectively, Q_S is the Q-marginal of S, and the infimum is over all Markov kernels $P_{\hat{W}|S}$.

This rate-distortion function $\mathfrak{RD}(Q,\epsilon)$ is the adaptation of the rate-distortion function emerged in the lossy source compression context [Ber75, CT06] to stochastic learning algorithms [SGRS22]. In the lossy source compression context [Ber75, CT06], this function quantifies the fundamental compression rate of a source $X \sim P_X$ to within some desired average distortion level ϵ. To this end, infinitely many i.i.d. instances of the source ($\{X_i\}_{i\in[m]}$, $X_i \sim P_X$, and $m \to \infty$) are compressed (quantized) simultaneously. The joint compression approach is known as the *block-coding* technique. In the learning context, this term quantifies the *lossy algorithm compressibility* [SGRS22], with the following intuition for the case where $\mathcal{Z}$ and $\mathcal{W}$ are finite sets. Each $P_{\hat{W}|S}$ denotes a learning algorithm and the distortion criterion $\mathbb{E}[\mathrm{gen}(S,W) - \mathrm{gen}(S,\hat{W})] \leqslant \epsilon$ guarantees that in average the difference between the generalization error of the original algorithm $P_{W|S}$ and that of the "compressed" algorithm $P_{\hat{W}|S}$ does not exceed ϵ. The rate-distortion term $\mathfrak{RD}(P_{S,W},\epsilon)$ quantifies the lossy compressibility of the algorithm $P_{W|S}$ in the following sense: for large $m \in \mathbb{N}$ and every admissible $P_{\hat{W}|S}$, a compressed hypothesis space of size $N_m \approx e^{mI(S;\hat{W})}$ can be found such that, with high probability, for each m i.i.d. instances of the original algorithm $P_{W|S}$ there exists at least one compressed hypothesis for which the difference between the average generalization error over the m i.i.d. instances of $P_{W|S}$ at hand and that of the found hypothesis does not exceed ϵ. Recall the tail bound on the generalization error of [SGRS22, Theorem 9].

Theorem 1 ([SGRS22, Theorem 9]). *Suppose that the learning algorithm $\mathcal{A}(S)$ induces $P_{S,W}$ and suppose that for all $\hat{w} \in \hat{\mathcal{W}}$, $\hat{\ell}(Z,\hat{w})$ is σ-subgaussian. Then, for any $\epsilon \in \mathbb{R}$ and $\delta \geqslant 0$, with probability at least $1-\delta$, $\mathrm{gen}(S,W) \leqslant \sqrt{2\sigma^2(R_p(\delta,\epsilon) + \log(1/\delta))/n} + \epsilon$, where $R_p(\delta,\epsilon) := \sup_{Q\colon D_{KL}(Q\|P_{S,W}) \leqslant \log(1/\delta)} \mathfrak{RD}(Q,\epsilon)$ and the supremum is over all possible distributions over $\mathcal{S} \times \mathcal{W}$.*

The above bound depends not only $\mathfrak{RD}(P_{S,W},\epsilon)$ but also on $\mathfrak{RD}(Q,\epsilon)$ terms for all distributions Q that are close enough to P. In other words, to guarantee a good generalization performance, the compressibility of the algorithm under every such small perturbation of $P_{S,W}$ needs to be considered.

Remark 1. *The reader may notice two (minor) differences between Theorem 1 as stated here and [SGRS22, Theorem 9]. First, we here express the rate-distortion terms with respect to the distance* $d(w;\hat{w}|s) \coloneqq \mathrm{gen}(s,w) - \mathrm{gen}(s,\hat{w})$ *instead of a possibly smaller distance* $d'(w;\hat{w}|s) \coloneqq \inf_{(s',w')\in\mathrm{supp}(Q)} \mathrm{gen}(s',w') - \mathrm{gen}(s,\hat{w})$ *that was considered in [], where* $\mathrm{supp}(Q)$ *denotes the support of* Q*. Even-though the latter could possibly lead to stronger results, we do not consider it here because it is less amenable to computations. Second, in Theorem 1 the loss function* $\hat{\ell}(z,\hat{\overline{w}})$ *is allowed to differ from the original* $\ell(z,w)$*; and this possibly leads to some (rather small) improvement of the result of [SGRS22, Theorem 9]. The mentioned two small differences, however, do not require any change in the proof of [SGRS22, Theorem 9].*

Problem setup. In this work, we consider a *homogeneous distributed learning* setup that exploits the participation of K clients, as described in the following. Each client $i \in [K]$ has access to a training dataset $S_i = \{Z_{i,1},\dots,Z_{i,n}\} \sim \mu^{\otimes n}$ of size n, drawn independently of each other and independently of other clients' training datasets from a same distribution μ. The local learning algorithm $\mathcal{A}_i$ at each client picks a hypothesis $\mathcal{A}_i(S_i) = W_i \in \mathcal{W}_i = \mathcal{W}$ according to $P_{W_i|S_i}$. The induced joint distribution of (S_i, W_i) is denoted by P_{S_i,W_i}. The server receives the hypotheses $W_{1:K}$ and picks the hypothesis $\overline{W}$ as

$$\overline{W} \coloneqq (W_1 + \cdots + W_K)/K.$$

We denote the distributed learning algorithm as $\mathcal{A}_{1:K}(S_{1:K})$. It induces the joint distribution $P_{S_{1:K},W_{1:K},\overline{W}} = P_{\overline{W}|W_{1:K}} \prod_{i\in[K]} P_{S_i,W_i}$, where $P_{S_i} = \mu^{\otimes n}$ and $P_{\overline{W}|W_{1:K}} = \mathbb{1}_{\{\overline{W}=(W_1+\cdots+W_K)/K\}}$. Similar to [YDVP20, BDP22], the population and empirical risks are defined as

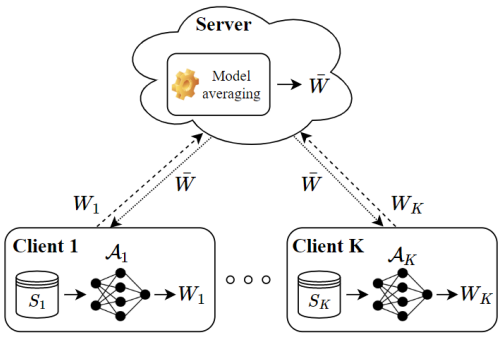

Figure 1: The considered distributed setup.

$$\mathcal{L}(\overline{w}) \coloneqq \mathbb{E}_{Z\sim\mu}[\ell(Z,\overline{w})], \quad \hat{\mathcal{L}}(s_{1:K},\overline{w}) \coloneqq \frac{1}{K}\sum\nolimits_{i\in[K]} \hat{\mathcal{L}}(s_i,\overline{w}). \tag{1}$$

The main goal of the paper is to upper bound the generalization error $\mathrm{gen}(s_{1:K},\overline{w}) \coloneqq \mathcal{L}(\overline{w}) - \hat{\mathcal{L}}(s_{1:K},\overline{w})$ for the described (one-round) distributed learning setting as well as multi-round extension of it, i.e., Federated Learning. Results for the latter are deferred to the supplemental material.

3 Information-theoretic bounds on the generalization error

In this section, we establish information-theoretic bounds on the generalization error of any stochastic distributed learning algorithm defined as in the previous section. It is important to note that the bounds hold for $\overline{W}$ being any stochastic function of $W_{1:K}$. For the ease of the exposition, however, we only focus on the deterministic average case. The bounds are applied later for the DSVM and federated SGLD, suggesting a decreasing generalization error behavior as number of clients increases.

3.1 Tail bound

A trivial approach that was already considered and used in [YDVP20] to establish in-expectation bounds for the problem studied therein consists in thinking a distributed learning algorithm that is composed of K local algorithms and an aggregator at the server as a single centralized algorithm, by considering the end-to-end system from the input at each local algorithm to the final aggregated model $\overline{W}$ generated as $P_{\overline{W}|S_{1:K}}$. The idea also applies to tail bounds. Hence, known tail and in-expectation bounds on the generalization error of centralized learning algorithms translate trivially into (generally loose) counter-part bounds for the distributed learning setting. Accordingly, the next theorem follows easily using the result of the above Theorem 1.

Theorem 2. *Suppose that the distributed learning algorithm* $\mathcal{A}_{1:K}(S_{1:K})$ *induces* $P_{S_{1:K},\overline{W}}$ *and suppose that for all* $\hat{\overline{w}} \in \hat{\mathcal{W}}$*,* $\hat{\ell}(Z,\hat{\overline{w}})$ *is* σ*-subgaussian. Then, for any* $\epsilon \in \mathbb{R}$ *and* $\delta \geqslant 0$*, with*

probability at least $1-\delta$, $\mathrm{gen}\big(S_{1:K}, \overline{W}\big) \leqslant \sqrt{2\sigma^2(R_p(\delta,\epsilon)+\log(1/\delta))/(nK)}+\epsilon$, *where* $R_p(\delta,\epsilon) \coloneqq \sup_{Q\colon D_{KL}(Q\|P_{S_{1:K},\overline{W}})\leqslant \log(1/\delta)} \mathfrak{RD}(Q,\epsilon)$ *and the supremum is over all possible distributions over* $\mathcal{W} \times \prod_{i\in[K]} \mathcal{S}_i$.

In the distributed setup, for every $i \in [K]$ only the hypothesis W_i depends on the dataset S_i, which has then the effect with a factor of $1/K$ on $\overline{W}$. The result of Theorem 2, however, does not explicitly take the structure of the distributed learning problem into account and, instead, it considers the joint compressibility of all local algorithms and $P_{\overline{W}|S_{1:K}}$. Thus, the dependency of the bound on the structure of the problem is considered only implicitly, via the conditional distribution $P_{\overline{W}|S_{1:K}}$. In this work, we establish an alternate bound which is tailored specifically for the considered distributed setup. First, for every $i \in [K]$ we let the *compressed* hypothesis $\hat{\overline{W}}_i$ of $\overline{W}$ to depend on both S_i and $W_{1:K\setminus i}$ (recall that $W_{1:K\setminus i}$ is independent from S_i). We then establish an upper bound on the generalization error in terms of the maximum (over all clients $i \in [K]$) of the minimum achievable compressibility using $P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$. The advantage can be exemplified as follows. For $i \in [K]$, consider only compressing the local algorithm $P_{W_i|S_i}$ and pick some $\hat{W}_i$ while keeping $W_{1:K\setminus i}$ un-compressed. Let $\hat{\overline{W}}_i$ be average of $\hat{W}_i$ and $W_{1:K\setminus i}$. As $\hat{W}_i$ has the effect with factor $1/K$ on $\hat{\overline{W}}$, in order to meet the distortion constraint $\mathbb{E}[\mathrm{gen}(S_i,\overline{W}) - \mathrm{gen}(S_i,\hat{\overline{W}}_i)] \leqslant \epsilon$ a more relaxed distortion criterion would be needed in compressing the local algorithm $P_{W_i|S_i}$. More precisely, for example when the loss is $\mathfrak{L}$-Lipschitz with L_2^2-norm, *i.e.,* $|\ell(z,w)-\ell(z,w')| \leqslant \mathfrak{L}\|w-w'\|^2$, a local distortion of $K^2\epsilon$ results into only ϵ distortion in the aggregated model.[1] Hence, this translates into smaller rate-distortion terms and, in turn, tighter bounds on the generalization error.

Now, we formally state the result. For a distribution Q defined over $\mathcal{W} \times \prod_{i\in[K]}(\mathcal{S}_i \times \mathcal{W}_i)$, let

$$\mathfrak{RD}_i(Q,\epsilon) \coloneqq \inf_{P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}} I(S_i;\hat{\overline{W}}_i|W_{1:K\setminus i}), \tag{2}$$

$$\text{s.t.} \quad \mathbb{E}\Big[\mathrm{gen}\big(S_i,\overline{W}\big) - \mathrm{gen}\Big(S_i,\hat{\overline{W}}_i\Big)\Big] \leqslant \epsilon, \tag{3}$$

where $\mathrm{gen}(S_i,\overline{W}) \coloneqq \mathcal{L}(\overline{W}) - \hat{\mathcal{L}}(S_i,\overline{W})$, $\hat{\overline{W}}_i \in \hat{\mathcal{W}}$, $\mathrm{gen}(S_i,\hat{\overline{W}})$ is defined with respect to a loss function $\hat{\ell}\colon \mathcal{Z}\times\hat{\mathcal{W}} \mapsto \mathbb{R}^+$ (possibly different that ℓ), the mutual information and expectations are with respect to $Q_{S_i,,W_{1:K\setminus i}}P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$ and $QP_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$, respectively, $Q_{S_i,W_{1:K\setminus i}}$ is the marginal distribution of $(S_i, W_{1:K\setminus i})$ in Q, and the infimum is over all Markov kernels $P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$.

Theorem 3. *Suppose that the distributed learning algorithm* $\mathcal{A}_{1:K}(S_{1:K})$ *induces* $P_{S_{1:K},W_{1:K},\overline{W}}$ *and suppose that for all* $\hat{\overline{w}} \in \hat{\mathcal{W}}$, $\hat{\ell}(Z,\hat{\overline{w}})$ *is* σ*-subgaussian. Then, for any* $\epsilon \in \mathbb{R}$ *and* $\delta \geqslant 0$, *with probability at least* $1-\delta$,

$$\mathrm{gen}\big(S_{1:K},\overline{W}\big) \leqslant \sqrt{2\sigma^2\big(\max_{i\in[K]} R_p(\delta,\epsilon,i) + \log(1/\delta)\big)/n} + \epsilon, \tag{4}$$

$$R_p(\delta,\epsilon,i) \coloneqq \sup_{Q\colon D_{KL}(Q\|P_{S_{1:K},W_{1:K}})\leqslant\log(1/\delta)} \mathfrak{RD}_i(Q,\epsilon), \tag{5}$$

where the supremum is over all possible distributions over $\prod_{i\in[K]}(\mathcal{S}_i \times \mathcal{W}_i)$.

The theorem is proved in Section 5.1. In the binary classification setup, the result also holds for the margin generalization error $\mathrm{gen}_\theta(S_{1:K},\overline{W})$ by letting $\mathrm{gen}_\theta\big(S_i,\overline{W}\big) \coloneqq \mathcal{L}(\overline{W}) - \hat{\mathcal{L}}_\theta(S_i,\overline{W})$.

While the advantage of this result was already stated in part above, it should be noted that the denominator in the result of Theorem 3 is n, rather than nK in Theorem 2. In fact, none of the two results of Theorem 2 and Theorem 3 outperforms the other in general. For the case of distributed SVM which will be considered in the next section it can be shown that the result of Theorem 3 results into a bound that decays with K faster than one that uses Theorem 2. Moreover, as it will become clearer from the sequel the bounds for DSVM require a strict lossy compression, i.e., $\epsilon \neq 0$, and do not seem to be obtainable with a lossless compression framework, illustrating the utility of our rate-distortion based approach in general.

[1] For non-Lipschitz losses, such as 0-1 loss, the analysis is less trivial. An example can be found in Section 5.2.

3.2 In expectation bound

Similar to Theorems 2 and 3, we establish upper bounds on the expectation of the generalization error of any distributed stochastic learning algorithm.

Theorem 4. *Suppose that the distributed learning algorithm $\mathcal{A}_{1:K}(S_{1:K})$ induces $P_{S_{1:K},W_{1:K},\overline{W}}$ and suppose that for all $\hat{\overline{w}} \in \hat{\mathcal{W}}$, $\hat{\ell}(Z,\hat{\overline{w}})$ is σ-subgaussian. Then, the following holds for every $\epsilon \in \mathbb{R}$.*

$$\mathbb{E}\big[\mathrm{gen}\big(S_{1:K},\overline{W}\big)\big]$$

$$\leqslant \frac{1}{n}\min\left\{\frac{1}{K}\sum_{j\in[n]}\sum_{i\in[K]}\sqrt{2\sigma^2\mathfrak{RD}(P_{Z_{i,j},\overline{W}},\epsilon)}+\epsilon,\ \sum_{j\in[n]}\sqrt{2\sigma^2\max_{i\in[K]}\mathfrak{RD}_i(P_{Z_{i,j},W_{1:K},\overline{W}},\epsilon)}+\epsilon\right\} \tag{6}$$

$$\leqslant \frac{1}{\sqrt{n}}\min\left\{\sqrt{2\sigma^2\mathfrak{RD}(P_{S_{1:K},\overline{W}},\epsilon)/K}+\epsilon,\ \sqrt{2\sigma^2\max_{i\in[K]}\mathfrak{RD}_i(P_{S_i,W_{1:K},\overline{W}},\epsilon)}+\epsilon\right\}. \tag{7}$$

The first terms of the minimization in (6) and (7) follow easily by an application similar to in Theorem 2. In particular, setting $\epsilon = 0$ in the first term of the minimization in (6) one recovers the result of [YDVP20, Thoerem 2] under the assumed subgaussianity. The second terms of the minimization in (6) and (7) are derived in a manner that is essentially similar to in the proof of Theorem 3. The details are omitted for brevity.

4 Distributed support vector machines

In this section, we establish upper bounds on the generalization error of *Support Vector Machines* (SVM) [Vap06, CV95] when applied in a distributed learning setting. The algoritm is called hereafter as *Distributed* SVM (DSVM). SVMs are popular and widely used for binary classification problems. They can be combined with different kernels in a computationally efficient way [BGV92, GKL20]. An easy application of SVM consists in finding a hyperplane based on a training dataset that separates the data with the smallest possible average margin error. Formally, let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = \{-1,+1\}$. The hypotheses are vectors $w \in \mathbb{R}^d$, which represent hyperplanes. For the simplicity of the exposition we only consider the case with zero bias. The hyperplane predicts a label of a data x according to the sign of the inner product $\langle x, w\rangle$. The 0-1 loss and margin loss functions are thus defined as $\ell_0(z,w) := \mathbb{1}_{\{y\langle x,w\rangle<0\}}$ and $\ell_\theta(z,w) := \mathbb{1}_{\{y\langle x,w\rangle<\theta\}}$, respectively. In the distributed learning setup, each client $i \in [K]$ has access to dataset S_i and picks the vector W_i; and the server node computes the aggregated model as $\overline{W} := (W_1 + \ldots + W_K)/K$ [Car20]. In this section, we study the generalization gap, which we denote hereafter as $\mathrm{gen}_\theta(S_{1:K},\overline{W})$, defined as the difference between the population risk $\mathcal{L}(\overline{w})$ calculated using ℓ_0 and the margin empirical risk $\frac{1}{K}\sum_{i\in[K]}\hat{\mathcal{L}}_\theta(S_i,\overline{W})$ calculated using ℓ_θ.

Theorem 5. *Let $d \in \mathbb{N}^+$ and $\mathbb{P}(\|X\| \leqslant B) = 1$ for some $B > 0$. Consider DSVM with K clients each using* any *arbitrary local learning algorithm such that $\mathbb{P}(\|W_i\| \leqslant 1) = 1$, $i \in [K]$.*

i) For any $\delta > 0$, with probability at least $1-\delta$,

$$\mathrm{gen}_\theta\big(S_{1:K},\overline{W}\big) \leqslant \mathcal{O}\left(\frac{1}{nK\sqrt{K}} + \sqrt{\frac{\big(\frac{B}{K\theta}\big)^2\log(nK\sqrt{K})\log\big(\max\big(\frac{K\theta}{B},2\big)\big)+\log(1/\delta)}{n}}\right).$$

ii) Also,

$$\mathbb{E}\big[\mathrm{gen}_\theta\big(S_{1:K},\overline{W}\big)\big] \leqslant \mathcal{O}\left(\frac{1}{nK\sqrt{K}} + \sqrt{\frac{B^2\log(nK\sqrt{K})\log\big(\max\big(\frac{K\theta}{B},2\big)\big)}{nK^2\theta^2}}\right).$$

This theorem is proved in Section 5.2, by bounding the corresponding rate-distortion terms in Theorem 3 (for the part i.) and in the second term of (7) (for part ii.). To establish such bounds, we show the existence of a proper $P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$ using techniques and results developed in [GKL20], and in particular by making use of the Johnson-Lindenstrauss transformation [JL84].

The above bound on the expectation of the generalization error of the DSVM decreases with K as $\log(K)/K$. Moreover, it is important to note that the in expectation bound for K clients each having n data samples is smaller than that of the counterpart centralized learning algorithm that has nK input data samples by a factor of order $\mathcal{O}(\sqrt{\log(K)/K})$. This also holds for the tail bound, as long as $\log(1/\delta)/n$ is not the dominant term in the square root. Note that in general $\sqrt{\log(1/\delta)/n}$ is very small, corresponding to the generalization error of an dataset-independent algorithm [XR17].

Remark 2. *The tail bound in Theorem 5 for $K = 1$ does not recover the best known upper bound to the margin generalization error of SVMs. More precisely, for a centralized setup with dataset of size nK, [GKL20, Theorem 2] states*

$$\mathrm{gen}_\theta(S, W) \leqslant \mathcal{O}\left(\frac{B^2 \log(nK)}{nK\theta^2} + \sqrt{\frac{\left(\frac{B}{\theta}\right)^2 \log(nK) + \log(1/\delta)}{nK}\hat{\mathcal{L}}_\theta(S, W)}\right). \tag{8}$$

This is particularly important for the separable training dataset, where $\hat{\mathcal{L}}_\theta(s, w) = 0$, which makes the bound of order $\mathcal{O}\big(B^2 \ln(nK)/(nK\theta^2)\big)$. For the non-separable case, this term is asymptotically lower bounded by $\hat{\mathcal{L}}_\theta(W^*)$, where W^* is the optimal population risk minimizer when μ is known. However, in the distributed learning setup, even if $\hat{\mathcal{L}}_\theta(s_i, w_i) = 0$, $\hat{\mathcal{L}}_\theta(s_i, \overline{w})$ is not necessarily zero.

5 Proofs

5.1 Proof of Theorem 3

Proof. Similar to [SGRS22, Theorem 8], we use a technique, known as *block-coding* in information theory. Consider $m \in \mathbb{N}$ independent realizations of the datasets and also hypotheses choices $\Big\{\{S_{i,j}, W_{i,j}\}_{i\in[K]}, \overline{W}_j\Big\}_{j\in[m]}$, *i.e.,* $S_{i,j} := \{Z_{i,j,1}, \dots, Z_{i,j,n}\}$, $Z_{i,j,l} \sim \mu$, independent of other $\{Z_{i',j',l'}\}_{(i',j',l')\neq(i,j,l)}$, $W_{i,j}$ is chosen according to $P_{W_i|S_i}(w_{i,j}|S_{i,j})$, and $\overline{W}_j := (W_{1,j} + \cdots + W_{K,j})/K$. Fix $\epsilon \in \mathbb{R}$ and $\delta \in \mathbb{R}^+$. Let $\Delta_i := \sqrt{2\sigma^2(R_p(\delta, \epsilon, i) + \log(1/\delta))/n} + \epsilon$ and $\Delta := \max_{i\in[K]} \Delta_i$. We consider the probability that the generalization error for all $j \in [m]$ instances of the algorithm $\mathcal{A}_{1:K}(S_{1:K})$ being greater than Δ, and upper bound this by the probability that average of them being greater than Δ. Then, we relate this to the sum of K terms, each only depending on the training dataset of the client $i \in [K]$. Specifically, we have

$$\begin{aligned}
\mathbb{P}\big(\mathrm{gen}(S_{1:K}, \overline{W}) \geqslant \Delta\big)^m &= \mathbb{P}\big(\forall j \in [K]\colon\ \mathrm{gen}(S_{1:K,j}, \overline{W}_j) \geqslant \Delta\big) \\
&\leqslant \mathbb{P}\left(\frac{1}{mK}\sum_{i\in[K]}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta\right) \\
&\overset{(a)}{\leqslant} K \max_{i\in[K]} \mathbb{P}\left(\frac{1}{m}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta_i\right),
\end{aligned} \tag{9}$$

where (a) holds since $\Delta \geqslant \Delta_i, i \in [K]$ and using the fact that for every pair of random variables $(X, Y) \in \mathbb{R}^2$ and constants $(c_1, c_2) \in \mathbb{R}^2$ it holds that $\mathbb{P}(X+Y \geqslant c_1+c_2) \leqslant \mathbb{P}(X \geqslant c_1)+\mathbb{P}(Y \geqslant c_2)$. Now, use the following claim the proof of which will follow,

$$\mathbb{P}\Big(\frac{1}{m}\sum\nolimits_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta_i\Big) \leqslant \delta^m, \qquad \text{for every } i \in [K]. \tag{10}$$

Combining this claim with (9) and taking the mth root of both sides, we get $\mathbb{P}\big(\mathrm{gen}(S_{1:K}, \overline{W}) \geqslant \Delta\big) \leqslant \delta e^{\log(K)/m}$. Finally, letting $m \to \infty$ completes the proof. Observe that the effect of considering m independent instances helped to diminish the effect of the multiplicative term K, and hence avoiding to have the failure probability as $K\delta$, rather than δ. It remains to show that (10) holds.

Fix $i \in [K]$. The rest of the proof is as follows. First, we show that $\frac{1}{m}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j)$ corresponds to the generalization error of an algorithm, which is a concatenation of m i.i.d. instances of another algorithm. Then, using results of [SGRS22], we establish a rate-distortion theoretic bound

on the generalization error of this concatenated algorithm. The rest of proof follows by showing that the rate-distortion function of the concatenated algorithm (with parameters ϵ and δ^m) is not bigger than the sum of the rate-distortion functions of each instances (with parameters ϵ and δ). A similar relation is known in the lossy source compression context, and here we show this hereafter for the adaptation of the rate-distortion functions for statistical leaning algorithms.

Formally, let $\mathbf{S} \coloneqq \{S_{i,j}\}_{j\in[m]}$ with $S_j \coloneqq S_{i,j}$, $\overline{\mathbf{W}} \coloneqq \{\overline{W}_j\}_{j\in[m]}$, and $\mathbf{U} \coloneqq \{W_{1:K\setminus i,j}\}_{j\in[m]}$ with $U_j \coloneqq W_{1:K\setminus i,j}$. Note that $\mathbf{U}$ is independent of $\mathbf{S}$ and $P_{\overline{\mathbf{W}}|\mathbf{S},\mathbf{U}} = \prod_{j\in[m]} P_{\overline{W}j|S_{i,j},W_{1:K\setminus i,j}}$. Finally, for $\mathbf{z} \coloneqq (z_1,\dots,z_m) \in \mathcal{Z}^m$, define $\ell(\mathbf{z},\overline{\mathbf{w}}) \coloneqq \frac{1}{m}\sum_{j\in[m]} \ell(z_j,\overline{w}_j)$. Now, consider a learning algorithm that has access to the dataset $\mathbf{S}$ and the randomness $\mathbf{U}$ and picks the hypothesis $\overline{\mathbf{W}}$according to $P_{\overline{\mathbf{W}}|\mathbf{S},\mathbf{U}}$. Then, $\mathrm{gen}\big(\mathbf{S},\overline{\mathbf{W}}\big) = \frac{1}{m}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j},\overline{W}_j)$. Hence, showing (10) is equivalent to showing $\mathbb{P}\big(\mathrm{gen}(\mathbf{S},\overline{\mathbf{W}}) \geqslant \Delta_i\big) \leqslant \delta^m$. Using [SGRS22, Extended Theorem 9 in Appendix D.2.] we get

$$\mathbb{P}\left(\mathrm{gen}\big(\mathbf{S},\overline{\mathbf{W}}\big) \geqslant \sqrt{\frac{2\sigma^2(R(\delta^m,\epsilon)+\log(1/\delta^m))}{mn}}\right) \leqslant \delta^m.$$

where $R(\delta^m,\epsilon) \coloneqq \sup_{Q\colon D_{KL}\left(Q\|P_{\mathbf{S},\overline{\mathbf{W}},\mathbf{U}}\right)\leqslant \log(1/\delta^m)} \inf_{P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}} I\left(\mathbf{S};\hat{\overline{\mathbf{W}}}|\mathbf{U}\right)$ and the infimum is over all Markov kernels $P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}$ such that

$$\mathbb{E}\Big[\mathrm{gen}(\mathbf{S},\overline{\mathbf{W}}) - \mathrm{gen}(\mathbf{S},\hat{\overline{\mathbf{W}}})\Big] \leqslant \epsilon. \tag{11}$$

Note that as explained in Remark 1 (and as noted in [SGRS22, Equation (48)]), the above constraint is stronger than the one needed in [SGRS22, Extended Theorem 9], and hence it is valid. Moreover, note that the denominator in the square root is mn, as $\mathbf{S}$ is composed of mn independent samples. More precisely, considering the second proof of [SGRS22, Extended Theorem 9 in Appendix D.2.] for $\mathrm{gen}(\mathbf{S},\overline{\mathbf{W}})$, "$\mathrm{gen}(\hat{W}_2,S)$" in the paragraph before (64) in [SGRS22] would be the sum of mn i.i.d. variables. Now,

$$\begin{aligned}\mathbb{P}\big(\mathrm{gen}\big(\mathbf{S},\overline{\mathbf{W}}\big) \geqslant \Delta_i\big) &= \mathbb{P}\left(\mathrm{gen}\big(\mathbf{S},\overline{\mathbf{W}}\big) \geqslant \sqrt{\frac{2\sigma^2(mR_p(\delta,\epsilon,i)+\log(1/\delta^m))}{mn}}\right)\\ &\leqslant \mathbb{P}\left(\mathrm{gen}\big(\mathbf{S},\overline{\mathbf{W}}\big) \geqslant \sqrt{\frac{2\sigma^2(R(\delta^m,\epsilon)+\log(1/\delta^m))}{mn}}\right) \leqslant \delta^m,\end{aligned}$$

where the last step holds by the next lemma the proof of which is deferred to the supplements.

Lemma 1. $R(\delta^m,\epsilon) \leqslant mR_p(\delta,\epsilon,i)$.

This completes the proof of Theorem 3. □

5.2 Proof of Theorem 5

Proof. We establish a proof based on Theorem 3. First, note that for binary classification problem we have $\sigma = 1$. Then, we use the following claim, the proof of which will follow: for every $m \in \mathbb{N}$ and non-negative triplet $(c_1,c_2,\nu) \in \mathbb{R}^3$ it holds that

$$R_p(\delta,\epsilon,i) \leqslant m\log((c_2+\nu)/\nu), \tag{12}$$

where

$$\epsilon \coloneqq 8e^{-\frac{m}{7}\left(\frac{K\theta}{4B}\right)^2} + \frac{m\nu^m}{\sqrt{\pi}}e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{4c_1\nu B}\right)^2} + 4e^{-0.21m(c_1^2-1)} + 4e^{-0.21m(c_2^2-1)}. \tag{13}$$

This claim together with Theorem 3 imply that with probability at least $1-\delta$ we have

$$\mathrm{gen}_\theta(S_{1:K},W) \leqslant \sqrt{\frac{2m\log(c_2/\nu+1)+2\log(1/\delta)}{n}} + \epsilon.$$

Letting $m = \lceil 112\big(\frac{B}{K\theta}\big)^2 \log(nK\sqrt{K})\rceil$, $c_1 = c_2 = \sqrt{\frac{K^2\theta^2}{20B^2}+1}$, $\nu = 1/(2c_1)$ completes the proof of part (i). The proof of the part (ii) follows similarly.

Now, we proceed to show the inequality (12). We show that under any distribution Q defined over $\mathcal{W} \times \prod_{i\in[K]}(\mathcal{S}_i \times \mathcal{W}_i)$, there exist proper choices of $P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$ and 0-1 loss function $\hat{\ell}(z,\hat{\overline{w}}_i)$ such that

$$I\Big(S_i;\hat{\overline{W}}_i|W_{1:K\setminus i}\Big) \leqslant m\log((c_2+\nu)/\nu) \tag{14}$$

$$\mathbb{E}\Big[\mathrm{gen}_\theta\big(S_i,\overline{W}\big) - \mathrm{gen}\Big(S_i,\hat{\overline{W}}_i\Big)\Big] \leqslant \epsilon, \tag{15}$$

where $\mathrm{gen}(S_i,\hat{\overline{W}}_i)$ is computed based on $\hat{\ell}(z,\hat{\overline{w}}_i)$. We borrow some ideas and results from [GKL20] and use them in a new context to show the claim. In particular, similar to [GKL20] we make use of Johnson-Lindenstrauss transformation [JL84] to make the *compression* in a space that has a smaller dimension. More specifically, fix $i \in [K]$ and a matrix $\mathsf{A} \in \mathbb{R}^{d\times m}$. Also, let $\hat{\mathcal{W}} = \mathbb{R}^{d(K-1)+m}$ and denote each $\hat{\overline{w}}_i \in \hat{\mathcal{W}}$ as $\hat{\overline{w}}_i := (\hat{w}_{i,1},\ldots,\hat{w}_{i,K})$, where $\hat{w}_{i,j} \in \mathbb{R}^d$ for $j \neq i$ and $\hat{w}_{i,i} \in \mathbb{R}^m$. First, define the loss function as

$$\hat{\ell}(z,\hat{\overline{w}}_i) := \mathbb{1}_{\{y\langle\!\langle x,\hat{\overline{w}}_i\rangle\!\rangle_{\mathsf{A}}<\theta/2\}}, \quad \text{where} \quad \langle\!\langle x,\hat{\overline{w}}_i\rangle\!\rangle_{\mathsf{A}} := \frac{\langle \mathsf{A}x,\hat{w}_{i,i}\rangle + \sum_{j\neq i}\langle x,\hat{w}_{i,j}\rangle}{K}.$$

Next, we define a proper choice of $P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$. We impose the Markov chain $S_i - W_{1:K} - \hat{\overline{W}}_i$, and define $P_{\hat{\overline{W}}_i|W_{1:K}}$ as follows. For $j \neq i$, we let $\hat{W}_{i,j} := W_j$, and if $\|\mathsf{A}W_i\| \leqslant c_2$, we choose $\hat{W}_{i,i}$ uniformly over the m-dimensional ball around $\mathsf{A}W_i$ with radius ν. Otherwise, we choose $\hat{W}_{i,i}$ uniformly over the m-dimensional ball around the origin $(0,\ldots,0) \in \mathbb{R}^m$ with radius ν. This gives

$$\begin{aligned} I\Big(S_i;\hat{\overline{W}}_i|W_{1:K\setminus i}\Big) &\overset{(a)}{\leqslant} I\Big(W_i;\hat{\overline{W}}_i|W_{1:K\setminus i}\Big) = h(\hat{W}_{i,i}) - h(\hat{W}_{i,i}|W_i) \\ &\overset{(b)}{\leqslant} \log(\mathrm{Vol}_m(c_2+\nu)) - \log(\mathrm{Vol}_m(\nu)) = m\log((c_2+\nu)/\nu), \end{aligned}$$

where (a) follows by using the Markov chain $S_i - W_{1:K} - \hat{\overline{W}}_i$ and the data processing inequality; (b) holds since by using the triangle inequality $\hat{W}_{i,i}$ is always within the m-dimensional ball with radius $(c_2+\nu)$ around the origin $(0,\ldots,0) \in \mathbb{R}^m$, the entropy term $h(\hat{W}_{i,i})$ is maximized by the uniform distribution over the ball; and $h(\hat{W}_{i,i}|W_i) = \log(\mathrm{Vol}_m(\nu))$ since conditionally given W_i the variable $\hat{W}_{i,i}$ is distributed uniformly over an m-dimensional ball with radius ν. Note that here $\mathrm{Vol}_m(r)$ denotes the volume of m-dimensional ball with radius r, which is equal to $\frac{\pi^{m/2}r^m}{\Gamma(m/2+1)}$ and $\Gamma(\cdot)$ is the Gamma function. This shows that (14) holds for any choice of A. Now, we proceed to show that there exists a proper choice of A such that (15) also holds, using techniques in [GKL20]. Fix a distribution Q such that $D_{KL}(Q\|P_{S_{1:K},W_{1:K},\overline{W}}) < \infty$. Note that the expectation in (15) is with respect to $QP_{\hat{\overline{W}}_i|W_{1:K}}$. This term is bounded in the following lemma, proved in the supplements.

Lemma 2. *We have* $\mathbb{E}\Big[\mathrm{gen}_\theta\big(S_i,\overline{W}\big) - \mathrm{gen}\Big(S_i,\hat{\overline{W}}_i\Big)\Big] \leqslant D_{\mathsf{A}}$*, where*

$$D_{\mathsf{A}} := \mathbb{E}_{(Z,W_i,\hat{W}_i)}\Big[\mathbb{1}_{\{\frac{1}{K}|\langle X,W_i\rangle - \langle \mathsf{A}X,\hat{W}_i\rangle|>\theta/2\}}\Big] + \mathbb{E}_{(S_i,W_i,\hat{W}_i)}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\frac{1}{K}|\langle X_{i,j},W_i\rangle - \langle \mathsf{A}X_{i,j},\hat{W}_i\rangle|>\theta/2\}}\right],$$

with $(Z,W_i,\hat{W}_i) \sim \mu Q_{W_i}P_{\hat{W}_i|W_i}$ *and* $(S_i,W_i,\hat{W}_i) \sim Q_{S_i,W_i}P_{\hat{W}_i|W_i}$.

Next, to show the existence of a proper A we consider generating it as in the Johnson-Lindenstrauss transformation [GKL20]. The following lemma is proved in the supplements.

Lemma 3. . *The expectation of* D_{A}*, when each entry of the matrix* A *is generated i.i.d. according to* $\mathcal{N}(0,1/m)$*, is upper bounded by*

$$\mathbb{E}_{\mathsf{A}}[D_{\mathsf{A}}] \leqslant 8e^{-\frac{m}{7}\left(\frac{K\theta}{4B}\right)^2} + \frac{m\nu^m}{\sqrt{\pi}}e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{4c_1\nu B}\right)^2} + 4e^{-0.21m(c_1^2-1)} + 4e^{-0.21m(c_2^2-1)} = \epsilon.$$

Finally, since the expectation of D_{A} is upper bounded by ϵ there exists at least an $\mathsf{A} \in \mathbb{R}^{d\times m}$ such that $D_{\mathsf{A}} \leqslant \epsilon$. Thus, (15) holds by Lemma 2. This completes the proof of Theorem 5. □

References

[AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[AZ19] Inaki Estella Aguerri and Abdellatif Zaidi. Distributed variational representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):120–138, 2019.

[BDP22] L. P. Barnes, Alex Dytso, and H. V. Poor. Improved information theoretic generalization bounds for distributed and federated learning, 2022.

[Ber75] Toby Berger. *Rate Distortion Theory and Data Compression*, pages 1–39. Springer Vienna, Vienna, 1975.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[BLGŞ21] Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[BSE+21] Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy tails in SGD and compressibility of overparametrized neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[BZV20] Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, May 2020.

[Car20] Robert Carlsson. Privacy-preserved federated learning: A survey of applicable machine learning algorithms in a federated environment, 2020.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[EPB+20] Anis Elgabli, Jihong Park, Amrit S Bedi, Mehdi Bennis, and Vaneet Aggarwal. Gadmm: Fast and communication efficient framework for distributed machine learning. *J. Mach. Learn. Res.*, 21(76):1–39, 2020.

[GKL20] Allan Grønlund, Lior Kamma, and Kasper Green Larsen. Near-tight margin-based generalization bounds for support vector machines. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3779–3788. PMLR, 13–18 Jul 2020.

[GM91] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[GR18] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

[HDMR21] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Daniel M. Roy. Towards a unified information-theoretic framework for generalization. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[HJTW21] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021.

[HNK+20] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

[HRVSG21] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

[HŞKM21] Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W Mahoney. Generalization properties of stochastic optimizers via trajectory analysis. *arXiv preprint arXiv:2108.00781*, 2021.

[JL84] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26:28, 1984.

[KKM+20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[KLG+21] Lorenz Kuhn, Clare Lyle, Aidan N. Gomez, Jonas Rothfuss, and Yarin Gal. Robustness to Pruning Predicts Generalization in Deep Neural Networks. *arXiv:2103.06002 [cs, stat]*, March 2021.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[KMA+19] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019.

[KMY+16] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[LSTS20] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[MSS19] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[MZ21] Matei Moldoveanu and Abdellatif Zaidi. In-network learning for distributed training and inference in networks. In *2021 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2021.

[NHD+20] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates, 2020.

[PJ92] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[PW14] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

[RCZ+20] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[RZ16] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR.

[SAN20] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

[SGRS22] Milad Sefidgaran, Amin Gohari, Gael Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. *arXiv:2203.02474*, 2022.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[ŞSDE20] Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc., 2020.

[SZ20] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020.

[Vap06] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[VWK+20] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *ACM Computing Surveys (CSUR)*, 53(2):1–33, 2020.

[WHGC21] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24222–24234. Curran Associates, Inc., 2021.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

[YAE+18] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

[YDVP20] Semih Yagli, Alex Dytso, and H. Vincent Poor. Information-theoretic bounds on the generalization error and privacy leakage in federated learning. In *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2020.

[YMNS22] Honglin Yuan, Warren Richard Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations*, 2022.

[YZR21] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Checklist

1. For all authors...
   i. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   ii. Did you describe the limitations of your work? [Yes] We establish bounds on the generalization suggesting a decreasing behavior with number of clients. However, for a rigorous analytical conclusion, a lower bound is also required, which is notoriously a hard problem.
   iii. Did you discuss any potential negative societal impacts of your work? [N/A] Our work deals with theoretical investigation of the generalization behavior fo the distributed case and their comparison with the corresponding centralized case. We believe this work does not have any potential negative social impact.
   iv. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   i. Did you state the full set of assumptions of all theoretical results? [Yes] Please see Sections 2-4.
   ii. Did you include complete proofs of all theoretical results? [Yes] The proofs of the main results are stated in Section 5, and the proofs of the used lemmas are stated in the supplements.
3. If you ran experiments...
   i. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provided these in the supplementary materials.
   ii. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These are also provided in the supplementary materials.
   iii. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] This item is not needed in our experiments, as explained in the supplements.
   iv. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provided these in the supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   i. If your work uses existing assets, did you cite the creators? [Yes]
   ii. Did you mention the license of the assets? [Yes]
   iii. Did you include any new assets either in the supplemental material or as a URL? [Yes] The code used for experiments in this paper was made accessible through supplementary materials.

iv. Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use publicly available and non-personal data only. Please see the supplementary materials.

v. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use publicly available and non-personal data only. Please see the supplementary materials.

5. If you used crowdsourcing or conducted research with human subjects...

i. Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] This item is not applicable for our research.

ii. Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] This item is not applicable for our research.

iii. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] This item is not applicable for our research.