Rate-Distortion Theoretic Bounds on Generalization Error for Distributed Learning

Milad Sefidgaran [†], **Romain Chor** [†], **Abdellatif Zaidi** [† †]
[†] Paris Research Center, Huawei Technologies France
[†] Université Paris-Est, Champs-sur-Marne 77454, France
{milad.sefidgaran2, romain.chor}@huawei.com, abdellatif.zaidi@u-pem.fr

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 leveraging 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, 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 $\mathcal{O}(\sqrt{\log(K)/K})$. This finding is validated also experimentally. A similar conclusion is obtained for a *multiple-round federated learning* setup where each client uses *stochastic gradient Langevin dynamics* (SGLD).

1 Introduction

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-parameterized 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 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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

emergence of new applications in which resources are constrained, data is distributed or the need to preserve privacy [VWK+20, MMR+17, KMY+16, KLN+11]. Reducing the computational complexity by offloading the model training using algorithms such as *parallel* stochastic gradient descent is another reason for the popularity of these algorithms. Generally, that results in extra communication costs, however [ZZ15, AGL+17, CCB+18, LM22, WLC22]. Example such 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 [XR17, BZV20]. 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 Bregman 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], 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 the 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 an effect with a factor of only $1/K$ on the aggregated model order-wise. This is shown to result in a 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 than that of the centralized one with a factor of $\sqrt{\log(K)/K}$. We conducted experiments on DSVM that confirm this finding. 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 (in Appendix C.1). In all 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 $t \in \mathbb{R}$, $\mathbb{E}[\exp(t(X - \mathbb{E}[X]))] \leqslant \exp(\sigma^2 t^2/2)$, *e.g.,* if $X \in [a, b]$, then X is $\frac{b-a}{2}$-subgaussian. 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 a few definitions and recent results associated with 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* $\mathcal{W} \subseteq \mathbb{R}^d$ 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 $\mathrm{gen}(s, w) \coloneqq \mathcal{L}(w) - \hat{\mathcal{L}}(s, w)$ where $\mathcal{L}(w) \coloneqq \mathbb{E}_{Z\sim\mu}[\ell(z, w)]$ denotes the population risk and $\hat{\mathcal{L}}(s, w) \coloneqq \frac{1}{n}\sum_{j\in[n]} \ell(z_j, w)$ denotes the empirical risk. Note that the generalization error depends on the loss function $\ell(z, w)$, underlying distribution μ, the sample size n, and also the learning algorithm $P_{W|S}$. 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 $\ell_0(z, w) \coloneqq \mathbb{1}_{\{yf(x,w)<0\}}$, 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 $\ell_\theta(z, w) \coloneqq \mathbb{1}_{\{yf(x,w)<\theta\}}$, while using 0-1 loss function for the population risk evaluation. We denote the corresponding empirical risk as $\hat{\mathcal{L}}_\theta(s, w) \coloneqq \frac{1}{n}\sum_{j\in[n]} \ell_\theta(z_j, w)$ and the generalization error as $\mathrm{gen}_\theta(s, w) \coloneqq \mathcal{L}(w) - \hat{\mathcal{L}}_\theta(s, w)$.

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 an 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 [XR17] 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 unrelated, approaches, while offering a unifying framework to them. The upper bounds of [SGRS22] are rate-distortion theoretic. Specifically, let $\hat{\mathcal{W}}$ be the alphabet of the *compressed hypothesis* and $\hat{\ell}\colon \mathcal{Z} \times \hat{\mathcal{W}} \mapsto \mathbb{R}^+$ a loss function (possibly different from ℓ). Accordingly, for $\hat{w} \in \hat{\mathcal{W}}$ and $s \in \mathcal{Z}^n$, let $\mathrm{gen}(s, \hat{w})$ be defined with respect to $\hat{\ell}$. For every distribution Q defined over $\mathcal{S} \times \mathcal{W}$, the rate-distortion function with respect to $\hat{\mathcal{W}}$ is defined as

$$\begin{aligned} &\mathfrak{RD}(Q, \epsilon) \coloneqq \inf_{P_{\hat{W}|S}} I(S; \hat{W}), \\ &\text{s.t.} \quad \mathbb{E}_{(S,W)\sim Q}[\mathrm{gen}(S, W)] - \mathbb{E}_{(S,\hat{W})\sim Q_S P_{\hat{W}|S}}[\mathrm{gen}(S, \hat{W})] \leqslant \epsilon. \end{aligned} \tag{1}$$

where Q_S is the Q-marginal of S, and the infimum is taken over all Markov kernels (conditional distributions) of a random variable $\hat{W} \in \hat{\mathcal{W}}$ given S. Note that $P_{S,W}$ induced by the algorithm $\mathcal{A}$ is a particular case of Q. In the following, we shortly discuss the related intuition and concept behind the above terms. The reader is referred to Appendix A for more details on this.

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 10].

Theorem 1 ([SGRS22, Theorem 10]). *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 fixed $\epsilon \in \mathbb{R}$ 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 $R_p(\delta,\epsilon) \coloneqq \sup_{Q\colon D_{KL}(Q\|P_{S,W}) \leqslant \log(1/\delta)} \mathfrak{RD}(Q,\epsilon)$ 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 10]. First, we here express the rate-distortion terms with respect to the distortion function $d(w,\hat{w};s) \coloneqq \mathrm{gen}(s,w) - \mathrm{gen}(s,\hat{w})$ instead of a possibly smaller distortion function $d'(w,\hat{w};s) \coloneqq \inf_{(s',w')\in \mathrm{supp}(Q)} \mathrm{gen}(s',w') - \mathrm{gen}(s,\hat{w})$ that was considered in [SGRS22], 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 10]. The mentioned two small differences, however, do not require any change in the proof of [SGRS22, Theorem 10].*

Remark 2. *Similar to [SGRS22, Theorem 10], most of our results that will follow in this paper require the subgaussianity assumption of the loss function to hold. In both cases the assumption is used to properly bound the moment generating function (MGF) of a specific zero-mean random variable using the Hoeffding inequality (see [SGRS22, Section E.6.2] for the details) . Alternatively, this MGF can be upper bounded using approaches similar to [BZV20].*

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 the 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

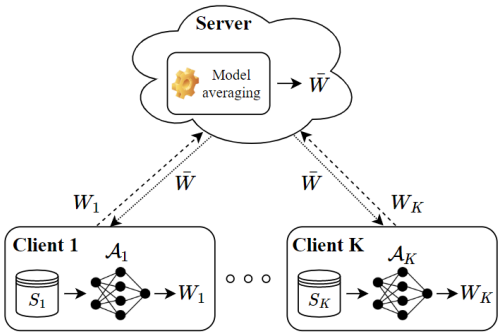


Figure 1: The considered distributed setup.

$$\mathcal{L}(\overline{w}) \coloneqq \mathbb{E}_{Z\sim\mu}[\ell(Z,\overline{w})], \quad \hat{\mathcal{L}}(s_{1:K},\overline{w}) \coloneqq \frac{1}{K}\sum\nolimits_{i\in[K]} \hat{\mathcal{L}}(s_i,\overline{w}). \tag{2}$$

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 stated in Section 5.

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 the 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 of a distributed learning algorithm that is composed of K local algorithms and an aggregator at the server as a single centralized algorithm. In other words, to consider the end-to-end system from the input at each local algorithm to the final aggregated model $\overline{W}$ 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 fixed $\epsilon \in \mathbb{R}$ 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 that 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, to formally state the result, we introduce some definitions. Let $\mathrm{gen}(s_i,\overline{w}) \coloneqq \mathcal{L}(\overline{w}) - \hat{\mathcal{L}}(s_i,\overline{w})$. Besides, as in the centralized case, let $\hat{\mathcal{W}}$ be the alphabet of the compressed hypothesis and $\hat{\ell}\colon \mathcal{Z}\times\hat{\mathcal{W}} \mapsto \mathbb{R}^+$ a loss function (possibly different from ℓ). Accordingly, for $\hat{\overline{w}} \in \hat{\mathcal{W}}$ and $s \in \mathcal{Z}^n$, let $\mathrm{gen}(s,\hat{\overline{w}})$ and $\mathrm{gen}(s_i,\hat{\overline{w}})$ be defined similarly with respect to $\hat{\ell}$. For a distribution Q defined over $\mathcal{W}\times\prod_{i\in[K]}(\mathcal{S}_i\times\mathcal{W}_i)$, let

$$\mathfrak{RD}_i(Q,\epsilon) \coloneqq \inf_{P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}} I(S_i;\hat{\overline{W}}_i|W_{1:K\setminus i}), \tag{3}$$

$$\text{s.t.}\quad \mathbb{E}\Big[\mathrm{gen}\big(S_i,\overline{W}\big) - \mathrm{gen}\Big(S_i,\hat{\overline{W}}_i\Big)\Big] \leqslant \epsilon, \tag{4}$$

where the infimum is taken over all Markov kernels (conditional distributions) of a random variable $\hat{\overline{W}}_i \in \hat{\mathcal{W}}$ given $(S_i, W_{1:K\setminus i})$. Note that the mutual information and expectations are with respect to

[1] For non-Lipschitz losses, such as 0-1 loss, the analysis is less trivial. An example can be found in Section D.2.

$Q_{S_i,,W_{1:K\setminus i}} P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$ and $Q P_{\hat{\overline{W}}_i|S_i,W_{1:K\setminus i}}$, respectively, where $Q_{S_i,W_{1:K\setminus i}}$ is the Q-marginal of $(S_i, W_{1:K\setminus i})$.

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 fixed $\epsilon \in \mathbb{R}$ and $\delta \geqslant 0$, with probability at least $1-\delta$,*

$$\operatorname{gen}\big(S_{1:K}, \overline{W}\big) \leqslant \sqrt{2\sigma^2\big(\max_{i\in[K]} R_p(\delta,\epsilon,i) + \log(1/\delta)\big)/n} + \epsilon, \tag{5}$$

$$R_p(\delta,\epsilon,i) \coloneqq \sup_{Q\colon D_{KL}(Q\|P_{S_{1:K},W_{1:K}})\leqslant \log(1/\delta)} \mathfrak{RD}_i(Q,\epsilon), \tag{6}$$

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 Appendix D.1. In the binary classification setup, the result also holds for the margin generalization error $\operatorname{gen}_\theta(S_{1:K}, \overline{W})$ by letting $\operatorname{gen}_\theta\big(S_i, \overline{W}\big) \coloneqq \mathcal{L}(\overline{W}) - \hat{\mathcal{L}}_\theta(S_i, \overline{W})$.

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 in 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.

Remark 3. *For the ease of the presentation, the results of this work, including Theorem 3, are stated for the homogeneous case where the underlying data distribution μ_i is same for all clients. However, the result can be extended straightforwardly to the heterogeneous case, by considering S_i and $\mathfrak{RD}_i(Q,\epsilon)$, $i \in [K]$, with respect to μ_i.*

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, for every fixed $\epsilon \in \mathbb{R}$ we have:*

$$\begin{aligned}
&\mathbb{E}\big[\operatorname{gen}\big(S_{1:K}, \overline{W}\big)\big] \\
&\leqslant \frac{1}{n}\min\left\{\frac{1}{K}\sum_{j\in[n]}\sum_{i\in[K]}\sqrt{2\sigma^2\mathfrak{RD}(P_{Z_{i,j},\overline{W}},\epsilon)} + \epsilon, \sum_{j\in[n]}\sqrt{2\sigma^2\max_{i\in[K]}\mathfrak{RD}_i(P_{Z_{i,j},W_{1:K},\overline{W}},\epsilon)} + \epsilon\right\} && (7)\\
&\leqslant \frac{1}{\sqrt{n}}\min\left\{\sqrt{2\sigma^2\mathfrak{RD}(P_{S_{1:K},\overline{W}},\epsilon)/K} + \epsilon, \sqrt{2\sigma^2\max_{i\in[K]}\mathfrak{RD}_i(P_{S_i,W_{1:K},\overline{W}},\epsilon)} + \epsilon\right\}. && (8)
\end{aligned}$$

The first terms of the minimization in (7) and (8) follow easily by an application similar to in Theorem 2. In particular, setting $\epsilon = 0$ in the first term of the minimization in (7) one recovers the result of [YDVP20, Thoerem 2] under the assumed subgaussianity. The second terms of the minimization in (7) and (8) are derived in a manner that is essentially similar to in the proof of Theorem 3. The details are omitted for brevity.

In Appendix C.2, it is shown that under the assumed subgaussianity, the upper bounds in [BDP22, Theorems 4 & 5] also can be recovered from Theorem 4. The results of [BDP22] are derived using the stability approach and by applying the *leave-one-out expansion* lemma.

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 algorithm 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. It should be noted that the results of [BDP22, Theorems 4 & 5] require the loss function to be either a Bregman divergence or Lipschitz of some order; and, as such, they are not applicable to the DVSM setup that we consider here. This is because the 0-1 loss function is neither a Bregman divergence, nor a Lipschitz loss.

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 $1 - \delta$,

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

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 Appendix D.2, by bounding the corresponding rate-distortion terms in Theorem 3 (for the part i.) and in the second term of (8) (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 with a rate of $\log(K)/K$. 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 4. *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*

$$\mathrm{gen}_\theta(S, W) \leqslant \mathcal{O}\left(\frac{B^2 \log(nK)}{nK\theta^2} + \sqrt{\frac{\left(\frac{B}{\theta}\right)^2 \log(nK) + \log(1/\delta)}{nK}\hat{\mathcal{L}}_\theta(S, W)}\right). \tag{9}$$

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.

Remark 5. *The proof of Theorem 5 shows how our generalization bounds exploit the particular topology of the distributed learning setup. In particular, we show that for a fixed non-zero distortion level ϵ in* (4) *an upper bound on the rates in* (4) *scales as $\mathcal{O}\big((\log(K)/K)^2\big)$. When $\epsilon = 0$, a case for which the rate-distortion approach reduces to the mutual information-based approach (see the discussion right after Theorem 4), that upper bound scales only as $\mathcal{O}(1)$. In part this explains the benefits brought up by the rate-distortion approach upon the mutual information-based one.*

5 Federated stochastic gradient Langevin dynamics

In this section, we consider a *homogeneous federated learning* setup, denoted as FSGLD, where local learning algorithms use *stochastic gradient Langevin dynamics* (SGLD) method [WHGC21]. For the case of the centralized learning algorithm, [WHGC21] has proposed a new tractable upper bound on the generalization error, that describes well the generalization behavior, by using results of [BZV20] and by connecting SGLD to the Gaussian channels.

We consider a multiple-round distributed learning algorithm. At beginning of each round $t \in [T]$, the central server sends the updated hypothesis at the end of the previous round $\overline{W}_{t-1} \in \mathbb{R}^d$ to the clients, where $\overline{W}_0$ is a randomly initialized hypothesis. Then, each client $i \in [K]$ performs locally one iteration of SGLD, as explained later, and outputs $W_{i,t}$. The central server upon receiving these hypotheses choices, let $\overline{W}_t \coloneqq (W_{1,t} + \cdots + W_{K,t})/K$. The final hypothesis $\overline{W}$ is chosen as a deterministic function of $\{\overline{W}_t\}_{t\in[T]}$. An example of such choices is Polyak averaging [PJ92], where $\overline{W} \coloneqq (\overline{W}_1 + \cdots + \overline{W}_T)/T$.

Now, we explain the local SGLD algorithm applied by each client [GM91, WT11, WHGC21]. Each client $i \in [K]$, partitions its dataset S_i into m disjoint mini-batches $\{S_{i,1}, \ldots, S_{i,m}\}$, each one having equal size b with elements $S_{i,j} = \{Z_{i,j,1}, \ldots, Z_{i,j,b}\}$. Then, client i at round t, by receiving $\overline{W}_{t-1}$ from the central node, let $W_{i,t}$ be

$$W_{i,t} = \overline{W}_{t-1} - \eta_t \nabla_w \hat{\ell}(S_{i,j_t}, \overline{W}_{t-1}) + \sqrt{\frac{2\eta_t}{\beta_t}} V.$$

Here, η_t is the learning rate, β_t the inverse temperature, V a d-dimensional random variable with distribution $\mathcal{N}(0, \mathrm{I}_d)$, where I_d is the $d \times d$ identity matrix, $j_t \in [m]$ is the mini-batch index, $\hat{\ell}\colon \mathcal{Z} \times \mathcal{W} \mapsto \mathbb{R}^+$ is a surrogate loss function, and

$$\nabla_w \hat{\ell}(S_{i,j_t}, \overline{W}_{t-1}) \coloneqq \frac{1}{b} \sum\nolimits_{l\in[b]} \nabla_w \hat{\ell}(Z_{i,j_t,l}, \overline{W}_{t-1}).$$

In the following we use the first term of (7) with $\epsilon = 0$ (which is reduced to the upper bound in [YDVP20, Thoerem 2]), to derive a bound on the expectation of the generalization error of FSGLD. Note that the first term of (7) is established by viewing the whole learning algorithm as a black-box, and hence it applies for the multiple-round federated learning setup as well, by considering $S_{1:K}$ as the training dataset and $\overline{W}$ as the chosen hypothesis.

Theorem 6. *Suppose that for each $w \in \mathcal{W}$, the loss $\ell(Z, w)$ is σ-subgaussian. The expected generalization error of FSGLD is upper bounded by*

$$\mathbb{E}\big[\mathrm{gen}\big(S_{1:K}, \overline{W}\big)\big] \leqslant \frac{\sqrt{2b}\sigma}{2nK\sqrt{K}} \sum_{j\in[m]} \sum_{i\in[K]} \sqrt{\sum_{t\in\mathcal{T}_{i,j}} \beta_t \eta_t \mathit{Var}\Big(\nabla_w \hat{\ell}(S_{i,j}, \overline{W}_{t-1})\Big)},$$

where the set $\mathcal{T}_{i,j}$ contains the indices t such that $j_t = j$ at client i, and

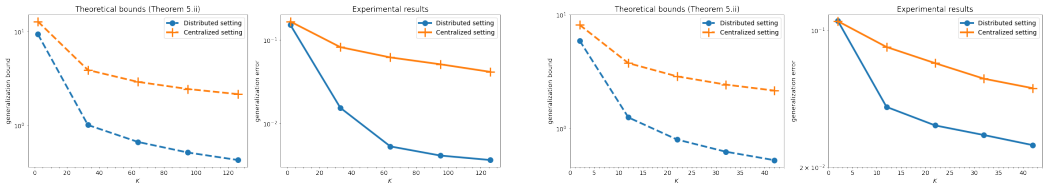
$$\mathit{Var}\Big(\nabla_w \hat{\ell}(S_{i,j}, \overline{W}_{t-1})\Big) \coloneqq \mathbb{E}\Big[\Big\|\nabla_w \hat{\ell}(S_{i,j}, \overline{W}_{t-1}) - e_i\Big\|^2\Big],$$

where $e_i \coloneqq \mathbb{E}\Big[\nabla_w \hat{\ell}(S_{i,j}, \overline{W}_{t-1})\Big]$.

This result for the case of $K = 1$ is reduced in [WHGC21, Theorem 1] and it can be proved for any $K \in \mathbb{N}$ along the same lines of the proof of [WHGC21, Theorem 1]. Indeed,

$$\overline{W}_t = \overline{W}_{t-1} - \frac{\eta_t}{K}\left(\sum_{i\in[K]} \nabla_w \hat{\ell}(S_{i,j_t}, \overline{W}_{t-1})\right) + \frac{1}{\sqrt{K}}\sqrt{\frac{2\eta_t}{\beta_t}} V', \tag{10}$$

where V' is a d-dimensional random variable with distribution $\mathcal{N}(0, \mathrm{I}_d)$. Now, proceeding similar to the proof of [WHGC21, Theorem 1] concludes the result. The reason behind the extra factor $1/\sqrt{K}$ in the bound is that, when K independent Gaussian noises are added, their variances are added up linearly with K. Thus, while in (10), the gradient is divided by K, the noise term is divided by $\sqrt{K}$,



(a) $n = 100$ (b) $n = 300$

Figure 2: The generalization error for distributed and centralized SVM

which results in an extra $\sqrt{K}$ term in the denominator of the upper bound. The proof is omitted for brevity.

The established bound is decreasing with K, as far as the variance of the gradient with training data size nK is not smaller than the variance of the gradient with training data size n, divided by K. Note that the variance of the gradient is correlated with the *sharpness* of the loss landscape and becomes small as the algorithm converges [WHGC21].

6 Experiments

This section aims at showing, through simulations, the evolution of the generalization error of the distributed SVM with the number K of clients and the size n of each individual training dataset. We will provide also comparison with the centralized setup, having training dataset of size nk, and also with the bounds on the generalization error as predicted by our analysis. We also provide further experiments and discussions on the comparison of the empirical and population risks of the distributed and centralized SVM. The details of the experiments are explained in Appendix B and here, we discuss our experimental findings.

On the generalization error: Figure 2 shows the evolution of the generalization error (with zero margin) of the distributed setup (K clients, each having n data samples) as a function of the number K of used clients for two values of n, $n = 100$ (Subfigure 2a) and $n = 300$ (Subfigure 2b). Also shown for comparison, the generalization performance of the associated centralized setup with a dataset of nK samples.

The figure also depicts the evolution of the bound of our Theorem 5 (computed using part ii) of the theorem and with the value of the parameter θ set to 0.2). For the centralized learning setting, the theoretical bound is obtained by applying the result of Theorem 5 (part ii) with the substitution $K_c = 1$ and $n_c = nK$.

Observe that, as predicted by our theoretical analysis of Section4, the system generalizes better in the distributed setup.

On the empirical and population risks: While considering the empirical risk $\hat{\mathcal{L}}(s_{1:K}, \overline{w}) \coloneqq \frac{1}{K}\sum_{i\in[K]} \hat{\mathcal{L}}(s_i, \overline{w})$ is motivated by previous works [YDVP20, BDP22], it differs from the average local empirical risks minimized at clients, *i.e.,*

$$\tilde{\mathcal{L}}(s_{1:K}, w_{1:K}) \coloneqq \frac{1}{K}\sum\nolimits_{i\in[K]} \hat{\mathcal{L}}(s_i, w_i). \tag{11}$$

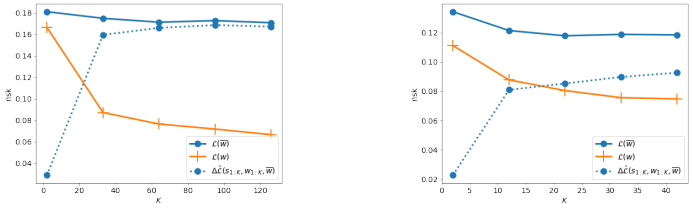
Denote the difference of these empirical risks as

$$\Delta\hat{\mathcal{L}}(s_{1:K}, w_{1:K}, \overline{w}) \coloneqq \hat{\mathcal{L}}(s_{1:K}, \overline{w}) - \tilde{\mathcal{L}}(s_{1:K}, w_{1:K}). \tag{12}$$

Then, the population risk can be written as

$$\mathcal{L}(\overline{w}) = \tilde{\mathcal{L}}(s_{1:K}, w_{1:K}) + \mathrm{gen}(s_{1:K}, \overline{w}) + \Delta\hat{\mathcal{L}}(s_{1:K}, w_{1:K}, \overline{w}). \tag{13}$$

The first term of the RHS of Eq. (13) can be made sufficiently small by minimizing the local empirical loss at each client. The second term is the generalization term considered in this paper, which is shown to decrease faster than the corresponding centralized case. In what follows, we study the evolution of the last term of the RHS of (13) with K.



(a) $n = 100$ (b) $n = 300$

Figure 3: The population risk and empirical risk differences for distributed and centralized SVM

Indeed, as shown in Fig 3, while the considered generalization error decreases with K, the population risk of the centralized setup is smaller than the distributed setup. Also, the latter decreases more slowly with K. This is caused by the increase of term $\Delta\hat{\mathcal{L}}(S_{1:K}, W_{1:K}, \overline{W})$ with K (note that this term is zero for the centralized setting). As it will become clearer from the rest of this section, in the distributed setup the term $\mathrm{gen}(S_{1:K}, \overline{W})$ decreases with K as specified by our Theorem 5 and the term $\Delta\hat{\mathcal{L}}(S_{1:K}, W_{1:K}, \overline{W})$ increases with K (the term $\tilde{\mathcal{L}}(S_{1:K}, W_{1:K}) + \mathrm{gen}(S_{1:K}, \overline{W})$ as defined by (11) can be made negligible for both centralized and distributed settings if local models are trained well enough, an assumption which will be made throughout hereafter). For the distributed case, the net effect is a decrease of the population risk with K, but which is slower than in the corresponding centralized setup. In what follows, we will show that as K becomes large, if the clients use the same algorithm, i.e., $P_{W_i|S_i} = P_{W_1|S_1}$ for all $i \in [K]$, and they are trained well enough to minimize their respective empirical risks $\hat{\mathcal{L}}(S_i, W_i)$ then the population risk in the distributed case tends to a constant that depends only on n and the used local algorithms, but not on K.

More formally, fix a training data size $n \in \mathbb{N}$ and suppose that all clients use the same local learning algorithm $P_{W_i|S_i} = P_{W_1|S_1}$. This empirical risks difference term is zero for $K = 1$, as $\overline{W} = W_1$. By increasing K, $\overline{W}$ departs further from W_i. In particular, as $K \to \infty$, $\overline{W} \to \mathbb{E}_{W_1}[W_1]$, where the expectation is with respect to the marginal distribution P_{W_1} (induced by $P_{W_1|S_1}\mu^{\otimes n}$) and the convergence is almost surely due to the strong law of large numbers. Note the for a fixed data distribution μ, the marginal distribution P_{W_1} is a function of n and local algorithms $P_{W_1|S_1}$. Hence,

$$\begin{aligned}\lim_{K\to\infty} \Delta\hat{\mathcal{L}}(S_{1:K}, W_{1:K}, \overline{W}) &\overset{a.s.}{=} \mathbb{E}_{S_1\sim\mu^{\otimes n}}\Big[\hat{\mathcal{L}}\big(S_1, \mathbb{E}_{W_1\sim P_{W_1}}[W_1]\big)\Big] - \mathbb{E}_{(S_1,W_1)\sim P_{S_1,W_1}}\Big[\hat{\mathcal{L}}(S_1, W_1)\Big] \\ &= \mathbb{E}_{Z\sim\mu}\big[\ell\big(Z, \mathbb{E}_{W_1\sim P_{W_1}}[W_1]\big)\big] - \mathbb{E}_{(S_1,W_1)\sim P_{S_1,W_1}}\Big[\hat{\mathcal{L}}(S_1, W_1)\Big] \\ &= \mathcal{L}(E_{W_1}[W_1]) - \mathbb{E}_{(S_1,W_1)\sim P_{S_1,W_1}}\Big[\hat{\mathcal{L}}(S_1, W_1)\Big]. \end{aligned} \tag{14}$$

The above is illustrated in Fig.3, where as visible from therein the population risk for the distributed setting (as computed experimentally) approaches the limit given by the RHS of (14) as K gets large.

7 Concluding remarks

In this work, we established rate-distortion theoretic tail and in-expectation bounds on the generalization error of the distributed learning algorithms. Unlike previous approaches to this problem, our bounds, which are more general comparatively, are tailored specifically for the distributed setup. In particular, when applied to distributed SVM and FSGLD our results suggest a decreasing behavior of the generalization error with respect to the number of clients. The conducted experiments on DSVM are in accordance with the analytical results. Also, we partly investigated the evolution of the population risk with the number of clients and the size of the dataset. The analysis revealed the presence of a bias term. Possible directions for future works include (i) further investigation of the effect of the aforementioned bias term, (ii) study of the tightness of the proposed general bounds (experimentally), (iii) extension of the results to the setting in which the data distribution is not identical across the clients or when the training data samples are not independent, (iv) multiple-round scenarios, and (v) study of associated computational and communication costs.

References

[AGL+17] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[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.

[Ari72] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

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

[BB05] Mayank Bakshi and Rakesh K. Bansal. On error exponent in lossy source coding, 2005.

[BDP22] L. P. Barnes, A. Dytso, and H. V. Poor. Improved information theoretic generalization bounds for distributed and federated learning. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1465–1470, 2022.

[Ber75] Toby Berger. *Rate Distortion Theory and Data Compression", bookTitle="Advances in Source Coding*, 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.

[Bla72] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

[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.

[CCB+18] Chia-Yu Chen, Jungwook Choi, Daniel Brand, Ankur Agrawal, Wei Zhang, and Kailash Gopalakrishnan. Adacomp: Adaptive residual gradient compression for data-parallel distributed training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[CK11] Imre Csiszàr and Jànos Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition, 2011.

[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.

[EGK11] Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, 2011.

[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 rˆd. *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.

[Han00] Te Sun Han. The reliability functions of the general source with fixed-length coding. *IEEE Transactions on Information Theory*, 46(6):2117–2132, 2000.

[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.

[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.

[Iri05] Kiminori Iriyama. Probability of error for the fixed-length lossy coding of general sources. *IEEE Transactions on Information Theory*, 51(4):1498–1507, 2005.

[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.

[LCB10] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. at&t labs, 2010.

[LM22] Rui Liu and Barzan Mozafari. Communication-efficient distributed learning for large batch optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13925–13946. PMLR, 17–23 Jul 2022.

[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.

[Mar74] Katalin Marton. Error exponent for source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 20(2):197–199, 1974.

[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.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

[RBK18] Erwin Riegler, Helmut Bölcskei, and Günther Koliander. Rate-distortion theory for general sets and measures. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 101–105. IEEE, 2018.

[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, Gaël Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Conference on Learning Theory*, pages 4416–4463. PMLR, 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.

[WLC22] Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22802–22838. PMLR, 17–23 Jul 2022.

[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.

[ZZ15] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1–9, Lille, France, 07–09 Jul 2015. PMLR.

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 2-5.
   ii. Did you include complete proofs of all theoretical results? [Yes] The proofs are stated in Appendix D.
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 provided in Appendix B.
   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 appendices.
   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 Appendix B.
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 Appendix B.
   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 Appendix B.
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.

Rate-Distortion Theoretic Bounds on Generalization Error for Distributed Learning

Appendices

The appendices are organized as follows:

- In Appendix A, we briefly recall the main concepts and definitions of the rate-distortion theory.
- In Appendix B, we provide the details of the considered simulation setups in our experiments.
- In Appendix C, we present the theoretical results that are not stated in the main document due to lack of space. The results are on
  - the generalization error of deterministic algorithms with Lipschitz loss,
  - the comparison of Theorem 4 with results of [BDP22].
- Finally, the proofs of the results are given in Appendix D.

A Background on rate-distortion theory

In this section, we recall the concepts and definitions related to the rate-distortion theory. We first present a brief review of classic lossy source coding [Ber75, CK11, PW14]. This will help understand better the adaptation of the concepts to the related topic of study of algorithm compressibility [SGRS22].

A.1 Lossy source compression

We start by describing the notion of *lossy compression* of a source $X \in \mathcal{X}$ in the information-theoretic context [Ber75, CK11, PW14]. The goal is to compress (or to quantize) the source $X \in \mathcal{X}$ as $\hat{X} \in \hat{\mathcal{X}}$ such that i) $\hat{X}$ requires a smaller number of bits to be saved, and ii) from $\hat{X}$, X (or a desired function of X) can be recovered within a certain allowed level of *distortion*.

To this end, we first fix a proper compressed alphabet $\hat{\mathcal{X}}$ and consider a suitable distortion function $d\colon \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, that measures discrepancies between the symbol x (or a desired function of it) and a reconstructed version of it. It is assumed that there exists a $\hat{x}_0 \in \hat{X}$ such that for every $\mathbb{E}[d(X, \hat{x}_0)] < \infty$ [Ber75]. Often the set $\hat{\mathcal{X}}$ can be chosen to be a subset of $\mathcal{X}$, but it needs not to be in general.

Next, we explain the compression scheme. A key idea is to encode and reconstruct block-wise, i.e, the so-called *block-coding* technique. Let $X^m = (X_1, \ldots, X_m) \in \mathcal{X}^m$ be a block of m i.i.d. realizations of the source $X \sim P_X$. The *block-coding* technique considers the joint compression of any realizations of x^m as a $\hat{x}^m = (\hat{x}_1, \ldots, \hat{x}_m) \in \hat{\mathcal{X}}^m$. The distortion between two vectors x^m and $\hat{x}^m$ is given by the average of the coordinate-wise distortions,

$$d(x^m, \hat{x}^m) := \frac{1}{m} \sum_{i \in [m]} d(x_i, \hat{x}_i). \tag{15}$$

Now, we define the (R, ϵ)-compressibility of a source $X \sim P_X$ with respect to the distortion function d, where R is the compression rate and ϵ is the maximum allowed distortion value. The source X is said to be (R, ϵ)-compressible if there exists a sequences of *codebooks* $\{\mathcal{C}_m\}_{m \in \mathbb{N}}$, $\mathcal{C}_m = \{\hat{\mathbf{x}}_{j;m} = (\hat{x}_{j,1}, \ldots, \hat{x}_{j,m}), j \in [l_m]\} \subseteq \hat{\mathcal{X}}^m$, such that $l_m \leqslant e^{mR}$ and

$$\lim_{m \to \infty} \mathbb{P}\left(\min_{j \in [l_m]} d(X^m, \hat{\mathbf{x}}_{j;m}) > \epsilon \right) = \lim_{m \to \infty} \mathbb{P}(\forall j \in [l_m] \colon d(X^m, \hat{\mathbf{x}}_{j;m}) > \epsilon) = 0, \tag{16}$$

where $X^m \sim {P_X}^{\otimes m}$. The (R, ϵ)-compressibility implies that for m sufficiently large one can find a codebook $\mathcal{C}_m$ such that (i) every element of $\mathcal{C}_m$ can be represented using $mR \log_2(e)$ bits and (ii) with high probability, for every X^m there exists at least one element of $\mathcal{C}_m$ for which the distortion from X^m does not exceed ϵ. For example ϵ-net covering can be seen as a special case of the above

compressibility definition with $\mathcal{C}_m$ being the m-times Cartesian product of some ϵ-net covering of $\mathcal{X}$. However, in ϵ-net covering the more stringent condition $\mathbb{P}\big(\min_{j\in[l_1]} d(X, \hat{\mathbf{x}}_{j;1}) > \epsilon\big) = 0$ is required. This condition yields $\mathbb{P}\big(\min_{j\in[l_m]} d(X^m, \hat{\mathbf{x}}_{j;m}) > \epsilon\big) = 0$ for every $m \in \mathbb{N}$, which is a stronger condition than (20). This explains why the ϵ-net covering number is an upper bound on the minimum achievable rate R.

It turns out [Ber75, CT06, PW14] that given a distortion threshold ϵ, the fundamental lossy compression rate R of the source $X \sim P_X$ is determined by the rate-distortion function $\mathfrak{RD}(P_X, \epsilon)$, defined as

$$\mathfrak{RD}(P_X, \epsilon) \coloneqq \inf_{P_{\hat{X}|X}} I(X; \hat{X}), \quad \text{s.t.} \quad \mathbb{E}_{(X,\hat{X})\sim P_X P_{\hat{X}|X}}[d(X, \hat{X})] \leqslant \epsilon, \tag{17}$$

where the infimum is taken over all possible Markov kernels (conditional distributions) of a random variable $\hat{X} \in \hat{\mathcal{X}}$ given X.

For certain source distributions, closed-form expression of the associated rate-distortion function can be found explicitly [Ber75, CT06, PW14]. For example, for a d-dimensional i.i.d. Gaussian vector $X \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$, the rate-distortion function with respect to the squared two-norm L_2^2 is given by $\frac{d}{2}\log(\max(\sigma^2/\epsilon, 1))$. Besides, for sources of finite cardinality alphabets $\mathcal{X}$, the rate-distortion function can be computed efficiently using the Blahut-Arimoto algorithm [Bla72, Ari72]. For continuous sources or sources with infinite-sized alphabets, the rate-distortion function can be estimated by first applying a suitable (fine) quantization technique [EGK11, Proof of Theorem 3.6] and then using the Blahut-Arimoto algorithm. For small distortion levels ϵ, optimal rate-distortion tradeoffs can be estimated using appropriate (variational) lower-bounds *e.g.,* [RBK18]. Finally, we mention that for sources with finite-sized sets an easy upper bound on the rate-distortion function is given by $\log(|\mathcal{X}|)$; whereas for sources whose support set $\mathcal{X} \subset \mathbb{R}^d$ is inside a d-dimensional ball of radius R the function $\mathfrak{RD}(P_X, \epsilon)$ with respect to the L_2 norm is upper bounded by the ϵ-net covering number of this d-dimensional ball.

The rate-distortion formulation in (17) is asymptotic with respect to the block length m; *i.e.,* as formulated in (20), the probability that there does not exist a proper $\hat{\mathbf{x}}_{j;m}$ within ϵ distortion of X^m goes to zero, as $m \to \infty$. This formulation does not require any condition on the rate of this convergence to zero. Marton [Mar74] showed that for the finite sets $\mathcal{X}$, to have this probability go exponentially fast to zero (for large values of m), *i.e.,* at least as fast as δ^m for some $0 < \delta < 1$, the needed rate R needs to be at least as large as $\sup_Q \mathfrak{RD}(Q, \epsilon)$, where the supremum is over all distribution Q defined over $\mathcal{X}$ such that $D_{KL}(Q\|P_X) \leqslant \log(1/\delta)$. As it can be observed, the minimum needed rate depends not only on the rate-distortion term with respect to the underlying distribution P_X but also with respect to all those distributions Q that are close to P_X in some sense. The result has been extended to countably infinite and continuous sets in [Han00, Iri05, BB05].

A.2 Lossy algorithm compression

Now, we are ready to present an adaptation of the rate-distortion concept to the related setting of the study of the compressibility of statistical learning algorithms which was developed and used for the first time in [SGRS22, Section 3.1]. Such adaptation is also used extensively in our present work; and, hence, hereafter we provide a short review for the sake of completeness.

Consider a learning algorithm $\mathcal{A}$ which, for a training data set $S = \{Z_1, \ldots, Z_n\} \in \mathcal{S} = \mathcal{Z}^n$ picks a hypothesis $W \in \mathcal{W}$ according to $P_{W|S}$. Similar to the source coding context, consider the set $\hat{\mathcal{W}}$ of *compressed hypotheses* and the following distortion measure between any w and $\hat{w}$[2]

$$d(w, \hat{w}; s) \coloneqq \mathrm{gen}(s, w) - \mathrm{gen}(s, \hat{w}).$$

In this work, we consider a slightly extended version of the compressibility framework of [SGRS22] by allowing the compressed hypothesis space to be assessed with respect to a potentially different loss function $\ell' \colon \mathcal{Z} \times \hat{\mathcal{W}} \to \mathbb{R}^+$. The generalization error $\mathrm{gen}(s, \hat{w})$ is then defined with respect to

[2]Observe that in contrast with classic source coding contexts the distortion measure here depends not only on the true value of the hypothesis and its estimated fit but also on the given training data set, its distribution and size n as well as the considered loss function. Also, the values of the incurred distortion are allowed to take negative values.

this loss function. However, all results of [SGRS22] hold trivially for this extended framework as well.

Next, we present an adaptation of the described block coding technique to the study of the generalization power of statistical learning algorithms. Let $S^m = (S_1, \dots, S_m)$ be a block of i.i.d. training datasets $S_i = \{Z_{i,1}, \dots, Z_{i,n}\} \sim \mu^{\otimes n}$, and $W^m = (W_1, \dots, W_m)$ be the block of corresponding chosen hypothesis, *i.e.,* $\mathcal{A}(S_i) = W_i$, $i \in [m]$. Similarly, denote the block of length m of compressed hypothesis $(\hat{w}_1, \dots, \hat{w}_m)$ by $\hat{w}^m$. While in [SGRS22], different distortion functions between w^m and $\hat{w}^m$ given s^m are considered, in this work we only consider the following distortion function:

$$d(w^m, \hat{w}^m; s^m) \coloneqq \frac{1}{m} \sum_{j \in [m]} d(w_j, \hat{w}_j; s_j). \tag{18}$$

As in the source coding context, we consider the joint *lossy compression* of (S^m, W^m) as follows. Fix a *hypothesis book* $\mathcal{H}_m \coloneqq \{\hat{\mathbf{w}}_{j;m} = (\hat{w}_{j,1}, \dots, \hat{w}_{j,m}), j \in [l_m]\} \subseteq \hat{\mathcal{W}}^m$. For this chosen hypothesis book, the optimal compression scheme $\hat{\mathcal{A}}^*_m \colon \mathcal{S}^m \times \mathcal{W}^m \to \mathcal{H}_m$ assigns for every (s^m, w^m) the compressed hypothesis vector $\hat{\mathbf{w}}_{j;m}$, $j \in [l_m]$ that has the smallest distortion, *i.e.,*

$$\hat{\mathcal{A}}^*_m(s^m, w^m) = \hat{\mathbf{w}}_{j;m} \quad \text{s.t.} \quad j = \operatorname{argmin}_{j \in [l_m]} d(w^m, \hat{\mathbf{w}}_{j;m}; s^m).$$

Now, we recall the definition of the algorithm compressibility as given in [SGRS22, Definitions 1 and 8]. The learning algorithm $\mathcal{A}$ is (R, ϵ)-compressible for some $R \in \mathbb{R}^+$ and $\epsilon \in \mathbb{R}$ if there exists sequences of hypothesis books $\{\mathcal{H}_m\}_{m \in \mathbb{N}}$, $\mathcal{H}_m = \{\hat{\mathbf{w}}_{j;m} = (\hat{w}_{j,1}, \dots, \hat{w}_{j,m}), j \in [l_m]\} \subseteq \hat{\mathcal{W}}^m$, such that $l_m \leqslant e^{mR}$ and

$$\lim_{m \to \infty} \mathbb{P}\left(\min_{j \in [l_m]} d(W^m, \hat{\mathbf{w}}_{j;m}; S^m) > \epsilon \right) = \lim_{m \to \infty} \mathbb{P}(\forall j \in [l_m] \colon d(W^m, \hat{\mathbf{w}}_{j;m}; S^m) > \epsilon) = 0, \tag{19}$$

where $(S^m, W^m) \sim {P_{S,W}}^{\otimes m}$. Furthermore, for $0 < \delta < 1$ $\mathcal{A}$ is (R, ϵ, δ)-exponentially compressible if in addition to (20) the above conditions, we have

$$\lim_{m \to \infty} \left[-\frac{1}{m} \log\left(\mathbb{P}\left(\min_{j \in [l_m]} d(W^m, \hat{\mathbf{w}}_{j;m}; S^m) > \epsilon \right) \right) \right] \geqslant \log(1/\delta). \tag{20}$$

In [SGRS22, Theorems 2 and 9] it is shown that the generalization error of $\mathcal{A}$ can be upper bounded in terms of the compressibility parameters of $\mathcal{A}$. Furthermore, in [SGRS22, Theorems 3, 4, and 10] it is shown that the compressibility of the algorithm, and consequently its generalization error, can be upper bounded in terms of an adaptation of the rate-distortion function for the learning algorithm, as stated in (1). Here, we re-state this: for every distribution Q defined over $\mathcal{S} \times \mathcal{W}$ and every distortion threshold $\epsilon \in \mathbb{R}$, the rate-distortion function for $\mathcal{A}$ is defined as

$$\mathfrak{RD}(Q, \epsilon) \coloneqq \inf_{P_{\hat{W}|S}} I(S; \hat{W}), \tag{21}$$

$$\text{s.t.} \quad \mathbb{E}_{(S,W) \sim Q}[\operatorname{gen}(S, W)] - \mathbb{E}_{(S, \hat{W}) \sim Q_S P_{\hat{W}|S}}[\operatorname{gen}(S, \hat{W})] \leqslant \epsilon. \tag{22}$$

Here, Q_S is the Q-marginal of S, and the infimum is taken over all Markov kernels (conditional distributions) of a random variable $\hat{W} \in \hat{\mathcal{W}}$ given S. Note that $\operatorname{gen}(S, \hat{W})$ is defined with respect to the loss function $\hat{\ell}(z, \hat{w})$.

Every conditional distribution $P_{\hat{W}|S}$ corresponds to a compressed learning algorithm that picks a compressed hypothesis $\hat{W} \in \hat{\mathcal{W}}$ for S according to $P_{\hat{W}|S}$. While this compressed learning algorithm does not depend directly on W, it needs to satisfy the distortion criterion (22) which clearly depends on W.

Finally, suppose that the loss function is Lipschitz, *i.e.,* there exist some constant $\mathfrak{L} \in \mathbb{R}^+$ and distortion function $\rho \colon \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$ such that for every $z \in \mathcal{Z}$, $w \in \mathcal{W}$ and $\hat{w} \in \hat{\mathcal{W}}$, it holds that $|\ell(z, w) - \hat{\ell}(z, \hat{w})| \leqslant \mathfrak{L}\rho(w, \hat{w})$. Under this Lipschitz assumption, in [SGRS22, Corollaries 6 and 11] it is shown that the rate-distortion function (21) can be upper bounded in terms of the classic rate-distortion function $\mathfrak{RD}(Q_W, 2\epsilon/\mathfrak{L})$ of W with respect to the distortion function ρ. Here, Q_W is the Q-marginal of W.

Hyperparameter	Symbol	Value
Initial learning rate	η_0	0.01
Regularization parameter	α	0.00001
Batch size	b	1
Kernel parameter	γ	0.01
Dimension of the kernel feature space	p	2000

Table 1: Hyperparameters

B Details of the experiments

In this section, we describe the details of the performed simulations.

B.1 Tasks description

In our experiments, we simulate a distributed learning setup on a single machine, where each "client" is an instance of the SVM model, equipped with a dataset and a model to train. We iteratively train each client's model. All tasks are performed on the same machine and we do not consider any communication constraint as it is not of interest in this paper. Once all clients' models are trained, their weights are averaged and used by another instance of the SVM model (meant to be the one at the central server). This is utilized for computing $\hat{\mathcal{L}}_\theta(S_{1:K}, \overline{W})$ and $\mathcal{L}(\overline{W})$. In particular, $\mathcal{L}(\overline{W})$ is estimated using a test set of fixed size.

For comparison, we also consider the associated centralized learning setup consisting of a single SVM model that needs to be trained on a dataset S of size $N = nK$. For the distributed learning setting, the (local) dataset of client # $i \in [K]$, denoted as S_i, is composed of n samples that are drawn uniformly at random from S.

Each experiment is repeated 10 times and only average values are reported.

B.2 Datasets

We consider a supervised learning setup, where the task is binary image classification. More precisely, in our experiment, we train the SVM model to distinguish digits 1 and 6 from the standard MNIST [LCB10]. Similar results, that are not reported here, were obtained for distinguishing other pairs of digits.

B.3 Model

Every client uses SVM with the Gaussian kernel and applies Stochastic Gradient Descent (SGD) for the optimization.

B.4 Training and hyperparameters

All models were trained with the same hyperparameters for up to 200 epochs until the empirical risk is made smaller than 0.001. The data is scaled and normalized to get zero-mean and unit variance. The used training hyperparameters are listed in Table 1 Note that we chose an adaptive learning rate, starting from η_0, which is then decreased by 0.2 if the training loss does not increase by more than 0.01 during 10 epochs.

B.5 Hardware and other resources

We performed our experiments on a server equipped with 56 CPUs Intel Xeon E5-2690v4 2.60GHz.

The experiments are conducted using Python language. The open source machine learning library scikit-learn [PVG+11] is used to implement SVM. In particular, we use *SGDClassifier* to implement SGD optimization and *RBFSampler* for Gaussian kernel feature map approximation. The code used for the experiments is available at `https://github.com/RomainChor/DataScience/tree/master/Generalization_NeurIPS2022`.

C Other results

Here, we present further analytical results on the generalization error of distributed learning setups.

C.1 Lipschitz loss

Here, we provide an upper bound on the expectation of the generalization error of a distributed learning algorithm with deterministic local learning algorithms. This upper bound, proved in Section D.3, has a decreasing behavior in comparison to the corresponding centralized algorithm.

Proposition 1. *Suppose the local learning algorithms $\mathcal{A}_i(S_i) = W_i \in \mathbb{R}^d$, $i \in [K]$, are deterministic with bounded hypothesis variance σ_W^2. Moreover, suppose that for every $w \in \mathcal{W}$, the loss $\ell(Z, w)$ is σ-subgaussian and further $\ell(z, w)$ is $\mathfrak{L}$-Lipschitz,* i.e., $|\ell(z, w) - \ell(z, w')| \leqslant \mathfrak{L}\|w - w'\|^2$. *Then,*

$$\mathbb{E}\big[\mathrm{gen}(S_{1:K}, \bar{W})\big] \leqslant \min\left(2\sqrt[3]{\frac{2\mathfrak{L}\sigma^2\sigma_W^2}{nK^2}}, \frac{2\mathfrak{L}\sigma_W^2}{K^2}\right).$$

C.2 Comparison with related prior art results

Here, we show that some of the existing results are particular cases of our established bounds.

C.2.1 Comparison with [YDVP20]

As we already mentioned in Section 3.2, under the assumed subgaussianity condition of the loss function our result of Theorem 4 recovers the upper bound of [YDVP20, Theorem 2]. More precisely, this is obtained by setting $\epsilon = 0$ in the first term of the minimization in (7).

C.2.2 Comparison with [BDP22]

Hereafter we show that under the assumed subgaussianity condition of the loss function the upper bounds of [BDP22, Theorems 4 & 5] can be recovered as special cases, and are subsumed by our Theorem 4. Specifically, the below is obtained by investigating, and comparing with, the second term of the minimization in (8).

i) **[BDP22, Theorem 4]:** First, we recall the result. We only focus on the upper bound in [BDP22, Theorem 4, part ii.] – a similar analysis follows for the part i of their theorem.
Let for $\mathcal{Z}, \mathcal{W} \in \mathbb{R}^d$ the loss function $\ell(Z, w)$ be σ-subgaussian for every $w \in \mathcal{W}$. Also, similar to in [BDP22], assume that the loss $\ell(z, w)$ is a Bregman divergence, *i.e.,*

$$\ell(z, w) := F(w) - F(z) - \langle \nabla F(z), w - z \rangle,$$

for some convex function $F\colon \mathbb{R}^d \mapsto \mathbb{R}$. Then, as per [BDP22], we have

$$\mathbb{E}\big[\mathrm{gen}(S_{1:K}, \overline{W})\big] \leqslant \frac{1}{K}\sqrt{\frac{2\sigma^2 \max_{i\in[K]} I(S_i; W_i)}{n}}.$$

Hereafter we show that this result can be obtained as a specific case of our Theorem 4. Specifically, let $\epsilon := \max_{i\in[K]} \epsilon_i$, where

$$\epsilon_i := \frac{1}{K}\sqrt{\frac{2\sigma^2 I(S_i; W_i)}{n}}, \quad i \in [K].$$

Also, let $\hat{\overline{W}}_i$ be a constant. Thus, $\mathbb{E}[\mathrm{gen}(S_i, \hat{\overline{W}}_i)] = 0$ and $I(S_i; \overline{W}_i | W_{1:K\setminus i}) = 0$. We have,

$$\begin{aligned}
&\mathbb{E}\Big[\,\mathrm{gen}(S_i, \overline{W}) - \mathrm{gen}(S_i, \hat{\overline{W}}_i)\Big] \\
&\quad = \mathbb{E}_{(S_i, \overline{W}) \sim P_{S_i, \overline{W}}}\big[\mathrm{gen}(S_i, \overline{W})\big] \\
&\quad = \mathbb{E}_{(S_i, \tilde{S}_i, \overline{W}) \sim P_{S_i, \overline{W}} P_{\tilde{S}_i}}\Big[\hat{\mathcal{L}}(\tilde{S}_i, \overline{W}) - \hat{\mathcal{L}}(S_i, \overline{W})\Big] \\
&\quad = \frac{1}{n}\sum_{j\in[n]} \mathbb{E}_{(Z_{i,j}, \tilde{Z}_{i,j}, \overline{W}) \sim P_{Z_{i,j}, \overline{W}} P_{\tilde{Z}_{i,j}}}\Big[\ell(\tilde{Z}_{i,j}, \overline{W}) - \ell(Z_{i,j}, \overline{W})\Big]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=}\frac{1}{n}\sum_{j\in[n]}\mathbb{E}_{(Z_{i,j},\tilde{Z}_{i,j},\overline{W})\sim P_{Z_{i,j},\overline{W}}P_{\tilde{Z}_{i,j}}}\Big[\langle\nabla F(Z_{i,j}),\overline{W}-Z_{i,j}\rangle-\langle\nabla F(\tilde{Z}_{i,j}),\overline{W}-\tilde{Z}_{i,j}\rangle\Big]\\
&\overset{(b)}{=}\frac{1}{n}\sum_{j\in[n]}\mathbb{E}_{(Z_{i,j},\tilde{Z}_{i,j},\overline{W})\sim P_{Z_{i,j},\overline{W}}P_{\tilde{Z}_{i,j}}}\Big[\langle\nabla F(Z_{i,j}),\frac{W_i}{K}-Z_{i,j}\rangle-\langle\nabla F(\tilde{Z}_{i,j}),\frac{W_i}{K}-\tilde{Z}_{i,j}\rangle\Big]\\
&\overset{(c)}{=}\frac{1}{n}\sum_{j\in[n]}\mathbb{E}_{(Z_{i,j},\tilde{Z}_{i,j},\overline{W})\sim P_{Z_{i,j},\overline{W}}P_{\tilde{Z}_{i,j}}}\Big[\langle\nabla F(Z_{i,j}),\frac{W_i}{K}\rangle-\langle\nabla F(\tilde{Z}_{i,j}),\frac{W_i}{K}\rangle\Big]\\
&=\frac{1}{nK}\sum_{j\in[n]}\mathbb{E}_{(Z_{i,j},\tilde{Z}_{i,j},\overline{W})\sim P_{Z_{i,j},\overline{W}}P_{\tilde{Z}_{i,j}}}\Big[\langle\nabla F(Z_{i,j}),W_i\rangle-\langle\nabla F(\tilde{Z}_{i,j}),W_i\rangle\Big]\\
&\overset{(d)}{=}\frac{1}{K}\mathbb{E}[\mathrm{gen}(S_i,W_i)]\\
&\overset{(e)}{=}\frac{1}{K}\sqrt{\frac{2\sigma^2 I(S_i;W_i)}{n}}\\
&=\epsilon_i,
\end{aligned}
$$

where $\tilde{S}_i := \{\tilde{Z}_{i,1},\dots,\tilde{Z}_{i,n}\}$ and $P_{\tilde{S}_i} = \mu^{\otimes n}$ which is identical to P_{S_i}, (a) holds since the loss is a Bregman loss and by using that $\mathbb{E}[F(Z_{i,j})] = \mathbb{E}[F(Z'_{i,j})]$, $i \in [K], j \in [n]$, (b) holds since $\overline{W} = (W_1 + \cdots + W_K)/K$ and $Z_{i,j}$ is independent of $W_{1:K\setminus i}$, and hence distributions of $(Z_i, W_{1:K\setminus i})$ and $(\tilde{Z}_i, W_{1:K\setminus i})$ are identical, (c) holds since the marginal distributions of $Z_{i,j}$ and $\tilde{Z}_{i,j}$ are identical, (d) follows by straightforward algebra using the fact that the loss is a Bregman loss, and finally (e) is due to [XR17, Theorem 1].

Moreover,

$$\mathfrak{RD}_i(P_{S_i,W_{1:K},\overline{W}},\epsilon) \leqslant I(S_i;\overline{W}_i|W_{1:K\setminus i}) = 0.$$

Now, using the second term of the minimization in (8) completes the proof.

ii) [BDP22, Theorem 5]: First, we recall the result. Suppose that $\mathcal{Z},\mathcal{W}\in\mathbb{R}^d$, for every $w\in\mathcal{W}$, the loss function $\ell(Z,w)$ is σ-subgaussian, and the loss is a Lipschitz continuous function of w with respect to L_2-norm, *i.e.,*

$$\big|\ell(z,w)-\ell(z,w')\big| \leqslant c\|w-w'\|, \quad \forall z\in\mathcal{Z}, w,w'\in\mathcal{W},$$

for some constant $c>0$. Moreover, suppose that for every $i\in[K]$,

$$\mathbb{E}[\|W_i-\mathbb{E}[W_i]\|] \leqslant \sigma_W,$$

where the expectations are with respect to the marginals P_{W_i}. Then, as per [BDP22] we have

$$\mathbb{E}\big[\mathrm{gen}(S_{1:K},\overline{W})\big] \leqslant \frac{2c\sigma_W}{K}. \tag{23}$$

We show that this result too can be obtained as a special case of the bound of our Theorem 4. Specifically, for every $i\in[K]$ let

$$\hat{\overline{W}}_i := \frac{\mathbb{E}[W_i]+\sum_{j\neq i}W_j}{K}.$$

Hence $I(S_i;\overline{W}_i|W_{1:K\setminus i}) = 0$. This, we have

$$
\begin{aligned}
\mathbb{E}\Big[\mathrm{gen}(S_i,\overline{W})-\mathrm{gen}(S_i,\hat{\overline{W}}_i)\Big] &\overset{(a)}{\leqslant} 2c\mathbb{E}\Big[\|\overline{W}-\hat{\overline{W}}_i\|\Big]\\
&=\frac{2c}{K}\mathbb{E}[\|W_i-E[W_i]\|]\\
&\overset{(b)}{\leqslant}\frac{2c\sigma_W}{K} =: \epsilon,
\end{aligned}
$$

where (a) holds by the assumed Lipschitzness of the loss function and (b) follows by using (23).

Moreover,

$$\mathfrak{RD}_i(P_{S_i,W_{1:K},\overline{W}},\epsilon) \leqslant I(S_i;\overline{W}_i|W_{1:K\setminus i}) = 0.$$

Finally, using the second term of the minimization in (8) completes the proof.

D Proofs

Here, we state the proofs of the results in the main text and the appendices.

D.1 Proof of Theorem 3

Proof. Similar to [SGRS22, Theorem 9], 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

$$\begin{aligned}
\mathbb{P}\big(\mathrm{gen}(S_{1:K}, \overline{W}) \geqslant \Delta\big)^m &= \mathbb{P}\big(\forall j \in [m]\colon\ \mathrm{gen}(S_{1:K,j}, \overline{W}_j) \geqslant \Delta\big) \\
&\leqslant \mathbb{P}\left(\frac{1}{mK}\sum_{i\in[K]}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta\right) \\
&\overset{(a)}{\leqslant} K \max_{i\in[K]} \mathbb{P}\left(\frac{1}{m}\sum_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta_i\right), \qquad (24)
\end{aligned}$$

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,

$$\mathbb{P}\Big(\frac{1}{m}\sum\nolimits_{j\in[m]} \mathcal{L}(\overline{W}_j) - \hat{\mathcal{L}}(S_{i,j}, \overline{W}_j) \geqslant \Delta_i\Big) \leqslant \delta^m, \qquad \text{for every } i \in [K]. \qquad (25)$$

Combining this claim with (24) 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 (25) 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 $\mathbf{S} := \{S_{i,j}\}_{j\in[m]}$ with $S_j := S_{i,j}$, $\overline{\mathbf{W}} := \big\{\overline{W}_j\big\}_{j\in[m]}$, and $\mathbf{U} := \big\{W_{1:K\setminus i,j}\big\}_{j\in[m]}$ with $U_j := W_{1:K\setminus i,j}$. 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 $\mathbf{z} := (z_1, \dots, z_m) \in \mathcal{Z}^m$, define $\ell(\mathbf{z}, \overline{\mathbf{w}}) := \frac{1}{m}\sum_{j\in[m]} \ell(z_j, \overline{w}_j)$. 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 (25) 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 10 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 $R(\delta^m, \epsilon) := \sup_{Q:\, D_{KL}\left(Q \| P_{\mathbf{S},\overline{\mathbf{W}},\mathbf{U}}\right) \leqslant \log(1/\delta^m)} \inf_{P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}} I\left(\mathbf{S}; \hat{\overline{\mathbf{W}}} | \mathbf{U}\right)$ and the infimum is over all Markov kernels $P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}$ such that

$$\mathbb{E}\Big[\mathrm{gen}(\mathbf{S}, \overline{\mathbf{W}}) - \mathrm{gen}(\mathbf{S}, \hat{\overline{\mathbf{W}}})\Big] \leqslant \epsilon. \tag{26}$$

Note that as explained in Remark 1 (and as noted in [SGRS22, Equation (49)]), the above constraint is stronger than the one needed in [SGRS22, Extended Theorem 10], 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 10 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 Appendix D.4.

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

This completes the proof of Theorem 3. □

D.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

$$R_p(\delta, \epsilon, i) \leqslant m \log((c_2 + \nu)/\nu), \tag{27}$$

where

$$\epsilon := 8e^{-\frac{m}{7}\left(\frac{K\theta}{4B}\right)^2} + \frac{2m\nu^m}{\sqrt{\pi}} e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{4c_1\nu B}\right)^2} + 4e^{-0.21m(c_1^2-1)} + 4e^{-0.21m(c_2^2-1)}. \tag{28}$$

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 (27). 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

$$I\Big(S_i; \hat{\overline{W}}_i | W_{1:K\setminus i}\Big) \leqslant m\log((c_2 + \nu)/\nu) \tag{29}$$

$$\mathbb{E}\Big[\mathrm{gen}_\theta\big(S_i, \overline{W}\big) - \mathrm{gen}\Big(S_i, \hat{\overline{W}}_i\Big)\Big] \leqslant \epsilon, \tag{30}$$

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 $(0,\dots,0) \in \mathbb{R}^m$ 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 $(0,\dots,0) \in \mathbb{R}^m$, 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 (29) holds for any choice of A. Now, we proceed to show that there exists a proper choice of A such that (30) 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 (30) is with respect to $QP_{\hat{\overline{W}}_i|W_{1:K}}$. This term is bounded in the following lemma, proved in Appendix D.5.

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 Appendix D.6.

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*

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

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, (30) holds by Lemma 2. This completes the proof of Theorem 5. □

D.3 Proof of Proposition 1

Proof. Let V_i be a d-dimensional random variable distributed according to $\mathcal{N}(0,\sigma_V^2\mathrm{I}_d)$, independent of all other random variables, where I_d is the $d\times d$ identity matrix. Let $\hat{\overline{W}}_i := \frac{1}{K}\Big(\hat{W}_i + \sum_{j\neq i}W_j\Big)$, where $\hat{W}_i := W_i + V_i$. Note that

$$\mathbb{E}\Big[\mathrm{gen}(S_i,\overline{W}) - \mathrm{gen}(S_i,\hat{\overline{W}}_i)\Big] \leqslant 2\mathfrak{L}\mathbb{E}\Big[\|\overline{W} - \hat{\overline{W}}_i\|^2\Big] = \frac{2\mathfrak{L}\sigma_V^2}{K^2} =: \epsilon.$$

Hence,

$$\begin{aligned} \mathfrak{RD}_i(P_{S_i,W_{1:K},\overline{W}},\epsilon) &\leqslant I(S_i;\overline{W}_i|W_{1:K\setminus i}) \\ &= I(S_i;\hat{W}_i) \\ &= I(S_i;\mathcal{A}_i(S_i)+V_i) \\ &\overset{(a)}{\leqslant} \frac{1}{2\sigma_V^2}\mathrm{Var}(W_i) \end{aligned}$$

$$\leqslant \frac{\sigma_W^2}{2\sigma_V^2},$$

where (a) is derived due to [WHGC21, Lemma 6].

Now, using Theorem 4, we have

$$\mathbb{E}\big[\mathrm{gen}(S_{1:K}, \bar{W})\big] \leqslant \sqrt{\frac{\sigma^2\sigma_W^2}{n\sigma_V^2}} + \frac{2\mathfrak{L}\sigma_V^2}{K^2}.$$

Letting $\sigma_V^2 := \sqrt[3]{\frac{\sigma^2\sigma_W^2 K^4}{4\mathfrak{L}^2 n}}$ completes the proof.

In a similar way, let $\hat{\overline{W}}_i := \frac{1}{K}\Big(\mathbb{E}[W_i] + \sum_{j\neq i} W_j\Big)$. Then,

$$\mathbb{E}\Big[\mathrm{gen}(S_i, \overline{W}) - \mathrm{gen}(S_i, \hat{\overline{W}}_i)\Big] \leqslant 2\mathfrak{L}\mathbb{E}\big[\|W_i - \mathbb{E}[W_i]\|^2\big] = \frac{2\mathfrak{L}\sigma_W^2}{K^2} =: \epsilon.$$

Moreover,

$$\mathfrak{RD}_i(P_{S_i, W_{1:K}, \overline{W}}, \epsilon) \leqslant I(S_i; \overline{W}_i | W_{1:K\setminus i}) = 0.$$

Hence,

$$\mathbb{E}\big[\mathrm{gen}(S_{1:K}, \bar{W})\big] \leqslant \frac{2\mathfrak{L}\sigma_W^2}{K^2}.$$

□

D.4 Proof of Lemma 1

Proof. Let Q be product of m i.i.d. Q_j, where Q_j is a distribution defined over $\mathcal{S}\times\mathcal{W}\times\mathcal{U}$. Hence, $D_{KL}(Q_j \| P_{S_j, \overline{W}_j, U_j}) = \frac{1}{m} D_{KL}\Big(Q \| P_{\mathbf{S}, \overline{\mathbf{W}}, \mathbf{U}}\Big) \leqslant \log(1/\delta)$. Moreover,

$$\begin{aligned}
R(\delta^m, \epsilon) \leqslant & \sup_{Q_j:\, D_{KL}\left(Q_j \| P_{S_j, \overline{W}_j, U_j}\right) \leqslant \log(1/\delta)} \inf_{P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}} I(\mathbf{S}; \hat{\overline{\mathbf{W}}} | \mathbf{U}) \\
\overset{(a)}{=} & \sum_{j\in[m]} \sup_{Q_j:\, D_{KL}\left(Q_j \| P_{S_j, \overline{W}_j, U_j}\right) \leqslant \log(1/\delta)} \inf_{P_{\hat{\overline{W}}_j|S_j, U_j}} I(S_j; \hat{\overline{W}}_j | U_j) \\
= & m R_p(\delta, \epsilon, i).
\end{aligned}$$

where (a) is deduced to the following claim: under the considered Q (where $\big\{(S_j, \overline{W}_j, U_j)\big\}_{j\in[m]}$ are *i.i.d.*), to study the infimum, it is sufficient to consider $\hat{\overline{\mathbf{W}}}$ that can be written as $(\hat{\overline{W}}_1, \ldots, \hat{\overline{W}}_m)$, such that $P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}} = \prod_{i\in[m]} P_{\hat{\overline{W}}_j|S_j,U_j}$ and $\mathbb{E}\Big[\mathrm{gen}(S_j, \overline{W}_j) - \mathrm{gen}(S_j, \hat{\overline{W}}_j)\Big] \leqslant \epsilon$. To show this claim, given a $P_{\hat{\overline{\mathbf{W}}}|\mathbf{S},\mathbf{U}}$, let $\tilde{\overline{\mathbf{W}}} := (\hat{\overline{W}}_1, \ldots, \hat{\overline{W}}_m)$, where $\hat{\overline{W}}_j$, $j \in [m]$ are chosen based on (S_j, U_j), according to $P_{\hat{\overline{W}}_j|S_j,U_j} := P_{\hat{\overline{\mathbf{W}}}|S_j,U_j}$, independent of $\{S_{j'}, U_{j'}, \hat{\overline{W}}_{j'}\}_{j'\neq j}$. Then,

$$\begin{aligned}
I(\mathbf{S}; \hat{\overline{\mathbf{W}}} | \mathbf{U}) &= \sum\nolimits_{j\in[m]} I(S_j; \hat{\overline{\mathbf{W}}} | \mathbf{U}, S_{1:j-1}) \\
&\overset{(a)}{\geqslant} \sum\nolimits_{j\in[m]} I(S_j; \hat{\overline{\mathbf{W}}} | U_j) \\
&= \sum\nolimits_{j\in[m]} I(S_j; \hat{\overline{W}}_j | U_j) \\
&= I(\mathbf{S}; \tilde{\overline{\mathbf{W}}} | \mathbf{U})
\end{aligned}$$

where we define $S_{1:0} :=$ Constant, (a) is derived since $H(S_j|\mathbf{U}, S_{1:j-1}) = H(S_j)$ and $H(S_j|\hat{\overline{\mathbf{W}}}, \mathbf{U}, S_{1:j-1}) \leqslant H(S_j|\hat{\overline{\mathbf{W}}}, U_j)$. Note that $\tilde{\overline{\mathbf{W}}}$ satisfies the constraint (26), and hence for studying the infimum it is sufficient to consider $\tilde{\overline{\mathbf{W}}}$. Finally, due to the symmetry, $\forall j \in [m]$, $\mathbb{E}\Big[\mathrm{gen}(S_j, \overline{W}_j) - \mathrm{gen}(S_j, \tilde{W}_j)\Big] \leqslant \epsilon$. This completes the proof. □

D.5 Proof of Lemma 2

Proof. Let $(Z, \overline{W}, W_{1:K}, \hat{\overline{W}}_i) \sim \mu Q_{\overline{W}, W_{1:K}} P_{\hat{\overline{W}}_i | W_{1:K}}$. Note that since $D_{KL}(Q \| P_{S_{1:K}, W_{1:K}, \overline{W}}) < \infty$, then under Q also $\overline{W} = (W_1 + \cdots + W_K)/K$. Hence, we have

$$
\begin{aligned}
\mathbb{E}_{\overline{W}}\big[\mathcal{L}(\overline{W})\big] =& \mathbb{E}_{(Z,\overline{W})}\Big[\mathbb{1}_{\{Y\langle X, \overline{W}\rangle \leqslant 0\}}\Big] \\
\leqslant& \mathbb{E}_{(Z,\overline{W},\hat{\overline{W}}_i)}\Big[\mathbb{1}_{\{Y\langle X, \overline{W}\rangle \leqslant 0, Y\langle\!\langle X, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} > \theta/2\}}\Big] + \mathbb{E}_{(Z,\overline{W},\hat{\overline{W}}_i)}\Big[\mathbb{1}_{\{Y\langle\!\langle X, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} \leqslant \theta/2\}}\Big] \\
=& \mathbb{E}_{(Z,\overline{W},\hat{\overline{W}}_i)}\Big[\mathbb{1}_{\{Y\langle X, \overline{W}\rangle \leqslant 0, Y\langle\!\langle X, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} > \theta/2\}}\Big] + \mathbb{E}_{\hat{\overline{W}}_i}\Big[\mathcal{L}\Big(\hat{\overline{W}}_i\Big)\Big] \\
\leqslant& \mathbb{E}_{(Z,\overline{W},\hat{\overline{W}}_i)}\Big[\mathbb{1}_{\{\left|\langle X, \overline{W}\rangle - \langle\!\langle X, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}}\right| > \theta/2\}}\Big] + \mathbb{E}_{\hat{\overline{W}}_i}\Big[\mathcal{L}\Big(\hat{\overline{W}}_i\Big)\Big] \\
=& \mathbb{E}_{(Z,W_i,\hat{W}_i)}\Big[\mathbb{1}_{\{\frac{1}{K}\left|\langle X, W_i\rangle - \langle \mathsf{A}X, \hat{W}_i\rangle\right| > \theta/2\}}\Big] + \mathbb{E}_{\hat{\overline{W}}_i}\Big[\mathcal{L}\Big(\hat{\overline{W}}_i\Big)\Big]. \qquad (31)
\end{aligned}
$$

Similarly, and by considering $(S_i, \overline{W}, W_{1:K}, \hat{\overline{W}}_i) \sim \mu Q_{S_i, \overline{W}, W_{1:K}} P_{\hat{\overline{W}}_i | W_{1:K}}$, with $S_i = \{Z_{i,1}, \ldots, Z_{i,n}\}$, we have

$$
\begin{aligned}
1 - & \mathbb{E}_{(S_i,\overline{W})}\Big[\hat{\mathcal{L}}_\theta(S_i, \overline{W})\Big] \\
=& \mathbb{E}_{(S_i,\overline{W})}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{Y_{i,j}\langle X_{i,j}, \overline{W}\rangle > \theta\}}\right] \\
\leqslant& \mathbb{E}_{(S_i,\overline{W},\hat{\overline{W}}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{Y_{i,j}\langle X_{i,j}, W\rangle > \theta, Y_{i,j}\langle\!\langle X_{i,j}, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} \leqslant \theta/2\}}\right] \\
&+ \mathbb{E}_{(S_i,\overline{W},\hat{\overline{W}}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{Y_{i,j}\langle\!\langle X_{i,j}, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} > \theta/2\}}\right] \\
=& \mathbb{E}_{(S_i,\overline{W},\hat{\overline{W}}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{Y_{i,j}\langle X_{i,j}, \overline{W}\rangle > \theta, Y_{i,j}\langle\!\langle X_{i,j}, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}} \leqslant \theta/2\}}\right] + 1 - \mathbb{E}_{(S_i,\hat{\overline{W}}_i)}\Big[\hat{\mathcal{L}}\Big(S_i, \hat{\overline{W}}_i\Big)\Big] \\
\leqslant& \mathbb{E}_{(S_i,\overline{W},\hat{\overline{W}}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\left|\langle X_{i,j}, \overline{W}\rangle - \langle\!\langle X_{i,j}, \hat{\overline{W}}_i\rangle\!\rangle_{\mathsf{A}}\right| > \theta/2\}}\right] + 1 - \mathbb{E}_{(S_i,\hat{\overline{W}}_i)}\Big[\hat{\mathcal{L}}\Big(S_i, \hat{\overline{W}}_i\Big)\Big] \\
=& \mathbb{E}_{(S_i,W_i,\hat{W}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\frac{1}{K}\left|\langle X_{i,j}, W_i\rangle - \langle \mathsf{A}X_{i,j}, \hat{W}_i\rangle\right| > \theta/2\}}\right] + 1 - \mathbb{E}_{(S_i,\hat{\overline{W}}_i)}\Big[\hat{\mathcal{L}}\Big(S_i, \hat{\overline{W}}_i\Big)\Big]. \qquad (32)
\end{aligned}
$$

Combining (31) and (32) completes the proof. □

D.6 Proof of Lemma 3

Proof. We start by further upper bounding D_{A} by sum of four terms:

$$
\begin{aligned}
D_{\mathsf{A}} =& \mathbb{E}_{(Z,W_i,\hat{W}_i)}\Big[\mathbb{1}_{\{\frac{1}{K}\left|\langle X, W_i\rangle - \langle \mathsf{A}X, \hat{W}_i\rangle\right| > \theta/2\}}\Big] + \mathbb{E}_{(S_i,W_i,\hat{W}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\frac{1}{K}\left|\langle X_{i,j}, W_i\rangle - \langle \mathsf{A}X_{i,j}, \hat{W}_i\rangle\right| > \theta/2\}}\right] \\
\leqslant& \mathbb{E}_{(Z,W_i)}\Big[\mathbb{1}_{\{\frac{1}{K}|\langle X, W_i\rangle - \langle \mathsf{A}X, \mathsf{A}W_i\rangle| > \theta/4\}}\Big] + \mathbb{E}_{(S_i,W_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\frac{1}{K}|\langle X_{i,j}, W_i\rangle - \langle \mathsf{A}X_{i,j}, \mathsf{A}W_i\rangle| > \theta/4\}}\right] \\
&+ \mathbb{E}_{(Z,W_i,\hat{W}_i)}\Big[\mathbb{1}_{\{\frac{1}{K}\left|\langle X, W_i\rangle - \langle \mathsf{A}X, \hat{W}_i - \mathsf{A}W_i\rangle\right| > \theta/4\}}\Big] \\
&+ \mathbb{E}_{(S_i,W_i,\hat{W}_i)}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\frac{1}{K}\left|\langle X_{i,j}, W_i\rangle - \langle \mathsf{A}X_{i,j}, \hat{W}_i - \mathsf{A}W_i\rangle\right| > \theta/4\}}\right] \\
\leqslant& D_{\mathsf{A},1} + D_{\mathsf{A},2} + D_{\mathsf{A},3} + D_{\mathsf{A},4},
\end{aligned}
$$

where

$$
\begin{aligned}
D_{\mathsf{A},1} :=& \mathbb{E}_{(Z,W_i)\sim\mu Q_{W_i}}\left[\mathbb{1}_{\left\{\frac{1}{K}|\langle X,W_i\rangle-\langle \mathsf{A}X,\mathsf{A}W_i\rangle|>\theta/4\right\}}\right] \\
&+ \mathbb{E}_{(S_i,W_i)\sim Q_{S_i,W_i}}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{\frac{1}{K}|\langle X_{i,j},W_i\rangle-\langle \mathsf{A}X_{i,j},\mathsf{A}W_i\rangle|>\theta/4\right\}}\right] \\
D_{\mathsf{A},2} :=& \mathbb{E}_{(Z,W_i,\hat{W}_i)\sim\mu Q_{W_i}P_{\hat{W}_i|W_i}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle \mathsf{A}X,\hat{W}_i-\mathsf{A}W_i\rangle\right|>\theta/4\right\}}\middle|\|\mathsf{A}X\|\leqslant c_1B, \|\mathsf{A}W_i\|\leqslant c_2\right] \\
&+ \mathbb{E}_{S_i\sim Q_{S_i}}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{(W_i,\hat{W}_i)\sim Q_{W_i|S_i}P_{\hat{W}_i|W_i}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle \mathsf{A}X_{i,j},\hat{W}_i-\mathsf{A}W_i\rangle\right|>\theta/4\right\}}\middle|\|\mathsf{A}X_{i,j}\|\leqslant c_1B, \|\mathsf{A}W_i\|\leqslant c_2\right] \\
D_{\mathsf{A},3} :=& \mathbb{P}_{Z\sim\mu}(\|\mathsf{A}X\|\geqslant c_1B) + \mathbb{E}_{S_i\sim Q_{S_i}}\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\|\mathsf{A}X_{i,j}\|\geqslant c_1B\}} \\
=& 2\mathbb{P}_{Z\sim\mu}(\|\mathsf{A}X\|\geqslant c_1B) \\
D_{\mathsf{A},4} :=& \mathbb{P}_{(Z,W_i)\sim\mu Q_{W_i}}(\|\mathsf{A}W_i\|\geqslant c_2) + \mathbb{E}_{S_i\sim Q_{S_i}}\frac{1}{n}\sum_{j=1}^{n}\mathbb{P}_{W_i\sim Q_{W_i|S_i}}(\|\mathsf{A}W_i\|\geqslant c_2) \\
=& 2\mathbb{P}_{W_i\sim Q_{W_i}}(\|\mathsf{A}W_i\|\geqslant c_2)
\end{aligned}
$$

The proof completes by showing

i. $\mathbb{E}_{\mathsf{A}}[D_{\mathsf{A},1}] \leqslant 8e^{-\frac{m}{7}\left(\frac{K\theta}{4B}\right)^2}$.

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

iii. $\mathbb{E}_{\mathsf{A}}[D_{\mathsf{A},3}] \leqslant 4e^{-0.21m(c_1^2-1)}$.

iv. $\mathbb{E}_{\mathsf{A}}[D_{\mathsf{A},4}] \leqslant 4e^{-0.21m(c_2^2-1)}$.

Now, we proceed to show each part of the above remaining claim.

Part i. First, for any $x, w_i \in \mathbb{R}^d$,

$$
\begin{aligned}
\mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}|\langle x,w_i\rangle-\langle \mathsf{A}x\mathsf{A}w_i\rangle|>\theta/4\right\}}\right] =& \mathbb{P}_{\mathsf{A}}\left(\frac{1}{K}|\langle x,w_i\rangle-\langle \mathsf{A}x\mathsf{A}w_i\rangle|>\theta/4\right) \\
\overset{(a)}{\leqslant}& 4e^{-\frac{m}{7}\left(\frac{K\theta}{4B}\right)^2},
\end{aligned}
$$

where (a) is due to [GKL20, Lemma 8, part 2.]. Part i is then proved due to the linearity of the expectation.

Part ii. Fix $\mathsf{A} \in \mathbb{R}^{d\times m}$ and $x, w_i \in \mathbb{R}^d$ such that $\|\mathsf{A}x\| \leqslant c_1B$ and $\|\mathsf{A}w_i\| \leqslant c_2$. The proof completes by showing that

$$
\mathbb{E}_{\hat{W}_i\sim P_{\hat{W}_i|w_i}}\left[\mathbb{1}_{\left\{\frac{1}{K}|\langle \mathsf{A}x,\hat{w}_i-\mathsf{A}w_i\rangle|>\theta/4\right\}}\right] \leqslant \frac{m\nu^m}{\sqrt{\pi}}e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{4c_1\nu B}\right)^2}.
$$

Note that since $\|\mathsf{A}w_i\| \leqslant c_2$, $P_{\hat{W}_i|w_i}$ is a uniform distribution over m-dimensional ball of radius ν with the center $\mathsf{A}w_i$. Denote m-dimensional ball of radius $r \in \mathbb{R}^+$ around the origin by $\mathcal{B}_r$ and the uniform distribution over this ball by $\mathcal{U}(\mathcal{B}_r)$.

$$
\begin{aligned}
\mathbb{E}_{\hat{W}_i\sim P_{\hat{W}_i|w_i}}\left[\mathbb{1}_{\left\{\frac{1}{K}|\langle \mathsf{A}x,\hat{w}_i-\mathsf{A}w_i\rangle|>\theta/4\right\}}\right] =& \mathbb{P}_{V\sim\mathcal{U}(\mathcal{B}_\nu)}\left(\frac{1}{K}|\langle \mathsf{A}x,V\rangle|>\theta/4\right) \\
=& \mathbb{P}_{V\sim\mathcal{U}(\mathcal{B}_\nu)}\left(\left|\left\langle\frac{\mathsf{A}x}{\|\mathsf{A}x\|},V\right\rangle\right|>\frac{K\theta}{4\|\mathsf{A}x\|}\right) \\
\overset{(a)}{\leqslant}& \mathbb{P}_{V\sim\mathcal{U}(\mathcal{B}_\nu)}\left(\left|\left\langle\frac{\mathsf{A}x}{\|\mathsf{A}x\|},V\right\rangle\right|>\frac{K\theta}{4c_1B}\right)
\end{aligned}
$$

$$\overset{(b)}{=} \mathbb{P}_{V\sim\mathcal{U}(\mathcal{B}_\nu)}\left(|\langle e_1, V\rangle| > \frac{K\theta}{4c_1 B}\right)$$

where $e_1 := (1,0,0,\ldots,0) \in \mathbb{R}^m$, (a) is derived since $\|\mathsf{A}x\| \leqslant c_1 B$, and (b) is derived due to symmetry of the uniform distribution. In other words, the above expectation is bounded by the probability that the absolute value of the projection of a uniformly chosen random variable over m-ball $\mathcal{B}_\nu$ on one of the axis becomes greater than $\frac{K\theta}{4c_1 B}$. Let $t := \frac{K\theta}{4c_1\nu B}$. If $t \geqslant 1$, then the above probability is zero, and the claim is then proved. Suppose that $t < 1$.

$$\begin{aligned}
\mathbb{P}_{V\sim\mathcal{U}(\mathcal{B}_\nu)}(|\langle e_1, V\rangle| > t) &= \frac{2\Gamma(m/2+1)}{\sqrt{\pi}\Gamma(m/2+1/2)}\int_{t\nu}^{\nu}(\nu^2 - x^2)^{(m-1)/2}\mathrm{d}x \\
&= \frac{2\Gamma(m/2+1)\nu^m}{\sqrt{\pi}\Gamma(m/2+1/2)}\int_{t}^{1}(1 - y^2)^{(m-1)/2}\mathrm{d}y \\
&\leqslant \frac{m\nu^m}{\sqrt{\pi}}(1-t)(1-t^2)^{(m-1)/2} \\
&\leqslant \frac{m\nu^m}{\sqrt{\pi}}(1-t^2)^{(m+1)/2} \\
&= \frac{m\nu^m}{\sqrt{\pi}}e^{(m+1)\log(1-t^2)/2} \\
&\overset{(a)}{\leqslant} \frac{m\nu^m}{\sqrt{\pi}}e^{-(m+1)t^2/2},
\end{aligned}$$

where (a) holds for $t < 1$.

Part iii and iv. These parts are direct consequence of [GKL20, Lemma 8, part 1.].

□