Lessons from Generalization Error Analysis of Federated Learning : You May Communicate Less Often!

Milad Sefidgaran [1] **Romain Chor** [1 2] **Abdellatif Zaidi** [1 2] **Yijun Wan** [1]

Abstract

We investigate the generalization error of statistical learning models in a Federated Learning (FL) setting. Specifically, we study the evolution of the generalization error with the number of communication rounds R between K clients and a parameter server (PS), *i.e.,* the effect on the generalization error of how often the clients' local models are aggregated at PS. In our setup, the more the clients communicate with PS the less data they use for local training in each round, such that the amount of training data per client is identical for distinct values of R. We establish PAC-Bayes and rate-distortion theoretic bounds on the generalization error that account explicitly for the effect of the number of rounds R, in addition to the number of participating devices K and individual datasets size n. The bounds, which apply to a large class of loss functions and learning algorithms, appear to be the first of their kind for the FL setting. Furthermore, we apply our bounds to FL-type Support Vector Machines (FSVM); and derive (more) explicit bounds in this case. In particular, we show that the generalization bound of FSVM increases with R, suggesting that more frequent communication with PS diminishes the generalization power. This implies that the population risk decreases less fast with R than does the empirical risk. Moreover, our bound *suggests* that the generalization error of FSVM decreases faster than that of centralized learning by a factor of $\mathcal{O}(\sqrt{\log(K)/K})$. Finally, we provide experimental results obtained using neural networks (ResNet-56) which show evidence that not only may our observations for FSVM hold more generally but also that the population risk may even start to increase beyond some value of R.

[1]Huawei Paris Research Center, Paris, France [2]Université Gustave Eiffel, Champs-sur-Marne, France. Correspondence to: Milad Sefidgaran <milad.sefidgaran2@huawei.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

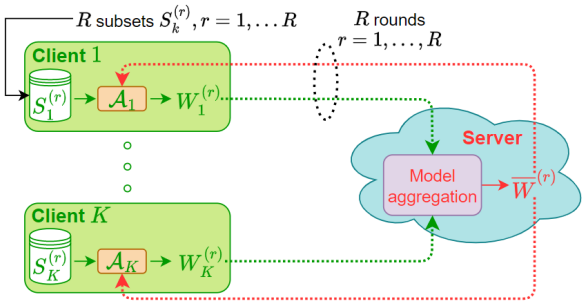


Figure 1. Multi-round Federated Learning

1. Introduction

A major focus of machine learning research over recent years has been the development of statistical learning algorithms that can be applied to spatially distributed data. In part, this is due to the emergence of new applications for which it is either not possible (due to lack of resources (Zinkevich et al., 2010; Kairouz et al., 2021)) or not desired (due to privacy concerns (Truex et al., 2019; Wei et al., 2020; Mothukuri et al., 2021)) to collect all the data at one point prior to applying a suitable machine learning algorithm to it (Konečnỳ et al., 2016; McMahan et al., 2017; Aguerri & Zaidi, 2019; Zaidi et al., 2020; Verbraeken et al., 2020; Moldoveanu & Zaidi, 2021; Kasiviswanathan et al., 2011; Sefidgaran et al., 2023). One popular such algorithm is Federated-Learning (FL) (McMahan et al., 2017; Yang et al., 2018; Kairouz et al., 2019; Li et al., 2020; Reddi et al., 2020; Karimireddy et al., 2020; Yuan et al., 2021). In FL, there are K clients or devices that hold each their own dataset and collaborate to collectively train a *(global) model* without sharing their data. The distributions of the clients' data can be identical (*homogeneous*) or different (*heterogeneous*). The model is depicted in Figure 1 and described in Section 2. An FL algorithm typically proceeds in $R \in \mathbb{N}^*$ *communication rounds*: at each round, every device produces a local model or hypothesis by means of a possibly different learning algorithm, *e.g.,* Stochastic Gradient Descent (SGD). Then, all clients' local models are sent to the *parameter server* (PS) which aggregates them into a (global) model. This aggregated model is sent back to

clients and used *typically* as an initialization point for the local algorithms in the next round, although more general forms of statistical dependencies are allowed.

The multi-round interactions between clients and PS are critical to the FL algorithm. Despite its importance, however, little is known about its effect on the performance of the algorithm. In fact, it was shown theoretically (Stich, 2019; Haddadpour et al., 2019; Qin et al., 2022), and also observed experimentally therein, that in FL-type algorithms the empirical risk generally decreases with the number of rounds. However, understanding the effect of the number of rounds on the *true* performance measure, the population risk (or test error), requires study of the evolution as a function of R of the generalization error, which is less amenable to analysis comparatively. In a recent work (Chor et al., 2023), it was shown that, in sharp contrast with the empirical risk, the generalization error may increase with the number of rounds. Their result, which is obtained by studying the evolution of a bound on the generalization error that they developed for their setup, relies strongly, however, on the taken assumptions, namely (i) linearity of the considered Bregman divergence loss with respect to the hypothesis, (ii) all the devices are constrained to run an SGD with mini-batch size one at every iteration and (iii) linearity of the aggregated model with respect to the devices' individual models. Also, even for the restricted setting considered therein, their bound on the generalization error (Chor et al., 2023, Theorem 1), which is essentially algebraic in nature, does not exploit fully the distributed architecture of the learning problem. Moreover, the dependence on the number of rounds is somewhat buried, at least in part, by decomposing the problem into several virtual one-client SGDs which are inter-dependent among clients and rounds through their initialization.

The effect of multi-round interactions on the generalization power of FL algorithms remains highly unexplored for general loss functions and algorithms, however. For example, with the very few exceptions that we discuss in this paper (most of which pertain to the specific case $R = 1$ (Yagli et al., 2020; Barnes et al., 2022a;b; Sefidgaran et al., 2022a), and with relatively strong assumptions on the loss function and algorithm or the model (Huang et al., 2021)) the existing literature crucially lacks true generalization bounds for FL-type algorithms, i.e., ones that bring the analysis of the distributed architecture and rounds into the bound, and even less that show some form of explicit dependence on R. One central mathematical difficulty in studying the behavior of the expected generalization error is caused by the interactions with the PS inducing statistical correlations among the devices' models which become stronger with R and are not easy to handle. For example, common tools that are generally applied in similar settings, such as the Leave-one-out Expansion Lemma of Shalev-Shwartz et al. (2010), do not apply easily in this case.

Contributions. In this paper, we study the problem of how the generalization error of FL-type algorithms evolves with the number of rounds R. Unless otherwise specified (for the specialization to Support Vector Machines in Section 4), we assume no particular form for the devices' individual algorithms or deterministic aggregation function at the PS. For this general setting:

- We establish PAC-Bayes bounds (Theorems 3.1 and 3.2) and rate-distortion theoretic bounds (Theorem 3.3 and Theorem D.1 in Appendix D.1) on the generalization error that account explicitly for the effect of the number of rounds R, in addition to the number of participating devices K and the size of the dataset n. These bounds appear to be the first of their kind for this problem. They reflect the structure of the distributed interactive learning algorithm in particular by capturing the contribution of each client at each round to the generalization error of the final model. The bounds are in terms of *averages* of such contributions among the clients and rounds. In a sense, this validates the intuition that, for a desired generalization error of the final model, some devices may be allowed to overfit during some or all of the rounds as long as other devices compensate for that overfitting. That is, the targeted generalization error level of the final model is suitably split among the devices and rounds. This intuition is also captured, but in a different way, by our *lossy* bounds of Theorems 3.2, 3.3, and D.1, in the form of a trade-off between the amounts of "lossy compressions" (or "distortion levels") of all clients across all rounds. Finally, we notice that the Kullback–Leibler divergence terms in our PAC-Bayes bounds have the advantage of involving priors that are possibly distinct across devices and rounds. This may be beneficial when these terms are used as regularizers during training. This direction, which can be seen as an extension of the centralized online setup of Haddouche & Guedj (2022) is left for future works.
- We apply our bounds to Federated Support Vector Machines (FSVM); and derive (more) explicit bounds on the generalization error in this case (Theorem 4.1). Interestingly, we show that the margin generalization bound of FSVM decreases with K and increases with R. In particular, this *suggests* that more frequent communication with the PS diminishes the generalization power of FSVM. Incidentally, this provides a theoretic justification of the experimentally observed slower decrease of the population risk with R (compared to that of empirical risk). Besides, our bound *suggests* that for any R, the generalization error of FSVM decreases faster than that of centralized learning by a factor of $\mathcal{O}(\sqrt{\log(K)/K})$, thereby generalizing recent findings in this direction (Sefidgaran et al., 2022a) for $R = 1$ to any arbitrary number of rounds.

- We validate our theoretical findings for FSVM through experiments. Moreover, we perform similar experiments using neural networks (ResNet-56) and observe that our findings for FSVM also hold more generally in this case. That is: (i) the generalization error increases with the number of rounds R, and (ii) not only the population risk decreases with R less fast than does the empirical risk, but may even start to increase beyond some threshold $R^\star$.

We hasten to mention that: (i) because in our experimental setup the amount of training data and SGD iterations are kept fixed regardless of the value of R (namely, $n = \alpha\,\tau \times R$, where τ is the number of per-round SGD steps for every client and α is a constant) the observed increase in the generalization error *cannot* be attributed to classic "overfitting". (ii) our experimental observation that the population risk may even start to increase beyond some value $R^\star$ was also observed experimentally in the recent work (Gu et al., 2023) (see Fig. 2 therein) in the context of the related setup of *LocalSGD* (Stich, 2019; Haddadpour et al., 2019; Qin et al., 2022).

Notations. We denote random variables (r.v.), their realizations, and their domains by upper-case, lower-case, and calligraphy fonts, *e.g.,* X, x, and $\mathcal{X}$. We denote the distribution of X by P_X and its support by $\operatorname{supp} P_X$. A r.v. X is called σ-subgaussian, if for all $t \in \mathbb{R}$, $\log \mathbb{E}[\exp(t(X - \mathbb{E}[X]))] \leqslant \sigma^2 t^2/2$, where $\mathbb{E}$ denote the expectation. As an example, if $X \in [a, b]$, then X is $\frac{b-a}{2}$-subgaussian. For two distributions P and Q with the Radon-Nikodym derivative $\mathrm{d}Q/\mathrm{d}P$ of Q with respect to P, the Kullback–Leibler (KL) divergence is defined as $D_{KL}(Q\|P) := \mathbb{E}_Q[\log(\mathrm{d}Q/\mathrm{d}P)]$ if $Q \ll P$ and ∞ otherwise. The mutual information between two r.v. (X, Y) with distribution $P_{X,Y}$ and marginals P_X and P_Y is defined as $I(X;Y) := D_{KL}(P_{X,Y}\|P_X P_Y)$. The notation $\{x_i\}_{i\in[m]}$ is used to denote a collection of m real numbers. The integer ranges $\{1, \ldots, K\} \subset \mathbb{N}^*$ and $\{K_1, \ldots, K_2\} \subset \mathbb{N}^*$ are denoted by $[K]$ and $[K_1 : K_2]$, respectively. Finally, for $k \in [K]$, we use the shorthand notation $[K]\backslash k := \{1, \ldots, K\}\backslash\{k\}$.

2. Formal problem setup

Consider the K-client federated learning model of Figure 1.

Datasets. For $k \in [K]$, let Z_k be some input data for client or device k distributed according to an unknown distribution μ_k over some data space $\mathcal{Z}_k = \mathcal{Z}$. For example, in supervised learning settings $Z_k := (X_k, Y_k)$ where X_k stands for a data sample at device k and Y_k its associated label. The distributions $\{\mu_k\}$ are allowed to be distinct across devices. [1] Each client is equipped with its own training dataset $S_k := \{Z_{k,1}, \ldots, Z_{k,n}\} \subseteq \mathcal{Z}^n$, consisting of n independent and identically distributed (i.i.d.) data points drawn according to the unknown distribution μ_k. We consider an R-round learning framework, $R \in \mathbb{N}^*$, where every sample of a training dataset can be used during only one round by the device that holds it, but possibly multiple times during that round. Accordingly, it is assumed that every device partitions its data S_k into R disjoint subsets such that $S_k = \bigcup_{r\in[R]} S_k^{(r)}$ where $S_k^{(r)}$ is the dataset used by client $k \in [K]$ during round $r \in [R]$. This setup, which has some similarity with "online Federated Learning" (Kamp et al., 2016; Kamp, 2019; Mitra et al., 2021; Chen et al., 2020; M. Ghari & Shen, 2022; Gogineni et al., 2022), encompasses many practical situations. The reader is referred to Appendix B for extensions of our results to a more general setup in which every data point may be used in multiple rounds and to Appendix F for further discussions on the studied setup. For ease of the exposition, we assume that R divides n and let $n_R := n/R$ and $S_k^{(r)} := \{Z_{k,1}^{(r)}, \ldots, Z_{k,n_R}^{(r)}\}$. Also, throughout we will often find it convenient to use the handy notation $P_{S_k} := \mu_k^{\otimes n}$, $P_{S_k^{(r)}} := \mu_k^{\otimes n_R}$ and, for $k \in [K]$ and $r \in [R]$,

$$S_{[K]}^{(r)} = S_1^{(r)}, \ldots, S_K^{(r)}, \qquad S_k^{[r]} = S_k^{(1)}, \ldots, S_k^{(r)}, \quad (1)$$
$$S_{[K]}^{[r]} = S_1^{[r]}, \ldots, S_k^{[r]} = S_{[K]}^{(1)}, \ldots, S_{[K]}^{(r)}, \qquad \mathbf{S} = S_{[K]}^{[R]}.$$

Similar notations will be used for other variables, e.g., $W_{[K]}^{(r)} = W_1^{(r)}, \ldots, W_K^{(r)}$ and $\overline{W}^{[R]} = \overline{W}^1, \ldots, \overline{W}^{(R)}$.

Overall algorithm. The devices collaboratively train a (global) model by performing both local computations and updates based on R-round interactions with the parameter server (PS). Let $\mathcal{A}_k$ denote the algorithm used by device $k \in [K]$. An example is the popular SGD where in round r every device k takes one or more gradient steps with respect to samples from the part $S_k^{(r)}$ of its local dataset S_k. It should be emphasized that the algorithms $\{\mathcal{A}_k\}$ may be identical or not. During round $r \in [R]$ the algorithm $\mathcal{A}_k$ produces a local model $W_k^{(r)}$. At the end of every round r, all the devices send their local models $W_k^{(r)}$ to the PS which aggregates them into a (global) model $\overline{W}^{(r)} \in \mathcal{W}$ and sends it back to them. The aggregated model is used by every device in the next round $(r + 1)$, together with the part $S_k^{(r+1)}$ of its dataset S_k in order to obtain a new local model $W_k^{(r+1)}$.

Local training at devices. Formally, the algorithm $\mathcal{A}_k$ is a possibly stochastic mapping $\mathcal{A}_k \colon \mathcal{Z}^{n_R} \times \mathcal{W} \to \mathcal{W}$; and, for $r \in [R]$, we have $W_k^{(r)} := \mathcal{A}_k(S_k^{(r)}, \overline{W}^{(r-1)})$ – for convenience, we set $\overline{W}^{(0)} = \varnothing$ or some default value. We denote the conditional distribution induced by $\mathcal{A}_k$ over $\mathcal{W}$

[1] Our framework and results can be extended to the case in which μ_k also depends on the running round $r = 1, \ldots, R$, as discussed in Appendix F; but, for the ease of exposition we restrict to μ_k independent of r.

at round r by $P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}$.

Model aggregation. The aggregation function at the PS is set to be deterministic and arbitrary, such that at round r this aggregation can be represented equivalently by a degenerate conditional distribution $P_{\overline{W}^{(r)}|W_{[K]}^{(r)}} = P_{\overline{W}^{(r)}|W_1^{(r)},\ldots,W_K^{(r)}}$. A common choice is the simple averaging $\overline{W}^{(r)} = \big(\sum_{k=1}^K W_k^{(r)}\big)/K$.

The above process repeats until all R rounds are completed, and yields a final (global) model $\overline{W}^{(R)}$. Let $\mathbf{W} = \big(W_{[K]}^{[R]}, \overline{W}^{[R]}\big)$ where the notation used here and throughout is similar to (1).

Induced probability distributions. The above-described algorithm, summarized in Algorithm 1 in the appendices, induces the conditional distribution

$$\bigotimes_{r\in[R]}\left\{\bigotimes_{k\in[K]}\Big(P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\Big)\,P_{\overline{W}^{(r)}|W_{[K]}^{(r)}}\right\}, \quad (2)$$

of models, denoted by $P_{\mathbf{W}|\mathbf{S}}$. We further denote their joint distribution by $\mathbf{S}$ by $P_{\mathbf{S},\mathbf{W}}$, which is equal to

$$\bigotimes_{r\in[R]}\left\{\bigotimes_{k\in[K]}\Big(P_{S_k^{(r)}}P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\Big)\,P_{\overline{W}^{(r)}|W_{[K]}^{(r)}}\right\}. \quad (3)$$

Hereafter, we will refer to the aforementioned algorithm for short as being an $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$*-FL model*.

Generalization error. Let $\ell\colon \mathcal{Z}\times\mathcal{W}\to\mathbb{R}^+$ be a given loss function. For a (global) model or hypothesis $\overline{w}^{(R)}\in\mathcal{W}$, its associated empirical and population risks are defined respectively as

$$\hat{\mathcal{L}}\left(\mathbf{s},\overline{w}^{(R)}\right) = \frac{1}{nK}\sum_{k=1}^K\sum_{i=1}^n \ell\big(z_{k,i},\overline{w}^{(R)}\big),$$
$$\mathcal{L}\big(\overline{w}^{(R)}\big) = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Z_k\sim\mu_k}\big[\ell\big(Z_k,\overline{w}^{(R)}\big)\big].$$

Note that by letting $\hat{\mathcal{L}}\left(s_k^{(r)},\overline{w}^{(R)}\right) = \frac{1}{n_R}\sum_{i=1}^{n_R}\ell\big(z_{k,i}^{(r)},\overline{w}^{(R)}\big)$, the empirical risk can be rewritten as

$$\hat{\mathcal{L}}\left(\mathbf{s},\overline{w}^{(R)}\right) = \frac{1}{KR}\sum_{r=1}^R\sum_{k=1}^K \hat{\mathcal{L}}\left(s_k^{(r)},\overline{w}^{(R)}\right).$$

The *generalization error* of the model $\overline{w}^{(R)}$ for dataset $\mathbf{s} = (s_1,\ldots,s_K)$, $s_k = \bigcup_{r=1}^R s_k^{(r)}$, is evaluated as

$$\begin{aligned}\mathrm{gen}\left(\mathbf{s},\overline{w}^{(R)}\right) &= \mathcal{L}\big(\overline{w}^{(R)}\big) - \hat{\mathcal{L}}\left(\mathbf{s},\overline{w}^{(R)}\right)\\ &= \frac{1}{KR}\sum_{r=1}^R\sum_{k=1}^K \mathrm{gen}\left(s_k^{(r)},\overline{w}^{(R)}\right), \quad (4)\end{aligned}$$

where $\mathrm{gen}\left(s_k^{(r)},\overline{w}^{(R)}\right) = \mathbb{E}_{Z_k\sim\mu_k}\big[\ell\big(Z_k,\overline{w}^{(R)}\big)\big] - \hat{\mathcal{L}}\left(s_k^{(r)},\overline{w}^{(R)}\right)$.

Example (FL-SGD). An important example is one in which every device runs *Stochastic Gradient Decent* (SGD) or variants of it, such as *mini-batch* SGD. In this latter case, denoting by e the number of *epochs* and by b the *mini-batch* size, at iteration $t\in[e\,n_R/b]$ client k updates its model as

$$W_{k,t}^{(r)} = \mathrm{Proj}\left(W_{k,t-1}^{(r)} + \frac{\eta_{r,t}}{b}\sum_{z\in\mathcal{B}_{k,r,t}}\nabla\tilde{\ell}\Big(z, W_{k,t-1}^{(r)}\Big)\right),$$

where $\tilde{\ell}\colon \mathcal{Z}\times\mathcal{W}\to\mathbb{R}^+$ is some *differentiable surrogate* loss function used for optimization, $\eta_{r,t}>0$ is the learning rate at iteration t of round r, $\mathcal{B}_{k,r,t}\in S_k^{(r)}$ is the mini-batch with size b chosen at iteration t, and $\mathrm{Proj}\big(w'\big) = \arg\min_{w\in\mathcal{W}}\|w-w'\|$. Also, in this case we let $W_k^{(r)} := W_{k,\tau}^{(r)}$, where $\tau := e\,n_R/b$. Besides, here the aggregation function at the PS is typically set to be the arithmetic average $\overline{W}^{(r)} = \big(\sum_{k\in[K]}W_k^{(r)}\big)/K$. This example will be analyzed in the context of Support Vector Machines in Section 4.

3. Generalization bounds for Federated Learning algorithms

In this section, we consider a (general) $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$*-FL algorithm*, as defined formally in Section 2; and study the generalization error of the final (global) hypothesis $\overline{W}^{(R)}$ as measured by (4). Note that the statistical properties of $\overline{W}^{(R)}$ are described by the induced distributions (2) and (3). We establish several bounds on the generalization error (4). The bounds, which are of PAC-Bayes type and rate-distortion theoretic, have the advantage of taking the structure of the studied multi-round interactive learning problem into account. Also, they account explicitly for the effect of the number of communication rounds R with the PS, in addition to the number of devices K and the size n of each individual dataset. To the best of our knowledge, they are the first of their kind for this problem.

3.1. PAC-Bayes bounds

For convenience, we start with two *lossless* bounds, which can be seen as distributed versions of those of McAllester (1998; 1999); Maurer (2004); Catoni (2003) tailored specifically for the multi-round interactive FL problem at hand.

Theorem 3.1. *Assume that the loss $\ell(Z_k,w)$ is σ-subgaussian for every $w\in\mathcal{W}$ and any $k\in[K]$. Also, let for every $k\in[K]$ and $r\in[R]$, $\mathsf{P}_{k,r}$ denote a conditional prior on $W_k^{(r)}$ given $\overline{W}^{(r-1)}$. Then we have:*

(i) *With probability at least $(1-\delta)$ over $\mathbf{S}$, for all $P_{\mathbf{W}|\mathbf{S}}$, $\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\big]$ is upper bounded by*

$$\sqrt{\frac{\frac{1}{KR}\sum\limits_{k,r}\mathbb{E}\Big[D_{KL}\big(P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\|\mathsf{P}_{k,r}\big)\Big]+\log\big(\frac{\sqrt{2n}}{\sqrt{R}\delta}\big)}{(2n/R-1)/(4\sigma^2)}},$$

where the summation is over $k \in [K]$ and $r \in [R]$ and expectation is with respect to $\overline{W}^{(r-1)} \sim P_{\overline{W}^{(r-1)}|S_{[K]}^{[r-1]}}$.

(ii) *For any FL-model $P_{\mathbf{W}|\mathbf{S}}$, with probability at least $(1-\delta)$ over $(\mathbf{S},\mathbf{W})\sim P_{\mathbf{S},\mathbf{W}}$, $\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})$ is upper bounded by*

$$\sqrt{\frac{\frac{1}{KR}\sum\limits_{k\in[K],r\in[R]}\log\left(\frac{\mathrm{d}P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}}{\mathrm{d}\mathsf{P}_{k,r}}\right)+\log(\frac{\sqrt{2n}}{\sqrt{R}\delta})}{(2n/R-1)/(4\sigma^2)}}.$$

The proof of Theorem 3.1, stated in Appendix G.1, judiciously extends the technique of a *variable-size* compressibility approach that was proposed recently in Sefidgaran & Zaidi (2023) in the context of establishing data-dependent PAC-Bayes bounds for a centralized learning setting, i.e., $K = 1$ and $R = 1$, to the more involved setting of FL. This extension is not trivial, however. For example, while the bound of the part (i) of Theorem 3.1 involves KL-divergence terms that may seem typical of classic PAC-Bayes bounds, the utility of the result is in expressing the bound in terms of average (over clients and rounds) of expected KL-divergence terms, where for every $r \in [R]$ the expectation is over $\overline{W}^{(r-1)}$. This is important, and non-intuitive, as the FL algorithm is *not* composed of $K\times R$ independent centralized algorithms. Indeed, a major technical difficulty in the analysis is to account properly for the problem's distributed nature as well as the statistical "couplings" among the devices' models, induced by the multi-round interactions. In part, these couplings are accounted for in the bound through the conditioning on $\overline{W}^{(r-1)}$. We refer the reader to a discussion after Theorem 3.2, which is a "lossy" version of the result of Theorem 3.1, on how the proof proceeds to break down the overall FL-algorithm into $K\times R$ *inter-dependent* "centralized"-like algorithms.

It should also be noted that, in fact, one could still consider the R-round FL problem end-to-end and view the entire system as a (virtual) centralized learning system with input the collection $\mathbf{S} = (S_1, \dots, S_K)$ of all devices' datasets and output the final aggregated model $\overline{W}^{(R)}$ and apply the results of Sefidgaran & Zaidi (2023) (or those of McAllester (1998; 1999); Maurer (2004); Catoni (2003); Seeger (2002); Tolstikhin & Seldin (2013)). The results obtained that way, however, do not account for the interactive and distributed structure of the problem. In contrast, note that for example the first bound of Theorem 3.1 involves, for each device k and round r the KL-divergence term $\mathbb{E}_{\overline{W}^{(r-1)}}\left[D_{KL}\left(P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\|\mathsf{P}_{k,r}\right)\right]$ which can be seen as accounting for (a bound on) the contribution of the model of that client at that round r to the overall generalization error. In a sense, the theorem says that only the average of those KL divergence terms matters for the bound, which could be interpreted as validating the intuition that some clients may be allowed to "overfit" during some or all of the rounds as long as the overall generalization error budget is controlled suitably by other devices. Also, as it can be seen from the result, the choice of priors is specifically tailored for the multi-round multi-client setting in the sense that the prior of client k at round r could depend on the aggregated model $\overline{W}^{(r-1)}$. The result also has implications on the design of practical learning FL systems: for example, on one aspect it suggests that the aforementioned KL-divergence term can be used as a round-dependent regularizer which account better for the variation of the quality of the training datasets across devices and rounds.

Finally, by viewing our studied learning setup with disjoint datasets along the clients and rounds as some form of distributed *semi-online* process, Theorem 3.1 may be seen as a suitable distributed version of the PAC-Bayes bound of Haddouche & Guedj (2022) established therein for a centralized *online-learning* setup. Note that a direct extension of that result to our distributed setup would imply considering the generalization error of client k at round r with respect to the intermediate aggregated model $\overline{W}_k^{(r)}$, not the final $\overline{W}_k^{(R)}$ as we do. In fact, in that case, the problem reduces to an easier virtual one-round setup that was also studied in (Barnes et al., 2022a;b) for Bregman divergences losses and linear and local Gaussian models; but at the expense of analyzing alternate quantity in place of the true generalization error (4) that we study.

We now present a more general, *lossy*, version of the bound of Theorem 3.1. The improvement is allowed by introducing a proper *lossy compression* that is defined formally below into the bound. This prevents the new bound from taking large values for deterministic algorithms with continuous hypothesis spaces.

Lossy compression. Consider a *quantization set* $\hat{\mathcal{W}} \subseteq \mathcal{W}$ and let $\{\hat{W}_k^{(r)}\}_{k\in[K],r\in[R]}$, $\hat{W}_k^{(r)} \subseteq \hat{\mathcal{W}}$, be a set of *lossy* versions of $\{W_k^{(r)}\}_{k\in[K],r\in[R]}$, defined via some conditional Markov kernels $p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}$, *i.e.,* we consider *lossy compression* of $W_k^{(r)}$ using only $S_k^{(r)}$ and the round-$(r-1)$ aggregated model $\overline{W}^{(r-1)}$. For a given *distortion level* $\epsilon \in \mathbb{R}^+$, $\{p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\}_{k\in[K],r\in[R]}$ is said to satisfy the ϵ-distortion criterion if following holds: for every $\mathbf{s} \in \mathcal{Z}^{nK}$,

$$\left|\mathbb{E}_{P_{\mathbf{W}|\mathbf{s}}}\left[\mathrm{gen}(\mathbf{s},\overline{W}^{(R)})\right]- \frac{1}{KR}\sum\nolimits_{k\in[K],r\in[R]}\mathbb{E}_{(Pp)_{k,r}}\left[\mathrm{gen}(s_k^{(r)},\hat{\overline{W}}^{(R)})\right]\right| \leqslant \epsilon/2, \tag{5}$$

$$\text{where } (Pp)_{k,r}=P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|s_{[K]}^{[r-1]},s_{[K]\setminus k}^{(r)}}\times p_{\hat{W}_k^{(r)}|s_k^{(r)},\overline{W}^{(r-1)}}P_{\hat{\overline{W}}^{(R)}|W_{[K]\setminus k}^{(r)},\hat{W}_k^{(r)},s_{[K]}^{[r+1:R]}}. \tag{6}$$

This condition can be simplified for Lipschitz losses, *i.e.,* when $\forall w, w' \in \mathcal{W}$, $|\ell(z,w) - \ell(z,w')| \leqslant \mathfrak{L}\rho(w,w')$, for some distortion function $\rho\colon \mathcal{W} \times \mathcal{W} \to \mathbb{R}^+$. Then, a sufficient condition for (5) is

$$\sum_{k,r} \mathbb{E}_{(Pp)_{k,r} P_{\mathbf{W}|\mathbf{S}}, \overline{W}^{(r-1)}, W^{(r)}_{[K]\setminus k}} \left[\rho(\overline{W}^{(R)}, \hat{\overline{W}}^{(R)}) \right] \leqslant \frac{KR\epsilon}{4\mathfrak{L}}.$$

Theorem 3.2. *Suppose that $\ell(z,w) \in [0,C] \subset \mathbb{R}^+$. Let for every $k \in [K]$ and $r \in [R]$, $\mathsf{P}_{k,r}$ be a conditional prior on $\hat{W}^{(r)}_k$ given $\overline{W}^{(r-1)}$. Fix any* distortion level $\epsilon \in \mathbb{R}^+$. *Consider any $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$-FL model and any choices of $\{p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}}\}_{k\in[K], r\in[R]}$ that satisfy the ϵ-distortion criterion. Then, with probability at least $(1-\delta)$ over $\mathbf{S} \sim P_{\mathbf{S}}$, $\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right]$ is upper bounded by*

$$\sqrt{\frac{\sum_{k,r} \left(\mathbb{E}\left[\left(D_{KL}\left(p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}} \| \mathsf{P}_{k,r} \right) \right] + \log(\frac{\sqrt{2n}}{\sqrt{R}\delta}) \right)}{(KR) \times (2n/R - 1)/C^2}} + \epsilon,$$

where the summation is over $k \in [K]$ and $r \in [R]$ and expectation is with respect to $\overline{W}^{(r-1)} \sim P_{\overline{W}^{(r-1)}|S^{[r-1]}_{[K]}}$.

A trivial extension of the lossy PAC-Bayes bounds for centralized algorithms could be done by considering the *quantization* of the final aggregated model $W^{(R)}$. Here, instead, for every $k \in [K]$ and round $r \in [R]$, we quantize the local model $W^{(r)}_k$ separately while keeping the other devices' local models at that round *i.e.,* the vector $W^{(r)}_{[K]\setminus k}$, fixed. This allows us to study the amount of the "propagated" distortion till round R. Note that by quantizing $W^{(r)}_k$ for a distortion constraint on the generalization error that is at most ν, the immediate aggregated model $\overline{W}^{(r)}_k$ is guaranteed to have a generalization error within a distortion level of at most ν/K from the true value. Also, interestingly, the distortion criterion (5) allows to "allocate" suitably the targeted total distortion $KR\epsilon/2$ into smaller constituent levels among all clients and across all rounds.

Proof outline: Theorem 3.2 is proved in Appendix G.2 by breaking down the FL algorithm into KR "centralized"-like algorithms, using the following high-level steps:

i. For every pair (k,r), we define a virtual FL algorithm that is equivalent to the original one except for round r of client k for which the "local quantized algorithm" $p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}}$ is considered instead of $P_{W^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}}$. This overall algorithm is denoted by $(Pp)_{k,r}$ as given by (6). Also, we define the overall algorithm with respect to the prior $\mathrm{P}_{k,r}$. These choices allow us to define $2KR$ 'virtual' algorithms, each of them differing in one distinct (client, round)-local algorithm from the original.

ii. Next, by Lemma G.3 we relate the probability that the generalization error exceeds a given threshold to the supremum of the expected average generalization error of all KR-virtual algorithms $(P\mathrm{P}_{k,r})_{k,r}$. The supremum is taken with respect to all joint distributions that are in the vicinity of $\mu^{\otimes nK}$.

iii. Finally, we apply a change of measure argument by two successive applications of Donsker-Varadhan's inequality, from $(P\mathrm{P}_{k,r})_{k,r}$ to $(P\mathrm{P}_{k,r})_{k,r}$ and from $\nu_{\mathbf{S}}$ to $\mu^{\otimes nK}$. Using the special form of $(P\mathrm{P}_{k,r})_{k,r}$, the KL-divergences $D_{KL}((P\mathrm{P}_{k,r})_{k,r} \| (\mathrm{P}\mathrm{P}_{k,r})_{k,r})$ lead to the desired KL-divergence terms. This allows in the last step of the proof to bound the cumulant generating function as needed.

3.2. Rate-distortion theoretic bounds

Define for $k \in [K]$, $r \in [R]$, and $\epsilon \in \mathbb{R}$, the rate-distortion function

$$\mathfrak{RD}(P_{\mathbf{S},\mathbf{W}}, k, r, \epsilon) = \inf I(S^{(r)}_k; \hat{W}^{(r)}_k | \overline{W}^{(r-1)}), \quad (7)$$

where the mutual information is evaluated with respect to $P_{S^{(r)}_k} P_{\overline{W}^{(r-1)}} p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}}$ and the infimum is over all conditional Markov kernels $p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}}$ that satisfy

$$\mathbb{E}_{\mathbf{S},\mathbf{W},\hat{\overline{W}}^{(R)}} \left[\mathrm{gen}(S^{(r)}_k, \overline{W}^{(R)}) - \mathrm{gen}(S^{(r)}_k, \hat{\overline{W}}^{(R)}) \right] \leqslant \epsilon, \quad (8)$$

where the joint distribution of $\mathbf{S}, \mathbf{W}, \hat{\overline{W}}^{(R)}$ factorizes as $P_{\mathbf{S},\mathbf{W}}\, p_{\hat{W}^{(r)}_k | S^{(r)}_k, \overline{W}^{(r-1)}} P_{\hat{\overline{W}}^{(R)} | W^{(r)}_{[K]\setminus k}, \hat{W}^{(r)}_k}$.

Theorem 3.3. *For any $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$-FL model with distributed dataset $\mathbf{S} \sim P_{\mathbf{S}}$, if $\ell(Z_k, w)$ is σ-subgaussian for every $w \in \mathcal{W}$ and any $k \in [K]$, then for every $\epsilon \in \mathbb{R}$, $\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right]$ is upper bounded by*

$$\sqrt{2\sigma^2 \sum\nolimits_{k\in[K], r\in[R]} \mathfrak{RD}(P_{\mathbf{S},\mathbf{W}}, k, r, \epsilon_{k,r})/(nK)} + \epsilon,$$

for any set of parameters $\{\epsilon_{k,r}\}_{k\in[K], r\in[R]} \subset \mathbb{R}$ which satisfy $\frac{1}{KR}\sum_{k\in[K]}\sum_{r\in[R]} \epsilon_{k,r} \leqslant \epsilon$.

Similar to the PAC-Bayes type bounds of Theorem 3.1 and Theorem 3.2, the bound of Theorem 3.3 also shows the "contribution" of each client's local model during each round to (a bound on) the generalization error as measured by (4). we refer the reader to Appendix G.3 for its proof, due to lack of space and the considerable needed technicality details. As it will become clearer from the below, an extended version of this theorem, stated in Appendix D.2, is particularly useful to study the *Federated Support Vector Machines* (FSVM) of the next section. For instance, an application of that result will yield a (more) explicit bound on the generalization error of FSVM in terms of the parameters K, R, and n.

Finally, a tail rate-distortion theoretic bound is derived in Appendix D.1. This result states loosely that having a good generalization performance with high probability requires the algorithm to be compressible not only under the true distribution $P_{\mathbf{S},\mathbf{W}}$, but also for all those distributions $\nu_{\mathbf{S},\mathbf{W}}$ that are in the vicinity of $P_{\mathbf{S},\mathbf{W}}$.

4. Federated Support Vector Machine (FSVM)

In this section, we study the generalization behavior of Support Vector Machines (SVM) (Cortes & Vapnik, 1995; Vapnik, 2006) when optimized in the FL setup using SGD. SVM is a popular model, mainly used for binary classification, and is particularly powerful when used with high-dimensional kernels. For ease of the exposition, however, we only developed results for linear SVMs. For setups with non-linear kernels, such as in (Kamp et al., 2016; Kamp, 2019), the results could possibly be extended, e.g., using finite-dimensional approximations of kernels and accounting for the caused approximation errors in that case.

Consider a binary classification problem in which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^d$ with $\mathcal{Y} = \{-1, +1\}$. Using SVM for this problem consists of finding a suitable hyperplane, represented by $w \in \mathbb{R}^d$, that properly separates the data according to their labels. For convenience, we only consider the case with zero bias. In this case, the label of an input $x \in \mathbb{R}^d$ is estimated using the sign of $\langle x, w\rangle$. The 0-1 loss $\ell_0 \colon \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ is then evaluated as $\ell_0(z, w) := \mathbb{1}_{\{y\langle x,w\rangle > 0\}}$. For the specific cases of centralized ($K = R = 1$) and "one-shot" FL ($K \geqslant 2, R = 1$) settings (Grønlund et al., 2020; Sefidgaran et al., 2022a), it was observed that if the algorithm finds a hyperplane that separates the data with some *margin* then it generalizes well. Motivated by this, we consider the 0-1 margin loss function $\ell_\theta \colon \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ for $\theta \in \mathbb{R}^+$ as $\ell_\theta(z, w) := \mathbb{1}_{\{y\langle x,w\rangle > \theta\}}$. Similar to previous studies, we consider the empirical risk with respect to the margin loss, *i.e.,* $\hat{\mathcal{L}}_\theta(\mathbf{s}, \overline{w}^{(R)}) = \frac{1}{nK}\sum_{k\in[K], i\in[n]} \ell_\theta(z_{k,i}, \overline{w}^{(R)})$ which is equal to $\frac{1}{KR}\sum_{k\in[K], r\in[R]} \hat{\mathcal{L}}_\theta(\mathbf{s}_k^{(r)}, \overline{w}^{(R)})$. The population risk is considered with respect to 0-1 loss function ℓ_0, *i.e.,* $\mathcal{L}(\overline{w}^{(R)}) = \frac{1}{K}\sum_{k\in[K]} \mathbb{E}_{Z_k\sim\mu_k}\big[\ell_0(Z_k, \overline{w}^{(R)})\big]$. The margin generalization error is then defined as $\mathrm{gen}_\theta(\mathbf{s}, \overline{w}^{(R)}) = \mathcal{L}(\overline{w}^{(R)}) - \hat{\mathcal{L}}_\theta(\mathbf{s}, \overline{w}^{(R)})$.

For the statement of the result that will follow, we make three assumptions on SGD. Due to lack of space, these assumptions are attentively described and discussed in Appendix C, and here we only state them informally. Consider an (K, R, n, e, b)-FL-SGD model. Let $\mathcal{W} = \mathfrak{B}_d(1)$, where $\mathfrak{B}_d(\nu)$ denotes the d-dimensional ball with origin center and radius $\nu > 0$.

- **Assumption 1:** There exists some $q_{e,b} \in \mathbb{R}^+$ such that for each client $k \in [K]$ at each round $r \in [2 : R]$, the local models $w_k^{(r)}$ and $w_k'^{(r)}$ attained by some initializations $\overline{w}^{(r-1)}$ and $\overline{w}'^{(r-1)}$, respectively, satisfy $\big\|w_k^{(r)} - w_k'^{(r)}\big\| \leqslant q_{e,b}\big\|\overline{w}^{(r-1)} - \overline{w}'^{(r-1)}\big\|$. If $q_{e,b} < 1$, SGD is called to be *contractive*.
- **Assumption 2:** There exists some $\alpha \in \mathbb{R}^+$ such that for each client $k \in [K]$ at each round $r \in [2 : R]$, the local models $w_k^{(r)}$ and $w_k'^{(r)}$ attained by some initializations $\overline{w}^{(r-1)}$ and $\overline{w}'^{(r-1)} := \overline{w}^{(r-1)} + \frac{w_\varepsilon}{K}$, satisfy $\big\|w_k'^{(r)} - \big(w_k^{(r)} + \frac{1}{K}\mathsf{D}_{k,r}\, w_\varepsilon\big)\big\| \leqslant \frac{\alpha}{K^2}\|w_\varepsilon\|^2$ for some matrix $\mathsf{D}_{k,r} \in \mathbb{R}^{d\times d}$, possibly dependent on $S_k^{(r)}$ and $\overline{w}^{(r-1)}$, whose spectral norm is bounded by $q_{e,b}$. Intuitively, this is an assumption on the first-order approximation error of the local steps of SGD for one client at one round.
- **Assumption 3:** The number of clients K is sufficiently large, *i.e.,* $K \geqslant f(q_{e,b}, \alpha, R)$ for some function f which is precised in Appendix C.

Now, we are ready to state our bound on the generalization error of FSVM.

Theorem 4.1. *For FSVM optimized using (K, R, n, e, b)-FL-SGD with $\mathcal{W} = \mathfrak{B}_d(1)$, $\mathcal{X} = \mathfrak{B}_d(B)$ and $\theta \in \mathbb{R}^+$, if Assumptions 1, 2, and 3 hold for some constants $q_{e,b}$ and α, then, $\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\big[\mathrm{gen}_\theta\big(\mathbf{S}, \overline{W}^{(R)}\big)\big]$ is upper bounded by*

$$\mathcal{O}\left(\sqrt{\frac{B^2\log(nK\sqrt{K})\sum_{r\in[R]} L_r}{nK^2\theta^2}}\right), \tag{9}$$

where

$$\begin{aligned} L_r &= \inf_{t\geqslant q^{(R-r)}}\left\{t^2\log\max\left(\frac{K\theta}{Bt}, 2\right)\right\} \\ &\leqslant q_{e,b}^{2(R-r)}\log\max\left(\frac{K\theta}{Bq_{e,b}^{(R-r)}}, 2\right). \end{aligned} \tag{10}$$

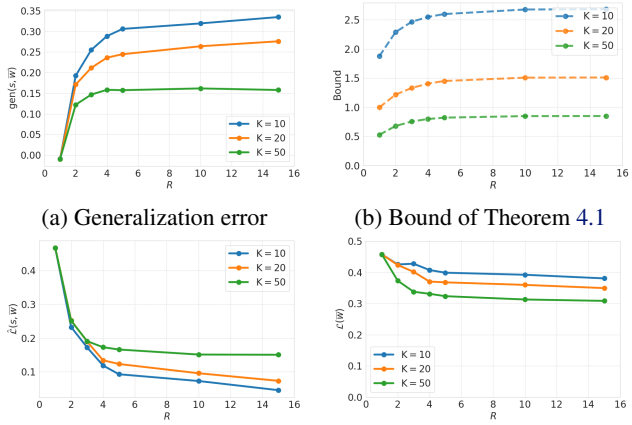
To the best of our knowledge, this result is the first of its kind for FSVM. We pause to state a few remarks that are in order, before discussing some key elements of its proof technique. First, the bound of Theorem 4.1 shows an explicit dependence on the number of communication rounds R, in addition to the number of participating devices K and the size of individual datasets n. In particular, the bound increases with R for fixed (n, K).[2] This *suggests* that the generalization power of FSVM may diminish by more frequent communication with the PS, as illustrated also numerically in Section 5. As a consequence, the population risk decreases less faster with R than the empirical risk (see Fig. 2). By taking into account the extra communication cost of larger R, this means that during the training phase of such systems, one might choose *deliberately* to stop before

[2]We refer the reader to Appendix F.3 for a discussion of the behavior with respect to R when n is not fixed and grows with R.

convergence (at some appropriate round $R^\star \leqslant R$), accounting for the fact that while interactions that are beyond $R^\star$ indeed generally contribute to diminishing the empirical risk further their net effect on the true measure of performance, which is the population risk, may be negligible. In the experiment section, we show such net effect when local models are ResNet-56 could be even negative. Second, for fixed (n, R) the bound improves (i.e., gets smaller) with K with a factor of $\sqrt{\log(K)/K}$. This behavior was previously observed in different contexts and under different assumptions in Sefidgaran et al. (2022a) and Barnes et al. (2022a;b), but in both works only for the "one-shot" FL setting, i.e., $R = 1$. In fact, it is easily seen that for the specific case $R = 1$, the bound recovers the result of Sefidgaran et al. (2022a, Theorem 5).

Proof outline: Theorem 4.1 is proved in Appendix G.4 using an extended version of Theorem 3.3, *i.e.,* Proposition D.2 (in Appendix D.2), and by bounding the appearing rate-distortion terms therein. Intuitively, rate-distortion terms are equal to the number of bits needed to represent an (optimal) quantized model given certain quantization precision. To establish an appropriate upper bound that is independent of the dimension d, we apply a quantization on a smaller d-independent dimension, using the Johnson-Lindenstrauss (JL) dimension-reduction transformation (Johnson & Lindenstrauss, 1984), which is inspired by Grønlund et al. (2020); Sefidgaran et al. (2022a); but with substantial extensions needed for the considered multiple-client multi-round setup. In particular, the main difficulty here is that while in the rate-distortion terms of Proposition D.2, the mutual information term (7) does not depend on $\overline{W}^{(R)}$, the distortion criterion (8) does. Thus, one needs to study the propagation of the distortion, induced by quantizing a local model, until the last round. The distortion propagation differs depending on the round r where the local model is located. Hence, for a fixed propagated distortion, the amount of dimension reduction should depend on r.

More precisely, for each $k \in [K]$ and $r \in [R]$ we first map the local models $W_k^{(r)}$ to a space with a smaller dimension of order $m_r = \mathcal{O}\left(\frac{B^2 \log(nK\sqrt{K}) \log(K/t)}{K^2\theta^2 t^2}\right)$, for some $t \geqslant q^{(R-r)}$, using JL transformation. As can be observed, we allow possibly different values of the dimensions for different values of r. For a contractive SGD, m_r decreases with r. Then, we define the quantized model subtly using this locally mapped model, as defined in equation (59). This quantization together with a newly defined loss functionin equation (61) allow to study the propagation of such quantization distortion in two steps: first, we show that the immediate aggregated model has K times smaller distortion level than that of client k. Next, we study the evolution of distortion along SGD iterations till the end of the round R,



(a) Generalization error (b) Bound of Theorem 4.1

(c) Empirical risk (d) Population risk

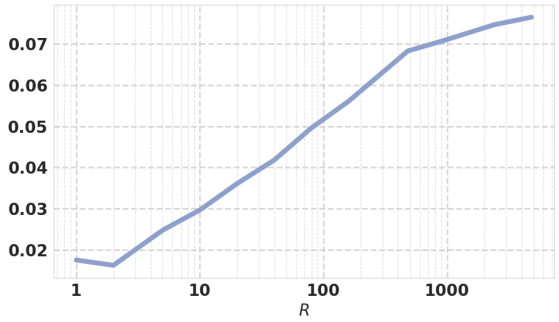
Figure 2. Generalization error and bound of Theorem 4.1 and empirical and population risks of FSVM and bound of Theorem 4.1 as functions of R, for $n = 100$

using induction on the rounds (see "Bounding (69)" in the proof) and by exploiting the properties of the JL transformation (see "Bounding (70)" & "Bounding (71)" in the proof). Intuitively, for a contractive SGD, the distortion *decreases*. This is why our m_r decreases with r. For a non-contractive SGD, the opposite holds.

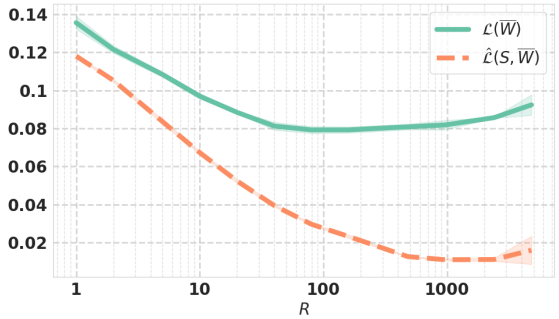
5. Experiments

FSVM. First, we run experiments which show that, as suggested by Theorem 4.1, the generalization error of FSVM increases with R. To this end, we consider a binary classification problem, with two extracted classes (digits 1 and 6) of MNIST dataset. Details and further experiments are provided in Appendix E.

Fig. 2a and Fig. 2b show the (estimated) expected generalization error $\mathbb{E}_{\mathbf{S},\mathbf{W}}[\text{gen}(\mathbf{S}, \overline{W}^{(R)}]$ and the computed bound of Theorem 4.1, respectively, versus the number of communication rounds R, for fixed $n = 100$ and for $K \in \{10, 20, 50\}$. The expectation is estimated over $M = 100$ Monte-Carlo simulations. As can be observed, for any value of K, the generalization error increases with R, as suggested by Theorem 4.1. Fig. 2c and Fig. 2d depict the evolution of the empirical risk $\hat{\mathcal{L}}(\mathbf{S}, \overline{W}^{(R)})$ and the (estimated) population risk $\mathcal{L}(\overline{W}^{(R)})$. As it can be observed therein, the population risk barely decreases beyond 5 rounds, while the empirical does. We emphasize that for Fig. 2 the total number of training data points and SGD iterations are kept fixed regardless of the number of running communication rounds R (for every value of R the devices perform en/R local SGD iterations per-round); and, hence, the increase in the generalization error *cannot* be attributed to the classical "overfitting" phenomenon. Appendix E.3 also reports similar behavior in



(a) Generalization error



(b) Empirical/population risks

Figure 3. Evolution of the performance of FL-SGD with ResNet-56 on CIFAR-10 as a function of R.

the context of *heterogeneous* data.

Additional experiments on generalization of FL. We also conducted additional experiments using ResNet-56 as local models and with the CIFAR-10 dataset. Fig. 3a shows the generalization error of the global model as a function of R while Fig. 3b shows the corresponding empirical and population risks. We provide average values over 5 runs and the shaded areas correspond to the standard deviation values. Experimental details are given in Appendix E.

As seen from Fig. 3a, the generalization error increases with R. This indicates that the behavior suggested by Theorem 4.1 and validated above experimentally for FSVM holds in this case as well. For instance, while the empirical risk decreases with R as expected the population risk is minimized for $R^* \simeq 100$ and may even start to increase beyond that value. In practice, this has far-reaching consequences, especially in resource-constrained applications of FL such as for wireless networks, as one can simply train less (i.e., save bandwidth) and get a better return in terms of population risk (test error).[3]

The code for reproducing the results can be found at https://github.com/RomainChor/Generalization_FL_ICML2024.

6. On population risk of Federated Learning algorithms

Hereafter, we discuss the evolution of the population risk with the number of communication rounds. To this end, hereafter we hold $n = b \times \tau \times R/e$ fixed and vary R (which therefore also implies varying the value of τ).

The population risk is given by the sum of the empirical risk and the generalization error. While the evolution of the empirical risk with R is rather well understood by now, and there is consensus that larger values of R induce smaller empirical risks, see *e.g.,* (Stich, 2019), little is known about the evolution of the generalization error with R.

The present paper provides the first generalization bounds for federated learning algorithms, with an emphasis on its behavior with respect to the number of communication rounds R. Our theoretical results for FSVM suggest that the generalization error *increases* with R. This directly implies that the population risk decays less rapidly with R than the empirical risk. Indeed, our experiments validate this result and show that it holds even beyond the FSVM setups. Interestingly, in some cases the population risk may even start to increase beyond a certain value of R, as shown in Fig. 3. It is worth emphasizing that while our results suggest that the choice of $R = 1$ is optimal for the generalization error, trivially it is not for the population risk (because the empirical loss is not small enough at that value $R = 1$). Thus, in practice, for the choice of a suitable value of number of communication rounds R with the parameter server one should strike a right balance between making the empirical risk as small as desired (larger values of R are better) and making the generalization error small enough (smaller values of R are better).

A consequence of the above is that, in practice, one may voluntarily choose to stop training before convergence (i.e., after only $R^* < R$ rounds), because while this can be non-optimal for empirical risk, the net effect may be positive from a population risk perspective. In more realistic settings in which communication costs need to be considered, this choice may become even more advantageous. We refer the reader to Appendix F.2 for a brief discussion of the estimation of the optimal number of communication rounds.

Impact statement

Essentially this work studies the performance of distributed learning algorithms from a theoretic angle. Henceforth, at this stage there are no clear direct societal implications of it.

[3]The question of estimating, *prior to training*, the optimal value R^* is an important one, with far-reaching consequences in practice, e.g., for the design of simultaneously communication-efficient and generalizable FL algorithms. This is left for future works.

References

Aguerri, I. E. and Zaidi, A. Distributed variational representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):120–138, 2019.

Ahn, K., Yun, C., and Sra, S. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17526–17535. Curran Associates, Inc., 2020.

Alquier, P. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

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

Barnes, L. P., Dytso, A., and Poor, H. V. Improved information-theoretic generalization bounds for distributed, federated, and iterative learning. *Entropy*, 24(9): 1178, 2022b.

Catoni, O. A pac-bayesian approach to adaptive classification. *preprint*, 840, 2003.

Chen, Y., Ning, Y., Slawski, M., and Rangwala, H. Asynchronous online federated learning for edge devices with non-iid data. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 15–24. IEEE, 2020.

Chor, R., Sefidgaran, M., and Zaidi, A. More communication does not result in smaller generalization error in federated learning. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 48–53, 2023.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains, 2018.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Gogineni, V. C., Werner, S., Huang, Y.-F., and Kuh, A. Communication-efficient online federated learning framework for nonlinear regression. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Grønlund, A., Kamma, L., and Larsen, K. G. Near-tight margin-based generalization bounds for support vector machines. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3779–3788. PMLR, 13–18 Jul 2020.

Gu, X., Lyu, K., Huang, L., and Arora, S. Why (and when) does local SGD generalize better than SGD? In *The Eleventh International Conference on Learning Representations*, 2023.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

Haddouche, M. and Guedj, B. Online pac-bayes learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25725–25738. Curran Associates, Inc., 2022.

Haghifam, M., Rodríguez-Gálvez, B., Thobaben, R., Skoglund, M., Roy, D. M., and Dziugaite, G. K. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pp. 663–706. PMLR, 2023.

Hu, X., Li, S., and Liu, Y. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

Huang, B., Li, X., Song, Z., and Yang, X. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pp. 4423–4434. PMLR, 2021.

Johnson, W. B. and Lindenstrauss, J. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26:28, 1984.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Kamp, M. *Black-box parallelization for machine learning*. PhD thesis, Universitäts-und Landesbibliothek Bonn, 2019.

Kamp, M., Bothe, S., Boley, M., and Mock, M. Communication-efficient distributed online learning with kernels. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16*, pp. 805–819. Springer, 2016.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Kozachkov, L., Wensing, P. M., and Slotine, J.-J. Generalization in supervised learning through riemannian contraction. *arXiv preprint arXiv:2201.06656*, 2022.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. at&t labs, 2010.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. doi: 10.1109/MSP.2020.2975749.

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

M. Ghari, P. and Shen, Y. Personalized online federated learning with multiple kernels. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33316–33329. Curran Associates, Inc., 2022.

Maurer, A. A note on the pac bayesian theorem. *arXiv preprint cs/0411099*, 2004.

McAllester, D. A. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 230–234, 1998.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mitra, A., Hassani, H., and Pappas, G. J. Online federated learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 4083–4090. IEEE, 2021.

Moldoveanu, M. and Zaidi, A. In-network learning for distributed training and inference in networks. In *2021 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, 2021.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Park, S., Simsekli, U., and Erdogdu, M. A. Generalization bounds for stochastic gradient descent via localized ε-covers. *Advances in Neural Information Processing Systems*, 35:2790–2802, 2022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227): 1–40, 2021.

Qin, T., Etesami, S. R., and Uribe, C. A. Faster convergence of local sgd for over-parameterized models. *arXiv preprint arXiv:2201.12719*, 2022.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Seeger, M. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

Sefidgaran, M. and Zaidi, A. Data-dependent generalization bounds via variable-size compressibility. *arXiv preprint arXiv:2303.05369*, 2023.

Sefidgaran, M., Chor, R., and Zaidi, A. Rate-distortion theoretic bounds on generalization error for distributed learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19687–19702. Curran Associates, Inc., 2022a.

Sefidgaran, M., Gohari, A., Richard, G., and Simsekli, U. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. *arXiv:2203.02474*, 2022b.

Sefidgaran, M., Zaidi, A., and Krasnowski, P. Minimum description length and generalization guarantees for representation learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Stich, S. U. Local sgd converges fast and communicates little. In *2019 International Conference on Learning Representations (ICLR)*, 2019.

Tolstikhin, I. O. and Seldin, Y. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., and Zhou, Y. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*, pp. 1–11, 2019.

Vapnik, V. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., and Rellermeyer, J. S. A survey on distributed machine learning. *ACM Computing Surveys (CSUR)*, 53 (2):1–33, 2020.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Wang, Z., Huang, S.-L., Kuruoglu, E. E., Sun, J., Chen, X., and Zheng, Y. Pac-bayes information bottleneck. *arXiv preprint arXiv:2109.14509*, 2021.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Yagli, S., Dytso, A., and Vincent Poor, H. 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)*, pp. 1–5, 2020. doi: 10.1109/SPAWC48557.2020.9154277.

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Yuan, H., Zaheer, M., and Reddi, S. Federated composite optimization. In *International Conference on Machine Learning*, pp. 12253–12266. PMLR, 2021.

Zaidi, A., Estella-Aguerri, I., and Shamai, S. On the information bottleneck problems: Models, connections, applications and information theoretic views. *Entropy*, 22(2):151, 2020.

Zhou, W., Veitch, V., Austern, M., Adams, R. P., and Orbanz, P. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Appendices

The appendices are organized as follows:

- Appendix A contains the algorithmic representation of the considered federated learning setup.
- Appendix B contains the discussion and some preliminary results on the case where the datasets used by each client during different rounds of the federated learning algorithm are not disjoint.
- Appendix C contains the rigorous statement and discussion of the Assumptions made for establishing the generalization bound of FSVM in Theorem 4.1.
- Appendix D contains further theoretical results. Specifically, this section contains
  - a rate-distortion theoretic tail bound in Appendix D.1,
  - extensions of all lossy bounds, *i.e.,* extensions of Theorems 3.2, 3.3, and D.1 in Appendix D.2.
- Appendix E contains experiments details and further experiments, including:
  - experimental details in Appendix E.1,
  - experiment on FSVM for different values of n in Appendix E.2,
  - experiment on FSVM in a heterogeneous data setting in Appendix E.3,
  - experimental verification of Assumption C.1 of Theorem 4.1 in Appendix E.4.
- Appendix F contains further discussions on the studied setups and established results and provides some potential future directions.
- Appendix G contains the deferred proofs of all our theoretical results.
  - Appendix G.1 contains the proof of Theorem 3.1,
  - Appendix G.2 contains the proof of Theorem 3.2,
  - Appendix G.3 contains the proof of Theorem 3.3,
  - Appendix G.4 contains the proof of Theorem 4.1,
  - Appendix G.5 contains the proof of Theorem B.1,
  - Appendix G.6 contains the proof of Theorem D.1,
  - Appendix G.7 contains the proof of Proposition D.2,
  - Appendices G.8, G.9, G.10, and G.11 contain the proofs of Lemmas G.1, G.2, G.3, and G.4, respectively.

A. Federated learning algorithm

Denote the deterministic aggregation function equivalent to the degenerate conditional distribution $P_{\overline{W}^{(r)}|W_{[K]}^{(r)}}$, $r \in [R]$, by the mapping $\mathcal{M}\colon \mathcal{W}^K \mapsto \mathcal{W}$, such that

$$\overline{W}^{(r)} = \mathcal{M}\Big(W_{[K]}^{(r)}\Big) \sim P_{\overline{W}^{(r)}|W_{[K]}^{(r)}}.$$

Then, the federated learning setup, described in Section 2 can be described alternatively as in Algorithm 1.

Algorithm 1 The federated learning algorithm

Inputs: number of rounds R, local algorithms $\mathcal{A}_k$ and dataset $S_k = \bigcup_{r\in[R]} S_k^{(r)}$ for $k \in [K]$, and the aggregation function $\mathcal{M}$
Output: $\overline{W}^{(R)}$
$\overline{W}^{(0)} = \varnothing$
for $r = 1, 2 \ldots, R$ **do**
  $\forall k \in [K]\colon W_k^{(r)} = \mathcal{A}_k(S_k^{(r)}, \overline{W}^{(r-1)})$
  $\overline{W}^{(r)} = \mathcal{M}\Big(W_{[K]}^{(r)}\Big)$
end for

B. Extension of the considered Federated Learning setup

In this paper, as a first step towards understanding the generalization behavior of Federated Learning algorithms, we considered a setup where in each round multiple "local epochs" $e \in \mathbb{N}^*$ can occur for each client's data but the used

datapoints are not reused in other rounds. That is, at round $r \in [R]$, client $k \in [K]$ computes updates using $S_k^{(r)} \subseteq S_k$ in order to minimize the empirical risk. This setup, coined as "online Federated Learning", have been already studied in the literature (Mitra et al., 2021; Chen et al., 2020; M. Ghari & Shen, 2022; Gogineni et al., 2022). Our setup can be also seen as the "without replacement" setup with one "global epoch", and when the number of local steps $e = 1$, as a "distributed online learning" setup. Therefore, our setup is a variant of the online learning framework that has been used before *e.g.,* in (Haddouche & Guedj, 2022) for a similar one-client multi-round setup.

Here, we show that our framework and proof techniques can be extended to include multiple epochs of with or without replacement setup, *i.e.,* when one data can be used in multiple rounds, as well. More precisely, assume $\bigcup_{r\in[R]} S_k^{(r)} \subseteq S_k$ where $S_k^{(r)}$ is the dataset used by client $k \in [K]$ during round $r \in [R]$ with size $n_{k,r}$. Here, in contrast to the rest of the paper, we allow the datasets $\{S_k^{(r)}\}_{r\in[R]}$ to have intersections for each $k \in [K]$, *i.e.,* a given $Z_{k,j} \in \mathcal{S}_k$, $j \in [n]$, may be used in multiple rounds and therefore belongs to multiple $S_k^{(r)}$, $r \in [R]$. We denote the elements of $S_k^{(r)}$ by $S_k^{(r)} = \{Z_{k,r,1}, \dots, Z_{k,r,n_{k,r}}\}$. Note that

$$\bigcup_{r\in[R]} \bigcup_{j\in[n_{k,r}]} \{Z_{k,r,j}\} \subseteq \{Z_{k,1}, \dots, Z_{k,n}\}.$$

The local training algorithm and the model aggregation are the same as before. In other words, $P_{\mathbf{W}|\mathbf{S}}$ is similar to (2) defined as below.

$$P_{\mathbf{W}|\mathbf{S}} = \bigotimes\nolimits_{r\in[R]} \left\{ \bigotimes\nolimits_{k\in[K]} \left(P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}} \right) P_{\overline{W}^{(r)}|W_{[K]}^{(r)}} \right\}. \tag{11}$$

Generalization error. For a (global) model or hypothesis $\overline{w}^{(R)} \in \mathcal{W}$, the population risk is defined as before:

$$\mathcal{L}\big(\overline{w}^{(R)}\big) = \frac{1}{K} \sum\nolimits_{k=1}^{K} \mathbb{E}_{Z_k\sim\mu_k}\big[\ell\big(Z_k, \overline{w}^{(R)}\big)\big].$$

We define the empirical risk as

$$\begin{aligned} \hat{\mathcal{L}}\left(\mathbf{s}, \overline{w}^{(R)}\right) &:= \frac{1}{KR} \sum\nolimits_{k=1}^{K} \sum\nolimits_{r=1}^{R} \hat{\mathcal{L}}\left(s_k^{(r)}, \overline{w}^{(R)}\right) \\ &= \frac{1}{KR} \sum\nolimits_{k=1}^{K} \sum\nolimits_{r=1}^{R} \frac{1}{n_{k,r}} \sum\nolimits_{j=1}^{n_{k,r}} \ell\big(z_{k,r,j}, \overline{w}^{(R)}\big). \end{aligned} \tag{12}$$

The *generalization error* of the model $\overline{w}^{(R)}$ for dataset $\mathbf{s} = s_{[K]}^{[R]}$ is evaluated as

$$\operatorname{gen}\left(\mathbf{s}, \overline{w}^{(R)}\right) = \mathcal{L}\big(\overline{w}^{(R)}\big) - \hat{\mathcal{L}}\left(\mathbf{s}, \overline{w}^{(R)}\right) = \frac{1}{KR} \sum\nolimits_{r=1}^{R} \sum\nolimits_{k=1}^{K} \operatorname{gen}\left(s_k^{(r)}, \overline{w}^{(R)}\right), \tag{13}$$

where $\operatorname{gen}\left(s_k^{(r)}, \overline{w}^{(R)}\right) = E_{Z_k\sim\mu_k}\big[\ell\big(Z_k, \overline{w}^{(R)}\big)\big] - \hat{\mathcal{L}}\left(s_k^{(r)}, \overline{w}^{(R)}\right)$.

For this setup, we derive an equivalent lossless version of Theorem 3.3. The equivalent results for the lossy version, as well as other theorems, can be established similarly, but with cumbersome notation and complex dependencies between each round's data for each client and their models at all rounds.

For every $k \in [K]$ and $1 \leqslant r \leqslant r' \leqslant R$, denote

$$W_k^{[r:r']} := W_k^{(r)}, W_k^{(r+1)}, \dots, W_k^{(r')},$$

with the convention that $W_k^{[r:r]} := W_k^{(r)}$.

Theorem B.1. *Assume that the loss $\ell(Z_k, w)$ is σ-subgaussian for every $w \in \mathcal{W}$ and any $k \in [K]$. Then,*

$$\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\left[\operatorname{gen}\left(\mathbf{S}, \overline{W}^{(R)}\right)\right] \leqslant \sqrt{\frac{2\sigma^2}{KR} \sum_{k\in[K],r\in[R]} \frac{1}{n_{k,r}} I\left(S_k^{(r)}; \overline{W}^{(r-1)}, W_k^{[r:R]}\right)}. \tag{14}$$

This result is proved in Section G.5.

Using chain rule, each of the the mutual information terms appearing in the bound can be re-written as

$$\begin{aligned} I\Big(S_k^{(r)}; \overline{W}^{(r-1)}, W_k^{[r:R]}\Big) =& I\Big(S_k^{(r)}; \overline{W}^{(r-1)}\Big) \\ &+ I\Big(S_k^{(r)}; W_k^{(r)} \big| \overline{W}^{(r-1)}\Big) \\ &+ I\Big(S_k^{(r)}; W_k^{(r+1)} \big| \overline{W}^{(r-1)}, W_k^{(r)}\Big) \\ &+ \dots \\ &+ I\Big(S_k^{(r)}; W_k^{(R)} \big| \overline{W}^{(r-1)}, W_k^{[r:R-1]}\Big). \end{aligned}$$

The term $I\Big(S_k^{(r)}; \overline{W}^{(r-1)}\Big)$ captures the dependence of $S_k^{(r)}$ on the aggregated model $\overline{W}^{(r-1)}$ and accounts for the appearance of elements of $S_k^{(r)}$ in the previous rounds. If $\{S_k^{(r)}\}_{r\in[R]}$ are disjoint, *i.e.,* one global epoch of the without replacement setup as considered throughout the paper, then this term is equal to zero. Furthermore, the terms

$$I\Big(S_k^{(r)}; W_k^{(r+1)} \big| \overline{W}^{(r-1)}, W_k^{(r)}\Big), \dots, I\Big(S_k^{(r)}; W_k^{(R)} \big| \overline{W}^{(r-1)}, W_k^{[r:R-1]}\Big),$$

capture the appearance of elements of $S_k^{(r)}$ in the following rounds. This term is also equal to zero for one global epoch of the without replacement setup. Thus, whenever $\{S_k^{(r)}\}_{r\in[R]}$ are disjoint, Theorem B.1 reduces to

$$\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\Big[\mathrm{gen}\Big(\mathbf{S}, \overline{W}^{(R)}\Big)\Big] \leqslant \sqrt{\frac{2\sigma^2}{KR} \sum_{k\in[K], r\in[R]} \frac{1}{n_{k,r}} I\Big(S_k^{(r)}; W_k^{(r)} \big| \overline{W}^{(r-1)}\Big)}, \tag{15}$$

which is the lossless version of Theorem 3.3. This result therefore *extends* the lossless version of Theorem 3.3.

Now, considering for example the round $(r+1)$ of client k, then only few samples of $S_k^{(r)}$ are used in this round (together with other samples from S_k) to generate $W_k^{(r+1)}$. Moreover, note that these used samples from $S_k^{(r)}$ in round $(r+1)$ had been already used before in round r to generate $W_k^{(r)}$, which is then aggregated by other client models and used as the initialization in round $r+1$. Thus, one *expects* $I\Big(S_k^{(r)}; W_k^{(r+1)} \big| \overline{W}^{(r-1)}, W_k^{(r)}\Big)$ to be smaller than $I\Big(S_k^{(r)}; W_k^{(r)} \big| \overline{W}^{(r-1)}\Big)$. In addition, the term $I\Big(S_k^{(r)}; \overline{W}^{(r-1)}\Big)$ captures the dependence of $S_k^{(r)}$ on the *aggregated* model $\overline{W}^{(r-1)}$. Again, if the overlap between datasets used in different rounds is small and if the number of clients K is sufficiently large, we can expect this term to be small. Thus, in the regime where the amount of data used from previous rounds is small and the number of clients is large, then one expects the RHS of (15) (which is the lossless version of Theorem 3.3) to be not very far from the RHS of (14). However, this *justification* requires rigorous analysis which is left for future work.

Finally, suppose that the setup considered in Section A is repeated for $e' \in \mathbb{N}^*$ epochs, *i.e.,* in total $e'R$ rounds. This means that for every $k \in [K]$ and $r \in [R]$, the sets $\{S_k^{(r)}\}_{r\in[R]}$ are disjoint and

$$S_k^{(r)} = S_k^{(R+r)} = \dots = S_k^{((e'-1)R+r)}.$$

Note that in this setup, the total number of rounds is $e'R$, and the final global model is $\overline{W}^{(e'R)}$. This setup for each client is called "SingleShuffle" without-replacement SGD with e'-epochs (Ahn et al., 2020). It can be shown that in this case, the bound of (15) can be further improved to

$$\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\Big[\mathrm{gen}\Big(\mathbf{S}, \overline{W}^{(e'R)}\Big)\Big] \leqslant \sqrt{\frac{2\sigma^2}{nK} \sum_{k\in[K], r\in[R]} \sum_{j\in[e']} I\Big(S_k^{(r)}; W_k^{((j-1)R+r)} \big| \overline{W}^{((j-1)R+r-1)}\Big)}. \tag{16}$$

Once again, intuitively, for each k and r, only the first few terms of the summation $\sum_{j\in[e']} I\Big(S_k^{(r)}; W_k^{((j-1)R+r)} \big| \overline{W}^{((j-1)R+r-1)}\Big)$ are expected to form the dominant term.

C. Assumptions for Generalization bound of FSVM

In this section, we state and discuss rigorously the Assumptions that are used for establishing the generalization bound of FSVM in Theorem 4.1 in Section 4.

Assumption C.1. Consider an (K, R, n, e, b)-FL-SGD model. For every $r \in [2 : R]$, every pair $(\overline{w}^{(r-1)}, \overline{w}'^{(r-1)})$ and every $k \in [K]$, the following holds for *any* pair $(w_k^{(r)}, w_k'^{(r)})$ that are both generated using $s_k^{(r)}$ with exactly similar mini-batches $\mathcal{B}_{k,r,t} \subseteq s_k^{(r)}$, $t \in [e\, n_R/b]$: there exists $q_{e,b} \in \mathbb{R}^+$[4] such that

$$\left\| w_k^{(r)} - w_k'^{(r)} \right\| \leqslant q_{e,b} \left\| \overline{w}^{(r-1)} - \overline{w}'^{(r-1)} \right\|. \tag{17}$$

If $q_{e,b} < 1$, we say that SGD is *contractive*. The contractivity property of SGD was previously studied and theoretically proved under some assumptions, e.g., when the surrogate loss function $\tilde{\ell}$ is smooth and strongly convex as in (Dieuleveut et al., 2018; Park et al., 2022; Kozachkov et al., 2022). Note that in general the value of $q_{e,b} < 1$ depends on the running round r and on the learning rate $\eta_{r,t}$. However, in our case, as stated in Section 2, for simplicity we constrain the learning rates $\eta_{r,t}$ to be identical across devices and rounds. This is assumed only for the sake of simplicity, and Theorem 4.1 can be extended straightforwardly to the case where $q_{e,b}$ could depend on r.

Theorem 4.1 holds for any value of $q_{e,b}$. However, the result becomes particularly interesting when $q_{e,b} \leqslant 1$. It can be shown that for a convex, Lipschitz, smooth (surrogate) loss function, $q_{e,b}$ is at most 1. For the case of strict inequality, i.e. $q_{e,b} < 1$, as mentioned above strongly convex loss functions (with some extra assumptions) indeed satisfy this condition. In FSVM, the considered surrogate loss, which is the hinge loss, is convex, which becomes strongly convex when L_2 regularizer is added. Even for the no-regularization case, for which the loss is not strongly convex and we only know theoretically $q_{e,b} \leqslant 1$, the numerical findings suggest that they are strictly less than one. Indeed, we computed the simulated values of $q_{e,b}$ with respect to R as displayed in Fig. 8. There, the values are computed for $K = 50$ clients and averaged over 10 runs. The standard deviation is extremely small. One can observe that all values are below 1.

Next, we state an assumption on the first-order approximation error of the local steps of SGD for one client at one round.

Assumption C.2. Consider the (K, R, n, e, b)-FL-SGD model with $\mathcal{W} = \mathfrak{B}_d(1)$. Consider any $k \in [K]$, $r \in [2 : R]$, and $\overline{w}^{(r-1)}$. Fix some dataset $s_k^{(r)}$ as well as the mini-batches $\mathcal{B}_{k,r,t} \subseteq s_k^{(r)}$, $t \in [e\, n_R/b]$. Then, there exists a constant $\alpha \in \mathbb{R}^+$, such that for any $w_\varepsilon \in \mathcal{W}$,

$$\left\| w_k'^{(r)} - \left(w_k^{(r)} + \frac{1}{K} \mathsf{D}_{(\overline{w}^{(r-1)}, \{\mathcal{B}_{k,r,t}\}_t)}\, w_\varepsilon \right) \right\| \leqslant \alpha \left\| w_\varepsilon \right\|^2 / K^2, \tag{18}$$

where $w_k^{(r)}$ and $w_k'^{(r)}$ are the local models of client k at round r, when initialized by $\overline{w}^{(r-1)}$ and $\overline{w}^{(r-1)} + \frac{w_\varepsilon}{K}$, respectively, and when the same mini-batches $\{\mathcal{B}_{k,r,t}\}_t$ are used. Moreover, the matrix $\mathsf{D}_{(\overline{w}^{(r-1)}, \{\mathcal{B}_{k,r,t}\}_t)} \in \mathbb{R}^{d \times d}$ represents the "overall gradient of SGD" of one client over one round, that depends on the initialization $\overline{w}^{(r-1)}$ and mini-batches $\{\mathcal{B}_{k,r,t}\}_t$. We further assume that the spectral norm of this matrix is bounded by $q_{e,b}$.

The inequality (18), which can be obtained using an easy expansion argument applied to the output of the perturbed initialization, is reasonable for moderate or large values of K, e.g., by assuming the boundedness of higher-order derivatives.

Moreover, it can be shown that

$$\mathsf{D}_{(\overline{w}^{(r-1)}, \{\mathcal{B}_{k,r,t}\}_t)} = \prod_{t \in [\tau]} \left(\mathrm{I}_d - \mathbf{H}^t_{(\overline{w}^{(r-1)}, \{\mathcal{B}_{k,r,t'}\}_{t' \in [t]})} \right), \tag{19}$$

where $\tau = e n_R/b$, I_d is the $d \times d$ identity matrix, and

$$\mathbf{H}^t_{(\overline{w}^{(r-1)}, \{\mathcal{B}_{k,r,t'}\}_{t' \in [t]})} := \mathrm{Hess}_{g_{\mathcal{B}_{k,r,t}}} \left(W_{k,t}^{(r)} \right).$$

Here $W_{k,t}^{(r)}$ is the model achieved at iteration t, when initialized by $\overline{w}^{(r-1)}$ and updated using the mini-batches $\{\mathcal{B}_{k,r,t'}\}_{t' \in [t]}$, $\mathrm{Hess}_g(\cdot)$ denotes the Hessian matrix, and

$$g_{\mathcal{B}}(w) := \frac{\eta}{b} \sum\nolimits_{z \in \mathcal{B}} \tilde{\ell}(z, w).$$

[4]Note that in general $q_{e,b}$ may depend on n_R as well. However, the dependence is dropped for brevity.

Lastly, we state an assumption on K being sufficiently large.

Assumption C.3. Suppose Assumptions 1 & 2 hold for some constants $q_{e,b}$ and α. Then, for $R \geqslant 2$,[5] the number of clients K satisfies at least one of the following conditions.

- **Condition 1:**

$$K^2 \geqslant \min_{\beta\in(0,1]} \max\left(\max_{r\in[R-2]} \frac{\alpha^2\left((\beta+1)q_{e,b}^2\right)^r}{\beta q_{e,b}^2}, \max_{r\in[R-1]} \frac{6\alpha B\Big(q_{e,b}^r - \left((\beta+1)q_{e,b}^2\right)^r\Big)}{\Big(q_{e,b} - (\beta+1)q_{e,b}^2\Big)\theta} \right), \tag{20}$$

- **Condition 2:** There exists some $\nu \geqslant 0$ such that $\frac{1}{K^2} \leqslant \frac{1-q_{e,b}^2}{\alpha} - \nu$ and

$$K^2 \geqslant \max_{r\in[R-1]} \frac{6\alpha B\Big(q_{e,b}^r - (1-\alpha\nu)^r\Big)}{(q_{e,b} - (1-\alpha\nu))\theta}. \tag{21}$$

This assumption is merely used for the simplification of the technical steps of the proof and more precisely the recursive steps in the part Bounding (69) in the proof of Theorem 4.1 in Appendix G.4. We *expect* Theorem 4.1 to hold for moderate values of K as well. It is insightful to note that in many cases the terms involving the maximization over r in (20) and (21) are either maximized for $r = 1$ or can be bounded by R, and hence the conditions can be simplified. As an example, whenever $q_{e,b} < 1$, then a sufficient condition to satisfy (20) is to have

$$K^2 \geqslant \min_{\beta\in(0,\min(1,1/q_{e,b}^2-1)]} \max\left(\alpha^2(\beta+1)/\beta, \max_{r\in[R-1]} \frac{6\alpha B r \max\Big(q_{e,b}, (\beta+1)q_{e,b}^2\Big)^{r-1}}{\theta} \right).$$

Note that the second term goes to zero as $r \to \infty$. The above condition can be further made loose to achieve the below sufficient condition

$$K^2 \geqslant \min_{\beta\in(0,\min(1,1/q_{e,b}^2-1)]} \max\big(\alpha^2(\beta+1)/\beta, 6\alpha BR/\theta\big).$$

D. Additional theoretical results

D.1. A Rate-distortion theoretic tail bound on FL

In this section, we state a rate-distortion theoretic tail bound. For this bound, let the randomness of the learning algorithm used by client $k \in [k]$ during round $r \in [R]$ be represented by some variable $U_k^{(r)}$, assumed to be independent of every other random variable. Denote $\mathbf{U} := U_{[K]}^{[R]}$. This assumption implies that $W_k^{(r)}$ is a *deterministic* function of $S_k^{(r)}$, $U_k^{(r)}$, and the initialization $\overline{W}^{(r-1)}$. For any fixed $\delta > 0$, denote $\mathcal{G}_{\mathbf{S},\mathbf{U},\mathbf{W}}^{\delta} := \{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \colon D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) \leqslant \log(1/\delta)\}$. Note that under ν, for every $k \in [K]$ and $r \in [R]$, $W_k^{(r)}$ is a *deterministic* function of $S_k^{(r)}$, $U_k^{(r)}$ and $\overline{W}^{(r-1)}$. Moreover, $\overline{W}^{(r)}$ is a deterministic function of $W_{[K]}^{(r)}$. Hence, $\nu_{\mathbf{W}|\mathbf{S},\mathbf{U}} = P_{\mathbf{W}|\mathbf{S},\mathbf{U}}$. Also, let for given $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}$

$$\mathfrak{RD}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}, k, r, \epsilon) = \inf_{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}} I(S_k^{(r)}; \hat{W}_k^{(r)} | U_k^{(r)}, \overline{W}^{(r-1)}), \tag{22}$$

where the mutual information is calculated with respect to $\nu_{S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}} p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}$ and the infimum is over all conditional Markov kernels $p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}$ that satisfy

$$\mathbb{E}_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}}\Big[\operatorname{gen}(S_k^{(r)}, \overline{W}^{(R)})\Big] - \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\Big[\operatorname{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\Big] \leqslant \epsilon, \tag{23}$$

with

$$(\nu p)_{k,r} = \nu_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)} | S_{[K]}^{[r]} \setminus S_k^{(r)}, U_{[K]}^{[r]} \setminus U_k^{(r)}} \, p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}} \, \nu_{\hat{\overline{W}}^{(R)}|\hat{W}_k^{(r)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, U_{[K]}^{[r+1:R]}}.$$

It can be easily seen that this definition is consistent with (7).

[5] No condition is needed for $R = 1$.

Theorem D.1. *Consider a $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$-FL model with distributed dataset $\mathbf{S} \sim P_{\mathbf{S}}$. Suppose that the loss $\ell(Z_k, w)$ is σ-subgaussian for every $w \in \mathcal{W}$ and any $k \in [K]$. Fix any* distortion level $\epsilon \in \mathbb{R}$. *Then, for any $\delta > 0$, with probability at least $(1-\delta)$, we have*

$$\operatorname{gen}(\mathbf{S}, \overline{W}^{(R)}) \leqslant \sqrt{\frac{\sup_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}} \sum_{k\in[K], r\in[R]} \mathfrak{RD}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}, k, r, \epsilon_{k,r}) + \log(1/\delta)}{nK/(2\sigma^2)}} + \epsilon,$$

for any $\{\epsilon_{k,r}\}_{k\in[K], r\in[R]} \subset \mathbb{R}$ such that

$$\frac{1}{KR} \sum_{k\in[K]} \sum_{r\in[R]} \epsilon_{k,r} \leqslant \epsilon. \tag{24}$$

In a sense, this result, proved in Appendix G.6, says that in order to have a good generalization performance (not only in-expectation as in Theorem 3.3 but also in terms of tails), the algorithm should be *compressible* not only under the true distribution $P_{\mathbf{S},\mathbf{W}}$ but also for all *all* those distributions $\nu_{\mathbf{S},\mathbf{U}}$ that are in the vicinity of $P_{\mathbf{S},\mathbf{U}}$.

D.2. Lossy compression bounds revisited

In all *lossy* bounds, *i.e.,* Theorems 3.2, 3.3, and D.1, we had three common assumptions on the *quantizations* of the local models:

1. $\hat{\mathcal{W}} \subseteq \mathcal{W}$,
2. the loss and generalization error definitions considered for the models $\hat{\mathcal{W}}$ are the same as the one considered for $\mathcal{W}$,
3. the *quantization* of the model $W_k^{(r)}$ is performed according to $p_{\hat{W}_k^{(r)}|\overline{W}^{(r-1)}, S_k^{(r)}}$.

These choices are natural and intuitive ones for the FL learning algorithms. However, it turns out that all these conditions can be relaxed, *i.e.,* in the following extended results, we consider

1. $\hat{\mathcal{W}}$ to be any arbitrary set in $\mathbb{R}^m$, for some arbitrary $m \in \mathbb{N}^*$,
2. the loss function $\hat{\ell}(z, \hat{w})\colon \mathcal{Z} \times \hat{\mathcal{W}} \to \mathbb{R}^+$ to be defined arbitrarily, possibly different than ℓ. The generalization error for $\hat{\mathcal{W}}$ is defined then with respect to this loss function,
3. for every $k \in [K]$ and $r \in [R]$, quantizing *the aggregated global model* $\overline{\hat{W}}^{(R)}$ according to

$$p_{\overline{\hat{W}}_k^{(R)}|\overline{W}^{(r-1)}, S_k^{(r)}, V_k^{(r)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, V_{[K]}^{[r+1:R]}}, \tag{25}$$

where $V_k^{(r)} \in \mathcal{V}$ presents some mutually independent available randomness used by client k during round r. Note that unlike the Section 3.2, they do not necessarily represent *all* the randomness used by client k during round r, *e.g.,* they can be set to some constants or they can be set to be $U_k^{(r)}$ which is all the used randomness. Denote $\mathbf{V} = V_{[K]}^{[R]}$. Note that

$$P_{\mathbf{S},\mathbf{V},\mathbf{W}} = P_{\mathbf{S}} P_{\mathbf{V}} P_{\mathbf{W}|\mathbf{S},\mathbf{V}}. \tag{26}$$

The new choice of the quantization distribution implies that for every $k \in [K]$ and $r \in [R]$, in a sense we fix $\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, V_{[K]}^{[r+1:R]}, S_{[K]}^{[r+1:R]}$, and consider *quantization* of the different global models obtained for different $S_k^{(r)}$.

Now, we state the extended results. For the following results, consider arbitrary set $\hat{\mathcal{W}} \in \mathbb{R}^m$, loss function $\hat{\ell}(z, \hat{w})$, and the generalization error that is defined with respect to this loss function. Next, we use the shorthand notation:

$$\mathfrak{V}_{k,r} = \Big(V_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, V_{[K]}^{[r+1:R]}\Big). \tag{27}$$

First, we state the extended version of Theorem 3.3, that is used in the proof of Theorem 4.1. The proofs of these extended results follow the same lines as the original results. As an example, we show the proof of Proposition D.2 in Appendix G.7. The rest of the proofs follow the proof of their corresponding theorems.

For $k \in [K]$, $r \in [R]$ and $\epsilon \in \mathbb{R}$, let $\mathcal{P}_{k,r}$ include all conditional Markov kernels

$$p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}, \tag{28}$$

that satisfy

$$\mathbb{E}_{\mathbf{S},\mathbf{V},\mathbf{W},\hat{\overline{W}}^{(R)}}\left[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)}) - \mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\right] \leqslant \epsilon, \tag{29}$$

where the joint distribution of $\mathbf{S},\mathbf{V},\mathbf{W},\hat{\overline{W}}^{(R)}$ factorizes as $P_{\mathbf{S},\mathbf{V},\mathbf{W}}p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}$.

Now, define the rate-distortion function

$$\mathfrak{RD}^{\star}(P_{\mathbf{S},\mathbf{W}},k,r,\epsilon) = \inf_{\mathcal{P}_{k,r}} I(S_k^{(r)};\hat{\overline{W}}_k^{(R)}|\mathfrak{V}_{k,r}), \tag{30}$$

where the mutual information is evaluated with respect to

$$P_{S_k^{(r)}}P_{\mathfrak{V}_{k,r}}\, p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}. \tag{31}$$

Proposition D.2. *For any $(P_{\mathbf{W}|\mathbf{S}},K,R,n)$-FL model with distributed dataset and randomness $(\mathbf{S},\mathbf{V}) \sim P_{\mathbf{S}}P_{\mathbf{V}}$, if the loss $\ell(Z_k,w)$ is σ-subgaussian for every $w \in \mathcal{W}$ and any $k \in [K]$, then for every $\epsilon \in \mathbb{R}$ it holds that*

$$\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right] \leqslant \sqrt{\frac{2\sigma^2\sum_{k\in[K],r\in[R]}\mathfrak{RD}^{\star}(P_{\mathbf{S},\mathbf{W}},k,r,\epsilon_{k,r})}{nK}} + \epsilon,$$

for any set of parameters $\{\epsilon_{k,r}\}_{k\in[K],r\in[R]} \subset \mathbb{R}$ which satisfy

$$\frac{1}{KR}\sum\nolimits_{k\in[K]}\sum\nolimits_{r\in[R]}\epsilon_{k,r} \leqslant \epsilon.$$

Next, we state the extension of Theorem 3.2.

Proposition D.3. *Suppose that $\ell(z,w) \in [0,C] \subset \mathbb{R}^+$. Consider any set of* priors $\{\mathsf{P}_{k,r}\}_{k\in[K],r\in[R]}$ *where $\mathsf{P}_{k,r}$ is a conditional prior on $\hat{\overline{W}}^{(R)}$ given $\mathfrak{V}_{k,r}$. Fix any* distortion level $\epsilon \in \mathbb{R}^+$. *Consider any $(P_{\mathbf{W}|\mathbf{S}},K,R,n)$-FL model and any choices of*

$$\left\{p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}\right\}_{k\in[K],r\in[R]}, \tag{32}$$

that for every $\mathbf{s} \in \mathcal{Z}^{nK}$ satisfy the ϵ-distortion criterion

$$\left|\mathbb{E}_{P_{\mathbf{W}|\mathbf{s}}}\Big[\mathrm{gen}(\mathbf{s},\overline{W}^{(R)})\Big] - \frac{1}{KR}\sum\nolimits_{k\in[K],r\in[R]}\mathbb{E}_{(Pp)^{\star}_{k,r}}\Big[\mathrm{gen}(s_k^{(r)},\hat{\overline{W}}^{(R)})\Big]\right| \leqslant \epsilon/2, \tag{33}$$

where

$$(Pp)^{\star}_{k,r} = P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|s_{[K]}^{[r-1]},s_{[K]\setminus k}^{(r)}}P_{V_k^{(r)},V_{[K]}^{[r+1:R]}}\, p_{\hat{\overline{W}}^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}.$$

Then, with probability at least $(1-\delta)$ over $\mathbf{S} \sim P_{\mathbf{S}}$, we have that $\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\big]$ is upper bounded by

$$\sqrt{\frac{\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}\left[D_{KL}\left(p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}\middle\|\mathsf{P}_{k,r}\right)\right] + \log(\frac{\sqrt{2n}}{\sqrt{R}\delta})}{(2n/R-1)/C^2}} + \epsilon,$$

where the expectations are with respect to $P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|s_{[K]}^{[r-1]},s_{[K]\setminus k}^{(r)}}P_{V_k^{(r)},V_{[K]}^{[r+1:R]}}$.

Finally, we present the extension of the rate-distortion theoretic tail bound. For the rest of this section, assume that $V_k^{(r)} = U_k^{(r)}$, for $k \in [K]$ and $r \in [R]$ and for a better clarity, denote $\mathfrak{V}_{k,r}$ by $\mathfrak{U}_{k,r}$ for this choice.

Consider any $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}$. For this distribution and $k \in [K]$ and $r \in [R]$, let the $\mathcal{P}_{\nu,k,r}$ include all conditional Markov kernels

$$p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{U}_{k,r}}, \tag{34}$$

that satisfy

$$\mathbb{E}_{\mathbf{S},\mathbf{U},\mathbf{W},\hat{\overline{W}}^{(R)}}\left[\mathrm{gen}(S_k^{(r)}, \overline{W}^{(R)}) - \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \leqslant \epsilon, \tag{35}$$

where the expectation is with respect to $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{U}_{k,r}}$. Now, let

$$\mathfrak{RD}^{\star}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}, k, r, \epsilon) = \inf_{\mathcal{P}_{\nu,k,r}} I(S_k^{(r)}; \hat{\overline{W}}_k^{(R)}|\mathfrak{U}_{k,r}), \tag{36}$$

where the mutual information is calculated with respect to $\nu_{S_k^{(r)},\mathfrak{U}_{k,r}} p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{U}_{k,r}}$.

Proposition D.4. *Consider a $(P_{\mathbf{W}|\mathbf{S}}, K, R, n)$-FL model with distributed dataset and randomness $(\mathbf{S}, \mathbf{U}) \sim P_{\mathbf{S}} P_{\mathbf{U}}$. Suppose that the loss $\ell(Z_k, w)$ is σ-subgaussian for every $w \in \mathcal{W}$ and any $k \in [K]$. Fix any* distortion level $\epsilon \in \mathbb{R}$. *Then, for any $\delta > 0$, with probability at least $(1-\delta)$, we have*

$$\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)}) \leqslant \sqrt{\frac{\sup_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}} \sum_{k\in[K], r\in[R]} \mathfrak{RD}^{\star}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}, k, r, \epsilon_{k,r}) + \log(1/\delta)}{nK/(2\sigma^2)}} + \epsilon,$$

for any $\{\epsilon_{k,r}\}_{k\in[K], r\in[R]} \subset \mathbb{R}$ such that

$$\frac{1}{KR} \sum_{k\in[K]} \sum_{r\in[R]} \epsilon_{k,r} \leqslant \epsilon. \tag{37}$$

E. Additional experimental results & details

In this section, we first provide experimental and implementation details that were omitted in the main text. Then, some additional numerical simulations are shown and described.

E.1. Experimental & implementation details

Tasks description In our experiments, we simulate a Federated Learning (FL) framework, according to the setup of this paper (see Section 1), on a single machine, where each "client" is an instance of the SVM model, equipped with a dataset. 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 parameter server). This is utilized for computing the empirical risk $\hat{\mathcal{L}}_{\theta}(\mathbf{S}, \overline{W})$ and population risk $\mathcal{L}(\overline{W})$. In particular, $\mathcal{L}(\overline{W})$ is estimated using a test set of fixed size N_{test}. Using these quantities, the generalization error $\mathrm{gen}(\mathbf{S}, \overline{W})$ is computed.

The (local) dataset of client # $k \in [K]$, denoted as S_k, is composed of n samples that are drawn uniformly without replacement from S. In other words, the dataset S is initially split into K subsets. Each local dataset S_k is then further split into $S_k^{(r)},\ r \in [R]$ *i.e.,* the training samples used at each round r by client k.

Datasets The learning task is binary image classification for the FSVM experiments. More precisely, we trained the SVM model to classify images of digits 1 and 6 from the standard MNIST (LeCun et al., 2010), which results in approximately 12000 training images and 2000 test images. Experiments were performed for different pairs of digits, giving similar results, and thus are not reported. For the additional experiments in Section 5, we considered the whole CIFAR-10 dataset (10-class classification) (Krizhevsky & Hinton, 2009) composed of 50000 training images and 10000 test images.

Hyperparameter	Symbol	Value
Local epochs	e	40
Learning rate	η	0.01
Batch size	b	1
Kernel parameter	γ	0.05
Dimension of the approximated kernel feature space	d	4000

(a) FSVM

Hyperparameter	Symbol	Value
Number of clients	K	16
Epochs	e	150
Learning rate	η	1.0
Batch size	b	128

(b) Additional experiments (ResNet-56 with CIFAR-10)

Table 1. Hyperparameters for experiments in Section 5

Model Each client is equipped with SVM with Radial Basis Function (RBF) kernel (FSVM experiments) or ResNet-56 with batch normalization (additional experiments with NNs). Stochastic Gradient Descent (SGD) is used for optimization.

Training and hyperparameters All local models were trained with the same hyperparameters, which are of common usage, and are given in Table 1. The data is scaled and normalized so as to get zero-mean and unit variance.

The training scheme of the SVM experiments follows the setup described in 4. All clients have the same fixed learning rate $\eta_t = \eta$. Each local model is trained for e local epoch at each round $r \in [R]$. A local epoch refers to a pass over the training samples at each round $r \in [R]$ *i.e.,* $S_k^{(r)}$. Before each local epoch, $S_k^{(r)}$ is shuffled. The training algorithm ends when each client # $k \in [K]$ has done e local epochs on each of the subsets $S_k^{(r)},\ r \in [R]$.

For the ResNet-56 simulations, we did not stick to the previous setup and used a more standard training scheme. Precisely, each client does only one pass over its data $S_k^{(r)}$ at each round $r = 1, \ldots, R$. When the R rounds are completed, a *global* epoch is achieved, and each client starts again to train with $S_k^{(1)}$ until another epoch is completed.

Hardware and implementation We performed our experiments on a server equipped with 56 CPUs Intel Xeon E5-2690v4 2.60GHz and 4 GPUs Nvidia Tesla P-100 PCIe 16GB.

Our implementations use the Python language. The open-source machine learning library scikit-learn (Pedregosa et al., 2011) is used to implement SVM. In particular, we use *SGDClassifier* to implement SGD optimization and *RBFSampler* for the Gaussian kernel feature map approximation. The deep learning library PyTorch is used for the CIFAR-10 experiments using ResNet-56.

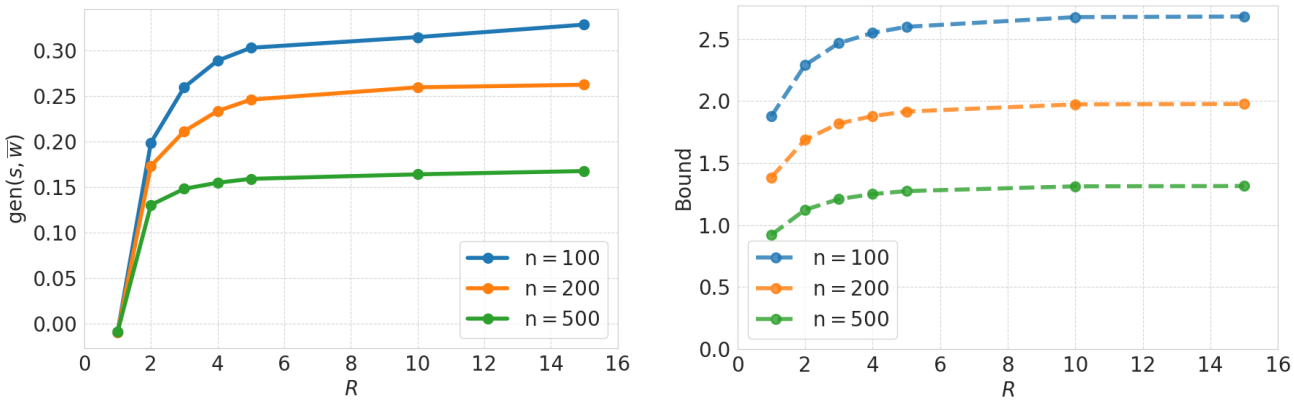
E.2. Generalization error of FSVM for different values of n

Fig. 4 shows the (expected) generalization error (Fig. 4a) of FSVM and the bound of Theorem 4.1 (Fig. 4b) with respect to R for $K = 10$ fixed and $n \in \{100, 200, 500\}$. Fig. 5 shows the empirical risk and the estimated population risk.

These plots suggest that our observed behaviors in Section 5 hold for fixed K and for various values of n. That is:

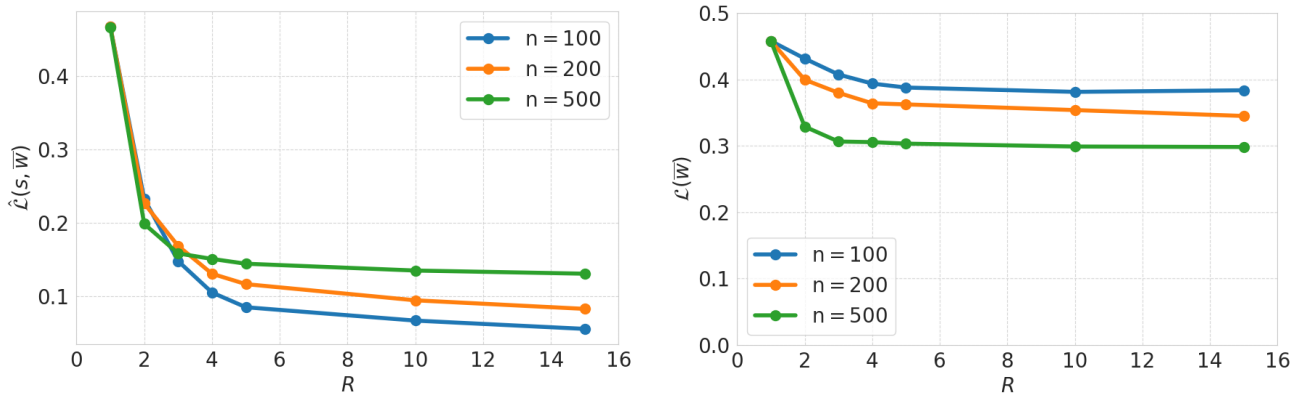
- The generalization error increases with R for any fixed n,
- For any fixed R, both the generalization error and bound improve as n increases,
- Above findings are compatible with the behavior predicted by the bound of Theorem 4.1,
- While the empirical risk keeps decreasing for $R \geqslant 5$, the population seems to have reached its minimum for $R = 5$.

Note that similar results were obtained for other values of n.



(a) (b)

Figure 4. Generalization error of FSVM and bound of Theorem 4.1 w.r.t. R, for $K = 10$



(a) (b)

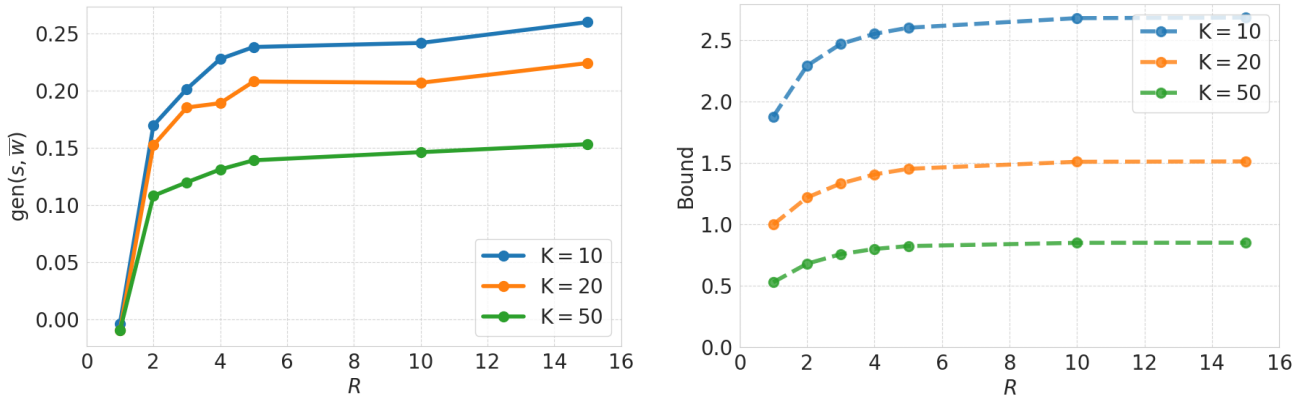
Figure 5. (a) Empirical risk and (b) population risk of FSVM w.r.t. R, for $K = 10$

E.3. Experiments for a heterogeneous data setting

We simulate a heterogeneous data setting (*i.e.,* non-i.i.d.) by adding Gaussian white noise with standard deviation $\sigma = 0.2$ to the training and testing images of a fraction $f = 0.2$ of the clients. Therefore, as in the setup described in Section 1, the data distributions of clients $\mu_k,\ k \in [K]$ are different. The hyperparameters remain, however, identical as we aim at comparing the numerical results with the homogeneous data case.

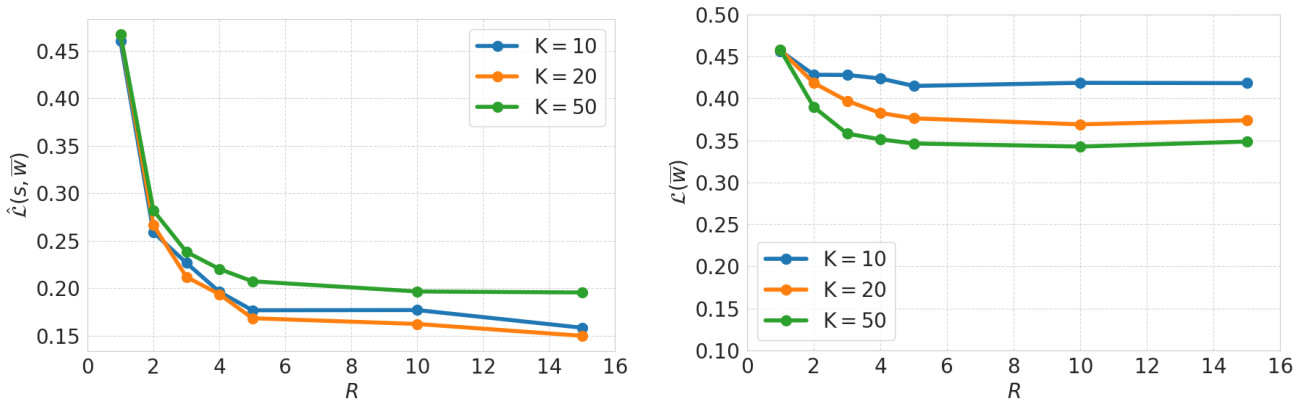
Similar to Section 5, we compute the generalization error of FSVM, as well as the bound of Theorem 4.1, with respect to R (see Fig. 6) and show the corresponding train and test risks (see Fig. 7).

The same observations can be made for the experiments in the homogeneous setting (Section 5). Remark that, on one hand, in comparison to Fig. 2c, the empirical risk values on Fig. 7 are larger. This is expected, as the final global model is a simple arithmetic average of local models and hence may struggle to achieve the optimum of each local objective function in the presence of data heterogeneity. Therefore, more communication rounds R are needed to achieve the same level of optimization as in the homogeneous case. On the other hand, the generalization error values (see Fig. 6a) are smaller than in the i.i.d. setup (see Fig. 2a). The global model is indeed less likely to overfit due to the "noise" that is injected into some of the local datasets.



(a) (b)

Figure 6. Generalization error of FSVM (non-iid setting) and bound of Theorem 4.1 w.r.t. R, for $n = 100$

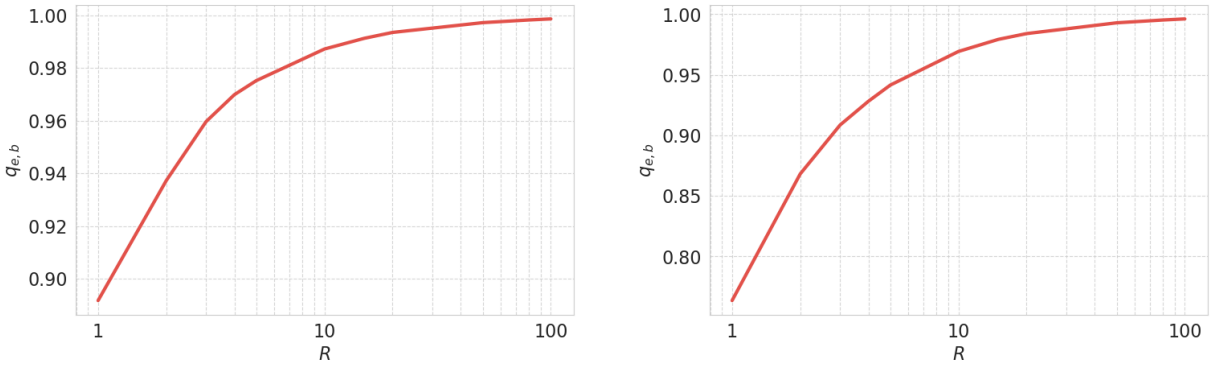


(a) (b)

Figure 7. Empirical risk and population risk for FSVM (non-iid setting) w.r.t. R, for $n = 100$

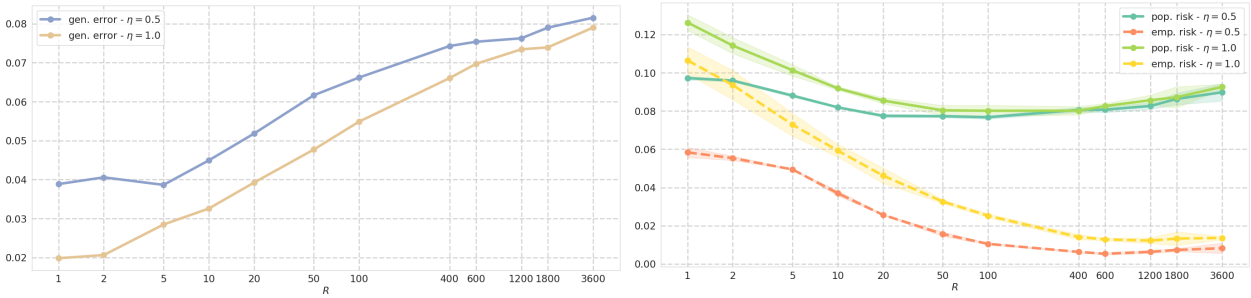
E.4. Experimental Verification of Assumption C.1 for Theorem 4.1

In this section, we show the computed values of the contraction coefficient $q_{e,b}$ in Assumption C.1 of Theorem 4.1 with respect to R and with $K = 50$ clients. The values are averaged over 10 runs. The standard deviation is extremely small. The



(a) $K = 20$ (b) $K = 50$

Figure 8. Estimated coefficient $q_{e,b}$ as function of R for FSVM



(a) Generalization error (b) Empirical and population risks

Figure 9. Performance of FL-SGD with ResNet-56 local models as a function of R for different learning rates

reader is referred to Appendix C for the discussion about the assumption.

E.5. Additional experiments for ResNet-56

The experiments on ResNet-56 presented on Fig. 3 were conducted for several sets of hyperparameters and showed similar behaviors. For example, we provide on Fig. 9 the generalization error and the empirical and population risks for two different values of the learning rate η. The other hyperparameters are similar to the experiment of Fig. 3 and are hence provided in Table 1b.

One can observe the same increasing trend of the generalization error with respect to R that was observed in the previous experiments. Also, the presence of a global minimizer R^* for the population risk is noticeable.

F. Further discussions and future works

In this section, we provide further clarification on some aspects related to the studied setup and the established bounds and suggest some future directions.

F.1. Regarding the setup

As discussed shortly before, the considered setup is in some aspects more simplified than the conventional FL setup. In fact, the studied setup is "closer" to the "online FL" setup. Here, we discuss in detail some of the limitations of the considered setup and point out the potentially interesting future directions. Besides, we show that some aspects of the studied setup are choices made for simplicity and ease of presentation – the results can easily be extended to cover such aspects.

Dataset splitting. Perhaps the main difference between the considered setup in the main part of the paper and the

conventional FL setup is that each data in the training dataset can be used during one communication round (possibly multiple times within that unique round).

This setup, which is closer to the Online FL setup, has potential applications in cases where clients have limited memory. Therefore, data from previous rounds may not be stored. In Appendix B, we partially showed how, using the established framework and proof techniques, our results can be extended to more general setups, but at the cost of more complicated terms appearing in the bound. Such extensions suggest that a similar type of bound holds when a data point can be used in multiple rounds. However, this requires further rigorous verification.

Unequal dataset sizes. In this paper, we have assumed that all clients have the same amount of training data for ease of presentation. However, the results of this paper can easily be extended to unequal dataset sizes. As an example, consider the setup where client $k \in [K]$ has a dataset of size n_k and uses n_k/R samples per round. Let $N = \sum_k n_k$. For this setup, a reasonable choice for the generalization error is $\mathrm{gen}\big(\mathbf{s}, \overline{w}^{(R)}\big) = \sum_{r,k} \frac{n_k}{NR} \mathrm{gen}\left(s_k^{(r)}, \overline{w}^{(R)}\right)$. Then the generalization bound of part i of 3.1 can be written as

$$\min_{0<\lambda<\min_k 2n_k/R} \sqrt{\frac{\sum_{k,r} \frac{n_k}{NR}\left(A(k,r) - \log(\delta\sqrt{1-1/(\lambda^2R^2/(2n_k^2))})\right)}{\lambda}},$$

where $A(k,r) = \mathbb{E}\Big[D_{KL}\big(p_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}} \| \mathsf{P}_{k,r}\big)\Big]$.

Similarly, the results can be easily extended if the number of data used by each client in each round is different.

Client participation. Another aspect of the FL that is overlooked in this work is the effect of unbalanced client participation. However, our framework can be used to study the effect of client participation or even nonparticipating clients on the generalization error. While, this requires dedicated work, to possibly combine with the ideas presented in (Hu et al., 2023), here, we provide a simple way where such an effect can be studied using our approach. For simplicity, assume that at each iteration M clients participate, and every client participates in exactly MR/K rounds with equal dataset size. With these assumptions, the generalization bound of the part i. of Theorem 3.1 becomes equal to

$$\sqrt{\frac{\frac{1}{MR}\sum_r \sum_{k\in\mathcal{I}_r} A(k,r) - \log(\delta\sqrt{MR/(2nK)})}{(2nK/(MR)-1)/(4\sigma^2)}},$$

where $A(k,r) = \mathbb{E}\Big[D_{KL}(p_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}} \| \mathsf{P}_{k,r})\Big]$ and $\mathcal{I}_r$ denotes the set of indices of the clients participating in round r.

Heterogeneity of the data. All established bounds in this work hold for any kind of heterogeneous data. However, as a future direction, it would be interesting to simplify the bounds for different forms of heterogeneity to observe their explicit impact on the generalization error.

Round-dependent distribution. As mentioned in Section 2, our framework and results can be extended to some cases where μ_k also depends on the running round $r = 1, \ldots, R$. Here, we briefly specify this direction.

The extension can be done in two ways. The first, which is quite straightforward, can be done by assuming that the distribution of the population data in the inference phase is equal to the average of the population distribution of the R training rounds. In other words, $\mu_k = \frac{1}{R}\sum_{r\in[R]} \mu_{k,r}$, for $k \in [K]$, where $\mu_{k,r}$ is the data distribution of client k during round r.

The second scenario, which is more realistic but also more complicated, is as follows. Assume that for each client $k \in [K]$, there exists a set (family) M_k of distributions μ_k, and consider a probability distribution ν_k over this set. Then, for each client during each round, as well as during the inference phase, a distribution μ_k from the set M_k is chosen in an i.i.d. manner (along the rounds) according to ν_k. It seems that our framework can be combined with the ideas presented in (Hu et al., 2023) (used there to study the effect of client participation) to address this extension. Alternatively, our framework can be combined with the approaches used in the "meta-learning" literature to address this extension. These directions are left for future work.

Online Federated Learning setup. Finally, we note that in the online federated learning setup, the generalization error for each round is often evaluated with respect to the immediately aggregated model at the end of that round (Kamp et al., 2016; Kamp, 2019). As discussed after Theorem 3.1, this scenario is a simpler one-round virtual setup that can be studied by our

framework. We recall that such setups have also been partially studied in (Kamp et al., 2016; Kamp, 2019; Barnes et al., 2022b; Haddouche & Guedj, 2022).

F.2. On PAC-Bayes bounds

In this paper, we proposed several PAC-Bayes bounds. Here, we discuss the future directions concerning their computation and their utilization as a regularizer term or to estimate the optimal R^*.

Computation. The data-dependent PAC-Bayes bounds of Theorems 3.1, 3.2, and D.3 does not have a closed-form in general. While PAC-Bayes bounds may become vacuous in some "extreme" cases, *e.g.,* (Haghifam et al., 2023), they are almost tight in many cases (Alquier, 2021). Therefore, as a future direction, it would be of particular interest to compute or estimate such bounds. In particular, the approaches discussed in (Alquier, 2021, Section 3.3.), based on (Dziugaite & Roy, 2017; Pérez-Ortiz et al., 2021), could potentially be applied to this goal. Alternatively, such bounds can be estimated by using the compressibility/pruning principle (Zhou et al., 2018; Lotfi et al., 2022).

Regularizer. The established PAC-Bayes bounds can also potentially be used for another purpose: to be used as a regularizer. This approach has already been used for the centralized setups *e.g.,* (Haddouche & Guedj, 2022; Wang et al., 2021).

Optimal R^*. If the obtained generalization bounds can be properly estimated, they can potentially be used to estimate the optimal number of communication rounds R^* that minimizes the population risk. More specifically, one can consider finding an R^* that minimizes $\text{emp}(R) + \lambda \text{gen}(R)$, where $\text{emp}(R)$ and $\text{gen}(R)$ are estimated bounds on the empirical risk and the generalization error in R rounds FL. This is an attempt to minimize the population risk. One can also consider the communication overhead ($C(R)$) and try to minimize $\text{emp}(R) + \lambda_1 \text{gen}(R) + \lambda_2 C(R)$, which is a tradeoff between population risk and communication overhead. Note that this goal also requires either a good estimate of the empirical risk or a guarantee that the empirical risk will be negligible.

F.3. FSVM

In Section 4, we established a generalization bound (in Theorem 4.1) that increases with the number of rounds R when the dataset size, *i.e.,* n, is fixed. In this section, we study the behavior of the bound of Theorem 4.1 with respect to the number of rounds R, but when the number of samples used in each communication round by each client is fixed. In particular, we show that in this case the upper bound 9 converges to zero when $R \to \infty$.

More precisely, let $n = R \times n_0$, where n_0 is a fixed constant that denotes the amount of dataset used by each client in each round. Note that in this setup, if $R \to \infty$, then $n \to \infty$ as well. Also consider fixed values of K, B, θ, and $q_{e,b} < 1$. For ease of presentation, denote $a := K\theta/B$.

With these notations, it is easy to see that the orderwise behavior of (9) with respect to R is

$$\mathcal{O}\left(\sqrt{\frac{\log(R)\sum_{r\in[R]} L_r}{R}}\right),$$

which can be further upper bounded by

$$\mathcal{O}\left(\sqrt{\frac{\log(R)\sum_{r\in[R]} q_{e,b}^{2r}\log\Big(\max(aq_{e,b}^{-r},2)\Big)}{R}}\right).$$

Now, since $\max(aq_{e,b}^{-r},2) \leqslant \max(a,1)(q_{e,b}^{-1}+2)^r$, the above orderwise behavior can be further upper bounded by

$$\mathcal{O}\left(\sqrt{\frac{\log(R)\sum_{r\in[R]}(r+1)q_{e,b}^{2r}}{R}}\right).$$

For the contractive SGD, the summation in the numerator of the square root converges to

$$\frac{q_{e,b}^2}{(1-q_{e,b}^2)^2} + \frac{q_{e,b}^2}{(1-q_{e,b}^2)},$$

as $R \to \infty$; and, hence, the upper bound converges to zero as $R \to \infty$ (and $n \to \infty$ consequently).

G. Proofs

In this section, we provide the proofs of all theoretical results established in the paper and the appendices, by the order of their appearances.

G.1. Proof of Theorem 3.1

Without loss of generality assume $\sigma = 1/2$, otherwise, consider the properly scaled loss function.

G.1.1. Part I.

Proof. The proof of this theorem is a particular case of Theorem 3.2, proved in Appendix G.2 for bounded losses and is similar for the σ-subgaussian case. We, however, provide the proof for the sake of completeness.

Similar to the proof of (Sefidgaran & Zaidi, 2023, Theorem 6), it is easy to see that for any choice of the prior $\mathsf{P}_{k,r}$, it suffices to only consider an arbitrary $P_{\mathbf{W}|\mathbf{S}}$. Now, let $\lambda := 2n/R - 1$. Let the RHS of the bound as $\sqrt{\Delta(\mathbf{S})}$, *i.e.,* let

$$\Delta(\mathbf{S}) := \frac{\frac{1}{KR}\sum_{k\in[K],r\in[R]} \mathbb{E}_{\overline{W}^{(r-1)}\sim P_{\overline{W}^{(r-1)}|S_{[K]}^{[r-1]}}}\left[D_{KL}\left(p_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\|\mathsf{P}_{k,r}\right)\right] + \log(\frac{\sqrt{2n}}{\sqrt{R}\delta})}{\lambda}.$$

Recall that

$$\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right] = \frac{1}{KR}\sum_{k\in[K],r\in[R]} \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})\right].$$

For any fixed $\delta > 0$, denote

$$\mathcal{G}_{\mathbf{S}}^{\delta} := \{\nu_{\mathbf{S}} \colon D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) \leqslant \log(1/\delta)\}, \tag{38}$$

as the set of all distributions $\nu_{\mathbf{S}}$ over $\mathcal{Z}^{nK}$, whose KL-divergence with $P_{\mathbf{S}} = \mu^{\otimes nK}$ does not exceed $\log(1/\delta)$. Then,

$$\begin{aligned}
&\log\mathbb{P}_{\mathbf{S}}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right] > \sqrt{\Delta(\mathbf{S})}\right)\\
&= \log\mathbb{P}_{\mathbf{S}}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]^2 > \Delta(\mathbf{S})\right)\\
&\overset{(a)}{\leqslant} \max\left(\log(\delta),\ \sup_{\nu_{\mathbf{S}}\in\mathcal{G}_{\mathbf{S}}^{\delta}}\left\{-D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S}}}\left[\Delta(\mathbf{S}) - \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]^2\right]\right\}\right),\\
&\overset{(b)}{\leqslant} \max\left(\log(\delta),\ \sup_{\nu_{\mathbf{S}}\in\mathcal{G}_{\mathbf{S}}^{\delta}}\left\{-D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S}}}\left[\Delta(\mathbf{S}) - \frac{1}{KR}\sum_{k,r}\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2\right]\right]\right\}\right),
\end{aligned} \tag{39}$$

where (a) is shown in the next lemma, proved in Appendix G.8, and (b) is due to the Jensen inequality.

Lemma G.1. *The inequality* (39) *holds.*

Now, showing that the second term of (39) is bounded by $\log(\delta)$ completes the proof. For any $\mathbf{s}$ and any $k \in [K]$ and $r \in [R]$, we have

$$\begin{aligned}
\lambda\mathbb{E}_{\nu_{\mathbf{S}}P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2\right] &\overset{(a)}{\leqslant} D_{KL}\left(\nu_{\mathbf{S}}P_{\mathbf{W}|\mathbf{S}}\|P_{\mathbf{S}}(P\mathsf{P}_{k,r})_{k,r}\right) + \log\mathbb{E}_{P_{\mathbf{S}}(P\mathsf{P}_{k,r})_{k,r}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)},\hat{\overline{W}}^{(R)})^2}\right]\\
&\overset{(b)}{\leqslant} D_{KL}\left(\nu_{\mathbf{S}}P_{\mathbf{W}|\mathbf{S}}\|P_{\mathbf{S}}(P\mathsf{P}_{k,r})_{k,r}\right) + \log(\sqrt{2n/R}) \qquad (40)\\
&= \mathbb{E}_{\nu_{\mathbf{S}}P_{\overline{W}^{(r-1)}|S_{[K]}^{[r-1]}}}\left[D_{KL}\left(P_{W_k^{(r)}|S_k^{(r)},W^{(r-1)}}\|\mathsf{P}_{W_k^{(r)}|\overline{W}^{(r-1)}}\right)\right]\\
&\quad + D_{KL}\left(\nu_{\mathbf{S}}\|P_{\mathbf{S}}\right) + \log(\sqrt{2n/R}), \qquad (41)
\end{aligned}$$

where (a) holds by Donsker-Varadhan's inequality, and where

$$(P\mathsf{P}_{k,r})_{k,r} := P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|S_{[K]}^{[r-1]},S_{[K]\setminus k}^{(r)}}\mathsf{P}_{W_k^{(r)}|\overline{W}^{(r-1)}}P_{\overline{W}^{(R)}|W_{[K]}^{(r)},S_{[K]}^{[r+1:R]}},$$

and (b) is derived by using (Wainwright, 2019) and the fact that $\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})$ is $1/\sqrt{4n/R}$-subgaussian. More precisely, using (Wainwright, 2019, Theorem 2.6.IV.), we have that for any $\lambda_1 \in [0,1)$,

$$\mathbb{E}\left[e^{\frac{\lambda_1 \,\mathrm{gen}\left(s_k^{(r)},\hat{\overline{W}}^{(R)}\right)^2}{(R/2n)}}\right] \leqslant \frac{1}{\sqrt{1-\lambda_1}}.$$

Now, letting $\lambda_1 = \frac{2n/R-1}{2n/R} = \frac{\lambda}{2n/R,}$, we derive that

$$\mathbb{E}\left[e^{\lambda\,\mathrm{gen}\left(s_k^{(r)},\hat{\overline{W}}^{(R)}\right)^2}\right] = \mathbb{E}\left[e^{\frac{\lambda_1 \,\mathrm{gen}\left(s_k^{(r)},\hat{\overline{W}}^{(R)}\right)^2}{(2R/n)}}\right] \leqslant \frac{1}{\sqrt{1-\lambda_1}} = \sqrt{2n/R}.$$

Combining (41) with (39) completes the proof. □

G.1.2. PART II.

Proof. Consider any FL-model $P_{\mathbf{W}|\mathbf{S}}$. Consider any set of *priors* $\{\mathsf{P}_{k,r}\}_{k\in[K],r\in[R]}$ such that $\mathsf{P}_{k,r}$ could depend on $\overline{W}^{(r-1)}$, *i.e.,* $\mathsf{P}_{k,r}$ is a conditional prior on $W_K^{(r)}$ given $\overline{W}^{(r-1)}$. Let $\lambda := 2n/R - 1$. Let the RHS of the bound as $\sqrt{\Delta(\mathbf{S},\mathbf{W})}$, *i.e.,* let

$$\Delta(\mathbf{S},\mathbf{W}) := \frac{\frac{1}{KR}\sum_{k\in[K],r\in[R]}\log\left(\frac{\mathrm{d}P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}}{\mathrm{d}\mathsf{P}_{k,r}r}\right) + \log(\frac{\sqrt{2n}}{\sqrt{R}\delta})}{\lambda}.$$

Further, recall that

$$\mathrm{gen}(\mathbf{S},\overline{W}^{(R)}) = \frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)}).$$

Denote

$$(P\mathsf{P})_{k,r} := P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|S_{[K]}^{[r-1]},S_{[K]\setminus k}^{(r)}}\mathsf{P}_{k,r}P_{\overline{W}^{(R)}|\overline{W}^{(r)},S_{[K]}^{[r+1:R]}}, \quad k\in[K], r\in[R].$$

For any fixed $\delta > 0$, denote

$$\mathcal{G}_{\mathbf{S},\mathbf{W}}^{\delta} := \{\nu_{\mathbf{S},\mathbf{W}} \colon D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|P_{\mathbf{S},\mathbf{W}}) \leqslant \log(1/\delta)\}, \tag{42}$$

as the set of all distributions $\nu_{\mathbf{S},\mathbf{W}}$ over $\mathcal{Z}^{nK}\times\mathcal{W}^{(n+1)K}$, whose KL-divergence with $P_{\mathbf{S},\mathbf{W}}$ does not exceed $\log(1/\delta)$. Then,

$$\mathbb{P}\Big(\mathrm{gen}(\mathbf{S},\overline{W}^{(R)}) \geqslant \sqrt{\Delta(\mathbf{S},\mathbf{W})}\Big) \tag{43}$$

$$\overset{(a)}{\leqslant}\mathbb{P}\Big(\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2 \geqslant \Delta(\mathbf{S},\mathbf{W})\Big)$$

$$\overset{(b)}{\leqslant}\max\left(\log(\delta), \sup_{\nu_{\mathbf{S},\mathbf{W}}\in\mathcal{G}_{\mathbf{S},\mathbf{W}}^{\delta}}\left\{-D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|P_{\mathbf{S},\mathbf{W}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\left[\Delta(\mathbf{S},\mathbf{W}) - \frac{1}{KR}\sum_{r,k}\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2\right]\right\}\right), \tag{44}$$

where (a) holds by Jensen inequality for the convex function $g(x) = x^2$ and (b) is due Lemma G.2, proved in Appendix G.9.

Lemma G.2. *The inequality* (44) *holds.*

Next, note that using Donsker-Varadhan's inequality we have

$$\begin{aligned}
&\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\Big[\lambda\,\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2\Big]\\
&\leqslant \mathbb{E}_{\nu_{\mathbf{S}}}\Big[D_{KL}\big(\nu_{\mathbf{W}|\mathbf{S}}\|(P\mathsf{P})_{k,r}\big)+\log\mathbb{E}_{(P\mathsf{P})_{k,r}}\Big[e^{\lambda\,\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2}\Big]\Big]\\
&\leqslant D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|\nu_{\mathbf{S}}(P\mathsf{P})_{k,r})+D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}})+\log\mathbb{E}_{P_{\mathbf{S}}(P\mathsf{P})_{k,r}}\Big[e^{\lambda\,\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2}\Big]\\
&\leqslant D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|\nu_{\mathbf{S}}(P\mathsf{P})_{k,r})+D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}})+\log\Big(\sqrt{2n/R}\Big)\\
&= D_{KL}\big(\nu_{\mathbf{S},\mathbf{W}}\|\nu_{\mathbf{S}}P_{\mathbf{W}|\mathbf{S}}\big)+\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\left[\log\left(\frac{\mathrm{d}P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}}{\mathrm{d}\mathsf{P}_{k,r}r}\right)\right]+D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}})+\log\Big(\sqrt{2n/R}\Big)\\
&= D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|P_{\mathbf{S},\mathbf{W}})+\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\left[\log\left(\frac{\mathrm{d}P_{W_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}}{\mathrm{d}\mathsf{P}_{k,r}r}\right)\right]+\log\Big(\sqrt{2n/R}\Big).
\end{aligned}$$

Putting everything together conclude that

$$-D_{KL}(\nu_{\mathbf{S},\mathbf{W}}\|P_{\mathbf{S},\mathbf{W}})-\lambda\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\left[\Delta(\mathbf{S},\mathbf{W})-\frac{1}{KR}\sum_{r,k}\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})^2\right]\leqslant\log(\delta).$$

This completes the proof.

□

G.2. Proof of Theorem 3.2

Proof. Similar to the proof of (Sefidgaran & Zaidi, 2023, Theorem 6), it is easy to see that for any choice of the prior $\mathsf{P}_{k,r}$, it suffices to only consider an arbitrary $P_{\mathbf{W}|\mathbf{S}}$. Furthermore, without loss of generality, we can assume $C=1$, otherwise, one can re-scale the loss function.

Let $\lambda := 2n/R-1$ and the RHS of the bound as $\sqrt{\Delta(\mathbf{S})}$, *i.e.,* let

$$\Delta(\mathbf{S}) := \frac{\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}_{\overline{W}^{(r-1)}\sim P_{\overline{W}^{(r-1)}|S_{[K]}^{[r-1]}}}\left[D_{KL}\left(p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}\|\mathsf{P}_{k,r}\right)\right]+\log(\frac{\sqrt{2n}}{\sqrt{R}\delta})}{\lambda}+\epsilon,$$

where $p_{\hat{\mathbf{W}}|\mathbf{S}}$ satisfy the ϵ-distortion criterion:

$$\left|\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big]-\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}_{\mathbf{W}\sim(Pp)_{k,r}}\Big[\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\Big]\right|\leqslant\epsilon/2, \tag{45}$$

where

$$(Pp)_{k,r} := P_{\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)}|S_{[K]}^{[r-1]},S_{[K]\setminus k}^{(r)}}p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}}P_{\hat{\overline{W}}^{(R)}|W_{[K]\setminus k}^{(r)},\hat{W}_k^{(r)},S_{[K]}^{[r+1:R]}}.$$

Further, recall that

$$\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big]=\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\Big[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})\Big].$$

First, note that

$$\mathbb{P}_{\mathbf{S}}\Big(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big]>\sqrt{\Delta(\mathbf{S})}\Big)=\mathbb{P}_{\mathbf{S}}\Big(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big]^2>\Delta(\mathbf{S})\Big). \tag{46}$$

Next, we state the following lemma, proved in Appendix G.10.

Lemma G.3. *For any $\delta > 0$,*

$$\log \mathbb{P}_{\mathbf{S}}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right]^2 > \Delta(\mathbf{S})\right) \leqslant \max\left(\log(\delta), \sup_{\nu_{\mathbf{S}}\in\mathcal{G}_{\mathbf{S}}^{\delta}}\left\{-D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S}}}\left[\Delta(\mathbf{S}) - \epsilon - \tilde{\Delta}(\mathbf{S})^2\right]\right\}\right), \tag{47}$$

where $\mathcal{G}_{\mathbf{S}}^{\delta}$ is defined in (38) and

$$\tilde{\Delta}(\mathbf{S}) := \frac{1}{KR}\sum_{r,k}\mathbb{E}_{\mathbf{W}\sim(Pp)_{k,r}}\left[\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})\right].$$

Now, for any $\mathbf{s}$, we have

$$\begin{aligned}
\lambda\tilde{\Delta}(\mathbf{S})^2 &\overset{(a)}{\leqslant} \frac{1}{KR}\sum_{r,k}\mathbb{E}_{\mathbf{W}\sim(Pp)_{k,r}}\left[\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2\right] \\
&\overset{(b)}{\leqslant} \frac{1}{KR}\sum_{r,k}\left(D_{KL}\left((Pp)_{k,r}\|(P\mathsf{P}_{k,r})_{k,r}\right) + \log\mathbb{E}_{\mathbf{W}\sim(P\mathsf{P}_{k,r})_{k,r}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2}\right]\right) \\
&= \frac{1}{KR}\sum_{r,k}\left(\mathbb{E}_{\overline{W}^{(r-1)}\sim P_{\overline{W}^{(r-1)}|S_{[K]}^{[r-1]}}}\left[D_{KL}\left(p_{\hat{W}_k^{(r)}|s_k^{(r)}, W^{(r-1)}}\|\mathsf{P}_{\hat{W}_k^{(r)}|\overline{W}^{(r-1)}}\right)\right]\right. \\
&\qquad\qquad \left. + \log\mathbb{E}_{\mathbf{W}\sim(P\mathsf{P}_{k,r})_{k,r}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2}\right]\right),
\end{aligned}$$

where (a) is derived by using Jensen inequality for the function $g(x) = x^2$, (b) using Donsker-Varadhan's inequality, and where

$$(P\mathsf{P}_{k,r})_{k,r} := P_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}|S_{[K]}^{[r-1]}, S_{[K]\setminus k}^{(r)}}\mathsf{P}_{\hat{W}_k^{(r)}|\overline{W}^{(r-1)}}P_{\hat{\overline{W}}^{(R)}|W_{[K]\setminus k}^{(r)}, \hat{W}_k^{(r)}, S_{[K]}^{[r+1:R]}}.$$

Hence, for any $\nu_{\mathbf{S}}$, we have

$$\begin{aligned}
&-D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S}}}\left[\Delta(\mathbf{s}) - \epsilon - \tilde{\Delta}(\mathbf{S})^2\right] - \log(\delta) \\
&\leqslant -\log\left(\frac{\sqrt{2n}}{\sqrt{R}}\right) - D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) + \frac{1}{KR}\sum_{r,k}\mathbb{E}_{\mathbf{S}\sim\nu_{\mathbf{S}}}\log\mathbb{E}_{\mathbf{W}\sim(P\mathsf{P}_{k,r})_{k,r}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2}\right] \\
&\overset{(a)}{\leqslant} -\log\left(\frac{\sqrt{2n}}{\sqrt{R}}\right) + \frac{1}{KR}\sum_{r,k}\log\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S}}(P\mathsf{P}_{k,r})_{k,r}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2}\right] \\
&= -\log\left(\frac{\sqrt{2n}}{\sqrt{R}}\right) + \frac{1}{KR}\sum_{r,k}\log\mathbb{E}_{\mathbf{S}\setminus S_k^{(r)},\mathbf{W}\sim P_{\mathbf{S}\setminus S_k^{(r)}}(P\mathsf{P}_{k,r})_{k,r}}\mathbb{E}_{S_k^{(r)}\sim\mu_k^{\otimes(n/R)}}\left[e^{\lambda\,\mathrm{gen}(s_k^{(r)}, \hat{\overline{W}}^{(R)})^2}\right] \\
&\leqslant 0,
\end{aligned}$$

where (a) holds due to Donsker-Varadhan's inequality. This completes the proof. □

G.3. Proof of Theorem 3.3

Proof. Consider any set of distributions $\left\{p_{\hat{W}_k^{(r)}|S_k^{(r)}, \overline{W}^{(r-1)}}\right\}_{k\in[K], r\in[R]}$ that satisfy the distortion criterion (8) for any $k \in [K]$ and $r \in [R]$. Then,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right] \\
&\leqslant \frac{1}{KR}\sum_{k\in[K], r\in[R]}\left(\mathbb{E}_{S_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, \hat{\overline{W}}^{(R)}\sim P_{S_k^{(r)}}(Pp)_{k,r}}\left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] + \epsilon_{k,r}\right)
\end{aligned}$$

$$\leqslant \frac{1}{KR} \sum_{k\in[K], r\in[R]} \mathbb{E}_{S_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, \hat{\overline{W}}^{(R)} \sim P_{S_k^{(r)}}(Pp)_{k,r}} \left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)}) \right] + \epsilon, \tag{48}$$

where

$$(Pp)_{k,r} := P_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}} p_{\hat{W}_k^{(r)}|S_k^{(r)}, \overline{W}^{(r-1)}} P_{\hat{\overline{W}}^{(R)}|W_{[K]\setminus k}^{(r)}, \hat{W}_k^{(r)}}.$$

Let $q_{\hat{W}_k^{(r)}|\overline{W}^{(r-1)}}$ be the marginal conditional distribution of $\hat{W}_k^{(r)}$ given $\overline{W}^{(r-1)}$ under $P_{S_k^{(r)}} p_{\hat{W}_k^{(r)}|S_k^{(r)}, \overline{W}^{(r-1)}}$, and denote similarly

$$(Pq)_{k,r} := P_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}} q_{\hat{W}_k^{(r)}|\overline{W}^{(r-1)}} P_{\hat{\overline{W}}^{(R)}|W_{[K]\setminus k}^{(r)}, \hat{W}_k^{(r)}}.$$

Then,

$$\begin{aligned}
&\lambda \mathbb{E}_{S_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, \hat{\overline{W}}^{(R)} \sim P_{\mathbf{S}}(Pp)_{k,r}} \left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)}) \right] \\
&\quad \overset{(a)}{\leqslant} D_{KL}\Big(P_{S_k^{(r)}}(Pp)_{k,r} \| P_{S_k^{(r)}}(Pq)_{k,r} \Big) + \log \mathbb{E}_{P_{S_k^{(r)}}(Pq)_{k,r}} \left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})} \right] \\
&\quad = I(S_k^{(r)}; \hat{W}_k^{(r)} | \overline{W}^{(r-1)}) + \log \mathbb{E}_{(Pq)_{k,r}} \mathbb{E}_{P_{S_k^{(r)}}} \left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})} \right] \\
&\quad \overset{(b)}{\leqslant} I(S_k^{(r)}; \hat{W}_k^{(r)} | \overline{W}^{(r-1)}) + \frac{\lambda^2 \sigma^2}{2n/R},
\end{aligned} \tag{49}$$

where (a) is deduced using Donsker-Varadhan's inequality and (b) using the fact that for any $w \in \mathcal{W}$, $\mathrm{gen}(S_k^{(r)}, w)$ is $\sigma/\sqrt{n/R}$-subgaussian.

Combining (48) and (49), and taking the infimum over all admissible choices of the conditional Markov kernels $\left\{ p_{\hat{W}_k^{(r)}|S_k^{(r)}, \overline{W}^{(r-1)}} \right\}_{k\in[K], r\in[R]}$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}, \mathbf{W} \sim P_{\mathbf{S}, \mathbf{W}}} \left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)}) \right] \\
&\leqslant \frac{1}{KR} \sum_{k\in[K], r\in[R]} \inf_{p_{\hat{W}_k^{(r)}|S_k^{(r)}, \overline{W}^{(r-1)}}} \left\{ \mathbb{E}_{S_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, \hat{\overline{W}}^{(R)} \sim P_{S_k^{(r)}}(Pp)_{k,r}} \left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)}) \right] \right\} + \epsilon \\
&\leqslant \frac{1}{KR} \sum_{k\in[K], r\in[R]} \mathfrak{RD}(P_{\mathbf{S}, \mathbf{W}}, k, r, \epsilon_{k,r})/\lambda + \frac{\lambda \sigma^2}{2n/R} + \epsilon \\
&\overset{(a)}{\leqslant} \sqrt{\frac{2\sigma^2 \sum_{k\in[K], r\in[R]} \mathfrak{RD}(P_{\mathbf{S}, \mathbf{W}}, k, r, \epsilon_{k,r})}{nK}} + \epsilon,
\end{aligned} \tag{50}$$

where the last step is established by letting

$$\lambda := \sqrt{\frac{2n/R}{\sigma^2 KR} \sum_{k\in[K], r\in[R]} \mathfrak{RD}(P_{\mathbf{S}, \mathbf{W}}, k, r, \epsilon_{k,r})}.$$

This completes the proof. □

G.4. Proof of Theorem 4.1

Proof. We use an extension of Theorem 3.3, that is Proposition D.2, to prove this theorem. Let $\epsilon_{k,r} = \epsilon$ for all $k \in [K]$ and $r \in [R]$. Fix a $k \in [K]$ and $r \in [R]$. We upper bound the rate-distortion term $\mathfrak{RD}^\star(P_{\mathbf{S}, \mathbf{W}}, k, r, \epsilon_{k,r})$, defined in (30). To

this end, let $V_k^{(r)}$ denote all the randomness in round r for client k, *i.e.,* given $s_k^{(r)}$, all mini-batches $\{\mathcal{B}_{k,r,t}\}_t$ will be fixed. We define a proper

$$p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}, \tag{51}$$

that satisfy (29), where

$$\mathfrak{V}_{k,r} = \Big(V_k^{(r)}, \overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, V_{[K]}^{[r+1:R]}\Big). \tag{52}$$

For simplicity, let

$$\mathfrak{U}_{k,r} = \Big(W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, V_{[K]}^{[r+1:R]}\Big). \tag{53}$$

To define $p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}$, first we define $p_{\hat{\overline{W}}_k^{(R)}|W_k^{(r)},\mathfrak{U}_{k,r}}$. Then, we let

$$p_{\hat{\overline{W}}_k^{(R)}|W_k^{(r)},\mathfrak{V}_{k,r},S_k^{(r)}} = p_{\hat{\overline{W}}_k^{(R)}|W_k^{(r)},\mathfrak{U}_{k,r}}. \tag{54}$$

Now, we have

$$p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}} = \mathbb{E}_{W_k^{(r)}\sim P_{W_k^{(r)}|S_k^{(r)},V_k^{(r)},\overline{W}^{(r-1)}}}\Big[p_{\hat{\overline{W}}_k^{(R)}|W_k^{(r)},\mathfrak{U}_{k,r}}\Big]. \tag{55}$$

Note that $P_{W_k^{(r)}|S_k^{(r)},\mathfrak{V}_{k,r}} = P_{W_k^{(r)}|S_k^{(r)},V_k^{(r)},\overline{W}^{(r-1)}}$. We proceed to define $p_{\hat{\overline{W}}_k^{(R)}|W_k^{(r)},\mathfrak{U}_{k,r}}$. Consider an integer $m \in \mathbb{N}^*$. Fix an $\mathfrak{U}_{k,r}$, *i.e.,* fix $w_{[K]\setminus k}^{(r)}$, $(s_{[K]}^{[r+1:R]}, v_{[K]}^{[r+1:R]})$, and consequently all mini-batches

$$\{\mathcal{B}_{j,r',t}\}_{j\in[K],r'\in[r+1:R],t\in[\tau]}.$$

Assume the aggregated model at round (r) to be

$$\overline{w}^{(r),\star} = \frac{1}{K}\sum_{j\neq k} w_j^{(r)}. \tag{56}$$

Then, with this aggregated model at step (r), and the considered fixed $\mathfrak{U}_{k,r}$, if $r < R$, consider all the induced $\mathsf{D}^\star_{(\overline{w}^{(r'-1)},\{\mathcal{B}_{j,r',t}\}_t)}$, for $j \in [K]$, $r' \in [r+1:R]$. Denote

$$\begin{aligned}\mathsf{D}^\star_{r'} &:= \frac{1}{K}\sum_{k\in[K]} \mathsf{D}^\star_{(\overline{w}^{(r'-1)},\{\mathcal{B}_{k,r',t}\}_t)}, \quad r' \in [r+1:R],\\ \mathsf{D}^\star &:= \prod_{r'\in[r+1:R]} \mathsf{D}^\star_{r'}.\end{aligned} \tag{57}$$

Denote also the resulting global model as $\overline{w}^{(R),\star}$. If $r = R$, let $\mathsf{D}^\star = \mathrm{I}_d$, where I_d is the identity matrix.

Consider the random matrix A, whose elements are distributed in an i.i.d. manner according to $\mathcal{N}(0, 1/m)$. The matrix is used for Johnson-Lindenstrauss transformation (Johnson & Lindenstrauss, 1984). Considering such matrix for dimension reduction in SVM for centralized and one-round distributed learning was previously considered in (Grønlund et al., 2020; Sefidgaran et al., 2022a).

Fix some $c_1, c_2, \nu > 0$. Let

$$\epsilon := 8e^{-\frac{m}{7}\left(\frac{K\theta}{6Bq_{e,b}^{R-r}}\right)^2} + 2e^{-0.21m(c_1^2-1)} + 2e^{-0.21m(c_2^2 q_{e,b}^{2(r-R)}-1)} + \frac{4m\nu^m}{\sqrt{\pi}} e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{6c_1\nu B}\right)^2}. \tag{58}$$

Now, for a given $w_k^{(r)}$, if $\|\mathsf{A}\mathsf{D}^\star w_k^{(r)}\| \leqslant c_2$, then let M be chosen uniformly over $\mathcal{B}_m(\mathsf{A}\mathsf{D}^\star w_k^{(r)}, \nu)$; otherwise let M be chosen uniformly over $\mathcal{B}_m(0, \nu)$. Define

$$\hat{\overline{W}}^{(R)} := \mathsf{A}\overline{W}^{(R),\star} + \frac{1}{K}M. \tag{59}$$

Note that we defined $\hat{\overline{W}}^{(R)} \in \mathbb{R}^m$, while $\overline{W}^{(R)} \in \mathbb{R}^d$. By this definition, we have

$$\begin{aligned}
I(S_k^{(r)}; \hat{\overline{W}}_k^{(R)} | \mathfrak{V}_{k,r}) &\overset{(a)}{\leqslant} I(W_k^{(r)}; \hat{\overline{W}}_k^{(R)} | \mathfrak{V}_{k,r}) \\
&\overset{(b)}{=} I(W_k^{(r)}; \hat{\overline{W}}_k^{(R)} | \mathfrak{U}_{k,r}) \\
&\overset{(c)}{=} I(W_k^{(r)}; M | \mathfrak{U}_{k,r}) \\
&= h(M|\mathfrak{U}_{k,r}) - h(M|W_k^{(r)}, \mathfrak{U}_{k,r}) \\
&= h(M|\mathfrak{U}_{k,r}) - \log(\mathrm{Vol}_m(\nu)) \\
&\overset{(d)}{\leqslant} \log(\mathrm{Vol}_m(c_2+\nu)) - \log(\mathrm{Vol}_m(\nu)) \\
&= m \log((c_2+\nu)/\nu),
\end{aligned} \tag{60}$$

where $\mathrm{Vol}_m(r)$ denote the volume of m-dimensional ball with radius r,

- (a) follows by data-processing inequality,
- (b) and (c) due to the construction of $\hat{\overline{W}}_k^{(R)}$ and since given $\mathfrak{U}_{k,r}$, $\mathsf{D}^\star$ is a fixed matrix,
- and (d) since $M \in \mathbb{R}^m$ is bounded always in the ball of radius $c_2 + \nu$.

Now, we investigate the distortion criterion (29). For this, first, we define the loss function $\hat{\ell}$. For fixed $\mathfrak{u}$ and A, let the 0-1 loss function $\hat{\ell}$ be,

$$\hat{\ell}_{\mathfrak{u},\mathsf{A},\theta}\left(z, \hat{\overline{w}}^{(R)}\right) := \mathbb{1}_{\{y \langle\!\langle x, \hat{\overline{w}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}} > \theta/2\}}, \tag{61}$$

where we define

$$\langle\!\langle x, \hat{\overline{w}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}} := \langle x, \overline{w}^{(R),\star} \rangle - \langle \mathsf{A}x, \mathsf{A}\overline{w}^{(R),\star} \rangle + \langle \mathsf{A}x, \hat{\overline{w}}^{(R)} \rangle. \tag{62}$$

Note that $\overline{w}^{(R),\star}$ and $\mathsf{D}^\star$ are deterministically determined by $\mathfrak{u}$.

It is easy to verify that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S},\mathbf{V},\mathbf{W},\hat{\overline{W}}^{(R)}}\left[\mathrm{gen}_\theta(S_k^{(r)}, \overline{W}^{(R)}) - \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \\
&= \mathbb{E}_{P_{S_k^{(r)}, W_k^{(r)}, \mathfrak{U}_{k,r}, \overline{W}^{(R)}}} \mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)} | W_k^{(r)}, \mathfrak{U}_{k,r}}}\left[\mathrm{gen}(S_k^{(r)}, \overline{W}^{(R)}) - \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \\
&= \mathbb{E}_{P_{W_k^{(r)}, \mathfrak{U}_{k,r}, \overline{W}^{(R)}}} \mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)} | W_k^{(r)}, \mathfrak{U}_{k,r}}} \mathbb{E}_{Z_k \sim \mu_k}\left[\ell_0(Z_k, \overline{W}^{(R)}) - \hat{\ell}_{\mathfrak{U},\mathsf{A},\theta}(Z_k, \hat{\overline{W}}_k^{(R)})\right] \\
&\quad - \mathbb{E}_{P_{S_k^{(r)}, W_k^{(r)}, \mathfrak{U}_{k,r}, \overline{W}^{(R)}}} \mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)} | W_k^{(r)}, \mathfrak{U}_{k,r}}}\left[\frac{1}{n_R}\sum_{i=1}^{n_R} \ell_\theta(Z_{k,r,i}, \overline{W}_k^{(R)}) - \hat{\ell}_{\mathfrak{U},\mathsf{A},\theta}(Z_{k,r,i}, \hat{\overline{W}}_k^{(R)})\right] \\
&\leqslant \mathbb{E}_{P_{W_k^{(r)}, \mathfrak{U}_{k,r}, \overline{W}^{(R)}}} \mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)} | W_k^{(r)}, \mathfrak{U}_{k,r}}} \mathbb{E}_{Z_k \sim \mu_k}\left[\mathbb{1}_{\left\{\left|\langle X_k, \overline{W}^{(R)} \rangle - \langle\!\langle X_k, \hat{\overline{W}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}}\right| > \theta/2\right\}}\right] \\
&\quad + \mathbb{E}_{P_{S_k^{(r)}, W_k^{(r)}, \mathfrak{U}_{k,r}, \overline{W}^{(R)}}} \mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)} | W_k^{(r)}, \mathfrak{U}_{k,r}}}\left[\frac{1}{n_R}\sum_{i=1}^{n_R} \mathbb{1}_{\left\{\left|\langle X_{k,r,i}, \overline{W}^{(R)} \rangle - \langle\!\langle X_{k,r,i}, \hat{\overline{W}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}}\right| > \theta/2\right\}}\right] \\
&:= \mathcal{E}_{\mathsf{A}}.
\end{aligned} \tag{63}$$

In the rest of the proof we show that

$$\begin{aligned}
\mathbb{E}_{\mathsf{A}}[\mathcal{E}_{\mathsf{A}}] &\leqslant 8e^{-\frac{m}{7}\left(\frac{K\theta}{6Bq_{e,b}^{R-r}}\right)^2} + 2e^{-0.21m(c_1^2-1)} + 2e^{-0.21m(c_2^2 q_{e,b}^{2(r-R)}-1)} + \frac{4m\nu^m}{\sqrt{\pi}} e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{6c_1\nu B}\right)^2} \\
&= \epsilon,
\end{aligned} \tag{64}$$

where ϵ is defined in (58). Then, this shows that there exists a least one A for which $\mathcal{E}_{\mathsf{A}} \leqslant \epsilon$. Combining this with (60), yields

$$\mathfrak{RD}^\star(P_{\mathbf{S},\mathbf{W}}, k, r, \epsilon) \leqslant m \log((c_2+\nu)/\nu).$$

Letting

$$\begin{aligned} m :=& \lceil 252 \left(\frac{Bt}{K\theta}\right)^2 \log(nK\sqrt{K}) \rceil, \\ c_1 :=& \sqrt{\frac{K^2\theta^2}{50B^2t^2} - 1}, \\ c_2 :=& q_{e,b}^{R-r} \sqrt{\frac{K^2\theta^2}{50B^2t^2} - 1}, \\ \nu :=& t/(2c_1), \end{aligned} \tag{65}$$

where $t \geqslant q_{e,n}^{R-r}$, and using Theorem 3.3 completes the proof.

Hence, it remains to show (64). A sufficient condition to show that is to prove for every $x \in \mathbb{R}^d$, and every $w_k^{(r)}$ and $\mathfrak{u}_{k,r}$, we have

$$\mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)}|w_k^{(r)},\mathfrak{u}_{k,r}}} \mathbb{E}_{\mathsf{A}} \left[\mathbb{1}_{\left\{ \left| \langle x, \overline{w}^{(R)} \rangle - \langle\!\langle x, \hat{\overline{W}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}} \right| > \theta/2 \right\}} \right] \leqslant \epsilon/2. \tag{66}$$

Note that $\overline{w}^{(R)}$ is a deterministic function of $w_k^{(r)}$ and $\mathfrak{u}_{k,r}$. Next, we decompose the difference of the inner products into three terms:

$$\begin{aligned} &\left| \langle x, \overline{w}^{(R)} \rangle - \langle\!\langle x, \hat{\overline{W}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}} \right| \\ &= \left| \langle x, \Delta \rangle + \frac{1}{K} \langle x, \mathsf{D}^\star w_k^{(r)} \rangle - \frac{1}{K} \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} \rangle + \frac{1}{K} \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M \rangle \right| \\ &\leqslant |\langle x, \Delta \rangle| + \frac{1}{K} \left| \langle x, \mathsf{D}^\star w_k^{(r)} \rangle - \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} \rangle \right| + \frac{1}{K} \left| \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M \rangle \right|, \end{aligned} \tag{67}$$

where

$$\Delta := \overline{w}^{(R)} - \left(\overline{w}^{(R),\star} + \frac{1}{K} \mathsf{D}^\star w_k^{(r)} \right). \tag{68}$$

Note that Δ is deterministic given $w_k^{(r)}$ and $\mathfrak{u}_{k,r}$.

Hence,

$$\begin{aligned} &\mathbb{E}_{p_{\hat{\overline{W}}_k^{(R)}|w_k^{(r)},\mathfrak{u}_{k,r}}} \mathbb{E}_{\mathsf{A}} \left[\mathbb{1}_{\left\{ \left| \langle x, \overline{w}^{(R)} \rangle - \langle\!\langle x, \hat{\overline{W}}^{(R)} \rangle\!\rangle_{\mathfrak{u},\mathsf{A}} \right| > \theta/2 \right\}} \right] \\ &\leqslant \mathbb{1}_{\{|\langle x, \Delta \rangle| > \theta/6\}} \qquad (69) \\ &\quad + \mathbb{E}_{\mathsf{A}} \left[\mathbb{1}_{\left\{ \frac{1}{K} \left| \langle x, \mathsf{D}^\star w_k^{(r)} \rangle - \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} \rangle \right| > \theta/6 \right\}} \right] \qquad (70) \\ &\quad + \mathbb{E}_{p_{M|w_k^{(r)}}} \mathbb{E}_{\mathsf{A}} \left[\mathbb{1}_{\left\{ \frac{1}{K} \left| \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M \rangle \right| > \theta/6 \right\}} \right]. \qquad (71) \end{aligned}$$

Now, we proceed to bound the probability that each of the three terms in the RHS of the above inequality.

Bounding (69). We show that with probability one, $|\langle x, \Delta\rangle| < \theta/6$. Note that for $r = R$, we have $\mathsf{D}^\star = \mathrm{I}_d$ and this term is zero. Now, for $R > 1$, a sufficient condition to prove $|\langle x, \Delta\rangle| < \theta/6$ is to show that for any $r < R$, we have $\|\Delta\| \leqslant \theta/(6B)$, when (20) holds. We show this if any of two conditions in Assumption C.3 hold.

For $r' \in [r+1 : R]$, define

$$\xi_{r'} := \overline{w}^{(r')} - \Big(\overline{w}^{(r'),\star} + \mathsf{D}^\star_{r'}\Big(\overline{w}^{(r'-1)} - \overline{w}^{(r'-1),\star}\Big)\Big). \tag{72}$$

1. **If Condition 1 holds:** Assume (20) holds. Fix some $\beta \in (0, 1]$. Then, first using induction we simultaneously show that for $r' = r+1, \dots, R$,

$$\|\xi_{r'}\| \leqslant \frac{\alpha((1+\beta)q_{e,b}^2)^{r'-r-1}}{K^2}, \tag{73}$$

$$\|\overline{w}^{(r'-1)} - \overline{w}^{(r'-1),\star}\|^2 \leqslant \frac{((1+\beta)q_{e,b}^2)^{r'-r-1}}{K^2}. \tag{74}$$

For $r' = r+1$, due to Assumption 2, we have

$$\|\xi_{r+1}\| \leqslant \alpha\|\overline{w}^{(r)} - \overline{w}^{(r),\star}\|^2/K^2 = \alpha\|w_k^{(r)}\|^2/K^2 \leqslant \alpha/K^2. \tag{75}$$

This proves the induction base for both (73) and (74).
Now, assume that (73) and (74) hold for $r' = r+1, \dots, j$, $j < R$. We show that they hold, for $j+1$ as well. First, we have

$$\begin{aligned}
\|\overline{w}^{(j)} - \overline{w}^{(j),\star}\|^2 &= \Big\|\xi_j + \mathsf{D}^\star_j\Big(\overline{w}^{(j-1)} - \overline{w}^{(j-1),\star}\Big)\Big\|^2 \\
&\overset{(a)}{\leqslant} \|\xi_j\|^2 + q_{e,b}^2\Big\|\overline{w}^{(j-1)} - \overline{w}^{(j-1),\star}\Big\|^2 \\
&\overset{(b)}{\leqslant} \frac{\alpha^2((1+\beta)q_{e,b}^2)^{2(j-r-1)}}{K^4} + q_{e,b}^2\frac{((1+\beta)q_{e,b}^2)^{j-r-1}}{K^2} \\
&\overset{(c)}{\leqslant} \frac{((1+\beta)q_{e,b}^2)^{j-r}}{K^2},
\end{aligned} \tag{76}$$

where (a) holds due to the triangle inequality and since spectral norm of $\mathsf{D}^\star_j$ is bounded by $q_{e,b}$, (b) using the assumption of the induction, and (c) holds when

$$K^2 > \frac{\alpha^2((1+\beta)q_{e,b}^2)^{j-r-1}}{\beta q_{e,b}^2},$$

which holds by (20). Hence, (74) holds for $r' = j+1$ as well. Now, we show (73) also holds for $r' = j+1$.

$$\begin{aligned}
\|\xi_{j+1}\| &:= \Big\|\overline{w}^{(j+1)} - \Big(\overline{w}^{(j+1),\star} + \mathsf{D}^\star_{j+1}\Big(\overline{w}^{(j)} - \overline{w}^{(j),\star}\Big)\Big)\Big\| \\
&\overset{(a)}{\leqslant} \alpha\Big\|\overline{w}^{(j)} - \overline{w}^{(j),\star}\Big\|^2 \\
&\overset{(b)}{\leqslant} \frac{\alpha((1+\beta)q_{e,b}^2)^{j-r}}{K^2},
\end{aligned} \tag{77}$$

where (a) holds due to Assumption 2 and (b) by (76). This completes the proof of the induction.
Now, for any $r' \in [r+1 : R-1]$, denote

$$\overline{\mathsf{D}}^\star_{r'} := \prod_{j=r'+1}^{R} \mathsf{D}^\star_j, \tag{78}$$

and $\overline{\mathsf{D}}_R^\star = \mathrm{I}_d$. Then,

$$\begin{aligned}\left\|\overline{w}^{(R)} - \left(\overline{w}^{(R),\star} + \frac{1}{K}\mathsf{D}^\star w_k^{(r)}\right)\right\| =& \left\|\sum_{r'=r+1}^{R} \overline{\mathsf{D}}_{r'}^\star \xi_{r'}\right\| \\ \leqslant& \sum_{r'=r+1}^{R} q_{e,b}^{R-r'} \|\xi_{r'}\| \\ \leqslant& \sum_{r'=r+1}^{R} q_{e,b}^{R-r'} \frac{\alpha((1+\beta)q_{e,b}^2)^{r'-r-1}}{K^2} \\ =& \frac{\alpha q_{e,b}^{R-r-1}}{K^2} \sum_{r'=r+1}^{R} ((1+\beta)q_{e,b})^{r'-r-1} \\ =& \frac{\alpha q_{e,b}^{R-r-1}\big(1-((1+\beta)q_{e,b})^{R-r}\big)}{K^2(1-(1+\beta)q_{e,b})}. \end{aligned} \tag{79}$$

This term is always less than $\theta/(6B)$, if

$$\frac{6\alpha B q_{e,b}^{R-r-1}\big(1-((1+\beta)q_{e,b})^{R-r}\big)}{\theta(1-(1+\beta)q_{e,b})} \leqslant K^2,$$

which holds by (20). This completes the proof of this step.

2. **If Condition 2 holds:** Note that similar to the previous case, it can be shown that for $r' = r+1, \ldots, R$

$$\|\xi_{r'}\| \leqslant \alpha \|\overline{w}^{(r'-1)} - \overline{w}^{(r'-1),\star}\|^2, \tag{80}$$

$$\|\overline{w}^{(r')} - \overline{w}^{(r'),\star}\|^2 \leqslant \|\xi_{r'}\|^2 + q_{e,b}^2 \|\overline{w}^{(r'-1)} - \overline{w}^{(r'-1),\star}\|^2. \tag{81}$$

Define the sequence $B_j \in \mathbb{R}^+$, $j = 0, \ldots, R-r-1$, recursively as

$$B_j = \alpha B_{j-1}^2 + q_{e,b}^2 B_{j-1},$$

where $B_0 = \frac{1}{K^2}$. Inequalities (80) and (81) conclude that

$$\|\xi_{r'}\| \leqslant \alpha B_{r'-1}. \tag{82}$$

We claim under Condition 2 in Assumption C.3, $B_j \leqslant \frac{1-q_{e,b}^2}{\alpha} - \nu$. Once this claim is shown, then we have

$$\begin{aligned} B_j =& \alpha B_{j-1}^2 + q_{e,b}^2 B_{j-1} \\ \leqslant& \alpha\left(\frac{1-q_{e,b}^2}{\alpha} - \nu\right) B_{j-1} + q_{e,b}^2 B_{j-1} \\ =& (1-\nu\alpha) B_{j-1}. \end{aligned}$$

Hence $B_j \leqslant \frac{(1-\alpha\nu)^j}{K^2}$, and $\|\xi_{r'}\| \leqslant \frac{\alpha(1-\alpha\nu)^{j-1}}{K^2}$. Finally, similar to (79),

$$\begin{aligned}\left\|\overline{w}^{(R)} - \left(\overline{w}^{(R),\star} + \frac{1}{K}\mathsf{D}^\star w_k^{(r)}\right)\right\| =& \left\|\sum_{r'=r+1}^{R} \overline{\mathsf{D}}_{r'}^\star \xi_{r'}\right\| \\ \leqslant& \sum_{r'=r+1}^{R} q_{e,b}^{R-r'} \|\xi_{r'}\| \\ \leqslant& \sum_{r'=r+1}^{R} q_{e,b}^{R-r'} \frac{\alpha(1-\alpha\nu)^{r'-r-1}}{K^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha q_{e,b}^{R-r-1}}{K^2} \sum_{r'=r+1}^{R} \left((1-\alpha\nu)/q_{e,b}\right)^{r'-r-1} \\
&= \frac{\alpha\left(q_{e,b}^{R-r} - (1-\alpha\nu)^{R-r}\right)}{K^2(q_{e,b} - (1-\alpha\nu))}.
\end{aligned} \tag{83}
$$

This term is less than $\theta/(6B)$ due to (21), which concludes this step. Hence, it remains to show the claim that $B_j \leqslant \frac{1-q_{e,b}^2}{\alpha} - \nu$. To show this, first note that

$$
B_0 = \frac{1}{K^2} \overset{(a)}{\leqslant} \frac{1-q_{e,b}^2}{\alpha} - \nu \leqslant \frac{1-q_{e,b}^2}{\alpha},
$$

where (a) holds by Assumption C.3. Now, recursively, if $B_{j-1} \leqslant \frac{1-q_{e,b}^2}{\alpha}$, then

$$
B_j = \alpha B_{j-1}^2 + q_{e,b}^2 B_{j-1} \leqslant B_{j-1}.
$$

Hence the sequence $\{B_j\}$ is non-increasing and

$$
B_j \leqslant B_0 = \frac{1}{K^2} \leqslant \frac{1-q_{e,b}^2}{\alpha} - \nu.
$$

This completes the proof of this part.

Bounding (70). Note that $\|\mathsf{D}^\star w_k^{(r)}\| \leqslant q_{e,b}^{R-r}$, due to the fact that spectral norm of all $\mathsf{D}_{r'}^*$ are bounded by $q_{e,b}$ due to Assumption 2. Then, for every $w_k^{(r)}$,

$$
\begin{aligned}
\mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle x, \mathsf{D}^\star w_k^{(r)}\rangle - \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)}\rangle\right| > \theta/6\right\}}\right] &= \mathbb{P}_{\mathsf{A}}\left(\frac{1}{K}\left|\langle x, \mathsf{D}^\star w_k^{(r)}\rangle - \langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)}\rangle\right| > \theta/6\right) \\
&\leqslant 4e^{-\frac{m}{7}\left(\frac{K\theta}{6Bq_{e,b}^{R-r}}\right)^2},
\end{aligned} \tag{84}
$$

where the last step is due to (Grønlund et al., 2020, Lemma 8, part 2.).

Bounding (71). Fix some $c_1, c_2 > 0$. Now, to bound this, we have

$$
\begin{aligned}
&\mathbb{E}_{p_{M|w_k^{(r)}}} \mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M\rangle\right| > \theta/6\right\}}\right] \\
&\quad \leqslant \mathbb{P}_{\mathsf{A}}(\|\mathsf{A}x\| > c_1 B) + \mathbb{P}_{\mathsf{A}}\left(\|\mathsf{AD}^\star w_k^{(r)}\| > c_2\right) \\
&\qquad + \mathbb{E}_{p_{M|w_k^{(r)}}} \mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M\rangle\right| > \theta/6\right\}} \middle| \|\mathsf{A}x\| \leqslant c_1 B, \|\mathsf{AD}^\star w_k^{(r)}\| \leqslant c_2\right] \\
&\quad = \mathbb{P}_{\mathsf{A}}(\|\mathsf{A}x\| > c_1 B) + \mathbb{P}_{\mathsf{A}}\left(\|\mathsf{AD}^\star w_k^{(r)}\| > c_2\right) \\
&\qquad + \mathbb{E}_{M\sim \mathrm{Unif}\left(\mathcal{B}_m(\mathsf{AD}^\star w_k^{(r)}, \nu)\right)} \mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}\left|\langle \mathsf{A}x, \mathsf{AD}^\star w_k^{(r)} - M\rangle\right| > \theta/6\right\}} \middle| \|\mathsf{A}x\| \leqslant c_1 B, \|\mathsf{AD}^\star w_k^{(r)}\| \leqslant c_2\right] \\
&\quad = \mathbb{P}_{\mathsf{A}}(\|\mathsf{A}x\| > c_1 B) + \mathbb{P}_{\mathsf{A}}\left(\|\mathsf{AD}^\star w_k^{(r)}\| > c_2\right) + \mathbb{E}_{U\sim \mathrm{Unif}(\mathcal{B}_m(\nu))} \mathbb{E}_{\mathsf{A}}\left[\mathbb{1}_{\left\{\frac{1}{K}|\langle \mathsf{A}x, U\rangle| > \theta/6\right\}} \middle| \|\mathsf{A}x\| \leqslant c_1 B\right] \\
&\quad \leqslant e^{-0.21m(c_1^2-1)} + e^{-0.21m(c_2^2 q_{e,b}^{2(r-R)}-1)} + \frac{2m\nu^m}{\sqrt{\pi}} e^{-\frac{(m+1)}{2}\left(\frac{K\theta}{6c_1\nu B}\right)^2},
\end{aligned} \tag{85}
$$

where the last step follows from (Grønlund et al., 2020, Lemma 8, part 1.) and (Sefidgaran et al., 2022a, Proof of Lemma 3), and since $\|\mathsf{D}^\star w_k^{(r)}\| \leqslant q_{e,b}^{R-r}$.

This completes the proof. □

G.5. Proof of Theorem B.1

Proof. We have

$$\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big] = \frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}_{S_k^{(r)},\overline{W}^{(r-1)},W_k^{[r:R]},\overline{W}^{(R)}\sim P_{S_k^{(r)}}P_{k,r}}\Big[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})\Big], \tag{86}$$

where

$$P_{k,r} := P_{\overline{W}^{(r-1)},W_k^{[r:R]}|S_k^{(r)}}P_{\overline{W}^{(R)}|\overline{W}^{(r-1)},W_k^{[r:R]}}.$$

Let $q_{\overline{W}^{(r-1)},W_k^{[r:R]}}$ be the marginal distribution of $(\overline{W}^{(r-1)},W_k^{[r:R]})$ under $P_{S_k^{(r)}}P_{\overline{W}^{(r-1)},W_k^{[r:R]}|S_k^{(r)}}$, and denote similarly

$$(Pq)_{k,r} := q_{\overline{W}^{(r-1)},W_k^{[r:R]}}P_{\overline{W}^{(R)}|\overline{W}^{(r-1)},W_k^{[r:R]}}.$$

Then for any $\lambda_{k,r}>0$,

$$\begin{aligned}
&\lambda_{k,r}\mathbb{E}_{S_k^{(r)},\overline{W}^{(r-1)},W_k^{[r:R]},\overline{W}^{(R)}\sim P_{S_k^{(r)}}P_{k,r}}\Big[\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})\Big]\\
&\overset{(a)}{\leqslant} D_{KL}\Big(P_{S_k^{(r)}}P_{k,r}\|P_{S_k^{(r)}}(Pq)_{k,r}\Big)+\log\mathbb{E}_{P_{S_k^{(r)}}(Pq)_{k,r}}\Big[e^{\lambda_{k,r}\,\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})}\Big]\\
&= I(S_k^{(r)};\overline{W}^{(r-1)},W_k^{[r:R]})+\log\mathbb{E}_{(Pq)_{k,r}}\mathbb{E}_{P_{S_k^{(r)}}}\Big[e^{\lambda_{k,r}\,\mathrm{gen}(S_k^{(r)},\overline{W}^{(R)})}\Big]\\
&\overset{(b)}{\leqslant} I(S_k^{(r)};\overline{W}^{(r-1)},W_k^{[r:R]})+\frac{\lambda_{k,r}^2\sigma^2}{2n_{k,r}},
\end{aligned} \tag{87}$$

where (a) is deduced using Donsker-Varadhan's inequality and (b) using the fact that for any $w\in\mathcal{W}$, $\mathrm{gen}(S_k^{(r)},w)$ is $\sigma/\sqrt{n_{k,r}}$-subgaussian.

Combining (86) and (87) we get

$$\begin{aligned}
\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\Big[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\Big] &\leqslant \frac{1}{KR}\sum_{k\in[K],r\in[R]}\Big(I(S_k^{(r)};\overline{W}^{(r-1)},W_k^{[r:R]})/\lambda_{k,r}+\frac{\lambda_{k,r}\sigma^2}{2n_{k,r}}\Big)\\
&\overset{(a)}{\leqslant}\sqrt{\frac{2\sigma^2}{KR}\sum_{k\in[K],r\in[R]}\frac{1}{n_{k,r}}I(S_k^{(r)};\overline{W}^{(r-1)},W_k^{[r:R]})},
\end{aligned} \tag{88}$$

where the last step is established by letting $\lambda_{k,r}:=n_{k,r}\lambda$ and

$$\lambda := \sqrt{\frac{2}{\sigma^2 KR}\sum_{k\in[K],r\in[R]}\frac{1}{n_{k,r}}I(S_k^{(r)};\overline{W}^{(r-1)},W_k^{[r:R]})}.$$

This completes the proof. □

G.6. Proof of Theorem D.1

Proof. We start by a lemma proved in Appendix G.11.

Lemma G.4. *For any $\delta>0$,*

$$\log\mathbb{P}\Big(\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})>\Delta\Big)\leqslant\max\Bigg(\log(\delta),\tag{89}$$

$$\sup_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}\in\mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}}\inf_{\Big\{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\in\mathcal{Q}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\Big\}_{k\in[K],r\in[R]}}\inf_{\lambda\geqslant 0}\Big\{-D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}\|P_{\mathbf{S},\mathbf{U},\mathbf{W}})-\frac{\lambda}{KR}\sum_{r,k}C_{k,r}(\nu,p)\Big\}\Bigg)$$

where

$$C_{k,r}(\nu,p) := \Delta - \epsilon - \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right],$$

with the notation

$$(\nu p)_{k,r} := \nu_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)} | S_{[K]}^{[r]} \setminus S_k^{(r)}, U_{[K]}^{[r]} \setminus U_k^{(r)}} \, p_{\hat{W}_k^{(r)} | S_k^{(r)}, U_k^{(r)}, \overline{W}^{(r-1)}} \, \nu_{\hat{\overline{W}}^{(R)} | \hat{W}_k^{(r)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, U_{[K]}^{[r+1:R]}},$$

and where $\mathcal{Q}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})$ *contains all distributions* $\left\{p_{\hat{W}_{k,r}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\right\}_{k\in[K],r\in[R]}$ *that satisfy*

$$\mathbb{E}_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right] - \frac{1}{KR}\sum_{k\in[K],r\in[R]} \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \leqslant \epsilon. \tag{90}$$

Next, consider the set $\tilde{\mathcal{Q}}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})$ that contains all distributions $\left\{p_{\hat{W}_{k,r}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\right\}_{k\in[K],r\in[R]}$ such that for any $k \in [k]$ and $r \in [R]$,

$$\mathbb{E}_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}}\left[\mathrm{gen}(S_k^{(r)}, \overline{W}^{(R)})\right] - \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \leqslant \epsilon_{k,r}, \tag{91}$$

where $\frac{1}{KR}\sum_{k\in[K],r\in[R]} \epsilon_{k,r} = \epsilon$. Trivially, $\tilde{\mathcal{Q}}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}) \subseteq \mathcal{Q}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})$, and hence, for a given $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}$, it suffices to bound

$$\inf_{\left\{p_{\hat{W}_k^{(r)}|S_k^{(r)},\overline{W}^{(r-1)}} \in \tilde{\mathcal{Q}}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\right\}_{k\in[K],r\in[R]}} \inf_{\lambda\geqslant 0}\left\{ - D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) - \frac{\lambda}{KR}\sum_{r,k} C_{k,r}(\nu,p)\right\}. \tag{92}$$

Fix a $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}$. Let $q_{\hat{W}_k^{(r)}|U_k^{(r)},\overline{W}^{(r-1)}}$ be the marginal conditional distribution of $\hat{W}_k^{(r)}$ given $\overline{W}^{(r-1)}$ and $U_k^{(r)}$ under $\nu_{S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}} p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}$, and denote similarly

$$(\nu q)_{k,r} := \nu_{\overline{W}^{(r-1)}, W_{[K]\setminus k}^{(r)} | S_{[K]}^{[r]} \setminus S_k^{(r)}, U_{[K]}^{[r]} \setminus U_k^{(r)}} \, q_{\hat{W}_k^{(r)} | U_k^{(r)}, \overline{W}^{(r-1)}} \, \nu_{\hat{\overline{W}}^{(R)} | \hat{W}_k^{(r)}, W_{[K]\setminus k}^{(r)}, S_{[K]}^{[r+1:R]}, U_{[K]}^{[r+1:R]}},$$

Note that

$$\begin{aligned}
&\lambda \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right] \\
&\overset{(a)}{\leqslant} \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}}\left[D_{KL}((\nu p)_{k,r} \| (\nu q)_{k,r}) + \log \mathbb{E}_{(\nu q)_{k,r}}\left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})}\right]\right] \\
&\overset{(b)}{\leqslant} \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}}[D_{KL}((\nu p)_{k,r} \| (\nu q)_{k,r})] + D_{KL}(\nu_{\mathbf{S},\mathbf{U}} \| P_{\mathbf{S},\mathbf{U}}) + \log \mathbb{E}_{P_{\mathbf{S},\mathbf{U}}(\nu q)_{k,r}}\left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})}\right] \\
&\overset{(c)}{\leqslant} D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) + \mathbb{E}_{\nu_{S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}}\left[D_{KL}\left(p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}} \| q_{\hat{W}_k^{(r)}|U_k^{(r)},\overline{W}^{(r-1)}}\right)\right] \\
&\quad + \log \mathbb{E}_{P_{\mathbf{S},\mathbf{U}}(\nu q)_{k,r}}\left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})}\right] \\
&= D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) + I(S_k^{(r)}; \hat{W}_k^{(r)} | U_k^{(r)}, \overline{W}^{(r-1)}) + \log \mathbb{E}_{P_{\mathbf{S},\mathbf{U}}(\nu q)_{k,r}} \mathbb{E}_{P_{S_k^{(r)}}}\left[e^{\lambda\, \mathrm{gen}(S_k^{(r)}, \hat{\overline{W}}^{(R)})}\right] \\
&\overset{(d)}{\leqslant} D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) + I(S_k^{(r)}; \hat{W}_k^{(r)} | U_k^{(r)}, \overline{W}^{(r-1)}) + \frac{\lambda^2\sigma^2}{2n/R},
\end{aligned} \tag{93}$$

where (a) and (b) are derived using Donsker-Varadhan's inequality, (c) by definitions of $(\nu p)_{k,r}$ and $(\nu q)_{k,r}$, and (d) using the fact that for any $w \in \mathcal{W}$, $\mathrm{gen}(S_k^{(r)}, w)$ is $\sigma/\sqrt{n/R}$-subgaussian.

Hence,

$$\begin{aligned}
&\inf_{\left\{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\in\mathcal{Q}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\right\}_{k\in[K],r\in[R]}} \inf_{\lambda\geqslant 0}\left\{-D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}\|P_{\mathbf{S},\mathbf{U},\mathbf{W}})-\frac{\lambda}{KR}\sum_{r,k}C_{k,r}(\nu,p)\right\}\\
\leqslant&\inf_{\left\{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\in\tilde{\mathcal{Q}}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\right\}_{k\in[K],r\in[R]}} \inf_{\lambda\geqslant 0}\left\{-D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}\|P_{\mathbf{S},\mathbf{U},\mathbf{W}})-\frac{\lambda}{KR}\sum_{r,k}C_{k,r}(\nu,p)\right\}\\
\leqslant&\inf_{\left\{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}}\in\tilde{\mathcal{Q}}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\right\}_{k\in[K],r\in[R]}} \inf_{\lambda\geqslant 0}\left\{-\lambda(\Delta-\epsilon)+\frac{1}{KR}\sum_{r,k}I(S_k^{(r)};\hat{W}_k^{(r)}|U_k^{(r)},\overline{W}^{(r-1)})+\frac{\lambda^2\sigma^2}{2n/R}\right\}\\
=&\inf_{\lambda\geqslant 0}\left\{-\lambda(\Delta-\epsilon)+\frac{1}{KR}\sum_{r,k}\mathfrak{RD}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}},k,r,\epsilon_{k,r})+\frac{\lambda^2\sigma^2}{2n/R}\right\}\\
\leqslant&\log(\delta),
\end{aligned}\tag{94}$$

where the last step is established by letting

$$\lambda:=\sqrt{\frac{(2n/R)}{\sigma^2}\left(\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathfrak{RD}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}},k,r,\epsilon_{k,r})+\log(1/\delta)\right)}.$$

This completes the proof. □

G.7. Proof of Proposition D.2

Proof. Recall that

$$\mathfrak{V}_{k,r}=\left(V_k^{(r)},\overline{W}^{(r-1)},W_{[K]\setminus k}^{(r)},S_{[K]}^{[r+1:R]},V_{[K]}^{[r+1:R]}\right).$$

Consider any set of distributions $\left\{p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}\right\}_{k\in[K],r\in[R]}$ that satisfy the distortion criterion (29) for any $k\in[K]$ and $r\in[R]$. Then,

$$\begin{aligned}
\mathbb{E}_{P_{\mathbf{S},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]=&\mathbb{E}_{P_{\mathbf{S},\mathbf{V},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]\\
\leqslant&\frac{1}{KR}\sum_{k\in[K],r\in[R]}\left(\mathbb{E}_{P_{S_k^{(r)},\mathfrak{V}_{k,r}}p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}}\left[\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\right]+\epsilon_{k,r}\right)\\
\leqslant&\frac{1}{KR}\sum_{k\in[K],r\in[R]}\mathbb{E}_{P_{S_k^{(r)},\mathfrak{V}_{k,r}}p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}}\left[\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\right]+\epsilon.
\end{aligned}\tag{95}$$

Let $q_{\hat{W}_k^{(R)}|\mathfrak{V}_{k,r}}$ be the marginal conditional distribution of $\hat{W}_k^{(R)}$ given $\mathfrak{V}_{k,r}$ under $P_{S_k^{(r)}}p_{\hat{W}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}$. Then,

$$\begin{aligned}
&\lambda\mathbb{E}_{P_{S_k^{(r)},\mathfrak{V}_{k,r}}p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}}\left[\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\right]\\
\overset{(a)}{\leqslant}&D_{KL}\left(P_{S_k^{(r)},\mathfrak{V}_{k,r}}p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}\middle\|P_{S_k^{(r)},\mathfrak{V}_{k,r}}q_{\hat{\overline{W}}_k^{(R)}|\mathfrak{V}_{k,r}}\right)+\log\mathbb{E}_{P_{S_k^{(r)},\mathfrak{V}_{k,r}}q_{\hat{\overline{W}}_k^{(R)}|\mathfrak{V}_{k,r}}}\left[e^{\lambda\,\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})}\right]\\
=&I(S_k^{(r)};\hat{W}_k^{(R)}|\mathfrak{V}_{k,r})+\log\mathbb{E}_{P_{\mathfrak{V}_{k,r}}q_{\hat{\overline{W}}_k^{(R)}|\mathfrak{V}_{k,r}}}\mathbb{E}_{P_{S_k^{(r)}}}\left[e^{\lambda\,\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})}\right]\\
\overset{(b)}{\leqslant}&I(S_k^{(r)};\hat{W}_k^{(R)}|\mathfrak{V}_{k,r})+\frac{\lambda^2\sigma^2}{2n/R},
\end{aligned}\tag{96}$$

where (a) is deduced using Donsker-Varadhan's inequality and (b) using the fact that for any $w \in \mathcal{W}$, $\mathrm{gen}(S_k^{(r)}, w)$ is $\sigma/\sqrt{n/R}$-subgaussian,

Combining (95) and (96), and taking the infimum over all admissible $\left\{p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}\right\}_{k\in[K],r\in[R]}$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S},\mathbf{W}\sim P_{\mathbf{S},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right] \\
&\leqslant \frac{1}{KR}\sum_{k\in[K],r\in[R]} \inf_{p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}} \left\{\mathbb{E}_{P_{S_k^{(r)},\mathfrak{V}_{k,r}} p_{\hat{\overline{W}}_k^{(R)}|S_k^{(r)},\mathfrak{V}_{k,r}}}\left[\mathrm{gen}(S_k^{(r)},\hat{\overline{W}}^{(R)})\right]\right\} + \epsilon \\
&\leqslant \frac{1}{KR}\sum_{k\in[K],r\in[R]} \mathfrak{RD}^{\star}(P_{\mathbf{S},\mathbf{W}},k,r,\epsilon_{k,r})/\lambda + \frac{\lambda\sigma^2}{2n/R} + \epsilon \\
&\overset{(a)}{\leqslant} \sqrt{\frac{2\sigma^2\sum_{k\in[K],r\in[R]}\mathfrak{RD}^{\star}(P_{\mathbf{S},\mathbf{W}},k,r,\epsilon_{k,r})}{nK}} + \epsilon,
\end{aligned} \tag{97}$$

where the last step is established by letting

$$\lambda := \sqrt{\frac{2n/R}{\sigma^2 KR}\sum_{k\in[K],r\in[R]}\mathfrak{RD}^{\star}(P_{\mathbf{S},\mathbf{W}},k,r,\epsilon_{k,r})}\,.$$

This completes the proof. □

G.8. Proof of Lemma G.1

Proof. This step is similar to (Sefidgaran et al., 2022b, Lemma 24) and (Sefidgaran & Zaidi, 2023, Lemma 12). Denote $\mathcal{B} := \left\{\mathbf{s}\in\mathrm{supp}(P_{\mathbf{S}})\colon \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\left[\mathrm{gen}(\mathbf{s},\overline{W}^{(R)})\right]^2 > \Delta(\mathbf{s})\right\}$. If $P_{\mathbf{S}}(\mathcal{B}) = 0$, then the lemma is proved. Assume then $P_{\mathbf{S}}(\mathcal{B}) > 0$. Consider the distribution $\nu_{\mathbf{S}}$ such that for any $\mathbf{s}\in\mathcal{B}$, $\nu_{\mathbf{S}}(\mathbf{s}) := P_{\mathbf{S}}(\mathbf{s})/P_{\mathbf{S}}(\mathcal{B})$, and otherwise $\nu_{\mathbf{S}}(\mathbf{s}) := 0$.

If $\nu_{\mathbf{S}} \notin \mathcal{G}_{\mathbf{S}}^{\delta}$, then it means that $D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) \geqslant \log(1/\delta)$. Hence,

$$\log(1/\delta) \leqslant D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) = -\log(P_{\mathbf{S}}(\mathcal{B})) = -\log\mathbb{P}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\left[\mathrm{gen}(\mathbf{s},\overline{W}^{(R)})\right]^2 > \Delta(\mathbf{s})\right),$$

Thus, $\log\mathbb{P}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\left[\mathrm{gen}(\mathbf{s},\overline{W}^{(R)})\right]^2 > \Delta(\mathbf{s})\right) \leqslant \log(\delta)$. This completes the proof of the lemma.

Otherwise, suppose that $\nu_{\mathbf{S}} \in \mathcal{G}_{\mathbf{S}}^{\delta}$. Then, for this distribution

$$\mathbb{E}_{\nu_{\mathbf{S}}}\left[\Delta(\mathbf{S}) - \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]^2\right] \overset{(a)}{\leqslant} 0,$$

where (a) holds by the way $\nu_{\mathbf{S}}$ is constructed. Therefore, for this $\nu_{\mathbf{S}}$, since $\lambda \geqslant 0$,

$$\begin{aligned}
-D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) - \lambda\mathbb{E}_{\nu_{\mathbf{S}}}&\left[\Delta(\mathbf{S}) - \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{S}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]^2\right] \\
&\geqslant -D_{KL}(\nu_{\mathbf{S}}\|P_{\mathbf{S}}) \\
&= \log(P_{\mathbf{S}}(\mathcal{B})) \\
&= \log\mathbb{P}\left(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\left[\mathrm{gen}(\mathbf{S},\overline{W}^{(R)})\right]^2 > \Delta(\mathbf{S})\right).
\end{aligned}$$

This proves the lemma. □

G.9. Proof of Lemma G.2

Proof. This step is similar to (Sefidgaran et al., 2022b, Lemma 24) and (Sefidgaran & Zaidi, 2023, Lemma 12).

Denote $f(\mathbf{s}, \overline{w}^{(R)}) := \frac{1}{KR}\sum_{k\in[K], r\in[R]} \mathrm{gen}(s_k^{(r)}, \overline{w}^{(R)})^2$ and define the set

$$\mathcal{B} := \Big\{(\mathbf{s}, \mathbf{w}) \in \mathrm{supp}(P_{\mathbf{S},\mathbf{W}})\colon f(\mathbf{s}, \overline{w}^{(R)}) > \Delta(\mathbf{s}, \mathbf{w})\Big\}.$$

If $P_{\mathbf{S},\mathbf{W}}(\mathcal{B}) = 0$, then the lemma is proved. Assume then $P_{\mathbf{S},\mathbf{W}}(\mathcal{B}) > 0$. Consider the distribution $\nu_{\mathbf{S},\mathbf{W}}$ such that for any $(\mathbf{s}, \mathbf{w}) \in \mathcal{B}$, $\nu_{\mathbf{S},\mathbf{W}}(\mathbf{s}, \mathbf{w}) := P_{\mathbf{S},\mathbf{W}}(\mathbf{s}, \mathbf{w})/P_{\mathbf{S},\mathbf{W}}(\mathcal{B})$, and otherwise $\nu_{\mathbf{S},\mathbf{W}}(\mathbf{s}, \mathbf{w}) := 0$.

If $\nu_{\mathbf{S},\mathbf{W}} \notin \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{W}}$, then it means that $D_{KL}(\nu_{\mathbf{S},\mathbf{W}} \| P_{\mathbf{S},\mathbf{W}}) \geqslant \log(1/\delta)$. Hence,

$$\log(1/\delta) \leqslant D_{KL}(\nu_{\mathbf{S},\mathbf{W}} \| P_{\mathbf{S},\mathbf{W}}) = -\log(P_{\mathbf{S},\mathbf{W}}(\mathcal{B})) = -\log \mathbb{P}\Big(f(\mathbf{s}, \overline{w}^{(R)}) > \Delta(\mathbf{s}, \mathbf{w})\Big),$$

Thus, $\log \mathbb{P}\big(f(\mathbf{s}, \overline{w}^{(R)}) > \Delta(\mathbf{s}, \mathbf{w})\big) \leqslant \log(\delta)$. This completes the proof of the lemma.

Otherwise, suppose that $\nu_{\mathbf{S},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{W}}$. Then, for this distribution, we have

$$\mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\Big[\Delta(\mathbf{S}, \mathbf{W}) - f(\mathbf{S}, \overline{W}^{(R)})\Big] \overset{(a)}{\leqslant} 0,$$

where (a) holds due to the way $\nu_{\mathbf{S},\mathbf{W}}$ is constructed. Therefore, it is optimal to let $\lambda = 0$ in (b) for this distribution. Thus, the infimum over $\lambda > 0$ of

$$-D_{KL}(\nu_{\mathbf{S},\mathbf{W}} \| P_{\mathbf{S},\mathbf{W}}) - \lambda \mathbb{E}_{\nu_{\mathbf{S},\mathbf{W}}}\Big[\Delta(\mathbf{S}, \mathbf{W}) - f(\mathbf{S}, \overline{W}^{(R)})\Big],$$

is equal to

$$-D_{KL}(\nu_{\mathbf{S},\mathbf{W}} \| P_{\mathbf{S},\mathbf{W}}) = \log(P_{\mathbf{S},\mathbf{W}}(\mathcal{B})) = \log \mathbb{P}\Big(f(\mathbf{S}, \overline{W}^{(R)}) > \Delta(\mathbf{S}, \mathbf{W})\Big).$$

This completes the proof of the lemma. □

G.10. Proof of Lemma G.3

Proof. This step is similar to (Sefidgaran et al., 2022b, Lemma 24) and (Sefidgaran & Zaidi, 2023, Lemma 12). Denote $\mathcal{B} := \Big\{\mathbf{s} \in \mathrm{supp}(P_{\mathbf{S}})\colon \mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\big]^2 > \Delta(\mathbf{s})\Big\}$. If $P_{\mathbf{S}}(\mathcal{B}) = 0$, then the lemma is proved. Assume then $P_{\mathbf{S}}(\mathcal{B}) > 0$. Consider the distribution $\nu_{\mathbf{S}}$ such that for any $\mathbf{s} \in \mathcal{B}$, $\nu_{\mathbf{S}}(\mathbf{s}) := P_{\mathbf{S}}(\mathbf{s})/P_{\mathbf{S}}(\mathcal{B})$, and otherwise $\nu_{\mathbf{S}}(\mathbf{s}) := 0$.

If $\nu_{\mathbf{S}} \notin \mathcal{G}^{\delta}_{\mathbf{S}}$, then it means that $D_{KL}(\nu_{\mathbf{S}} \| P_{\mathbf{S}}) \geqslant \log(1/\delta)$. Hence,

$$\log(1/\delta) \leqslant D_{KL}(\nu_{\mathbf{S}} \| P_{\mathbf{S}}) = -\log(P_{\mathbf{S}}(\mathcal{B})) = -\log \mathbb{P}\Bigg(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\Big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\Big]^2 > \Delta(\mathbf{s})\Bigg)$$

Thus, $\log \mathbb{P}\Bigg(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\big]^2 > \Delta(\mathbf{s})\Bigg), \leqslant \log(\delta)$. This completes the proof of the lemma.

Otherwise, suppose that $\nu_{\mathbf{S}} \in \mathcal{G}^{\delta}_{\mathbf{S}}$. Then, for this distribution

$$\begin{aligned}
\mathbb{E}_{\nu_{\mathbf{S}}}\Big[\Delta(\mathbf{S}) - \epsilon - \tilde{\Delta}(\mathbf{S})^2\Big] &\overset{(a)}{\leqslant} \mathbb{E}_{\nu_{\mathbf{S}}}\Bigg[\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\Big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\Big]^2 - \epsilon - \tilde{\Delta}(\mathbf{S})^2\Bigg] \\
&\overset{(b)}{\leqslant} \mathbb{E}_{\nu_{\mathbf{S}}}\Bigg[2\Big|\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\Big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\Big] - \tilde{\Delta}(\mathbf{S})\Big| - \epsilon\Bigg] \\
&\overset{(c)}{\leqslant} 0,
\end{aligned}$$

where (a) holds by the way $\nu_{\mathbf{S}}$ is constructed, (b) holds since the loss function is bounded by one, and (c) due to (45).

Therefore, for this $\nu_{\mathbf{S}}$, since $\lambda \geqslant 0$,

$$\begin{aligned}
-D_{KL}(\nu_{\mathbf{S}} \| P_{\mathbf{S}}) - \lambda \mathbb{E}_{\nu_{\mathbf{S}}}\Big[\Delta(\mathbf{S}) - \epsilon - \tilde{\Delta}(\mathbf{S})^2\Big] &\geqslant -D_{KL}(\nu_{\mathbf{S}} \| P_{\mathbf{S}}) \\
&= \log(P_{\mathbf{S}}(\mathcal{B})) \\
&= \log \mathbb{P}\Bigg(\mathbb{E}_{\mathbf{W}\sim P_{\mathbf{W}|\mathbf{s}}}\Big[\mathrm{gen}(\mathbf{s}, \overline{W}^{(R)})\Big]^2 > \Delta(\mathbf{s})\Bigg).
\end{aligned}$$

This proves the lemma. □

G.11. Proof of Lemma G.4

Proof. Denote $\mathcal{B} := \big\{(\mathbf{s}, \mathbf{u}, \mathbf{w}) \in \mathrm{supp}(P_{\mathbf{S},\mathbf{U},\mathbf{W}})\colon \mathrm{gen}(\mathbf{s}, \overline{w}^{(R)}) > \Delta\big\}$. If $P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathcal{B}) = 0$, then the lemma is proved. Assume then $P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathcal{B}) > 0$. Consider the distribution $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}$ such that for any $(\mathbf{s}, \mathbf{u}, \mathbf{w}) \in \mathcal{B}$, $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathbf{s}, \mathbf{u}, \mathbf{w}) := P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathbf{s}, \mathbf{u}, \mathbf{w})/P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathcal{B})$, and otherwise $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathbf{s}, \mathbf{u}, \mathbf{w}) := 0$.

If $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \notin \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}$, then it means that $D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) \geqslant \log(1/\delta)$. Hence,

$$\log(1/\delta) \leqslant D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) = -\log(P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathcal{B})) = -\log \mathbb{P}\Big(\mathrm{gen}(\mathbf{s}, \overline{w}^{(R)}) > \Delta\Big),$$

Thus, $\log \mathbb{P}\big(\mathrm{gen}(\mathbf{s}, \overline{w}^{(R)}) > \Delta\big) \leqslant \log(\delta)$. This completes the proof of the lemma.

Otherwise, suppose that $\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \in \mathcal{G}^{\delta}_{\mathbf{S},\mathbf{U},\mathbf{W}}$. Then, for this distribution and any set of $\Big\{p_{\hat{W}_k^{(r)}|S_k^{(r)},U_k^{(r)},\overline{W}^{(r-1)}} \in \mathcal{Q}_{k,r}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}})\Big\}_{k\in[K],r\in[R]}$, we have

$$\begin{aligned}
&\frac{1}{KR}\sum_{k,r} C_{k,r}(\nu, p) \\
&= \Delta - \epsilon - \frac{1}{KR}\sum_{k\in[K],r\in[R]} \left(\mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}_{k,r}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right]\right) \\
&\overset{(a)}{\leqslant} \mathbb{E}_{\nu_{\mathbf{S},\mathbf{U},\mathbf{W}}}\left[\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)})\right] - \frac{1}{KR}\sum_{k\in[K],r\in[R]} \left(\mathbb{E}_{\nu_{\mathbf{S},\mathbf{U}}(\nu p)_{k,r}}\left[\mathrm{gen}_{k,r}(S_k^{(r)}, \hat{\overline{W}}^{(R)})\right]\right) - \epsilon \\
&\overset{(b)}{\leqslant} 0,
\end{aligned}$$

where (a) holds due to the way $\nu_{\mathbf{S},\mathbf{W}}$ is constructed and (b) by distortion criterion (90).

Therefore, it is optimal to let $\lambda = 0$ in (89)for this distribution. Thus, the infimum over $\lambda > 0$ of

$$-D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) - \frac{\lambda}{KR}\sum_{k,r} C_{k,r}(\nu, p),$$

is equal to

$$-D_{KL}(\nu_{\mathbf{S},\mathbf{U},\mathbf{W}} \| P_{\mathbf{S},\mathbf{U},\mathbf{W}}) = \log(P_{\mathbf{S},\mathbf{U},\mathbf{W}}(\mathcal{B})) = \log \mathbb{P}\Big(\mathrm{gen}(\mathbf{S}, \overline{W}^{(R)}) > \Delta\Big).$$

This completes the proof. □