entropy

MDPI

Article

Improved Information-Theoretic Generalization Bounds for Distributed, Federated, and Iterative Learning [†]

Leighton Pate Barnes [1,*], Alex Dytso [2] and Harold Vincent Poor [1]

[1] Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA
[2] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA
[*] Correspondence: leightonbarnes@gmail.com

[†] This paper is an extended version of our paper published in Proceedings of the 2022 IEEE International Symposium on Information Theory, Espoo, Finland, 26 June–1 July 2022.

Abstract: We consider information-theoretic bounds on the expected generalization error for statistical learning problems in a network setting. In this setting, there are K nodes, each with its own independent dataset, and the models from the K nodes have to be aggregated into a final centralized model. We consider both simple averaging of the models as well as more complicated multi-round algorithms. We give upper bounds on the expected generalization error for a variety of problems, such as those with Bregman divergence or Lipschitz continuous losses, that demonstrate an improved dependence of $1/K$ on the number of nodes. These "per node" bounds are in terms of the mutual information between the training dataset and the trained weights at each node and are therefore useful in describing the generalization properties inherent to having communication or privacy constraints at each node.

Keywords: generalization error; information-theoretic bounds; distribution and federated learning

check for updates

Citation: Barnes, L.P.; Dytso, A.; Poor, H.V. Improved Information-Theoretic Generalization Bounds for Distributed, Federated, and Iterative Learning. *Entropy* **2022**, *24*, 1178. https://doi.org/10.3390/e24091178

Academic Editors: Chao Tian and Lizhong Zheng

Received: 25 July 2022
Accepted: 20 August 2022
Published: 24 August 2022

1. Introduction

A key feature of machine learning systems is their ability to generalize new and unknown data. Such a system is trained on a particular set of data but must then perform well even on new data points that have not previously been considered. This ability, deemed generalization, can be formulated in the language of statistical learning theory by considering the generalization error of an algorithm (i.e., the difference between the population risk of a model trained on a particular dataset and the empirical risk for the same model and dataset). We say that a model generalizes well if it has a small generalization error, and because models are often trained by minimizing empirical risk or some regularized version of it, a small generalization error also implies a small population risk, which is the average loss over new samples taken randomly from the population. It is therefore of interest to find an upper bound on the generalization error and understand which quantities control it so that we can quantify the generalization properties of a machine learning system and offer guarantees about its performance.

In recent years, it has been shown that information-theoretic measures such as mutual information can be used for generalization error bounds under the assumption of the tail of the distribution of the loss function [1–4]. In particular, when the loss function is sub-Gaussian, the expected generalization error can scale at most with the square root of the mutual information between the training dataset and the model weights [2]. Such bounds offer an intuitive explanation for generalization and overfitting: if an algorithm uses only limited information from its training data, then this will bound the expected generalization error and prevent overfitting. Conversely, if an algorithm uses all of the information from its training set, in the sense that the model is a deterministic function of the training set, then this mutual information can be infinite, and there is the possibility of overfitting.

Another modern focus of machine learning systems has been that of distributed and federated learning [5–7]. In these systems, data are generated and processed in a distributed network of machines. The main differences between the distributed and centralized settings are the information constraints imposed by the network. There has been considerable interest in understanding the impact of both communication constraints [8,9] and privacy constraints [10–13] on the performance of machine learning systems, as well as designing protocols that efficiently train the systems under these constraints.

Since both communication and local differential privacy constraints can be thought of as special cases of mutual information constraints, they should pair naturally with some form of information theoretic generalization bounding in order to induce control over the generalization error of the distributed machine learning system. The information constraints inherent to the network can themselves give rise to tighter bounds on generalization error and thus provide better guarantees against overfitting. Along these lines, in a recent work [14], a subset of the present authors introduced the framework of using information theoretic quantities for bounding both the expected generalization error and a measure of privacy leakage in distributed and federated learning systems. The generalization bounds in this work, however, are essentially the same as those obtained by thinking of the entire system, from the data at each node in the network to the final aggregated model, as a single, centralized algorithm. Any improved generalization guarantees from these bounds would remain implicit in the mutual information terms involved.

In this work, we develop improved bounds on the expected generalization error for distributed and federated learning systems. Instead of leaving the differences between these systems and their centralized counterparts implicit in the mutual information terms, we bring analysis of the structure of the systems directly to the bounds. By working with the contribution from each node separately, we are able to derive upper bounds on the expected generalization error that scale with the number of nodes K as $O\left(\frac{1}{K}\right)$ instead of $O\left(\frac{1}{\sqrt{K}}\right)$. This improvement is shown to be tight for certain examples, such as learning the mean of a Gaussian distribution with quadratic loss. We develop bounds that apply to distributed systems in which the submodels from K different nodes are averaged together, as well as bounds that apply to more complicated multi-round stochastic gradient descent (SGD) algorithms, such as in federated learning. For linear models with Bregman divergence losses, these "per node" bounds are in terms of the mutual information between the training dataset and the trained weights at each node and are therefore useful in describing the generalization properties inherent to having communication or privacy constraints at each node. For arbitrary nonlinear models that have Lipschitz continuous losses, the improved dependence of $O\left(\frac{1}{K}\right)$ can still be recovered but without a description in terms of mutual information. We demonstrate the improvements given by our bounds over the existing information theoretic generalization bounds via simulation of a distributed linear regression example. A preliminary conference version of this paper was presented in [15]. The present paper completes the work by including all of the missing proof details as well as providing new bounds for noisy SGD in Corollary 4.

Technical Preliminaries

Suppose we have independent and identically distributed (i.i.d.) data $Z_i \sim \pi$ for $i = 1, \ldots, n$, and let $S = (Z_1, \ldots, Z_n)$. Suppose further that $W = \mathcal{A}(S)$ is the output of a potentially stochastic algorithm. Let $\ell(W, Z)$ be a real-valued loss function and define

$$L(w) = \mathbb{E}_\pi[\ell(w, Z)]$$

to be the population risk for weights (or model) w. We similarly define

$$L_s(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$$

to be the empirical risk on dataset s for model w. The generalization error for dataset s is then

$$\Delta_{\mathcal{A}}(s) = L(\mathcal{A}(s)) - L_s(\mathcal{A}(s))$$

In addition, the expected generalization error is

$$\mathbb{E}_{S\sim\pi^n}[\Delta_{\mathcal{A}}(S)] = \mathbb{E}_{S\sim\pi^n}[L(\mathcal{A}(S)) - L_S(\mathcal{A}(S))] \tag{1}$$

where the expectation is also over any randomness in the algorithm. Below, we present some standard results for the expected generalization error that will be needed:

Theorem 1 (Leave-One-Out Expansion; Lemma 11 in [16])**.** *Let $S^{(i)} = (Z_1, \dots, Z_i', \dots, Z_n)$ be a version of S with Z_i replaced by an i.i.d. copy Z_i'. Denote $S' = (Z_1', \dots, Z_n')$. Then, we have*

$$\mathbb{E}_{S\sim\pi^n}[\Delta_{\mathcal{A}}(S)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,S'}[\ell(\mathcal{A}(S), Z_i') - \ell(\mathcal{A}(S^{(i)}), Z_i')] \,.$$

Proof. Observe that

$$\mathbb{E}_{S\sim\pi^n}[L(\mathcal{A}(S))] = \mathbb{E}_{S,S'}[\ell(\mathcal{A}(S), Z_i')] \tag{2}$$

for each i and that

$$\begin{aligned}\mathbb{E}_{S\sim\pi^n}[L_S(\mathcal{A}(S))] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S\sim\pi^n}[\ell(\mathcal{A}(S), Z_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,S'\sim\pi^n}\left[\ell(\mathcal{A}(S^{(i)}), Z_i')\right] \,. \end{aligned} \tag{3}$$

Putting Equations (2) and (3) together with (1) yields the result. □

In many of the results in this paper, we will use one of the two following assumptions:

Assumption 1. *The loss function $\ell(\widetilde{W}, \widetilde{Z})$ satisfies*

$$\log \mathbb{E}\left[\exp\left(\lambda\left(\ell(\widetilde{W}, \widetilde{Z}) - \mathbb{E}[\ell(\widetilde{W}, \widetilde{Z})]\right)\right)\right] \le \psi(-\lambda)$$

for $\lambda \in (b, 0]$, $\psi(0) = \psi'(0) = 0$, where $\widetilde{W}$ and $\widetilde{Z}$ are taken independently from the marginals for W and Z, respectively,

The next assumption is a special case of the previous one with $\psi(\lambda) = \frac{R^2\lambda^2}{2}$:

Assumption 2. *The loss function $\ell(\widetilde{W}, \widetilde{Z})$ is sub-Gaussian with parameter R^2 in the sense that*

$$\log \mathbb{E}\left[\exp\left(\lambda\left(\ell(\widetilde{W}, \widetilde{Z}) - \mathbb{E}[\ell(\widetilde{W}, \widetilde{Z})]\right)\right)\right] \le \frac{R^2\lambda^2}{2} \,.$$

Theorem 2 (Theorem 2 in [3]). *Under Assumption 1, we have*

$$\mathbb{E}_{S\sim\pi^n}[\Delta_{\mathcal{A}}(S)] \le \frac{1}{n}\sum_{i=1}^{n} \psi^{*-1}(I(W;Z_i))$$

where $\psi^{*-1}(y) = \inf_{\lambda\in[0,b)}\left(\frac{y+\psi(\lambda)}{\lambda}\right)$.

For a continuously differentiable and strictly convex function $F:\mathbb{R}^m \to \mathbb{R}$, we define the associated Bregman divergence [17,18] between two points $p,q\in\mathbb{R}^m$ to be

$$D_F(p,q) = F(p) - F(q) - \langle \nabla F(q), p-q\rangle\,,$$

where $\langle\cdot,\cdot\rangle$ denotes the usual inner product.

2. Distributed Learning and Model Aggregation

Now suppose that there are K nodes each having n samples. Each node $k=1,\dots,K$ has a dataset $S_k=(Z_{1,k},\dots,Z_{n,k})$, with $Z_{i,k}$ taken i.i.d. from π. We use $S=(S_1,\dots,S_K)$ to denote the entire dataset of size nK. Each node locally trains a model $W_k=\mathcal{A}_k(S_k)$ with algorithm $\mathcal{A}_k$. After each node locally trains its model, the models W_k are then combined to form the final model $\widehat{W}$ using an aggregation algorithm $\widehat{W}=\widehat{\mathcal{A}}(W_1,\dots,W_K)$ (see Figure 1). In this section, we will assume that $W_k\in\mathbb{R}^d$ and that the aggregation is performed by simple averaging (i.e., $\widehat{W}=\frac{1}{K}\sum_{k=1}^{K}W_k$). Define $\mathcal{A}$ to be the total algorithm from the data S to the final weights $\widehat{W}$ such that $\widehat{W}=\mathcal{A}(S)$. In this section, if we say that Assumption 1 or 2 holds, we mean that it holds for each algorithm $\mathcal{A}_k$. As in Theorem 1, we use $S^{(i,k)}$ to denote the entire dataset S with sample $Z_{i,k}$ replaced by an independent copy $Z'_{i,k}$, and similarly, we use $S_k^{(i)}$ to refer to the sub-dataset at node k, with sample $Z_{i,k}$ replaced by an independent copy $Z'_{i,k}$:

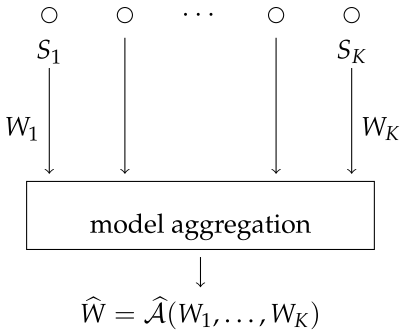

Figure 1. The distributed learning setting with model aggregation.

Theorem 3. *Suppose that* $\ell(\cdot,z)$ *is a convex function of* $w\in\mathbb{R}^d$ *for each* z *and that* $\mathcal{A}_k$ *represents the empirical risk minimization algorithm on local dataset* S_k *in the sense that*

$$W_k=\mathcal{A}_k(S_k)=\operatorname*{argmin}_{w}\sum_{i=1}^{n}\ell(w,Z_{i,k})\,.$$

Then, we have

$$\Delta_{\mathcal{A}}(s) \le \frac{1}{K}\sum_{k=1}^{K}\Delta_{\mathcal{A}_k}(s_k)\,.$$

Proof.

$$\begin{aligned}\Delta_{\mathcal{A}}(s) &= \mathbb{E}_{Z\sim\pi}[\ell(\mathcal{A}(s),Z)] - \frac{1}{nK}\sum_{i,k}\ell(\mathcal{A}(s),z_{i,k}) \\ &= \mathbb{E}_{Z\sim\pi}\left[\ell\left(\frac{1}{K}\sum_{k=1}^{K} w_k, Z\right)\right] - \frac{1}{nK}\sum_{i,k}\ell(\mathcal{A}(s),z_{i,k}) \\ &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{Z\sim\pi}[\ell(w_k,Z)] - \frac{1}{nK}\sum_{i,k}\ell(\mathcal{A}(s),z_{i,k}) \qquad (4) \\ &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{Z\sim\pi}[\ell(w_k,Z)] - \frac{1}{K}\sum_{k=1}^{K}\min_{w}\frac{1}{n}\sum_{i=1}^{n}\ell(w,z_{i,k}) \qquad (5) \\ &= \frac{1}{K}\sum_{k=1}^{K}\Delta_{\mathcal{A}_k}(s_k).\end{aligned}$$

In the above display, Equation (4) follows by the convexity of ℓ via Jensen's inequality, and Equation (5) follows by minimizing the empirical risk over each node's local dataset, which exactly corresponds to what each node's local algorithm $\mathcal{A}_k$ does. □

While Theorem 3 seems to be a nice characterization of the generalization bounds for the aggregate model (in that the aggregate generalization error cannot be any larger than the average generalization errors over each node), it does not offer any improvement in the expected generalization error that one might expect when given nK total samples instead of just n samples. A naive application of the generalization bounds from Theorem 2, followed by the data processing inequality $I(\widehat{W};Z_{i,k}) \leq I(W_k;Z_{i,k})$, runs into the same problem.

2.1. Improved Bounds

In this subsection, we demonstrate bounds on the expected generalization error that remedy the above shortcomings. In particular, we would like to demonstrate the following two properties:

(1) The bound should decay with the number of nodes K in order to take advantage of the total dataset from all K nodes.

(2) The bound should be in terms of the information theoretic quantities $I(W_k;S_k)$, which can represent (or be bounded from above by) the capacities of the channels over which the nodes are communicating. This can, for example, represent a communication or local differential privacy constraint for each node.

At a high level, we will improve on the bound from Theorem 3 by taking into account the fact that a small change in S_k will only change $\widehat{W}$ by a fraction $\frac{1}{K}$ of the amount that it will change W_k. In the case where W is a linear or location model, and the loss ℓ is a Bregman divergence, we can obtain an upper bound on the expected generalization error that satisfies properties (1) and (2) as follows:

Theorem 4 (Linear or Location Models with Bregman Loss)**.** *Suppose the loss ℓ takes the form of one of the following:*

(i) $\ell(w,(x,y)) = D_F(w^T x, y)$;

(ii) $\ell(w,z) = D_F(w,z)$.

In addition, assume that Assumption 1 holds. Then, we have

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] = \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{S_k\sim\pi^n}[\Delta_{\mathcal{A}_k}(S_k)]$$

and

$$\begin{aligned}\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] &\leq \frac{1}{nK^2}\sum_{i,k}\psi^{*-1}(I(W_k;Z_{i,k})) \\ &\leq \frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\frac{I(W_k;S_k)}{n}\right).\end{aligned}$$

Proof. Here, we restrict our attention to case (ii), but the two cases have nearly identical proofs. Using Theorem 1, we have

$$\begin{aligned}&\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] \\ &= \frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\left[\ell(\mathcal{A}(S),Z'_{i,k}) - \ell(\mathcal{A}(S^{(i,k)}),Z'_{i,k})\right] \\ &= \frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\Big[F(\mathcal{A}(S)) - F(Z'_{i,k}) - \langle\nabla F(Z'_{i,k}),\mathcal{A}(S)-Z'_{i,k}\rangle \\ &\qquad - F(\mathcal{A}(S^{(i,k)})) + F(Z'_{i,k}) + \langle\nabla F(Z'_{i,k}),\mathcal{A}(S^{(i,k)})-Z'_{i,k}\rangle\Big]\end{aligned}$$

$$= \frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\left[\langle\nabla F(Z'_{i,k}),\mathcal{A}(S^{(i,k)})-\mathcal{A}(S)\rangle\right] \tag{6}$$

$$= \frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\left[\langle\nabla F(Z'_{i,k}),\frac{1}{K}W_k^{(i)}+\frac{1}{K}\sum_{j\neq k}W_j-\frac{1}{K}\sum_j W_j\rangle\right]$$

$$= \frac{1}{nK^2}\sum_{i,k}\mathbb{E}_{S,S'}\left[\langle\nabla F(Z'_{i,k}),W_k^{(i)}-W_k\rangle\right]. \tag{7}$$

In Equation (7), we use $W_k^{(i)}$ to denote $\mathcal{A}_k(S_k^{(i)})$. Equation (6) follows the linearity of the inner product and cancels the higher order terms $F(\mathcal{A}(S))$ and $F(\mathcal{A}(S^{(i,k)}))$, which have the same expected values. The key step in Equation (7) then follows by noting that $\mathcal{A}(S^{(i,k)})$ only differs from $\mathcal{A}(S)$ in the submodel coming from node k, which is multiplied by a factor of $\frac{1}{K}$ when averaging all of the submodels. By backing out of Equation (6) and re-adding the appropriate canceled terms, we get

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] = \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{S_k\sim\pi^n}[\Delta_{\mathcal{A}_k}(S_k)].$$

By applying Theorem 2, this yields

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] \leq \frac{1}{nK^2}\sum_{i,k}\psi^{*-1}(I(W_k;Z_{i,k})).$$

Then, by noting that ψ^{*-1} is non-decreasing and concave, we have

$$\frac{1}{nK^2}\sum_{i,k}\psi^{*-1}(I(W_k;Z_{i,k})) \leq \frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\sum_{i=1}^{n}\frac{I(W_k;Z_{i,k})}{n}\right).$$

Using the property that conditioning decreases entropy yields

$$\sum_{i=1}^{n}I(W_k;Z_{i,k}) \leq I(W_k;S_k),$$

and we have

$$\frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\sum_{i=1}^{n}\frac{I(W_k;Z_{i,k})}{n}\right) \leq \frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\frac{I(W_k;S_k)}{n}\right)$$

as desired. □

The result in Theorem 4 is general enough to apply to many problems of interest. For example, if $F(p) = \|p\|_2^2$, then the Bregman divergence D_F gives the ubiquitous squared ℓ^2 loss (i.e., $D_F(p,q) = \|p-q\|_2^2$). For a comprehensive list of realizable loss functions, the interested reader is referred to [19]. Using F above, Theorem 4 can be applied to ordinary least squares regression, which we will examine in greater detail in Section 4. Other regression models such as logistic regression have loss functions that cannot be described with a Bregman divergence without the inclusion of additional nonlinearity. However, the result in Theorem 4 is agnostic to the algorithm that each node uses to fit its individual model. In this way, each node could fit a logistic model to its data, and the total aggregate model would then be an average over these logistic models. Theorem 4 would still control the expected generalization error for the aggregate model with the extra $\frac{1}{K}$ factor. However, critically, the upper bound would only be for the generalization error that is with respect to a loss of the form $D_F(w^T x, y)$, such as quadratic loss.

In order to show that the dependence on the number of nodes K from Theorem 4 is tight for certain problems, consider the following example from [3]. Suppose that $Z \sim \pi = \mathcal{N}(\mu, \sigma^2 I_d)$ and $\ell(w,z) = \|w-z\|_2^2$ so that we are trying to learn the mean μ of a Gaussian distribution. An obvious algorithm for each node to use is simple averaging of its dataset:

$$w_k = \mathcal{A}_k(s_k) = \frac{1}{n}\sum_{i=1}^{n} z_{i,k}\ .$$

For this algorithm, it can be shown that

$$I(\widehat{W};Z_{i,k}) = \frac{d}{2}\log\frac{nK}{nK-1}$$

and

$$\psi^{*-1}(y) = 2\sqrt{d\left(1+\frac{1}{nK}\right)^2\sigma^4 y}$$

See Section IV.A. in [3] for further details. If we apply the existing information theoretic bounds from Theorem 2 in an end-to-end way, such as in the approach from [14], we would get

$$\begin{aligned}\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] &\leq \sigma^2 d\sqrt{2\left(1+\frac{1}{nK}\right)^2 \log\frac{nK}{nK-1}} \\ &= O\left(\frac{1}{\sqrt{nK}}\right)\ .\end{aligned}$$

However, for this choice of algorithm at each node, the true expected generalization error can be computed to be

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] = \frac{2\sigma^2 d}{nK}\ .$$

By applying our new bound from Theorem 4, we get

$$\begin{aligned}\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] &\leq \frac{\sigma^2 d}{K}\sqrt{2\left(1+\frac{1}{n}\right)^2\log\frac{n}{n-1}}\\ &\leq O\left(\frac{1}{K\sqrt{n}}\right)\end{aligned}$$

which shows the correct dependence on K and improves upon the $O\left(\frac{1}{\sqrt{K}}\right)$ result from prior information theoretic methods.

2.2. General Models and Losses

In this section, we briefly describe some results that hold for more general classes of models and loss functions, such as deep neural networks and other nonlinear models:

Theorem 5 (Lipschitz Continuous Loss). *Suppose that $\ell(w,z)$ is Lipschitz continuous as a function of w in the sense that*

$$|\ell(w,z)-\ell(w',z)| \leq C\|w-w'\|_2$$

for any z and that $\mathbb{E}[\|W_k-\mathbb{E}[W_k]\|_2] \leq \sigma_0$ for each k. Then, we have

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] \leq \frac{2C\sigma_0}{K}.$$

Proof. Starting with Theorem 1, we have

$$\begin{aligned}&\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)]\\ &\quad=\frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\Big[\ell(\mathcal{A}(S),Z'_{i,k})-\ell(\mathcal{A}(S^{(i,k)}),Z'_{i,k})\Big]\\ &\quad\leq\frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\left[C\left\|\mathcal{A}(S)-\mathcal{A}(S^{(i,k)})\right\|_2\right] &(8)\\ &\quad=\frac{1}{nK^2}\sum_{i,k}\mathbb{E}_{S,S'}\left[C\left\|W_k-W_k^{(i)}\right\|_2\right]\\ &\quad\leq\frac{C}{nK^2}\sum_{i,k}\mathbb{E}_{S,S'}[\|W_k-\mathbb{E}[W_k]\|_2]+\mathbb{E}_{S,S'}\left[\left\|W_k^{(i)}-\mathbb{E}[W_k]\right\|_2\right] &(9)\\ &\quad\leq\frac{2C\sigma_0}{K}, &(10)\end{aligned}$$

where Equation (8) follows from Lipschitz continuity, Equation (9) uses the triangle inequality, and Equation (10) is assumed. □

The bound in Theorem 5 is not in terms of the information theoretic quantities $I(W_k;S_k)$, but it does show that the $O\left(\frac{1}{K}\right)$ upper bound can be shown for much more general loss functions and arbitrary nonlinear models.

2.3. Privacy and Communication Constraints

Both communication and local differential privacy constraints can be thought of as special cases of mutual information constraints. Motivated by this observation, Theorem 4 immediately implies corollaries for these types of systems:

Corollary 1 (Privacy Constraints)**.** *Suppose each node's algorithm $\mathcal{A}_k$ is an ε-local, differentially private mechanism in the sense that $\frac{p(w_k|s_k)}{p(w_k|s'_k)} \le e^{\varepsilon}$ for each w_k, s_k, s'_k. Then, for losses ℓ of the form in Theorem 4, and under Assumption 2, we have*

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] \le \frac{1}{K}\sqrt{\frac{2R^2\min\{\varepsilon,(e-1)\varepsilon^2\}}{n}}\,.$$

Proof. Note that

$$\begin{aligned} I(W_k;S_k) &= \sum_{w_k,s_k} p(w_k,s_k)\log\frac{p(w_k|s_k)}{\sum_{s'_k} p(w_k|s'_k)p(s'_k)} \\ &\le \sum_{w_k,s_k} p(w_k,s_k)\log\frac{p(w_k|s_k)}{\inf_{s'_k} p(w_k|s'_k)} \\ &\le \sum_{w_k,s_k} p(w_k,s_k)\varepsilon = \varepsilon\,. \end{aligned}$$

Similarly, it is true that

$$\begin{aligned} I(W_k;S_k) &= \mathsf{KL}(P_{W_kS_k}\|P_{S_k}P_{W_k}) \\ &\le \mathsf{KL}(P_{W_kS_k}\|P_{S_k}P_{W_k}) + \mathsf{KL}(P_{S_k}P_{W_k}\|P_{W_kS_k}) \\ &= \sum_{w_k,s_k} p(w_k)p(s_k)\left(\frac{p(w_k|s_k)}{p(w_k)}-1\right)\log\frac{p(w_k|s_k)}{p(w_k)} \\ &\le \sum_{w_k,s_k} p(w_k)p(s_k)(e^{\varepsilon}-1)\varepsilon \le (e-1)\varepsilon^2 \end{aligned}$$

where the last inequality is only true for $\varepsilon \le 1$. Putting these two displays together gives $I(W_k;S_k) \le \min\{\varepsilon,(e-1)\varepsilon^2\}$, and the result follows from Theorem 4. □

Corollary 2 (Communication Constraints)**.** *Suppose each node can only transit B bits of information to the model aggregator, meaning that each W_k can only take 2^B distinct possible values. Then, for losses ℓ of the form in Theorem 4, and under Assumption 2, this yields*

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_{\mathcal{A}}(S)] \le \frac{1}{K}\sqrt{\frac{2(\log 2)R^2B}{n}}\,.$$

Proof. The corollary follows immediately from Theorem 4 and

$$I(W_k;S_k) \le H(W_k) \le (\log 2)B\,.$$

□

3. Iterative Algorithms

We now turn to considering more complicated multi-round and iterative algorithms. In this setting, after T rounds, there is a sequence of weights $W^{(T)} = (W^1,\ldots,W^T)$, and the final model $\widehat{W}_T = f_T(W^{(T)})$ is a function of that sequence, where f_T gives a linear combination of the T vectors $W^1,\ldots,W^T$. The function f_T could represent, for example, averaging over the T iterates, choosing the last iterate W^T or some weighted average over the iterates. For each round t, each node k produces an updated model W_k^t based on its local dataset S_k and the previous timestep's global model W^{t-1}. The global model is then updated via an average over all K updated submodels:

$$W^t = \frac{1}{K}\sum_{k=1}^{K} W_k^t\,.$$

The particular example that we will consider is that of a distributed SGD, where each node constructs its updated model W_k^t by taking one or more gradient steps starting from W^{t-1} with respect to random minibatches of its local data. Our model is general enough to account for multiple local gradient steps, as are used in so-called federated learning [5–7], as well as noisy versions of SGDs, such as in [20,21]. If only one local gradient step is taken for each iteration, then the update rule for this example could be written as

$$W_k^t = W^{t-1} - \eta_t \nabla_w \ell(W^{t-1}, Z_{t,k}) + \xi_t \tag{11}$$

where $Z_{t,k}$ is a data point (or minibatch) sampled from S_k on timestep t, η_t is the learning rate, and ξ_t is some potential added noise. We assume that the data points $Z_{t,k}$ are sampled without replacement so that the samples are distinct across different values of t. We will also assume, for notational simplicity, that $\widehat{W}_T = W^T$, although the more general result follows in a straightforward manner.

For this type of iterative algorithm, we will consider the following timestep-averaged empirical risk quantity:

$$\frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\ell(W^t, Z_{t,k}) \ ,$$

and the corresponding generalization error, expressed as

$$\Delta_{\mathsf{sgd}}(S) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{Z\sim\pi}[\ell(W^t, Z)] - \frac{1}{K}\sum_{k=1}^{K}\ell(W^t, Z_{t,k})\right) . \tag{12}$$

Note that Equation (12) is slightly different from the end-to-end generalization error that we would get from considering the final model W^T and whole dataset S. It is instead an average over the generalization error we would get from each model, stopping at iteration t. We perform this so that when we apply the leave-one-out expansion from Theorem 1, we do not have to account for the dependence of W_k^t on past samples $Z_{t',k'}$ for $t' < t$ and $k' \neq k$. Since we expect the generalization error to decrease as we use more samples, this quantity should result in a more conservative upper bound and be a reasonable surrogate object to study. The next bound follows as a corollary to Theorem 4:

Corollary 3. *For losses ℓ of the form in Theorem 4, and under Assumption 2 (for each W_k^t), we have*

$$\mathbb{E}\big[\Delta_{\mathsf{sgd}}(S)\big] \le \frac{1}{T}\sum_{t=1}^{T}\frac{1}{K^2}\sum_{k=1}^{K}\sqrt{2R^2 I(W_k^t; Z_{t,k})} \ .$$

In the particular example described in Equation (11), where Gaussian noise $\xi_t \sim \mathcal{N}(0, I_d\sigma_t^2)$ is added to each iterate, Corollary 3 yields the following. As in [20], we assume that the updates are magnitude-bounded (i.e., $\sup_{w,x}\|\nabla_w \ell(w,z)\|_2 \le L$), the stepsizes satisfy $\eta_t = \frac{c}{t}$ for a constant $c > 0$, and that $\sigma_t = \sqrt{\eta_t}$:

Corollary 4. *Under the assumptions above, we have*

$$\mathbb{E}\big[\Delta_{\mathsf{sgd}}(S)\big] \le \frac{2RL}{K}\sqrt{\frac{c}{T}} \ .$$

Proof. The mutual information terms in Corollary 3 satisfy

$$\begin{aligned} I(W_k^t; Z_{t,k}) &\le I(W_k^t, W^{t-1}; Z_{t,k}) && (13)\\ &= I(W_k^t; Z_{t,k}|W^{t-1}) + I(W^{t-1}; Z_{t,k}) && (14)\\ &= I(W_k^t; Z_{t,k}|W^{t-1}) && (15)\\ &\le \frac{d}{2}\log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right) && (16)\\ &\le \frac{\eta_t^2 L^2}{2\sigma_t^2} = \frac{cL^2}{2t}\,. && (17) \end{aligned}$$

Equation (13) follows from the data-processing inequality, Equation (14) is the chain rule for mutual information, and Equation (15) follows from the independence of $Z_{t,k}$ and W^{t-1}. Equation (16) is due to the capacity of the additive white Gaussian noise channel, and Equation (17) just uses the approximation $\log(1+x) \le x$. Thus, we have

$$\mathbb{E}\left[\Delta_{\mathsf{sgd}}(S)\right] \le \frac{1}{TK}\sum_{t=1}^{T} RL\sqrt{\frac{c}{t}} \le \frac{2RL}{K}\sqrt{\frac{c}{T}}\,.$$

□

4. Simulations

We simulated a distributed linear regression example in order to demonstrate the improvement in our bounds over the existing information-theoretic bounds. To accomplish this, we generated $n = 10$ synthetic datapoints at each of K different nodes for various values of K. Each datapoint consisted of a pair (x, y), where $y = xw_0 + n$ with $x, n \sim \mathcal{N}(0,1)$, and $w_0 \sim \mathcal{N}(0,1)$ was the randomly generated true weight that was common to all datapoints. Each node constructed an estimate $\widehat{w}_k$ of w_0 using the well-known normal equations which minimize the quadratic loss (i.e., $\widehat{w}_k = \operatorname{argmin}_w \sum_{i=1}^{n}(wx_{i,k} - y_{i,k})^2$). The aggregate model was then the average $\widehat{w} = \frac{1}{K}\sum_{k=1}^{K}\widehat{w}_k$. In order to estimate the old and new information-theoretic generalization bounds (i.e., the bounds from Theorems 2 and 4, respectively), this procedure was repeated $M = 10^6$ times, and the datapoint and model values were binned in order to estimate the mutual information quantities. The value of M was increased until the mutual information estimates were no longer particularly sensitive to the number and widths of the bins. In order to estimate the true generalization error, the expectations for both the population risk and the dataset were estimated by Monte Carlo experimentation, with 10^4 trials each. The results can be seen in Figure 2, where it is evident that the new information theoretic bound is much closer to the true expected generalization error and decays with an improved rate as a function of K.

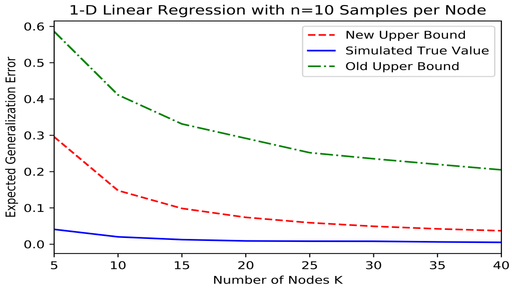

Figure 2. *Cont.*

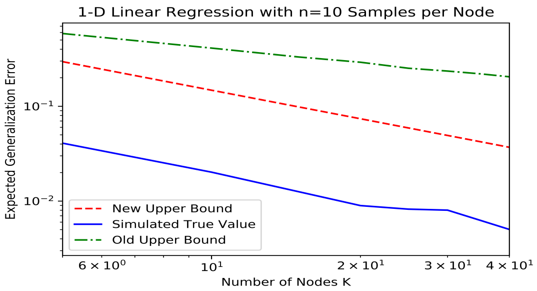

Figure 2. Information-theoretic upper bounds and expected generalization error for a simulated linear regression example in linear (**top**) and log (**bottom**) scales.

Author Contributions: Conceptualization, L.P.B., A.D. and H.V.P.; Formal analysis, L.P.B., A.D. and H.V.P.; Investigation, L.P.B., A.D. and H.V.P.; Methodology, L.P.B., A.D. and H.V.P.; Supervision, H.V.P.; Writing—original draft, L.P.B.; Writing—review & editing, L.P.B., A.D. and H.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation grant number CCF-1908308.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Russo, D.; Zou, J. How Much Does Your Data Exploration Overfit? Controlling Bias via Information Usage. *IEEE Trans. Inf. Theory* **2020**, *66*, 302–323. [CrossRef]
2. Xu, A.; Raginsky, M. Information-Theoretic Analysis of Generalization Capability of Learning Algorithms. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2521–2530.
3. Bu, Y.; Zou, S.; Veeravalli, V.V. Tightening Mutual Information-Based Bounds on Generalization Error. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 121–130. [CrossRef]
4. Aminian, G.; Bu, Y.; Wornell, G.W.; Rodrigues, M.R. Tighter Expected Generalization Error Bounds via Convexity of Information Measures. In Proceedings of the 2022 IEEE International Symposium on Information Theory (ISIT), Espoo, Finland, 26 June–1 July 2022.
5. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017.
6. Konecný, J.; McMahan, H.B.; Ramage, D.; Richtárik, P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv* **2016**, arXiv:1610.02527.
7. Konečný, J.; McMahan, H.B.; Yu, F.X.; Richtarik, P.; Suresh, A.T.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv* **2016**, arXiv:1610.05492.
8. Lin, Y.; Han, S.; Mao, H.; Wang, Y.; Dally, W.J. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In Proceedings of the 6th International Congress on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
9. Barnes, L.P.; Inan, H.A.; Isik, B.; Ozgur, A. rTop-k: A Statistical Estimation Approach to Distributed SGD. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 897–907. [CrossRef]
10. Warner, S.L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *J. Am. Stat. Assoc.* **1965**, *60*, 63–69. [CrossRef] [PubMed]
11. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*; Halevi, S., Rabin, T., Eds.; Springer: Berlin/Heidelberg, Geramny, 2006.
12. Kasiviswanathan, S.P.; Lee, H.K.; Nissim, K.; Raskhodnikova, S.; Smith, A. What Can We Learn Privately? *SIAM J. Comput.* **2011**, *40*, 793–826. [CrossRef]
13. Cuff, P.; Yu, L. Differential Privacy as a Mutual Information Constraint. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 43–54.
14. Yagli, S.; Dytso, A.; Poor, H.V. Information-Theoretic Bounds on the Generalization Error and Privacy Leakage in Federated Learning. In Proceedings of the 2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Atlanta, GA, USA, 26–29 May 2020; pp. 1–5. [CrossRef]

15. Barnes, L.P.; Dytso, A.; Poor, H.V. Improved Information Theoretic Generalization Bounds for Distributed and Federated Learning. *arXiv* **2022**, arXiv:2202.02423.
16. Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; Sridharan, K. Learnability, Stability and Uniform Convergence. *J. Mach. Learn. Res.* **2010**, *11*, 2635–2670.
17. Bregman, L.M. The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]
18. Dytso, A.; Fauß, M.; Poor, H.V. Bayesian Risk With Bregman Loss: A Cramér–Rao Type Bound and Linear Estimation. *IEEE Trans. Inf. Theory* **2022**, *68*, 1985–2000. [CrossRef]
19. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J.; Lafferty, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
20. Pensia, A.; Jog, V.; Loh, P.L. Generalization Error Bounds for Noisy, Iterative Algorithms. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 546–550.
21. Wang, H.; Gao, R.; Calmon, F.P. Generalization Bounds for Noisy Iterative Algorithms Using Properties of Additive Noise Channels. *arXiv* **2021**, arXiv:2102.02976.