ECE543 Project
Generalization Bounds for Uniformly Stable Algorithms

Katherine Tsai

April 2019

1 Abstract

In this project, we introduce the work [3] that shows the generalization bound of an uniform stable algorithm using differential privacy algorithm and stability argument [8] to bound the tail distribution. This method substantially improve the generalization bound [1] using the stability assumption and McDiarmid's inequality. In particular, the authors prove that the generalization error of a γ-uniformly stable algorithm can be improved from $O((\gamma+1/n)\sqrt{n\ln(1/\delta)})$ with probability at least $1-\delta$ to $O(\sqrt{(\gamma+1/n)\ln(1/\delta)})$ with probability at least $1-\delta$ without any further assumptions. In addition, this work also show a tighter second moment that is in the order of $O(\gamma+1/n)$ compared to $O(\gamma+1/n)$in the previous work [1].

2 Introduction

Generalization bound can be used to show the consistency of the strongly convex ERM algorithm [7], the uniform stability of the gradient descent algorithm on smooth convex function [5], and the estimation of the prediction error of deferentially private prediction [2]. The generalization bound is a useful tool to show the learnability of the algorithm without showing the uniform convergence of empirical means (UCEM) property, which is the sufficient condition for consistency of the learning algorithm but not a necessary condition for all learning algorithms.

An common approach to proving the generalization bound is to analyze the stability of the learning algorithm with respect to a change in the dataset. In this work, the author takes this approach to show the generalization bound. Furthermore, the generalization bound proved in this work matches the order of that using second moment estimation and Chebyshev's inequality in [1]. At the same time, the generalization bound produced by the second moment bound introduced in this paper matches the one produced by the first moment estimation [7] and the Markov inequality in the strongly convex function setting. The techniques here are to use the property of differential privacy to show the upper bound and the lower bound of the expectation of the generalization error. Since the learning algorithm $A(S)$ may not be an ϵ-differentially private algorithm, the authors introduce a new learning algorithm $G(S)$ that takes the input of a multidataset $Z^{m\times n}$ which is composed of m rows and n columns, the expected value of the generalization error of each row as scoring functions. Then the algorithm $G(S)$ outputs an index $k \in [m]$ with probability proportional $e^{\epsilon g_k(S)/2\Delta}$, where $g_k(S)$ is the scoring function of row k ϵ is the parameter that is positive and Δ is the stability of the expectation of the generalization error. It can be shown that $G(S)$ is an ϵ-differentially private algorithm. Putting all the ingredients together, the upper bound and the lower bound of the expectation of the generalization error using learning algorithm $A(S)$ can be shown. In the following stage, the stability argument [8] is employed to bound the tail distribution of the generalization error to show the high probability result.

3 Preliminaries

In this section, we introduce some basic properties that are later used in the proofs of the generalization bound.

Definition 3.1. A learning algorithm $A : Z^n \to \mathcal{F}$ is γ-uniformly stable if for all $S, S' \in Z^n$, where S, S' differs in only one sample and $z \in Z$, $\big|\ell\big(A(S), z\big) - \ell\big(A(S'), z\big)\big| \leq \gamma$

Definition 3.2. A function $f : Z^n \to \mathbb{R}$ has sensitivity at most γ if for all $S, S' \in Z^n$, where S, S' differs in only one sample, $\big|f(S) - f(S')\big| \leq \gamma$

Definition 3.3. A learning algorithm $A : Z^n \to \mathcal{F}$ is ϵ-differentially private if for any measurable subset E of $\mathcal{F}$for all $S, S' \in Z^n$, where S, S' differs in only one sample, $P\{A(S) \in E\} \leq e^{\epsilon} P\{A(S') \in E\}$

Lemma 3.1. Let $A : Z^n \to \mathcal{F}$ and $\ell : \mathcal{F} \times Z \to [0, 1]$ with uniform stability γ. $\mathcal{P}$ denotes the probability distribution over Z, a multi-dataset $S \in Z^{m \times n}$, where S_k, $\forall k \in [m]$ is the k-th row of the dataset, then we define a scoring function $g_k(S)$

$$g_k(S) = E_{z \sim \mathcal{P}}\Big[\ell\big(A(S_k), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n} \ell\big(A(S_k), S_{kj}\big)$$

$g_k(S)$ has sensitivity $2\gamma + \frac{1}{n}$

The following property will be used in bounding the tail of the generalization error in replacement of the McDiarmid's inequality.

Lemma 3.2. [8] Let Q be a probability distribution over real number $\mathbb{R}$. Then,

$$P_{v \sim Q}\Big\{v \geq 2E_{v_1,..,v_m \sim Q}\big[\max\{0, v_1, v_2, .., v_m\}\big]\Big\} \leq \frac{\ln 2}{m}$$

Here $v_1, v_2, .., v_m$ are drawn independent and identically from the distribution Q.

Check [8] for the detailed proof.

4 Algorithm

4.1 Generalization Bound

To construct a ϵ-differentially private algorithm onto the existing learning framework $A(S)$, we consider the following setting:

Theorem 4.1. [6] Let $g_1, g_2, ..., g_m : Z^n \to \mathbb{R}$ be m scoring functions such that $|g_i(S) - g_i(S')| \leq \Delta = 2\gamma + \frac{1}{n}$, $\forall i \in [m]$, where $S, S' \in Z^n$ differ from only one sample. Given inputs $S \in Z^n$ and a parameter $\epsilon > 0$, G is the algorithm that outputs an index $k \in [m]$ with probability $P\{G(S) = k\} = \frac{\exp(\frac{\epsilon}{2\Delta} g_k(S))}{\sum_{i \in [m]} \exp(\frac{\epsilon}{2\Delta} g_i(S))}$. Then G is ϵ-differentially private algorithm and for every $S \in Z^n$:

$$E_{P\{G(S)=k\}}\big[g_k(S)\big] \geq \max_{k \in [m]} \big\{g_k(S)\big\} - \frac{2\Delta}{\epsilon} \ln m$$

Proof. First show that $G(S)$ is ϵ-differentially private algorithm.

$$\begin{aligned}
P\{G(S) = k\} &= \frac{\exp\left(\frac{\epsilon}{2\Delta} g_k(S)\right)}{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta} g_i(S)\right)},\ P\{G(S') = k\} = \frac{\exp\left(\frac{\epsilon}{2\Delta} g_k(S')\right)}{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta} g_i(S')\right)} \\
\frac{P\{G(S) = k\}}{P\{G(S') = k\}} &= \frac{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta} g_i(S')\right)}{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta} g_i(S)\right)} \exp\left(\frac{\epsilon}{2\Delta}\big(g_k(S) - g_k(S')\big)\right) \\
&\leq \frac{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta}\big|g_i(S') - g_i(S)\big| + g_i(S)\right)}{\sum_{i \in [m]} \exp\left(\frac{\epsilon}{2\Delta} g_i(S)\right)} \exp\left(\frac{\epsilon}{2\Delta}\big|g_k(S) - g_k(S')\big|\right) \\
&\leq \exp(\frac{\epsilon}{2}) \exp(\frac{\epsilon}{2}) = \exp(\epsilon)
\end{aligned}$$

Let $C = \sum_{i\in[m]} \exp\big(\frac{\epsilon}{2\Delta} g_i(S)\big)$, then $g_k(S) = \frac{2\Delta}{\epsilon}\big(\ln C + \ln P\{G(S) = k\}\big)$

$$\begin{aligned} E[g_k(S)] &= \sum_{i\in[m]} P\{G(S) = i\}\frac{2\Delta}{\epsilon}\big(\ln C + \ln P\{G(S) = k\}\big) \\ &= \frac{2\Delta}{\epsilon}\Big(\ln C - H\big(G(S)\big)\Big) \end{aligned}$$

Here $H\big(G(S)\big)$ is the entropy of $G(S)$ and $H(G(S)) \le \ln(m)$. Moreover, $\frac{2\Delta}{\epsilon}\ln C \ge \max_{k\in[m]} g_k(S)$. Here completes the proof. □

Now, since $G(S)$ is an ϵ-differentially private algorithm, we can use this property to show the following bound.

Lemma 4.2. Let $A : Z^n \to \mathcal{F}$ and $\ell : \mathcal{F} \times Z \to [0,1]$ with uniform stability γ. Let $G : Z^{m\times n} \to [m]$ be an ϵ-differentially private algorithm. Then for any distribution $\mathcal{P}$ over Z,

$$V_s = E_{S\sim\mathcal{P}^{mn}, P\{G(S)=k\}}\Big[\frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S_k), S_{kj}\big)\Big]$$

$$e^{-\epsilon}V_s - \gamma \le E_{S\sim\mathcal{P}^{mn}, z\sim\mathcal{P}, P\{G(S)=k\}}\Big[\ell\big(A(S_k), z\big)\Big] \le e^{\epsilon}V_s + \gamma$$

We are now able to prove the following bound using above ingredients.

$$P\Big\{S : Z^n : E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S), S_j\big) \ge \sqrt{\big(2\gamma + \frac{1}{n}\big)\ln\frac{8}{\delta}}\Big\} \le \delta \tag{1}$$

Proof. To prove (1), we utilize Lemma 3.1 and Lemma 4.2 to find the upper bound of $E_{S\sim\mathcal{P}^{(m+1)n}}\Big[E_{P\{k=G(S)\}}\big[g_k(S)\big]\Big]$ and then use Theorem 4.1 and Lemma 3.2 to bound the tail of the (1). Consider the following setting: choose $m = \frac{\ln 2}{\delta}$ and consider an extra scoring function $g_{m+1}(S)$ such that $g_{m+1}(S) = 0$ for any $S \in Z^n$. The setting of an extra scoring function is to guarantee that $\max_{k\in[m+1]} g_k(S)$ is always equal or greater than zero, then we can use Lemma 3.2 to bound the tail. By Lemma 4.2,

$$\begin{aligned} E_{S\sim\mathcal{P}^{(m+1)n}}\Big[E_{P\{k=G(S)\}}\big[g_k(S)\big]\Big] &= E_{S\sim\mathcal{P}^{(m+1)n}, P\{k=G(S)\}}\Big[E_{z\sim\mathcal{P}}\big[\ell(A(S_k), z)\big] - \frac{1}{n}\sum_{j=1}^{n}\ell(A(S_k), S_{kj})\Big] \\ &\le e^{\epsilon} - 1 + \gamma \end{aligned}$$

Plugging into Theorem 4.1,

$$E_{S\sim\mathcal{P}^{mn}}\Big[\max\Big\{0, \max_{k\in[m]} E_{z\sim\mathcal{P}}\big[\ell(A(S_k), z)\big] - \frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S_k), S_{kj}\big)\Big\}\Big] = E_{S\sim\mathcal{P}^{(m+1)n}}\Big[\max_{k\in[m+1]} g_k(S)\Big]$$

$$\le E_{S\sim\mathcal{P}^{(m+1)n}}\Big[E_{P\{k=G(S)\}}\big[g_k(S)\big]\Big] + \frac{2(2\gamma + \frac{1}{n})}{\epsilon} \le e^{\epsilon} - 1 + \gamma + \frac{4\gamma + \frac{2}{n}}{\epsilon}\ln(m+1)$$

Choose $\epsilon = \sqrt{(2\gamma + \frac{1}{n})\ln(m+1)} = \sqrt{(2\gamma + \frac{1}{n})\ln(e\ln(2)/\delta)}$. Therefore, the bound is at least 2ϵ. Moreover, $(e^{\epsilon} - 1) \le 2\epsilon$ for $0 \le \epsilon \le 1.26$ and $\gamma \le \sqrt{\gamma}$ for $0 \le \gamma \le 1$. Then, the following inequality holds,

$$4\sqrt{\big(2\gamma + \frac{1}{n}\big)\ln\frac{2}{\delta}} + \gamma \le 4\sqrt{\big(2\gamma + \frac{1}{n}\big)\ln\frac{8}{\delta}}$$

Then, plugging into Lemma 3.2, we can obtain the following upper bound:

$$\begin{aligned} &P_{S\sim\mathcal{P}^n}\Big\{E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell(A(S), S_j) \ge 8\sqrt{\big(2\gamma + \frac{1}{n}\big)\ln\frac{8}{\delta}}\Big\} \\ &\le P_{S\sim\mathcal{P}^n}\Big\{E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell(A(S), S_j) \ge 2E_{S\sim\mathcal{P}^{(m+1)n}}\Big[\max_{k\in[m+1]} g_k(S_k)\Big]\Big\} \le \delta \end{aligned}$$

□

Since $g_k(S), \forall k \in [m], S \in Z^n$ has a bounded difference $c = 4\gamma + \frac{1}{n}$ and $E_{S\sim\mathcal{P}^n}[g_k(S)] \leq 2\gamma \forall k$, we can also obtian the upper bound by using McDiarmid's inequality,

$$P\Big\{S : Z^n : E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S), S_j\big) \geq \Big(4\gamma + \frac{1}{n}\Big)\sqrt{\frac{n\ln(\frac{1}{\delta})}{2}} + 2\gamma\Big\} \leq \delta \tag{2}$$

Comparing equation (1) and (2), the first states that the generalization error is within $O\Big(\sqrt{(\gamma + \frac{1}{n})\ln\frac{1}{\delta}}\Big)$ with probability $1-\delta$ and the latter shows that the error is within $O\Big((\gamma + \frac{1}{n})\sqrt{n\ln\frac{1}{\delta}}\Big)$ with probability $1-\delta$.

4.2 Second Moment Estimation

Lemma 4.3. Let $L : Z^n \times Z \to [-1, 1]$ be a function with uniform stability γ_L and $S^{(j)}$ be a copy of the dataset $S \in Z^n$ with the j-th component replaced by $z \sim \mathcal{P}$. Then,

$$E_{S\sim\mathcal{P}^n}\Big[\big(\frac{1}{n}\sum_{j=1}^{n}L(S^{(j)}, S_j)\big)^2\Big] \leq \gamma_L^2 + \frac{1}{n}$$

$$E_{S\sim\mathcal{P}^n}\left[\Big(E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S), S_j\big)\Big)^2\right] \leq 16\gamma^2 + \frac{2}{n} \tag{3}$$

Proof. To show that equation (3) holds true, we first define $L(S, z) := \ell\big(A(S), z\big) - E_{z\sim\mathcal{P}}\big[\ell(A(S), z)\big]$ $L \in [-1, 1]$, which is an unbiased estimator of $\mathcal{P}$. Here, L has stability 2γ and $E_{z\sim\mathcal{P}}\Big[\ell\big(A(S), z\big)\Big] - \frac{1}{n}\sum_{j=1}^{n}\ell\big(A(S), S_j\big) = -\frac{1}{n}\sum_{j=1}^{n}L(S, S_j)$. Therefore, the second moment of the LHS is equivalent to $E_{S\sim\mathcal{P}^n}\Big[\big(\frac{1}{n}\sum_{j=1}^{n}L(S, S_j)\big)^2\Big]$.

$$\Big|\frac{1}{n}\sum_{j=1}^{n}L(S, S_j) - E_{z\sim\mathcal{P}}\big[\frac{1}{n}\sum_{j=1}^{n}L(S^{(j)\leftarrow z}, S_j)\big]\Big| \leq \frac{1}{n}\sum_{j=1}^{n}E_{z\sim\mathcal{P}}\big[\big|L(S, S_j) - L(S^{(j)\leftarrow z}, S_j)\big|\big] \leq 2\gamma$$

From Lemma 4.3, we know that $E_{S\sim\mathcal{P}^n}\Big[\big(\frac{1}{n}\sum_{j=1}^{n}L(S^{(j)}, S_j)\big)^2\Big] \leq (2\gamma)^2 + \frac{1}{n}$. Then,

$$\begin{aligned}
E_{S\sim\mathcal{P}^n}\left[\Big(\frac{1}{n}\sum_{j=1}^{n}L(S, S_j)\Big)^2\right] &= E_{S\sim\mathcal{P}^n}\left[\Big(\frac{1}{n}\sum_{j=1}^{n}L(S, S_j) - L(S^{(j)}, S_j) + L(S^{(j)}, S_j)\Big)^2\right] \\
&\leq 2E_{S\sim\mathcal{P}^n}\Big[\big(\frac{1}{n}\sum_{j=1}^{n}L(S^{(j)}, S_j)\big)^2\Big] + 2E_{S\sim\mathcal{P}^n}\Big[\big(\frac{1}{n}\sum_{j=1}^{n}L(S, S_j) - L(S^{(j)}, S_j)\big)^2\Big] \\
&\leq 2\Big((2\gamma)^2 + \frac{1}{n}\Big) + 2(2\gamma)^2 = 16\gamma^2 + \frac{2}{n}
\end{aligned}$$

□

This give a tight bound of the second moment $O(\gamma^2 + \frac{1}{n})$ compared with the former work [1] which demonstrates that the second moment is $O(\gamma + \frac{1}{n})$. Then, an alternative bound for the tail distribution is to use Chebyshev's inequality which requires the estimation of second moment.

5 Application

In this section, the bounds (1), (3) shown in the previous section are applied to application examples discussed in class.

5.1 Learning without Uniform Convergence

Theorem 5.1. [4] Let $\mathcal{F}$ is a convex subset of a Hilbert space $\mathcal{H}$, and $\ell(f, z),\ z \in Z\ f \in \mathcal{F}$ is a m-strongly convex and L-Lipschitz function, which has stability at most $\frac{2L^2}{mn}$. Then, with probability $1-\delta$

$$E_{z\sim\mathcal{P}}\big[\ell(\hat{f}_n, z)\big] - E_{z\sim\mathcal{P}}\big[\ell(f^*, z)\big] \leq \frac{2L^2}{\delta mn}$$

where $\hat{f}_n$ is the empirical risk minimizer and f^* is the true risk minimizer.

In the setting of Theorem 5.1, we can apply Chebyshev's inequality to equation (3) and obtain the following bounds.

$$P_{s\sim\mathcal{P}^n}\left\{S \in Z^n : E_{z\sim\mathcal{P}}\left[\ell(\hat{f}_n, z)\right] - E_{z\sim\mathcal{P}}\left[\ell(f^*, z)\right] \geq c_1\left(\frac{L^2}{\sqrt{\delta}mn} + \frac{1}{\sqrt{\delta n}}\right)\right\} \leq \delta \tag{4}$$

Since $\frac{1}{\sqrt{\delta}}\sqrt{16\left(\frac{2L^2}{mn}\right)^2 + \frac{2}{n}} \leq \frac{1}{\sqrt{\delta}}\left(\frac{8L^2}{mn} + \sqrt{\frac{2}{n}}\right)$, choosing $c_1 \geq 8$ should be suffice. Similarly, we can obtain the upper bound by equation (1) and get the following:

$$P_{s\sim\mathcal{P}^n}\left\{S \in Z^n : E_{z\sim\mathcal{P}}\left[\ell(\hat{f}_n, z)\right] - E_{z\sim\mathcal{P}}\left[\ell(f^*, z)\right] \geq c_2 L\sqrt{\frac{\ln\frac{1}{\delta}}{mn}}\right\} \leq \delta \tag{5}$$

If $\frac{m}{L^2} < 1$, here c_2 48 is sufficient. When the loss function ℓ is not m-strongly convex, we can add a regularizer term $\frac{\lambda}{2}\|f\|^2$ to make it λ-strongly convex. In this case, choosing $\lambda = \frac{c}{\sqrt{\sqrt{\delta}n}}$ using the second moment and $\lambda = \frac{c}{n^{\frac{2}{3}}}$ using the high probability result leads to following result:

$$P_{s\sim\mathcal{P}^n}\left\{S \in Z^n : E_{z\sim\mathcal{P}}\left[\ell(\hat{f}_n, z)\right] - E_{z\sim\mathcal{P}}\left[\ell(f^*, z)\right] \geq c_1\left(\frac{L^2}{\delta^{\frac{1}{4}}\sqrt{n}}\right)\right\} \leq \delta \tag{6}$$

$$P_{s\sim\mathcal{P}^n}\left\{S \in Z^n : E_{z\sim\mathcal{P}}\left[\ell(\hat{f}_n, z)\right] - E_{z\sim\mathcal{P}}\left[\ell(f^*, z)\right] \geq c_2 L\frac{\sqrt{\ln\frac{1}{\delta}}}{n^{\frac{1}{3}}}\right\} \leq \delta \tag{7}$$

6 Conclusion

This paper prove a better upper bound of the generalization error. Previous work [1] shows the generalization error of γ-uniformly stable algorithm lies in the interval of order $O((\gamma + 1/n)\sqrt{n\ln(1/\delta)})$ which will not be a meaningful bound when $\gamma \geq 1/\sqrt{n}$ and works optimal when γ is in the order of $O(1/n)$. This work greatly improve the bound to $O(\sqrt{(\gamma + 1/n)\ln(1/\delta)})$ using differentially private algorithm and stability argument.

References

[1] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[2] Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *arXiv preprint arXiv:1803.10266*, 2018.

[3] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

[4] Bruce Hajek and Maxim Raginsky. Ece 543: Statistical learning theory.

[5] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

[6] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *(FOCS)*, pages 94–103, 2007.

[7] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[8] Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *arXiv preprint arXiv:1701.03493*, 2017.