Privacy Amplification for Episodic Training Methods

Vandy Tombs[1], Olivera Kotevska[1] and Steven Young[1]

[1] *Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830*

Abstract

It has been shown that differential privacy bounds improve when subsampling within a randomized mechanism. Episodic training, utilized in many standard machine learning techniques, uses a multistage subsampling procedure which has not been previously analyzed for privacy bound amplification. In this paper, we focus on improving the calculation of privacy bounds in episodic training by thoroughly analyzing privacy amplification due to subsampling with a multi-stage subsampling procedure. The newly developed bound can be incorporated into existing privacy accounting methods.

1. Introduction

As more data is being utilized by algorithms and machine learning techniques, rigorously maintaining the privacy of this data has become important. Cyber security, health, and census data collection are all examples of fields that are seeing increased scrutiny for ensuring the privacy of data, and it is well known that just anonymizing the data by removing features such as name is not sufficient to guarantee privacy due to vulnerabilities such as re-identification attacks, especially in the case when an adversary has access to auxiliary knowledge or data (see e.g. [1, 2]).

Differential privacy, first introduced by Dwork, is one technical definition of privacy that has been studied widely in the literature [3, 4]. This definition provides rigorous guarantees for the privacy of data that is utilized by an algorithm and has several nice properties like robustness to post processing and strong composition theorems.

Machine learning practitioners initially integrated differential privacy by naively applying these composition theorems algorithm by assuming that the algorithm accessed the entire training set on each step of training. Abadi et al. [5] noticed the data is subsampled into batches, so only a subset of the data is utilized for each step of training. This allowed for improved privacy bounds; however, they assumed that batches were created using Poisson sampling. Later authors showed improved bounds for creating batches using simple random sampling without replacement [6]. And most recently, Balle et al. [7] provided a fully unified theory for determining the privacy amplification due to subsampling as well as a complete analysis for Poisson and simple random subsampling both with and without replacement subsampling methods.

The subsampling methods analyzed previously include many of the subsampling methods utilized by machine learning; however, the methods does not capture batches formed by algorithms that use episodic training. Episodic training methods are utilized by a variety of machine learning algorithms, such as meta learning (e.g., [8, 9]) or metric learning (e.g., [10, 11]) algorithms. Domain generalization algorithms have also frequently utilized episodic training [12].

In this paper, we analyze the privacy amplification due to the subsampling method utilized in an episodic training regime. Specifically, we notice forming batches in episodic training is a multistage subsampling method, and we provide a complete analysis of the improved differential privacy bounds when applying a mechanism to a sample drawn using multistage subsampling. The resulting theorem can be easily applied to episodic training methods and integrated with privacy accounting methods such as the moment's accountant [5]. This bound can also be utilized by practitioners of other domains that use multistage subsampling within their algorithms.

2. Background and Related Works

2.1. Multistage Subsampling

In a multistage sampling procedure, the universe from which samples are drawn is partitioned. These partitions may contain the examples we are ultimately interested in sampling or may contain one or several levels of partitions. The subsampling procedure is to sample partitions at each level until examples are sampled. For example, if we are interested in the demographics of students at a school, we could partition students by teacher, sample some number of teachers and then sample students from each sampled teacher.

To see that episodic training is a multistage subsam-

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).
Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy.

pling procedure, consider how training batches are formed in Algorithm 2 of [8]. In this work, a subset of tasks are sampled from a collection of tasks, then the examples are sampled and provided to the training algorithm. This is a 2-stage sampling procedure since the training data is only partitioned into two levels: tasks and examples. In multistage subsampling, the first level of partitions are the *primary sampling units* and the final level is called the *ultimate sampling units* and this final level contains the examples we are ultimately interested in sampling. For more details on multistage subsampling, see e.g., [13].

2.2. Differential Privacy

Since our analysis utilizes the tools of Balle et al. [7], we introduce the necessary notations and definitions from it. Let $\mathcal{U}$ be an input space equipped with a binary symmetric relation $\simeq_{\mathcal{U}}$ that describes the concept of neighboring inputs. For our purposes, $\mathcal{U}$ is a universe that the training data is drawn from and the relation will be the add-one/remove-one relation, thus two training sets are related if they differ by the addition or removal of one element.

Given a randomized algorithm or *mechanism* $\mathcal{M} : X \to \mathbb{P}(Z)$, where $\mathbb{P}(Z)$ is the set of probability measures on the output space Z, $\mathcal{M}$ is (ε, δ)-*differentially private* w.r.t $\simeq_{\mathcal{U}}$ if for every pair $\mathcal{T} \simeq_{\mathcal{U}} \mathcal{T}'$ and every measurable subset $E \subseteq Z$,

$$\Pr[\mathcal{M}(\mathcal{T}) \in E] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{T}') \in E] + \delta.$$

Utilizing the tools from [7] requires expressing differential privacy in terms of α-*divergence* given by

$$D_{\alpha}(\mu || \mu') := \sup_{E} (\mu(E) - \alpha \mu(E))$$

of two probability measures $\mu, \mu' \in \mathbb{P}(Z)$, where E ranges over all measurable subsets of Z. Differential privacy can then be stated in terms of α-divergence; specifically, a mechanism $\mathcal{M}$ is (ε, δ)-differentially private if and only if $D_{e^{\epsilon}}(\mathcal{M}(\mathcal{T}) || \mathcal{M}(\mathcal{T}')) \leq \delta$ for every adjacent datasets $\mathcal{T} \simeq_{\mathcal{U}} \mathcal{T}'$.

We can now define the *privacy profile* of a mechanism $\mathcal{M}$ as $\delta_{\mathcal{M}} = \sup_{\mathcal{T} \simeq_{\mathcal{U}} \mathcal{T}'} D_{e^{\epsilon}}(\mathcal{M}(\mathcal{T}) || \mathcal{M}(\mathcal{T}'))$, which associates each privacy parameter $\alpha = e^{\varepsilon}$ with a bound on the α-divergence between the results of the mechanism on two adjacent datasets.

Two theorems from [7] are important in our analysis. The first is Advanced Joint Convexity, which we restate in terms of $\alpha = e^{\varepsilon}$ since we are interested in applying this theorem to improve the privacy bounds due to multistage subsampling.

Theorem 1. ([7], Advanced Joint Convexity of $D_{e^{\varepsilon}}$) Let $\mu, \mu' \in \mathbb{P}(Z)$ be measures satisfying $\mu = (1 - \eta)\mu_0 + \eta \mu_1$ and $\mu' = (1 - \eta)\mu_0 + \eta \mu'_1$ for some $\eta, \mu_0, \mu_1, \mu'_1$. Given $\varepsilon > 0$, let $\varepsilon' = log(1 + \eta(e^{\varepsilon} - 1))$ and $\beta = e^{\varepsilon} / e^{\varepsilon'}$, the following holds:

$$D_{e^{\varepsilon'}}(\mu || \mu') = \eta D_{e^{\varepsilon}}(\mu_1 || (1 - \beta)\mu_0 + \beta \mu'_1)$$

The final theorem provides the concrete privacy amplification that we need for our analysis. Before presenting this, we need to define when two distributions $\upsilon, \upsilon' \in \mathbb{P}(Y)$ are d_Y-*compatible.* Let π be a coupling of υ, υ', define $d_Y(y, y') = d_Y(y, \text{supp}(\upsilon'))$ where $(y, y') \in \text{supp}(\pi)$ and the distance between a point y and $\text{supp}(\upsilon')$ is defined to be the distance between y and the closest point in $\text{supp}(\upsilon')$.

Theorem 2. Let $C(\upsilon, \upsilon')$ be the set of all couplings between υ and υ' and for $k \geq 1$ let $Y_k = \{y \in \text{supp}(\upsilon) : d_Y(y, \text{supp}(\upsilon')) = k\}$. If υ and υ' are d_Y-compatible, then the following holds:

$$\min_{\pi \in C(\upsilon, \upsilon')} \sum_{y, y'} \pi_{y,y'} \delta_{\mathcal{M}, d_Y(y,y')}(\varepsilon) = \sum_{k \geq 1} \upsilon(Y_k) \delta_{\mathcal{M},k}(\varepsilon)$$

We are now equipped to begin an analysis of the privacy amplification due to multistage subsampling.

3. OUR APPROACH: Privacy Bounds for Multistage Sampling Analysis

We will begin the analysis with an example. Through this example, we will introduce the notation necessary for the general analysis.

Example 3.1. Let $\mathcal{U}$ be a universe of 18 examples from which the database or training data is drawn from. Suppose we can categorize the data from the universe at 3 different levels, so we will perform a 3-stage sampling. Let

$$\begin{aligned} \mathcal{U} &= U_1 \cup U_2 \\ &= (U_{11} \cup U_{12} \cup U_{13}) \cup (U_{21} \cup U_{22}) \\ &= \big(\{u_{111}, u_{112}, u_{113}, u_{114}\} \cup \{u_{121}, u_{122}\} \\ &\quad \cup \{u_{131}, u_{132}, u_{133}\}\big) \cup \big(\{u_{211}, u_{212}, u_{213}, u_{214}\} \\ &\quad \cup \{u_{221}, u_{222}, u_{223}, u_{224}, u_{225}\}\big) \end{aligned}$$

In this example, U_{i_1} for $i \in \{1, 2\}$ are the primary sampling units, the $U_{i_1 i_2}$ are the ultimate sampling units and the $u_{i_1 i_2 i_3}$ are the examples that would be provided to a training algorithm.

In general, let $\mathcal{U}$ be a universe from which the training data is drawn and suppose a finite number of levels, N_L, partition this universe. Define U_i be the primary sampling units and let $U_{i_1 i_2 \cdots i_{N_{L-1}}}$ be the sampling units of the $U_{i_1 i_2 \cdots i_{N_{L-2}}}$ unit. $U_{i_1 i_2 \cdots i_{N_{L-1}}}$ is an ultimate sampling unit which contain the examples we are interest in sampling. Note that we require that each sampling unit be of finite size except the ultimate sampling units, which may be infinite. The multistage sampling procedure can be described by Algorithm 1: Multistage Sampling. Most episodic training procedures only use 2- or 3-stage sampling but we analyze the general case; which may have applications to other scientific domains (e.g. medical domains) where multistage sampling may have more levels.

Algorithm 1: Multistage Sampling

```
Set $PrevLevel := \bigcup U_i$
Set $CurrentLevel := \emptyset$
Given $n_j$: the number of units to be sampled at
  each level $(1 \leq j \leq N_L)$
for $j \in \{1, ...., N_L\}$ do
    for $S \in PrevLevel$ do
        sample without replacement $n_j$ elements
          from $S$
        add sampled elements to $CurrentLevel$
    end
    $PrevLevel = CurrentLevel$
    $CurrentLevel := \emptyset$
end
```

Now, let $\mathcal{T} \subset \mathcal{U}$ be the training data or database we are analyzing. We will require that the training data has at least one element from each sampling unit described above. Thus we only allow the ultimate sampling units of the training data $T_{i_1 i_2 \cdots i_{N_{L-1}}} \subset U_{i_1 i_2 \cdots i_{N_{L-1}}}$, to be a non-empty finite subset of the ultimate sampling units with at least n_{N_L} elements (i.e. at least the number of units that will be sampled from the ultimate sampling units). All other sampling units defined for the universe will remain the same for the training set.

We want to analyze the privacy bound on algorithms that use a multistage subsampling procedure on $\mathcal{T}$. To do this, we will apply the theorems from [7] and will analyze this sampling procedure under the add-one/remove one relation. We begin by defining a probability measure for this sampling procedure. We can do this by simply defining

$$\mu(t_{i_1 i_2 \cdots i_{N_L}}) = \frac{\prod_{j=1}^{N_L} n_j}{|U_{i_1}||U_{i_1 i_2}| \cdots |T_{i_1 i_2 \cdots i_{N_{L-1}}}|}$$

where $t_{i_1 i_2 \cdots i_{N_L}}$ is in the ultimate unit $T_{i_1 i_2 \cdots i_{N_{L-1}}}$.

Now consider $\mathcal{T}'$ created by removing one element from $\mathcal{T}$, say without loss of generality, $t_{i_1 i_2 \cdots i_{N_{L-1}} 1}$ for some $i_1, i_2, ..., i_{N_{L-1}}$. The probability measure μ' for sampling from $\mathcal{T}'$ can be defined similar to above. We wish to compute the total variational distance between these two measure so that we can apply the Advanced Coupling Theorem from [7]. We just need to compute:

$$TV(\mu, \mu') = 1 - \sum_{u \in U} \min\left(\mu(u), \mu'(u)\right)$$

Note we can easily extend our probability measures μ, μ' to the entire universe by setting the inclusion probability to 0 for any element not in T or T'. For all elements $t \in \mathcal{T}' \setminus T_{i_1 i_2 \cdots i_{N_{L-1}}}$, we have $\min(\mu(t), \mu'(t)) = \mu(t) = \mu'(t)$. Since $t_{i_1 i_2 \cdots i_{N_{L-1}} 1} \notin \mathcal{T}'$, we also have $\min(\mu(t_{i_1 i_2 \cdots i_{N_{L-1}} 1}), \mu'(t_{i_1 i_2 \cdots i_{N_{L-1}} 1})) = 0$. So we just need to consider the elements of the ultimate unit from which we removed an element. Since, we removed an element from this unit, the probability $\mu'(t) > \mu(t)$ since $T'_{i_1 i_2 \cdots i_{N_{L-1}}}$ (the ultimate unit missing an element in T') has fewer elements than $T_{i_1 i_2 \cdots i_{N_{L-1}}}$, therefore for all $t_{i_1 i_2 \cdots i_{N_{L-1}} i} \in T'_{i_1 i_2 \cdots i_{N_{L-1}}}$ and $i \neq 1$, we have $\mu(t_{i_1 i_2 \cdots i_{N_{L-1}} i}) < \mu'(t'_{i_1 i_2 \cdots i_{N_{L-1}} i})$ where

$$\mu(t_{i_1 i_2 \cdots i_{N_{L-1}} i}) = \frac{\prod_{j=1}^{N_L} n_j}{|U_{i_1}||U_{i_1 i_2}| \cdots |T_{i_1 i_2 \cdots i_{N_{L-1}}}|}$$

$$\mu'(t'_{i_1 i_2 \cdots i_{N_{L-1}} i}) = \frac{\prod_{j=1}^{N_L} n_j}{|U_{i_1}||U_{i_1 i_2}| \cdots |T'_{i_1 i_2 \cdots i_{N_{L-1}}}|}.$$

Thus

$$\sum_{u \in \mathcal{U}} \min\left(\mu(u), \mu'(u)\right) = \sum_{t \in \mathcal{T}'} \mu(t) = 1 - \mu(t_{i_1 i_2 \cdots i_{N_{L-1}} 1}).$$

Hence the total variational distance is just the inclusion probability of the element we removed. Determining the total variational distance when adding an element from $\mathcal{U}$ to $\mathcal{T}$ is similar to the above argument.

We can now provide an amplified privacy bound for multistage subsampling.

Theorem 3. Let $\mathcal{M}'$ be a subsampled mechanism on $\mathcal{T}$ described by Algorithm 1 and let $m_1 m_2 \ldots m_{N_L - 1}$ be the index of the penultimate sampling unit that satisfies

$$\min_{i_1, i_2, \ldots, i_{N_{L-1}}} (|U_{i_1}||U_{i_1 i_2}| \cdots |T_{i_1 i_2 \cdots i_{N_{L-1}}}|).$$

Then, for any $\epsilon \geq 0$, we have that $\delta_{\mathcal{M}'}(\epsilon') \leq \eta \delta_{\mathcal{M}'}(\epsilon)$ for and $\eta = \frac{\prod_{j=1}^{N_L} n_j}{|U_{m_1}||U_{m_1 m_2}| \cdots |T_{m_1 m_2 \cdots m_{N_{L-1}}}|}$ and $\epsilon' = log(1 + \eta(e^{\epsilon} - 1))$ under the add-one/remove-one relation.

To fully complete the proof, let $\mathcal{T}, \mathcal{T}'$ be training sets drawn from $\mathcal{U}$ with $\mathcal{T} \simeq_r \mathcal{T}'$ under the add-one/remove-one relation $\simeq_r$ and let $S_\gamma(\mathcal{T})$ denote the subsampling mechanism described by Algorithm 1 for $\gamma = TV(\mu, \mu')$. Let $\mathcal{T}_0 = \mathcal{T} \cap \mathcal{T}'$, then by definition of $\simeq_r$, $\mathcal{T}_0 = \mathcal{T}$ or $\mathcal{T}_0 = \mathcal{T}'$. Let $\omega_0 = S_\gamma(\mathcal{T}_0)$, $\mu = S_\gamma(\mathcal{T})$ and $\mu' =$

$S_\gamma(\mathcal{T}')$. Then the decompositions of μ and μ' induced by their maximal coupling have that $\mu_1 = \omega_0$ when $\mathcal{T}_0 = \mathcal{T}$ or $\mu_1' = \omega_0$ when $\mathcal{T}_0 = \mathcal{T}'$. We only need to consider $\mathcal{T}_0 = \mathcal{T}'$ since this is when the maximum is obtained in applying advanced joint convexity. Finally, we note that one can easily create a $d_{\simeq_r}$-compatible pair according to the definition provided in [7] by first sampling y from μ and building y' by adding v (which may be empty) to y. Thus for each dataset pair, by Theorem 7 of [7], we have $\delta_{M'}(\varepsilon') \leq \gamma\delta_M(\varepsilon)$. In order to get a bound for all possible training set pairs, we need to take $\eta = min_{(\mathcal{T},\mathcal{T}')}(\gamma_{\mathcal{T}\simeq_r\mathcal{T}'})$. This occurs exacty when we remove an element from the penultimate unit with index $m_1 m_2 \cdot m_{N_L-1}$ which completes the proof. □

We briefly mention how one might incorporate this new bound into a privacy accounting method. Many accounting methods, like the moments accountant [5], use the moment generating function in conjunction with the Gaussian mechanism to calculate the privacy bounds while a machine learning algorithm is training. Using Theorem 4 from [7] with our new bound one can easily derive a subsampled Gaussian that can be utilized in algorithms like those described in [5, 14].

4. Conclusion

This paper completely analyzes the privacy amplification due to multistage subsampling. This provides the correct privacy bounds for any algorithm that utilizes multistage subsampling, such as machine learning algorithms that use episodic training. Our future goal is to perform experiments to better understand privacy in machine learning algorithms that use episodic training like meta-learning algorithms. We hope our presented approach and discussion will prove useful to other researchers wanting to apply privacy bounds on multistage sampling in other studies and applications.

References

[1] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in: 2008 IEEE Symposium on Security and Privacy (sp 2008), 2008, pp. 111–125. doi:10.1109/SP.2008.33.

[2] L. Rocher, J. M. Hendrickx, Y.-A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models 10 (????) 3069. URL: https://doi.org/10.1038/s41467-019-10933-3. doi:10.1038/s41467-019-10933-3.

[3] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (2014) 211–407. URL: https://doi.org/10.1561/0400000042. doi:10.1561/0400000042.

[4] C. Dwork, Differential privacy: A survey of results, in: M. Agrawal, D. Du, Z. Duan, A. Li (Eds.), Theory and Applications of Models of Computation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 1–19.

[5] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 308–318. URL: https://doi.org/10.1145/2976749.2978318. doi:10.1145/2976749.2978318.

[6] Y.-X. Wang, B. Balle, S. Kasiviswanathan, Subsampled rényi differential privacy and analytical moments accountant, Journal of Privacy and Confidentiality 10 (2021). URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/723. doi:10.29012/jpc.723.

[7] B. Balle, G. Barthe, M. Gaboardi, Privacy amplification by subsampling: Tight analyses via couplings and divergences, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 6280–6290.

[8] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1126–1135. URL: https://proceedings.mlr.press/v70/finn17a.html.

[9] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: ICLR, 2017.

[10] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, D. Wierstra, Matching networks for one shot learning, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

[11] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[12] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, T. Hospedales, Episodic training for domain generalization, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1446–1455. doi:10.1109/ICCV.2019.00153.

[13] C.-E. Särndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer-Verlag, 2003, p. 124–162.

[14] I. Mironov, K. Talwar, L. Zhang, R\'enyi differential privacy of the sampled gaussian mechanism (????). URL: http://arxiv.org/abs/1908.10530. arXiv:1908.10530.