PRIVATE MULTI-TASK LEARNING: FORMULATION AND METHODS

Shengyuan Hu, Zhiwei Steven Wu, Virginia Smith
Carnegie Mellon University
shengyua@andrew.cmu.edu
{zstevenwu, smithv}@cmu.edu

ABSTRACT

Many problems in machine learning rely on *multi-task learning (MTL)*, in which the goal is to solve multiple related machine learning tasks simultaneously. MTL is particularly relevant for privacy-sensitive applications in areas such as healthcare, finance, and IoT computing, where sensitive data from multiple, varied sources are shared for the purpose of learning. In this work, we formalize notions of multi-task privacy via *joint differential privacy* (JDP), a relaxation of Differential Privacy (DP) for mechanism design and distributed optimization. We then propose a differentially private algorithm for the commonly-used mean-regularized MTL objective. We analyze our objective and solver, providing certifiable guarantees on both privacy and utility. Our initial work provides groundwork for privacy-preserving multi-task learning and highlights several interesting directions of future study.

1 INTRODUCTION

Multi-task learning (MTL) aims to solve multiple learning tasks simultaneously, while exploiting similarities/differences across tasks (Caruana, 1997). Multi-task learning is commonly used in applications that warrant strong privacy guarantees. For example, MTL has been explored in healthcare applications, as a way to learn over diverse populations or between multiple institutions (Baytas et al., 2016; Suresh et al., 2018; Harutyunyan et al., 2019); in financial forecasting, to combine knowledge from multiple indicators or across organizations (Ghosn & Bengio, 1997; Cheng et al., 2020); and in IoT computing, as an approach for learning in federated networks of heterogeneous devices (Smith et al., 2017). While MTL can significantly improve accuracy in these applications, there is a dearth of work studying the privacy implications of learning in multi-task settings.

In this work, we develop and theoretically analyze methods for differentially private multi-task learning. Motivated by applications including federated learning and personalization, we focus on providing *task-level* privacy guarantees by incorporating the notion of *differential privacy* (DP) Dwork et al. (2006). Informally, an algorithm satisfies DP if its output is insensitive to the change of the any single individual's data. In the context of MTL, task-level DP directly would require the entire collection of predictive models to be insensitive to any change of the private data in any single task. Such a privacy requirement becomes too stringent, as it implies that the predictive model for task k must have little dependence on the training data for task k, which prevents usefulness of the model. To circumvent this, we leverage *joint differential privacy* (JDP) (Kearns et al., 2014), which requires that for each task k, the set of output predictive models for all other tasks except k is insensitive to k's private data.

Using these definitions of privacy, we then develop new learning algorithms for MTL with rigorous privacy and utility guarantees. Specifically, we propose DP Mean-Regularized MTL, a simple framework to learn multiple tasks privately. We show that our method achieves (ϵ, δ)-JDP (Section 3.3). We also analyze the convergence of our method on convex objectives, which demonstrates a tradeoff between privacy and utility.

2 Background and Related Work

Multi-task learning. In multi-task learning, the goal is to solve multiple related ML tasks simultaneously. Our work focuses on the general and widely-used formulation of multi-task relationship learning, as formalized in Section 3.1. MTL is particularly useful in privacy-sensitive applications where datasets are shared between multiple, heterogeneous entities (Baytas et al., 2016; Smith et al., 2017; Ghosn & Bengio, 1997). In these cases, it is natural to view each data source (e.g., financial institution, hospital, mobile phone) as a separate 'task' that is learned in union with the other tasks. This allows data to be shared, but the models to be personalized to each data silo. In the setting of federated learning, MTL is particularly helpful to train personalized model for each local device. Several examples of MTL application in personalized FL are training a mean regularized objective (Hanzely & Richtárik, 2020; Hanzely et al., 2020), clustering (Ghosh et al., 2020; Sattler et al., 2020), interpolate between the local and global model (Deng et al., 2020; Mansour et al., 2020), etc.

Differentially private MTL has also been explored in a few prior works with different problem settings. Wu et al. (2020) proposed a MTL setting where a common feature representation shared by all tasks is first learned, followed by task specific models on top of this private representation. Xie et al. (2017) proposed a method to preserve data-level privacy by representing the parametrized model for each task as the sum of a public weight that is shared by all tasks and a task specific weight that is only updated locally. However, both differ from our focus on multi-task relationship learning (stated in Equation 1). Geyer et al. (2017) studied differentially private multitask relationship learning for the mean estimation problem. Their method provides a guarantee that model of each task is differentially private. In other words, this requires the model to be insensitive to changes in its own task's training set. In this work, we focus on task-level privacy, i.e. protecting the private data of one task from other tasks. Hence, instead of making every task-specific model differentially private, we propose and analyze a method to learn a model for each task such that each model is insensitive to changes in any other task's training set.

Joint differential privacy. Differential privacy has been widely studied to provide certifiable guarantee for privacy in algorithms (Dwork & Roth, 2014; Dwork et al., 2010). In practice, differential privacy could be achieved by noise perturbation methods like Laplace mechanism, Gaussian mechanism, etc (Dwork & Roth, 2014). However, multi-task settings are not directly compatible with the standard formulation of differential privacy. Our work aims to remedy this by providing meaningful privacy formulations using joint differential privacy (JDP) (Kearns et al., 2014).

3 DPMTL: Differentially Private Multi-Task Learning

In this section, we first formalize our multi-task learning objective, which is a form of mean-regularized multi-task learning (Section 3.1). We then present DPMTL, a method for performing differentially-private MTL (Section 3.2). We provide both a privacy guarantee (Section 3.3) and utility guarantee (Section 3.4) for our approach.

3.1 Problem Setup

In the classical setting of multi-task relationship learning (Zhang & Yang, 2017; Zhang & Yeung, 2010), there are m different task learners with their own task-specific data. The aim is to solve:

$$\min_{W,\Omega}\left\{F(W,\Omega)=\left\{\sum_{k=1}^{m}\sum_{i=1}^{n_k} l_k(w_k^T x_i, y_i)+\mathcal{R}(W,\Omega)\right\}\right\}, \tag{1}$$

where w_k is the parametrization for the model of task learner k; $W=[w_1;\cdots;w_m]$; and $\Omega\in\mathbb{R}^{m\times m}$ characterizes the relationship between every pair of task learners. In this paper, we focus on studying the mean-regularized multi-task learning objective (Evgeniou & Pontil, 2004): a special case of (1) where $\Omega=(\mathbf{I_{m\times m}}-\mathbf{\frac{1}{m}1_m 1_m^T})^{\mathbf{2}}$. Here $\mathbf{I_{m\times m}}$ is the identity matrix of size $m\times m$ and $\mathbf{1_m}\in\mathbb{R}^{\mathbf{m}}$ is the vector with all entries equal to 1. In this case we set $\mathcal{R}$ to be

$$\mathcal{R}(W,\Omega)=\lambda_1\text{tr}(W\Omega W^T)+\lambda_2\|W\|_F^2.$$

Algorithm 1 DP Mean-Regularized MTL

```
Input: $m$, $T$, $\lambda$, $\eta_t$, $\eta_l$, $\{w_1^0, \cdots, w_m^0\}$, $\widetilde{w}^0 = \frac{1}{m}\sum_{k=1}^{m} w_k^0$
for $t = 0, \cdots, T-1$ do
    Global Learner broadcasts the mean weight $\widetilde{w}^t$
    for $k = 1, \cdots, m$ in parallel do
        Each task updates its weight $w_k$ for some $E$ iterations
            $w_k^{t+1} = \Pi_{\mathcal{B}}(w_k^t - \eta_t(\nabla_{w_k^t} l_k(w_k^t) + \lambda(w_k^t - \widetilde{w}^t)))$
        Each task sends $w_k^{t+1}$ back to the global learner.
    end for
    Global Learner computes a noisy aggregator of the weights
        $\widetilde{w}^{t+1} = \frac{1}{m}\sum_{k=1}^{m} w_k^{t+1} + \mathcal{N}(0, \sigma^2 \mathbf{I}_{\mathbf{d}\times\mathbf{d}})$
end for
for $k = 1, \cdots, m$ in parallel do
    Each task assigns $w_k = w_k^T$ and runs local finetuning for some iterations
        $w_k = w_k - \eta_l(\nabla_{w_k} l_k(w_k) + \lambda(w_k - \widetilde{w}^T))$
end for
return $w_1, \cdots, w_m$ as differentially private personalized models
```

Choosing $\lambda_1 = \frac{\lambda}{2}$ and $\lambda_2 = 0$, we can rewrite the objective as:

$$\min_{W} \left\{ F(W) = \left\{ \sum_{k=1}^{m} \frac{\lambda}{2} \|w_k - \bar{w}\|^2 + \sum_{i=1}^{n_k} l_k(w_k^T x_i, y_i) \right\} \right\}, \tag{2}$$

where $\bar{w}$ is the average of task-specific models: $\bar{w} = \frac{1}{m}\sum_{i=1}^{m} w_k$. Note that $\bar{w}$ is shared across all tasks, and each w_k is kept locally for task learner k. During optimization, each task learner k solves:

$$\min_{w_k} \left\{ f_k(w_k; \bar{w}) = \frac{\lambda}{2} \|w_k - \bar{w}\|^2 + \sum_{i=1}^{n_k} l_k(w_k^T x_i, y_i) \right\}. \tag{3}$$

Despite the prevalence of this simple form of multi-task learning and its recent use in applications such as federated learning with strong privacy motivations (e.g., Hanzely & Richtárik, 2020; Hanzely et al., 2020; Dinh et al., 2020), we are unaware of prior work that has formalized differential privacy in the context of solving this objective.

3.2 Algorithm

In this section we provide our DPMTL algorithm. Our optimization method consists of two steps: (i) each task learner solves its own local objective inexactly and sends the updated model to the global learner; (ii) the global learner adds random Gaussian noise after aggregating the local updates and broadcasts the aggregated mean. The major difference between our solver and a non-private solver for MTL is the noise added to the aggregation rule. We assume that we have a trusted global learner, i.e., it is safe for the global learner to learn about task-specific data. However, since the global model is a linear combination of all task specific models and is shared among all task learners, any single task learner could infer information about other tasks from the global model. For example, in the scenario where there are only two task learners whose models are parametrized by w_1 and w_2 respectively, task learner one could simply retrieve w_2 by subtracting $\frac{1}{2}w_1$ from $\bar{w}$ for every communication round. To overcome this privacy risk, we propose applying the Gaussian Mechanism (Dwork & Roth, 2014) during global aggregation. In this case, each task learner receives a noisy aggregated global model, making it hard for any task to leak private information to the other tasks. On the other hand, in order to bound the ℓ_2-sensitivity of the task training process, we require that the task model being communicated lies in $\mathcal{B} = \{w | \|w\| \leq B\}$. Hence, we use Projected Gradient Descent (Bubeck, 2014) for local updates. We formalize the privacy guarantee of Algorithm 1 in Section 3.3.

3.3 Privacy Analysis

In this section, we rigorously explore the privacy guarantee Algorithm 1 provides. As mentioned previously, since the output of MTL is a collection of models, we use the notion of Joint Differential Privacy (JDP) (Kearns et al., 2014) to study the privacy guarantee of Algorithm 1. JDP requires that for each task k, the set of output predictive models for all other tasks except k is insensitive to k's private data. Here, we provide a formal definition of JDP:

Definition 1 (Joint Differential Privacy (JDP))**.** *(Kearns et al., 2014) An algorithm $\mathcal{M} : \mathcal{U}^m \to \mathcal{R}^m$ is (ϵ, δ)-joint differentially private if for every i, for every pair of neighboring datasets that only differ in index i: $D, D' \in \mathcal{U}^m$ and for every set of subsets of outputs $S \subset \mathcal{R}^m$,*

$$Pr(\mathcal{M}(D)_{-i} \in S) \leq e^{\epsilon} Pr(\mathcal{M}(D')_{-i} \in S) + \delta, \tag{4}$$

where $\mathcal{M}(D)_{-i}$ represents the vector $\mathcal{M}(D)$ with the i-th entry removed.

JDP allows the predictive model for task k to depend on the private data of k, while still providing a strong guarantee: even if all the users from all the other tasks collude and share their information, they still will not be able to learn much about the private data in the task k.

In our optimization scheme, for each task k, at the end of each communication round, a shared model is received, then the task specific model is updated. We formalize this local task learning process as $h_k : \mathcal{D}_k \times \mathcal{W} \to \mathcal{W}$. Here we simply assume $\mathcal{W} \subset \mathbb{R}^d$. Define the mechanism for communication round t to be

$$\mathcal{M}^{1:t}(\{D_i\}, \widetilde{w}^t, \{h_i(\cdot)\}, \sigma) = \frac{1}{m} \sum_{i=1}^{m} h_i(D_i, \widetilde{w}^t) + \alpha^t. \tag{5}$$

where $\alpha^t \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_{d \times d})$. Hence, the whole joint optimization process described in line 2-9 of Algorithm 1 could be characterized by $\mathcal{M}^{1:T}(\{D_i\}, \widetilde{w}^T, \{h_i(\cdot)\}, \sigma)$.

Theorem 1. *The mechanism $\mathcal{M}^{1:T}$ is $(\alpha, 2\alpha \sum_{t=1}^{T} (\max_j \Delta_2 h_j^t)^2 / (m\sigma)^2)$-Rényi Differentially Private.*

We provide a detailed proof of Theorem 1 in Appendix A.1. Theorem 1 provides a provable privacy guarantee on the learned global model. In the next step of the training process, each task k optimizes its local objective given the private global model. We formally define this process as $h'_k : \mathcal{D}_k \times \mathcal{W} \to \mathcal{W}$. Note that this is not exactly the same as h_k because the number of iterations to perform a local update could differ between the two processes. We now present our main theorem of the JDP guarantee provided by Algorithm 1:

Theorem 2. *For any $\epsilon > 0$ and $0 < \delta < 1$, let $\alpha = \frac{4 \log(1/\delta)}{\epsilon}$ and $\sigma = \frac{8B\sqrt{T \log(1/\delta)}}{\epsilon m}$. Algorithm 1 that outputs $h'_k(D_k, \mathcal{M}^{1:T})$ for each task is (ϵ, δ)-joint differentially private.*

From Theorem 2, we can see that for any fixed δ, the more tasks involved in the learning process, the smaller σ we need in order to keep the privacy parameter ϵ the same. In other words, less noise is required for the global model to keep the task specific data private. When we have infinitely many tasks ($m \to \infty$), we have $\sigma \to 0$, in which case the global model only needs to add negligible amount of noise to make itself private to all tasks. We provide a detailed proof in Appendix A.1.

3.4 Convergence Analysis

As discussed in Section 3.1, we are interested in the following task-specific objective:

$$f_k(w_k; \widetilde{w}) = l_k(w_k) + \frac{\lambda}{2} \|w_k - \widetilde{w}\|_2^2 \,, \tag{6}$$

where $\widetilde{w} = \bar{w} + \alpha$, where $\alpha \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{\mathbf{d} \times \mathbf{d}})$, $\bar{w} = \frac{1}{m} \sum_{i=1}^{m} w_i$.

Here, we analyze the convergence behavior in the setting where every task participates in the optimization process at every communication round. Further, we assume the number of local optimization steps $E = 1$. We present the following convergence result:

Theorem 3 (Convergence under convex loss)**.** *Let f_k be both convex and $(L+\lambda)$-smooth. Further let $f_k^* = \min_{w,\bar{w}} f_k(w;\bar{w})$. If we use a fixed learning rate $\eta_t = \eta = \frac{1}{L+\lambda}$, we have*

$$\mathbb{E}[f_k(w_k^T;\widetilde{w}^T) - f_k^*] \leq \frac{3(L+\lambda)\|w_k^0 - w_k^*\|^2 + f_k(w_k^0;\widetilde{w}^0) - f_k^*}{T} + \lambda\Gamma \tag{7}$$

where

$$\Gamma = \mathcal{O}((B + \sqrt{d}\sigma)^2). \tag{8}$$

Denote $C = 3(L+\lambda)\|w_k^0 - w_k^*\|^2 + f_k(w_k^0;\widetilde{w}^0) - f_k^*$. Let α and σ chosen as we set in Theorem 2. Take $T = \frac{\sqrt{C}\epsilon m}{8B\sqrt{2\lambda d\log(1/\delta)}}$, the right hand side is bounded by

$$\mathbb{E}[f_k(w_k^T;\widetilde{w}^T) - f_k^*] \leq \frac{16B\sqrt{C\lambda d\log(1/\delta)}}{\epsilon m} + \mathcal{O}(\lambda B^2) \tag{9}$$

By Theorem 2, given fixed ϵ, σ^2 grows linearly with respect to T. Hence, there exists an optimal T that minimizes the upper bound in Equation 7, as shown in Equation 9. Note that the upper bound consists of $\mathcal{O}(\frac{1}{m\epsilon})$, which means when there are more tasks, the upper bound becomes smaller while the privacy parameter stays the same. On the other hand, Theorem 3 also shows a privacy-utility tradeoff using our Algorithm 1: the upper bound grows inversely proportional to the privacy parameter ϵ.

4 Conclusion and Future work

We propose a simple method for differentially private mean-regularized multi-task learning. Theoretically, we provide both privacy and utility guarantees for our approach. In future work, we plan to further explore the privacy-utility tradeoff of our objective empirically, and analyze/evaluate the performance of our method on non-convex objectives. We are also interested in extending our results to more general forms of MTL, e.g., the family of objectives in (1) with arbitrary matrix Ω.

References

Inci M Baytas, Ming Yan, Anil K Jain, and Jiayu Zhou. Asynchronous multi-task learning. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 11–20. IEEE, 2016.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12), 2020.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Canh T Dinh, Nguyen H Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pp. 265–284, Berlin, Heidelberg, 2006. Springer.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pp. 51–60, 2010.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Conference on Knowledge Discovery and Data Mining*, 2004.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *arXiv preprint arXiv:2006.04088*, 2020.

Joumana Ghosn and Yoshua Bengio. Multi-task learning for stock selection. In *Advances in neural information processing systems*, pp. 946–952, 1997.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM J. Comput.*, 45(6), 2016.

Michael J. Kearns, Mallesh M. Pai, Aaron Roth, and Jonathan R. Ullman. Mechanism design in large games: incentives and privacy. In Moni Naor (ed.), *Innovations in Theoretical Computer Science, ITCS'14, Princeton, NJ, USA, January 12-14, 2014*, pp. 403–410. ACM, 2014.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Virginia Smith, Chaokai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *NeurIPS*, 2017.

Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 802–810, 2018.

Huiwen Wu, Cen Chen, and Li Wang. A Theoretical Perspective on Differentially Private Federated Multi-task Learning. *arXiv e-prints*, art. arXiv:2011.07179, November 2020.

Liyang Xie, Inci M Baytas, Kaixiang Lin, and Jiayu Zhou. Privacy-preserving distributed multi-task learning with asynchronous updates. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1195–1204, 2017.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

A Appendix

A.1 Privacy Analysis: Proof for Theorem 1 and 2

Definition 2 (ℓ_2-sensitivity)**.** *Let $f : \mathcal{U} \to \mathbb{R}^d$ be some arbitrary function, the ℓ_2-sensitivity of f is defined as*

$$\Delta_2 f = \max_{\text{adjacent } D, D' \in \mathcal{U}} \|f(D) - f(D')\|_2 \tag{10}$$

Definition 3 (Rényi Divergence)**.** *(Mironov, 2017) Let P, Q be two probability distribution over the same probability space, and let p, q be the respective probability density function. The Rényi Divergence with finite order $\alpha \neq 1$ is:*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \ln \int_{\mathcal{X}} q(x) \left(\frac{p(x)}{q(x)}\right)^\alpha dx \tag{11}$$

Definition 4 ((α, ϵ)-Rényi Differential Privacy)**.** *(Mironov, 2017) A randomized mechanism $f : \mathcal{D} \to \mathcal{R}$ is said to have (α, ϵ)-Rényi Differential Privacy if for all adjacent $D, D' \in \mathcal{D}$ it holds that:*

$$D_\alpha(f(D)\|f(D')) \leq \epsilon. \tag{12}$$

Lemma 1. *(Mironov, 2017) The Gaussian mechanism is $(\alpha, \alpha(2(\Delta_2 f)^2/\sigma^2))$-Renyi Differentially Private.*

Proof for Theorem 1. Define $H^t : \prod_{i=1}^m \mathcal{D}_i \times \mathcal{W} \to \mathcal{W}$ as

$$H^t(\{D_i\}, \widetilde{w}) = \frac{1}{m} \sum_{i=1}^m h_i^t(D_i, \widetilde{w}^t). \tag{13}$$

By Lemma 1, $\mathcal{M}^{t:t+1}$ is $(\alpha, 2\alpha(\Delta_2 H^t)^2/d\sigma^2)$-Renyi Differentially Private. Note that

$$(\Delta_2 H^t)^2 = \max_j \max_{\text{adjacent } D_j, D_j' \in \mathcal{D}_j} \left\|H^t(\{D_1, \cdots, D_j, \cdots, D_m\}, \widetilde{w}^t) - H^t(\{D_1, \cdots, D_j', \cdots, D_m\}, \widetilde{w}^t)\right\|^2 \tag{14}$$

$$= \max_j \max_{\text{adjacent } D_j, D_j' \in \mathcal{D}_j} \left\|\frac{1}{m} h_j^t(D_j, \widetilde{w}^t) - \frac{1}{m} h_j^t(D_j', \widetilde{w}^t)\right\|^2 \tag{15}$$

$$= \frac{1}{m^2} \max_j \max_{\text{adjacent } D_j, D_j' \in \mathcal{D}_j} \left\|h_j^t(D_j, \widetilde{w}^t) - h_j^t(D_j', \widetilde{w}^t)\right\|^2 \tag{16}$$

$$= \frac{1}{m^2} \max_j (\Delta_2 h_j^t)^2. \tag{17}$$

Hence, by sequential composition of Rényi Differential Privacy, $\mathcal{M}^{1:T}$ is $(\alpha, \sum_{i=1}^T 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2)$-RDP. □

Lemma 2. *(Mironov, 2017) If f is (α, ϵ)-RDP, then it is $(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP for all $\delta > 0$.*

Lemma 3. *(Hsu et al., 2016) Suppose $\mathcal{M} : \mathcal{U}^m \to \mathcal{R}$ is (ϵ, δ)-DP. Consider any set of functions: $f_i : \mathcal{U} \times \mathcal{R} \to \mathcal{R}'$. Then the mechanism $\mathcal{M}'$ that outputs $f_i(D_i, \mathcal{M}(D))$ for each each i satisfies (ϵ, δ)-JDP.*

Proof for Theorem 2. From Theorem 1, Lemma 2, and Lemma 3, we know that Algorithm 1 is $(\sum_{i=1}^T 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-JDP.

Plug in $\alpha = \frac{4\log(1/\delta)}{\epsilon}$, $\sigma = \frac{8B\sqrt{T\log(1/\delta)}}{\epsilon m}$, we have

$$
\begin{aligned}
\sum_{i=1}^{T} 2\alpha \max_j (\Delta_2 h_j^t)^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha - 1} &\leq \sum_{i=1}^{T} 2\alpha 4B^2/m^2\sigma^2 + \frac{\log(1/\delta)}{\alpha-1} && (18)\\
&= \frac{8\frac{4\log(1/\delta)}{\epsilon}B^2}{m^2(\frac{8B\sqrt{T\log(1/\delta)}}{\epsilon m})^2} + \frac{\log(1/\delta)}{\frac{4\log(1/\delta)}{\epsilon} - 1} && (19)\\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} && (20)\\
&= \epsilon. && (21)
\end{aligned}
$$

Hence, Algorithm 1 is (ϵ, δ)-JDP. □

A.2 Convergence Analysis: Proof for Theorem 3

Lemma 4. *(Bubeck, 2014) Let f be a convex and L-smooth function on $\mathcal{X}$. Let $x, y \in \mathcal{X}$, $x^+ = \Pi_{\mathcal{X}}(x - \frac{1}{L}\nabla f(x))$, and $g_{\mathcal{X}}(x) = L(x - x^+)$. Then the following holds true:*

$$
f(x^+) - f(y) \leq g_{\mathcal{X}}(x)^T(x-y) - \frac{1}{2L}\|g_{\mathcal{X}}(x)\|^2. \tag{22}
$$

Proof for Theorem 3. Let $w_k^* = \arg\min_w f_k(w; \bar{w}^*)$. From Lemma 4, let the learning rate be a constant $\eta_t = \eta = \frac{1}{L+\lambda}$, we know

$$
\mathbb{E}[f_k(w_k^{t+1}; \widetilde{w}^t) - f_k(w_k^t; \widetilde{w}^t)] \leq -\frac{1}{2(L+\lambda)}\|(L+\lambda)(w_k^t - w_k^{t+1})\|^2 \tag{23}
$$

and

$$
\begin{aligned}
\mathbb{E}[f_k(w_k^{t+1}; \widetilde{w}^t) - f_k(w_k^*; \bar{w}^*)] &\leq \mathbb{E}[f_k(w_k^{t+1}; \widetilde{w}^t) - f_k(w_k^*; \widetilde{w}^t)] + \mathbb{E}[f_k(w_k^*; \widetilde{w}^t) - f_k(w_k^*; \bar{w}^*)] && (24)\\
&\leq \|(L+\lambda)(w_k^t - w_k^{t+1})\|\|w_k^t - w_k^*\| + \underbrace{\mathbb{E}[f_k(w_k^*; \widetilde{w}^t) - f_k(w_k^*; \bar{w}^*)]}_{\text{A}}. && (25)
\end{aligned}
$$

Note that

$$
\begin{aligned}
\text{A} &= \mathbb{E}\left[\frac{\lambda}{2}\|w_k^* - \widetilde{w}^t\|^2 - \frac{\lambda}{2}\|w_k^* - \bar{w}^*\|^2\right] && (26)\\
&= \frac{\lambda}{2}\mathbb{E}[\|\widetilde{w}^t - \bar{w}^*\|\|2w_k^* - \widetilde{w}^t - \bar{w}^*\|] && (27)\\
&= \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \bar{w}^*\|^2]}\sqrt{\mathbb{E}[\|(\bar{w}^* - \widetilde{w}^t) + 2(w_k^* - \bar{w}^*)\|^2]} && (28)\\
&\leq \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \bar{w}^*\|^2]}\sqrt{\mathbb{E}[\|\bar{w}^* - \widetilde{w}^t\|^2] + 4\mathbb{E}[\|w_k^* - \bar{w}^*\|^2] + 4\mathbb{E}[\|\bar{w}^* - \widetilde{w}^t\|\|w_k^* - \bar{w}^*\|]} && (29)\\
&\leq \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \bar{w}^*\|^2]}\sqrt{\mathbb{E}[\|\bar{w}^* - \widetilde{w}^t\|^2] + 4\mathbb{E}[\|w_k^* - \bar{w}^*\|^2] + 4\sqrt{\mathbb{E}[\|\bar{w}^* - \widetilde{w}^t\|^2]\mathbb{E}[\|w_k^* - \bar{w}^*\|^2]}} && (30)\\
&\leq \frac{\lambda}{2}(2B + \sqrt{d}\sigma)(2B + \sqrt{d}\sigma + 4B). && (31)
\end{aligned}
$$

Denote the right hand side as γ_1. Hence, we have

$$
\mathbb{E}[f_k(w_k^{t+1}; \widetilde{w}^{t+1}) - f_k(w_k^t; \widetilde{w}^t)] \leq \underbrace{\mathbb{E}[f_k(w_k^{t+1}; \widetilde{w}^{t+1}) - f_k(w_k^{t+1}; \widetilde{w}^t)]}_{\text{B}} - \frac{1}{2(L+\lambda)}\|(L+\lambda)(w_k^t - w_k^{t+1})\|^2 \tag{32}
$$

and

$$\mathbb{E}[f_k(w_k^{t+1};\widetilde{w}^{t+1}) - f_k(w_k^*;\bar{w}^*)] \leq \|(L+\lambda)(w_k^t - w_k^{t+1})\|\|w_k^t - w_k^*\| + \mathbb{E}[f_k(w_k^{t+1};\widetilde{w}^{t+1}) - f_k(w_k^{t+1};\widetilde{w}^t)] + \gamma_1. \tag{33}$$

It suffices to bound B:

$$\begin{aligned}
\mathrm{B} &= \mathbb{E}\left[\frac{\lambda}{2}\|w_k^{t+1} - \widetilde{w}^{t+1}\|^2 - \frac{\lambda}{2}\|w_k^{t+1} - \widetilde{w}^t\|^2\right] && (34)\\
&= \frac{\lambda}{2}\mathbb{E}[\|\widetilde{w}^t - \widetilde{w}^{t+1}\|\|2w_k^{t+1} - \widetilde{w}^t - \widetilde{w}^{t+1}\|] && (35)\\
&\leq \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \widetilde{w}^{t+1}\|^2]\mathbb{E}[\|2w_k^{t+1} - \widetilde{w}^t - \widetilde{w}^{t+1}\|^2]} && (36)\\
&= \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \widetilde{w}^{t+1}\|^2]}\sqrt{\mathbb{E}[\|(\widetilde{w}^{t+1} - \widetilde{w}^t) + 2(w_k^{t+1} - \widetilde{w}^{t+1})\|^2]} && (37)\\
&\leq \frac{\lambda}{2}\sqrt{\mathbb{E}[\|\widetilde{w}^t - \widetilde{w}^{t+1}\|^2]}\sqrt{\mathbb{E}[\|\widetilde{w}^{t+1} - \widetilde{w}^t\|^2] + 4\mathbb{E}[\|w_k^{t+1} - \widetilde{w}^{t+1}\|^2] + 4\mathbb{E}[\|\widetilde{w}^{t+1} - \widetilde{w}^t\|\|w_k^{t+1} - \widetilde{w}^{t+1}\|]} && (38)\\
&\leq \frac{\lambda}{2}\sqrt{\underbrace{\mathbb{E}[\|\widetilde{w}^t - \widetilde{w}^{t+1}\|^2]}_{\mathrm{C}_1}}\sqrt{\mathbb{E}[\|\widetilde{w}^{t+1} - \widetilde{w}^t\|^2] + 4\underbrace{\mathbb{E}[\|w_k^{t+1} - \widetilde{w}^{t+1}\|^2]}_{\mathrm{C}_2} + 4\sqrt{\mathbb{E}[\|\widetilde{w}^{t+1} - \widetilde{w}^t\|^2]\mathbb{E}[\|w_k^{t+1} - \widetilde{w}^{t+1}\|^2]}} && (39)
\end{aligned}$$

where the first and third inequality follows from Cauchy-Schwartz Inequality: $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. It suffices to find the upper bound of C_1 and C_2.

$$\begin{aligned}
\mathrm{C}_1 &= \mathbb{E}[\|(\bar{w}^t - \bar{w}^{t+1}) - (\alpha^t - \alpha^{t+1})\|^2] && (40)\\
&\leq \mathbb{E}[\|(\bar{w}^t - \bar{w}^{t+1})\|^2] + \mathbb{E}[\|\alpha^t - \alpha^{t+1}\|^2] + 2\sqrt{\mathbb{E}[\|(\bar{w}^t - \bar{w}^{t+1})\|^2]\mathbb{E}[\|\alpha^t - \alpha^{t+1}\|^2]} && (41)\\
&= \mathbb{E}[\|(\bar{w}^t - \bar{w}^{t+1})\|^2] + 2d\sigma^2 + 2\sqrt{\mathbb{E}[\|(\bar{w}^t - \bar{w}^{t+1})\|^2]2d\sigma^2} && (42)\\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{k=1}^m(w_k^t - w_k^{t+1})\right\|^2\right] + 2d\sigma^2 + 2\sqrt{2d}\sigma\sqrt{\mathbb{E}\left[\left\|\frac{1}{m}\sum_{k=1}^m(w_k^t - w_k^{t+1})\right\|^2\right]} && (43)\\
&\leq \frac{1}{m}\sum_{k=1}^m\mathbb{E}[\|w_k^t - w_k^{t+1}\|^2] + 2d\sigma^2 + 2\sqrt{2d}\sigma\sqrt{\frac{1}{m}\sum_{k=1}^m\mathbb{E}[\|w_k^t - w_k^{t+1}\|^2]} && (44)\\
&\leq 4B^2 + 2d\sigma^2 + 2\sqrt{2d}\sigma(2B) && (45)\\
&= (2B + \sqrt{2d}\sigma)^2. && (46)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathrm{C}_2 &= \mathbb{E}[\|w_k^t - \bar{w}^t + \alpha\|^2] && (47)\\
&\leq \mathbb{E}[\|w_k^t - \bar{w}^t\|^2] + \mathbb{E}[\|\alpha\|^2] + 2\mathbb{E}[\|w_k^t - \bar{w}^t\|\|\alpha\|] && (48)\\
&\leq \mathbb{E}[\|w_k^t - \bar{w}^t\|^2] + \mathbb{E}[\|\alpha\|^2] + 2\sqrt{\mathbb{E}[\|w_k^t - \bar{w}^t\|^2]\mathbb{E}[\|\alpha\|^2]} && (49)\\
&\leq (2B)^2 + d\sigma^2 + 2\sqrt{4B^2d\sigma^2} && (50)\\
&= (2B + \sqrt{d}\sigma)^2. && (51)
\end{aligned}$$

Plug the bounds for C_1 and C_2 into B:

$$\mathrm{B} \leq \frac{\lambda}{2}(2B + \sqrt{2d}\sigma)(2B + \sqrt{2d}\sigma + 2(2B + \sqrt{d}\sigma)). \tag{52}$$

Denote the right hand side as γ_2, we have

$$\mathbb{E}[f_k(w_k^{t+1};\widetilde{w}^{t+1}) - f_k(w_k^t;\widetilde{w}^t)] \leq \gamma_2 - \frac{1}{2(L+\lambda)}\|(L+\lambda)(w_k^t - w_k^{t+1})\|^2 \tag{53}$$

and

$$\mathbb{E}[f_k(w_k^{t+1};\widetilde{w}^{t+1}) - f_k(w_k^*;\bar{w}^*)] \leq \|(L+\lambda)(w_k^t - w_k^{t+1})\|\|w_k^t - w_k^*\| + \gamma_2 + \gamma_1. \tag{54}$$

Let $\delta_t = \mathbb{E}[f_k(w_k^t;\widetilde{w}^t) - f_k(w_k^*;\bar{w}^*)]$, we have

$$\delta_{t+1} \leq \delta_t + \gamma_2 - \frac{1}{2(L+\lambda)}\|(L+\lambda)(w_k^t - w_k^{t+1})\|^2 \tag{55}$$

$$\leq \delta_t + \gamma_2 - \frac{1}{2(L+\lambda)}\frac{1}{\|w_k^t - w_k^*\|^2}(\delta_{t+1} - \gamma_1 - \gamma_2)^2. \tag{56}$$

From the update rule of PGD, we know that

$$\|w_k^{t+1} - w_k^*\|^2 = \|w_k^t - \frac{1}{L+\lambda}((L+\lambda)(w_k^{t+1} - w_k^t)) - w_k^*\|^2 \tag{57}$$

$$\leq \|w_k^t - w_k^*\|^2 - \frac{2}{L+\lambda}((L+\lambda)(w_k^{t+1} - w_k^t))^T(w_k^t - w_k^*) + \frac{1}{(L+\lambda)^2}((L+\lambda)(w_k^{t+1} - w_k^t))^2 \tag{58}$$

$$\leq \|w_k^t - w_k^*\|^2 \tag{59}$$

$$\leq \|w_k^1 - w_k^*\|^2. \tag{60}$$

Hence, we have:

$$\delta_{t+1} - \gamma_1 - \gamma_2 \leq (\delta_t - \gamma_1 - \gamma_2) - \frac{1}{2(L+\lambda)}\frac{1}{\|w_k^1 - w_k^*\|^2}(\delta_{t+1} - \gamma_1 - \gamma_2)^2 + \gamma_2. \tag{61}$$

It can be shown by induction that

$$\delta_t - \gamma_1 - \gamma_2 \leq \frac{3(L+\lambda)\|w_k^1 - w_k^*\|^2 + f_k(w_k^1;\widetilde{w}^1) - f_k(w_k^*;\bar{w}^*)}{t} + \gamma_2. \tag{62}$$

□