Model-Protected Multi-Task Learning

Supplementary Material

Jian Liang*, Ziqi Liu†, Jiayu Zhou‡, Xiaoqian Jiang§, Changshui Zhang*, *Fellow*, *IEEE*, Fei Wang¶

*Department of Automation, Tsinghua University, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, P.R. China.

†Department of Computer Science, Xi'an Jiaotong University, Xi'an, P.R. China.

‡Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA.

§Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA.

¶Department of Healthcare Policy and Research, Weill Cornell Medical College, New York City, NY, USA.

APPENDIX B
WISHART DISTRIBUTION

Definition 8 (Gupta and Nagar [4]). *A $d \times d$ random symmetric positive definite matrix $\mathbf{E}$ is said to have a Wishart distribution $\mathbf{E} \sim W_d(\nu, \mathbf{V})$ if its probability density function is*

$$p(\mathbf{E}) = \frac{|\mathbf{E}|^{(\nu-d-1)/2}\exp(-\mathrm{tr}(\mathbf{V}^{-1}\mathbf{E})/2)}{2^{\frac{\nu d}{2}}|\mathbf{V}|^{1/2}\Gamma_d(\nu/2)},$$

where $\nu > d-1$ and $\mathbf{V}$ is a $d \times d$ positive definite matrix.

APPENDIX C
MODEL-DECOMPOSED MP-MTL METHODS

In this section, we consider the extension of our MP-MTL framework for MTL methods using the decomposed parameter/model matrix. Specifically, we focus on the following problem, where the trace norm is used for knowledge sharing across tasks and the $\|\cdot\|_1$ norm (sum of the ℓ_1 norm for vectors) is used for entry-wise outlier detection, as described in Algorithm 4.

$$\begin{aligned}&\min_{\mathbf{W}} \sum\nolimits_{i=1}^{m} \mathcal{L}_i(\mathbf{X}_i\mathbf{w}_i, \mathbf{y}_i) + \lambda_1\|\mathbf{P}\|_* + \lambda_2\|\mathbf{Q}\|_1 \\ &s.t.\ \mathbf{W} = \mathbf{P} + \mathbf{Q},\end{aligned} \tag{24}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{d\times m}$.

We note that in Algorithm 4, the role of $\mathbf{P}$ is the same as the role of $\mathbf{W}$ in Algorithm 2, and the additional procedures introduced to update $\mathbf{Q}$ are still STL algorithms. As such, we have the result in Corollary 2.

Corollary 2. *Algorithm 4 is an (ϵ, δ) - iterative MP-MTL algorithm.*

Remark 3. *Based on Algorithm 4, this will result in a similar procedure and identical theoretical results with respect to privacy by replacing the trace norm with the $\ell_{2,1}$ norm to force group sparsity in $\mathbf{P}$ or by replacing the $\|\cdot\|_1$ norm with the $\ell_{1,2}$ norm (sum of the ℓ_2 norm of column vectors) or $\|\cdot\|_F^2$ (square of the Frobenius norm).*

APPENDIX D
MP-MTL FRAMEWORK WITH SECURE MULTI-PARTY COMPUTATION

Pathak et al. [10] considered the demand for secure multi-party computation (SMC): protecting data instances from leaking to the curator and leaking between tasks during joint learning. However, by Proposition 3, the method of Pathak et al. [10] may introduce excess noise to protect both the data instances and the models simultaneously. To avoid unnecessary noise, we consider a divide-and-conquer strategy to ensure privacy for a single data instance and the model separately. Specifically, in each iteration of the Iterative MP-MTL algorithms, we perform private sharing after introducing the perturbation to the parameter matrix to protect a single data instance, as described in Algorithm 5, where a noise vector is added in Step 5 to the model vector based on sensitivity of replacing a single data instance.

Algorithm 4 Model-Protected Low-Rank and SParse (MP-LR-SP) Estimator

Input: Datasets $(\mathbf{X}^m, \mathbf{y}^m) = \{(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_m, \mathbf{y}_m)\}$, where $\forall i \in [m]$, $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}$ and $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$. Privacy loss $\epsilon, \delta \geq 0$. Number of iterations T. Step size η. Regularization parameter $\lambda_1, \lambda_2 > 0$. Norm clipping parameter $K > 0$. Acceleration parameters $\{\beta_t\}$. Initial models of tasks $\mathbf{W}^{(0)}$.

Output: $\widehat{\mathbf{W}}^{(1:T)}$.

For $t = 1, \ldots, T$, set ϵ_t such that $\tilde{\epsilon} \leq \epsilon$, where $\tilde{\epsilon}$ is defined in (7).
Let $\mathbf{P}^{(0)} = \mathbf{Q}^{(0)} = \widehat{\mathbf{Q}}^{(0)} = \mathbf{W}^{(0)}$.
for $t = 1 : T$ **do**
  Norm clipping: $\tilde{\mathbf{p}}_i^{(t-1)} = \mathbf{p}_i^{(t-1)}/\max(1, \frac{\|\mathbf{p}_i^{(t-1)}\|_2}{K})$ for all $i \in [m]$. Let $\widehat{\mathbf{P}}^{(0)} = \widetilde{\mathbf{P}}^{(0)}$.
  Compute sensitivity: $s_i^{(t-1)} = 2$ for all $i \in [m]$.
  $\widetilde{\mathbf{\Sigma}}^{(t)} = \widetilde{\mathbf{P}}^{(t-1)}(\widetilde{\mathbf{P}}^{(t-1)})^{\mathrm{T}}$.
  $\mathbf{\Sigma}^{(t)} = \widetilde{\mathbf{\Sigma}}^{(t-1)} + \mathbf{E}$, where $\mathbf{E} \sim W_d(d+1, \frac{\max_i s_i^{(t-1)}}{2\epsilon_t}\mathbf{I}_d)$ is a sample of the Wishart distribution.
  Perform SVD decomposition: $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{\Sigma}^{(t)}$.
  Let $\mathbf{S}_{\eta\lambda_1}$ be a diagonal matrix and let $\mathbf{S}_{\eta\lambda_1, ii} = \max\{0, 1 - \eta\lambda_1/\sqrt{\mathbf{\Lambda}_{ii}}\}$ for $i = 1, \ldots, \min\{d, m\}$.
  Let $\hat{\mathbf{p}}_i^{(t)} = \mathbf{U}\mathbf{S}_{\eta\lambda_1}\mathbf{U}^{\mathrm{T}}\mathbf{p}_i^{(t-1)}$ for all $i \in [m]$.
  Let $\hat{\mathbf{q}}_i^{(t)} = \mathrm{sign}(\mathbf{q}_i^{(t-1)}) \circ \max\{0, |\mathbf{q}_i^{(t-1)}| - \eta\lambda_2\}$ for all $i \in [m]$, where $\circ$ denotes the entry-wise product.
  Let $\widehat{\mathbf{W}}^{(t)} = \widehat{\mathbf{P}}^{(t)} + \widehat{\mathbf{Q}}^{(t)}$.
  Let $\boldsymbol{z}_{i,p}^{(t)} = \hat{\mathbf{p}}_i^{(t)} + \beta_t(\hat{\mathbf{p}}_i^{(t)} - \hat{\mathbf{p}}_i^{(t-1)})$ for all $i \in [m]$.
  Let $\boldsymbol{z}_{i,q}^{(t)} = \hat{\mathbf{q}}_i^{(t)} + \beta_t(\hat{\mathbf{q}}_i^{(t)} - \hat{\mathbf{q}}_i^{(t-1)})$ for all $i \in [m]$.
  $\mathbf{p}_i^{(t)} = \boldsymbol{z}_{i,p}^{(t)} - \eta\frac{\partial\mathcal{L}_i(\mathbf{X}_i(\boldsymbol{z}_{i,p}^{(t)} + \boldsymbol{z}_{i,q}^{(t)}), \mathbf{y}_i)}{\partial\hat{\mathbf{p}}_i^{(t)}}$ for all $i \in [m]$.
  $\mathbf{q}_i^{(t)} = \boldsymbol{z}_{i,q}^{(t)} - \eta\frac{\partial\mathcal{L}_i(\mathbf{X}_i(\boldsymbol{z}_{i,p}^{(t)} + \boldsymbol{z}_{i,q}^{(t)}), \mathbf{y}_i)}{\partial\hat{\mathbf{q}}_i^{(t)}}$ for all $i \in [m]$.
end for

The results in Proposition 4 show that we can simultaneously protect a single data instance and the model using such a divide-and-conquer strategy. Because it is not necessary to protect all the data instances in each task using data-protected algorithms, the perturbation for data-instance protection can be reduced.

Proposition 4. *Use Lemma 4 and Theorem 1. Algorithm 5 is an $(\epsilon_{mp}, \delta_{mp})$ - iterative MP-MTL algorithm and an $(\epsilon_{dp}, \delta_{dp})$ - iterative DP-MTL algorithm.*

Algorithm 5 MP-MTL framework with Secure Multi-party Computation (SMC)

Input: Datasets $(\mathbf{X}^m, \mathbf{y}^m) = \{(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_m, \mathbf{y}_m)\}$, where $\forall i \in [m]$, $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ and $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$. Privacy loss for model protection $\epsilon_{\text{mp}}, \delta_{\text{mp}} \geq 0$. Privacy loss for single data instance protection $\epsilon_{\text{dp}} \geq 0$. Number of iterations T. Shared information matrices $\mathbf{M}^{(0)}$. Initial models of tasks $\mathbf{W}^{(0)}$.

Output: $\widehat{\mathbf{W}}^{(1:T)}$.

1: For $t = 1, \ldots, T$, set $\epsilon_{\text{mp},t}$ such that $\tilde{\epsilon}_{\text{mp}} \leq \epsilon_{\text{mp}}$, where $\tilde{\epsilon}_{\text{mp}}$ is defined in (7), taking $\epsilon_t = \epsilon_{\text{mp},t}, \epsilon = \epsilon_{\text{mp}}, \delta = \delta_{\text{mp}}$.

2: For $t = 1, \ldots, T$, set $\epsilon_{\text{dp},t}$ such that $\tilde{\epsilon}_{\text{dp}} \leq \epsilon_{\text{dp}}$, where $\tilde{\epsilon}_{\text{dp}}$ is defined in (7), taking $\epsilon_t = \epsilon_{\text{dp},t}, \epsilon = \epsilon_{\text{dp}}, \delta = \delta_{\text{dp}}$.

3: **for** $t = 1 : T$ **do**

4: Compute the sensitivity vector $\tilde{\mathbf{s}}^{(t-1)} = [\tilde{s}_1^{(t-1)}, \ldots, \tilde{s}_m^{(t-1)}]^{\mathrm{T}}$, which is defined for all $i \in [m]$,

$$\tilde{s}_i^{(t-1)} = \max_{(\mathbf{w}_i')^{(t-1)}} \|\mathbf{w}_i^{(t-1)} - (\mathbf{w}_i')^{(t-1)}\|_2,$$

where $(\mathbf{w}_i')^{(t-1)}$ is assumed to be generated using $\mathcal{D}_i'$, which differ with $\mathcal{D}_i$ in a single data instance.

5: $\tilde{\mathbf{w}}_i^{(t-1)} = \mathbf{w}_i^{(t-1)} + \mathbf{b}_i$, where $\mathbf{b}_i$ is a sample with the density function of

$$p(\mathbf{b}_i) \propto \exp\left(-\frac{\tilde{s}_i^{(t-1)}}{\epsilon_{\text{dp},t}}\|\mathbf{b}_i\|_2\right),$$

for all $i \in [m]$.

6: Compute the sensitivity vector $\mathbf{s}^{(t-1)} = [s_1^{(t-1)}, \ldots, s_m^{(t-1)}]^{\mathrm{T}}$, which is defined for all $i \in [m]$,

$$s_i^{(t-1)} = \max_{(\mathbf{w}_i')^{(t-1)}} |\|\tilde{\mathbf{w}}_i^{(t-1)}\|_2^2 - \|(\tilde{\mathbf{w}}_i')^{(t-1)}\|_2^2|,$$

where $(\tilde{\mathbf{w}}_i')^{(t-1)}$ is assumed to be generated arbitrarily.

7: $\widetilde{\mathbf{\Sigma}}^{(t)} = \widetilde{\mathbf{W}}^{(t-1)}(\widetilde{\mathbf{W}}^{(t-1)})^{\mathrm{T}}$ (or $\widetilde{\mathbf{\Sigma}}^{(t)} = (\widetilde{\mathbf{W}}^{(t-1)})^{\mathrm{T}}\widetilde{\mathbf{W}}^{(t-1)}$).

8: $\mathbf{\Sigma}^{(t)} = \widetilde{\mathbf{\Sigma}}^{(t)} + \mathbf{E}$, where $\mathbf{E} \sim W_d(d+1, \frac{\max_i s_i^{(t-1)}}{2\epsilon_{\text{mp},t}}\mathbf{I}_d)$ (or $\mathbf{E} \sim W_m(m+1, \text{diag}(\mathbf{s}^{(t-1)}/2\epsilon_{\text{mp},t}))$) is a sample of the Wishart distribution.

9: Perform an arbitrary mapping $f : \mathbf{\Sigma}^{(1:t)} \to \mathbf{M}^{(t)}$.

10: $\hat{\mathbf{w}}_i^{(t)} = \mathcal{A}_{\text{st},i}(\mathbf{M}^{(t)}, \tilde{\mathbf{w}}_i^{(0:t-1)}, \mathbf{X}_i, \mathbf{y}_i)$ for all $i \in [m]$, where $\mathbf{w}_i^{(0:t-1)}$ are for the initialization.

11: Set the input for the next iteration: $\mathbf{W}^{(t)} = \widehat{\mathbf{W}}^{(t)}$.

12: **end for**

APPENDIX E
RESULTS OF UTILITY ANALYSES UNDER OTHER TWO SETTINGS

Here we consider the other two settings of $\{\epsilon_t\}$.

A. Setting No.1

In this setting, we have

$$\epsilon = \sum_{t=1}^{T} \epsilon_t.$$

Theorem 7 (Low rank - Convexity - Setting No.1). *Consider Algorithm 2. For an index $k \leq q$ that suffices the definition in Lemma 2 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LK\sqrt{m})$, assume $\epsilon_t \leq 4Kk^2 d(\log d)/q^2$ for $t \in [T]$.*

***No acceleration**: If we set $\beta_t = 0$ for $t \in [m]$, then setting*

$$T = \Theta\left(\left[\frac{(\alpha/2-1)^2|\alpha+1|\sqrt{m}\epsilon}{kd\log d}\right]^{\phi(\alpha)}\right)$$

for $\mathcal{E} = f(\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{W}}^{(t)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K^2 L\left[\frac{kd\log d}{(\alpha/2-1)^2|\alpha+1|\sqrt{m}\epsilon}\right]^{\phi(\alpha)}\right), \tag{25}$$

where

$$\phi(\alpha) = \begin{cases} 1/(\alpha+1), & \alpha > 2; \\ 1/3, & -1 < \alpha < 2; \\ 1/(2-\alpha), & \alpha < -1. \end{cases} \tag{26}$$

***Use acceleration**: If we set $\beta_t = (t-1)/(t+2)$ for $t \in [m]$, then setting*

$$T = \Theta\left(\left[\frac{(\alpha/2-2)^2|\alpha+1|\sqrt{m}\epsilon}{kd\log d}\right]^{\phi(\alpha)/2}\right)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K^2 L\left[\frac{kd\log d}{(\alpha/2-2)^2|\alpha+1|\sqrt{m}\epsilon}\right]^{\phi(\alpha)}\right), \tag{27}$$

where

$$\phi(\alpha) = \begin{cases} 2/(\alpha+1), & \alpha > 4; \\ 2/5, & -1 < \alpha < 4; \\ 2/(4-\alpha), & \alpha < -1. \end{cases} \tag{28}$$

Theorem 8 (Group sparse - Convexity - Setting No.1). *Consider Algorithm 3. For an index $k \leq d$ that suffices the definition in Lemma 3 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LKd\sqrt{m})$, assume $\epsilon_t \leq k^2\log(d)/4Kd(d-k)^2 m$ for $t \in [T]$.*

***No acceleration**: If we set $\beta_t = 0$ for $t \in [m]$, then setting*

$$T = \Theta\left(\left[\frac{(\alpha/2-1)^2|\alpha+1|Km\epsilon}{k\log d}\right]^{\phi(\alpha)}\right).$$

for $\mathcal{E} = f(\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{W}}^{(t)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K^2 L\left[\frac{k\log d}{(\alpha/2-1)^2|\alpha+1|Km\epsilon}\right]^{\phi(\alpha)}\right), \tag{29}$$

where $\phi(\alpha)$ is defined in (26).

***Use acceleration**: If we set $\beta_t = (t-1)/(t+2)$ for $t \in [m]$, then setting*

$$T = \Theta\left(\left[\frac{(\alpha/2-2)^2|\alpha+1|Km\epsilon}{k\log d}\right]^{\phi(\alpha)/2}\right).$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K^2 L\left[\frac{k\log d}{(\alpha/2-2)^2|\alpha+1|Km\epsilon}\right]^{\phi(\alpha)}\right), \tag{30}$$

where $\phi(\alpha)$ is defined in (28).

Now we further assume that $mf(\mathbf{W})$ is μ-strongly convex and has L-Lipschitz-continuous gradient, where $\mu < L$. We set $\epsilon_t = \Theta(Q^{-t})$ for $Q > 0$ and $t \in [T]$ for this case.

Theorem 9 (Low rank - Strong convexity - Setting No.1). *Consider Algorithm 2. For an index $k \leq q$ that suffices the definition in Lemma 2 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LK\sqrt{m})$, assume $\epsilon_t \leq 4Kk^2 d(\log d)/q^2$ for $t \in [T]$.*

***No acceleration**: If we set $\beta_t = 0$ for $t \in [m]$, then denoting $Q_0 = 1 - \mu/L$ and setting*

$$T = \Theta\left(\log_{1/\psi(Q,Q_0^2)}\left[\frac{(Q_0/\sqrt{Q}-1)^2|1-Q|\sqrt{m}\epsilon}{kd\log d}\right]\right)$$

for $\mathcal{E} = \frac{1}{\sqrt{m}}\|\widehat{\mathbf{W}}^{(T)} - \mathbf{W}_*\|_F$, we have with high probability,

$$\mathcal{E} = O\left(K\left[\frac{kd\log d}{(Q_0/\sqrt{Q}-1)^2|1-Q|\sqrt{m}\epsilon}\right]^{\log_{\psi(Q,Q_0^2)} Q_0}\right), \tag{31}$$

where $\psi(\cdot,\cdot)$ is defined in (19).
Use acceleration*: If we set $\beta_t = (1-\sqrt{\mu/L})/(1+\sqrt{\mu/L})$ for $t \in [m]$, then denoting $Q_0' = 1-\sqrt{\mu/L}$ and setting*

$$T = \Theta\bigg(\log_{1/\psi(Q,Q_0')}\bigg[\frac{(\sqrt{Q_0'}/\sqrt{Q}-1)^2|1-Q|\sqrt{m}\epsilon}{kd\log d}\bigg]\bigg)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K\bigg[\frac{kd\log d}{(\sqrt{Q_0'}/\sqrt{Q}-1)^2|1-Q|\sqrt{m}\epsilon}\bigg]^{\log_{\psi(Q,Q_0')}Q_0'}\bigg), \quad (32)$$

where $\psi(\cdot,\cdot)$ is defined in (19).

Theorem 10 (Group sparse - Strong convexity - Setting No.1)**.** *Consider Algorithm 3. For an index $k \le d$that suffices the definition in Lemma 3 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LKd\sqrt{m})$, assume $\epsilon_t \le k^2\log(d)/4Kd(d-k)^2m$ for $t \in [T]$.*
No acceleration*: If we set $\beta_t = 0$ for $t \in [m]$, then denoting $Q_0 = 1-\mu/L$ and setting*

$$T = \Theta\bigg(\log_{1/\psi(Q,Q_0^2)}\bigg[\frac{(Q_0/\sqrt{Q}-1)^2|1-Q|Km\epsilon}{k\log d}\bigg]\bigg)$$

for $\mathcal{E} = \frac{1}{\sqrt{m}}\|\widehat{\mathbf{W}}^{(T)} - \mathbf{W}_*\|_F$, we have with high probability,

$$\mathcal{E} = O\bigg(K\bigg[\frac{k\log d}{(Q_0/\sqrt{Q}-1)^2|1-Q|Km\epsilon}\bigg]^{\log_{\psi(Q,Q_0^2)}Q_0}\bigg), \quad (33)$$

where $\psi(\cdot,\cdot)$ is defined in (19).
Use acceleration*: If we set $\beta_t = (1-\sqrt{\mu/L})/(1+\sqrt{\mu/L})$ for $t \in [m]$, then denoting $Q_0' = 1-\sqrt{\mu/L}$ and setting*

$$T = \Theta\bigg(\log_{1/\psi(Q,Q_0')}\bigg[\frac{(\sqrt{Q_0'}/\sqrt{Q}-1)^2|1-Q|Km\epsilon}{k\log d}\bigg]\bigg)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K\bigg[\frac{k\log d}{(\sqrt{Q_0'}/\sqrt{Q}-1)^2|1-Q|Km\epsilon}\bigg]^{\log_{\psi(Q,Q_0')}Q_0'}\bigg), \quad (34)$$

where $\psi(\cdot,\cdot)$ is defined in (19).

Then we optimize the utility bounds with respect to the respective budget allocation strategies.

Theorem 11 (Budget allocation - Setting No.1)**.** *Consider Algorithm 2 and Algorithm 3.*

For convex *f, use Theorem 7 and Theorem 8.*

(1) No acceleration*: Both the bounds in (25) and (29) achieve their respective minimums w.r.t. α at $\alpha = 0$. Meanwhile, $\phi(\alpha) = 1/3$.*

(2) Accelerated*: Both the bounds in (27) and (30) achieve their respective minimums w.r.t. α at $\alpha = 2/3$. Meanwhile, $\phi(\alpha) = 2/5$.*

For strongly convex *f, use Theorem 9 and Theorem 10.*

(1) No acceleration*: Both the bounds in (31) and (33) achieve their respective minimums w.r.t. Q at $Q = Q_0^{2/3}$. Meanwhile, $\log_{\psi(Q,Q_0^2)} Q_0 = 1/2$.*

(2) Accelerated*: Both the bounds in (32) and (34) achieve their respective minimums w.r.t. Q at $Q = (Q_0')^{1/3}$. Meanwhile, $\log_{\psi(Q,Q_0')} Q_0' = 1$.*

B. Setting No.2

In this setting, we have

$$\epsilon = \sum_{t=1}^{T}\frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^{T}2\epsilon_t^2\log\Big(\frac{1}{\delta}\Big)}.$$

Theorem 12 (Low rank - Convexity - Setting No.2)**.** *Consider Algorithm 2. For an index $k \le q$ that suffices the definition in Lemma 2 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LK\sqrt{m})$, assume $\epsilon_t \le 4Kk^2d(\log d)/q^2$ for $t \in [T]$.*
No acceleration*: If we set $\beta_t = 0$ for $t \in [m]$, then setting*

$$T = \Theta\bigg(\bigg[\frac{(\alpha/2-1)^2\sqrt{|2\alpha+1|}\sqrt{m}\epsilon}{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}\bigg]^{\phi(\alpha)}\bigg)$$

for $\mathcal{E} = f(\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{W}}^{(t)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K^2L\bigg[\frac{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\alpha/2-1)^2\sqrt{|2\alpha+1|}\sqrt{m}\epsilon}\bigg]^{\phi(\alpha)}\bigg), \quad (35)$$

where

$$\phi(\alpha) = \begin{cases} 2/(2\alpha+1), & \alpha > 2; \\ 2/5, & -1/2 < \alpha < 2; \\ 1/(2-\alpha), & \alpha < -1/2. \end{cases} \quad (36)$$

Use acceleration*: If we set $\beta_t = (t-1)/(t+2)$ for $t \in [m]$, then setting*

$$T = \Theta\bigg(\bigg[\frac{(\alpha/2-2)^2\sqrt{|2\alpha+1|}\sqrt{m}\epsilon}{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}\bigg]^{\phi(\alpha)/2}\bigg)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K^2L\bigg[\frac{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\alpha/2-2)^2\sqrt{|2\alpha+1|}\sqrt{m}\epsilon}\bigg]^{\phi(\alpha)}\bigg), \quad (37)$$

where

$$\phi(\alpha) = \begin{cases} 4/(2\alpha+1), & \alpha > 4; \\ 4/9, & -1/2 < \alpha < 4; \\ 2/(4-\alpha), & \alpha < -1/2. \end{cases} \quad (38)$$

Theorem 13 (Group sparse - Convexity - Setting No.2)**.** *Consider Algorithm 3. For an index $k \le d$ that suffices the definition in Lemma 3 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LKd\sqrt{m})$, assume $\epsilon_t \le k^2\log(d)/4Kd(d-k)^2m$ for $t \in [T]$.*
No acceleration*: If we set $\beta_t = 0$ for $t \in [m]$, then setting*

$$T = \Theta\bigg(\bigg[\frac{(\alpha/2-1)^2\sqrt{|2\alpha+1|}Km\epsilon}{k\log d\sqrt{\log(1/\delta)+2\epsilon}}\bigg]^{\phi(\alpha)}\bigg).$$

for $\mathcal{E} = f(\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{W}}^{(t)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K^2L\bigg[\frac{k\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\alpha/2-1)^2\sqrt{|2\alpha+1|}Km\epsilon}\bigg]^{\phi(\alpha)}\bigg), \quad (39)$$

where $\phi(\alpha)$ is defined in (36).
Use acceleration*: If we set $\beta_t = (t-1)/(t+2)$ for $t \in [m]$, then setting*

$$T = \Theta\bigg(\bigg[\frac{(\alpha/2-2)^2\sqrt{|2\alpha+1|}Km\epsilon}{k\log d\sqrt{\log(1/\delta)+2\epsilon}}\bigg]^{\phi(\alpha)/2}\bigg).$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\bigg(K^2L\bigg[\frac{k\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\alpha/2-2)^2\sqrt{|2\alpha+1|}Km\epsilon}\bigg]^{\phi(\alpha)}\bigg), \quad (40)$$

where $\phi(\alpha)$ is defined in (38).

Now we further assume that $mf(\mathbf{W})$ is μ-strongly convex and has L-Lipschitz-continuous gradient, where $\mu < L$. We set $\epsilon_t = \Theta(Q^{-t})$ for $Q > 0$ and $t \in [T]$ for this case.

Theorem 14 (Low rank - Strong convexity - Setting No.2)**.** *Consider Algorithm 2. For an index $k \le q$ that suffices the definition in Lemma 2 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LK\sqrt{m})$, assume $\epsilon_t \le 4Kk^2d(\log d)/q^2$ for $t \in [T]$.*

***No acceleration**: If we set $\beta_t = 0$ for $t \in [m]$, then denoting $Q_0 = 1 - \mu/L$ and setting*

$$T = \Theta\left(\log_{1/\psi(Q,Q_0^2)}\left[\frac{(Q_0/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}\sqrt{m}\epsilon}{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}\right]\right)$$

for $\mathcal{E} = \frac{1}{\sqrt{m}}\|\widehat{\mathbf{W}}^{(T)} - \mathbf{W}_*\|_F$, we have with high probability,

$$\mathcal{E} = O\left(K\left[\frac{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}{(Q_0/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}\sqrt{m}\epsilon}\right]^{\log_{\psi(Q,Q_0^2)} Q_0}\right), \tag{41}$$

where $\psi(\cdot,\cdot)$ is defined in (19).
***Use acceleration**: If we set $\beta_t = (1-\sqrt{\mu/L})/(1+\sqrt{\mu/L})$ for $t \in [m]$, then denoting $Q_0' = 1 - \sqrt{\mu/L}$ and setting*

$$T = \Theta\left(\log_{1/\psi(Q,Q_0')}\left[\frac{(\sqrt{Q_0'}/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}\sqrt{m}\epsilon}{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}\right]\right)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K\left[\frac{kd\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\sqrt{Q_0'}/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}\sqrt{m}\epsilon}\right]^{\log_{\psi(Q,Q_0')} Q_0'}\right), \tag{42}$$

where $\psi(\cdot,\cdot)$ is defined in (19).

Theorem 15 (Group sparse - Strong convexity - Setting No.2)**.** *Consider Algorithm 3. For an index $k \le d$that suffices the definition in Lemma 3 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LKd\sqrt{m})$, assume $\epsilon_t \le k^2\log(d)/4Kd(d-k)^2 m$ for $t \in [T]$.*
***No acceleration**: If we set $\beta_t = 0$ for $t \in [m]$, then denoting $Q_0 = 1 - \mu/L$ and setting*

$$T = \Theta\left(\log_{1/\psi(Q,Q_0^2)}\left[\frac{(Q_0/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}Km\epsilon}{k\log d\sqrt{\log(1/\delta)+2\epsilon}}\right]\right)$$

for $\mathcal{E} = \frac{1}{\sqrt{m}}\|\widehat{\mathbf{W}}^{(T)} - \mathbf{W}_*\|_F$, we have with high probability,

$$\mathcal{E} = O\left(K\left[\frac{k\log d\sqrt{\log(1/\delta)+2\epsilon}}{(Q_0/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}Km\epsilon}\right]^{\log_{\psi(Q,Q_0^2)} Q_0}\right), \tag{43}$$

where $\psi(\cdot,\cdot)$ is defined in (19).
***Use acceleration**: If we set $\beta_t = (1-\sqrt{\mu/L})/(1+\sqrt{\mu/L})$ for $t \in [m]$, then denoting $Q_0' = 1 - \sqrt{\mu/L}$ and setting*

$$T = \Theta\left(\log_{1/\psi(Q,Q_0')}\left[\frac{(\sqrt{Q_0'}/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}Km\epsilon}{k\log d\sqrt{\log(1/\delta)+2\epsilon}}\right]\right)$$

for $\mathcal{E} = f(\widehat{\mathbf{W}}^{(T)}) - f(\mathbf{W}_*)$, we have with high probability,

$$\mathcal{E} = O\left(K\left[\frac{k\log d\sqrt{\log(1/\delta)+2\epsilon}}{(\sqrt{Q_0'}/\sqrt{Q}-1)^2\sqrt{|1-Q^2|}Km\epsilon}\right]^{\log_{\psi(Q,Q_0')} Q_0'}\right), \tag{44}$$

where $\psi(\cdot,\cdot)$ is defined in (19).

Then we optimize the utility bounds with respect to the respective budget allocation strategies.

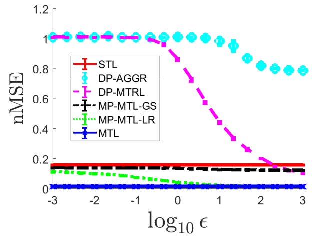
Theorem 16 (Budget allocation - Setting No.2)**.** *Consider Algorithm 2 and Algorithm 3.*

For convex f*, use Theorem 12 and Theorem 13.*

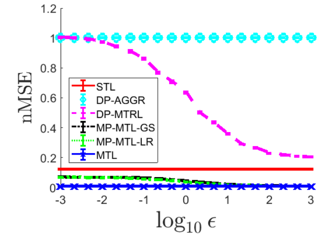
(1) No acceleration*: Both the bounds in (35) and (39) achieve their respective minimums w.r.t. α at $\alpha = 0$. Meanwhile, $\phi(\alpha) = 2/5$.*

(2) Accelerated*: Both the bounds in (37) and (40) achieve their respective minimums w.r.t. α at $\alpha = 2/5$. Meanwhile, $\phi(\alpha) = 4/9$.*

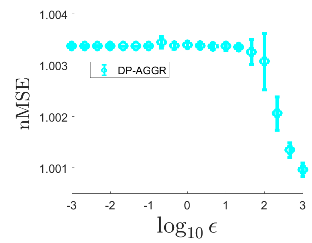
For strongly convex f*, use Theorem 14 and Theorem 15.*



(a) Low-rank pattern



(b) Group-sparse pattern



(c) Group-sparse pattern

Figure 1. Privacy-accuracy tradeoff on synthetic datasets. For (a), the data that associated with the low-rank model matrix were used; for (b) and (c), the data that associated with the group-sparse model matrix were used. In (c), the plot shows the same performances of DP-AGGR as in (b) with a finer vertical axis. MP-MTL-LR denotes Algorithm 2, MP-MTL-GS denotes Algorithm 3, and STL denotes the ℓ_2-norm-penalized method. In both panels, STL and MTL denote non-private methods.

(1) No acceleration*: Both the bounds in (41) and (43) achieve their respective minimums w.r.t. Q at $Q = Q_0^{2/5}$. Meanwhile, $\log_{\psi(Q,Q_0^2)} Q_0 = 1/2$.*

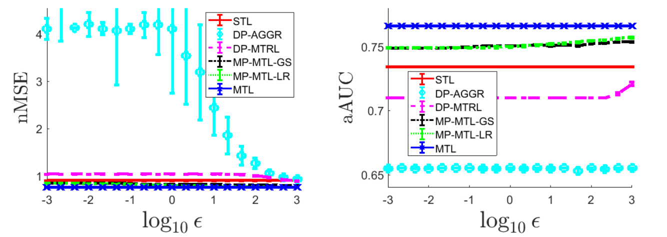
(2) Accelerated*: Both the bounds in (42) and (44) achieve their respective minimums w.r.t. Q at $Q = (Q_0')^{1/5}$. Meanwhile, $\log_{\psi(Q,Q_0')} Q_0' = 1$.*

Appendix F
Detailed Privacy-Accuracy Tradeoff for Baseline Methods on Synthetic Datasets

In Fig. 1, the detailed performances of both of DP-MTRL and DP-AGGR are shown. Note that we have tuned the regularization parameters for both DP-MTRL and DP-AGGR for acceptable accuracies. Our Algorithm 2 outperforms DP-MTRL and DP-AGGR. In Fig. 1 (a), DP-MTRL outperforms the STL method and our Algorithm 3 when ϵ is large, because it suits the true model matrix, in which the relatedness among tasks is modeled by a graph. However, the true model matrix is not group-sparse, hence our Algorithm 3 underperforms comparing with Algorithm 2 and DP-MTRL when ϵ is large. By contrast, in Fig. 1 (b), the true model matrix is group-sparse and is not suitable for DP-MTRL, hence DP-MTRL underperforms comparing with the STL method even when ϵ is large. Fig. 1 (c) is used to show that the accuracy of DP-AGGR grows with ϵ under the same setting as in Fig. 1 (b). As we discussed, DP-AGGR only performs model-averaging, which is not suitable for the true model matrices in both settings of Fig. 1 (a) and (b), hence the accuracies of DP-AGGR are much worse than those of the respective STL methods.

Appendix G
Detailed Privacy-Accuracy Tradeoff for DP-AGGR on Real-World Datasets

In Fig. 2, the detailed performances of DP-AGGR are shown. Because the dimension is large and the number of tasks is not sufficient, the accuracy of DP-AGGR barely grows with ϵ; other



(a) School Data (b) Patient Survey Data

Figure 2. Privacy-accuracy tradeoff on real-world datasets. In both panels, MTL denotes the method with the best performance among the four non-private MTL methods proposed by Ji and Ye [6], Liu et al. [9], Zhang and Yeung [13] and DP-AGGR without perturbations; MP-MTL-LR denotes Algorithm 2, whereas MP-MTL-GS denotes Algorithm 3; STL denotes the method with the better performance between the ℓ_1- and ℓ_2-regularized methods.

private-preserving methods, such as DP-MTRL and our algorithms, grow slowly with ϵ as well.

Appendix H
Varying Training-Data Percentage

Since the MTL behavior may change when the training-data percentage (the size of the training data divided by the size of the entire dataset) changes, we evaluated the methods on both real-world datasets at different training-data percentages. Here, we present the results mostly for our low-rank algorithm (denoted by MP-MTL-LR) because it always outperforms our group-sparse algorithm (MP-MTL-GS) in the above experiments. The results corresponding to School Data are shown in Fig. 3; the results corresponding to Patient Survey Data are shown in Fig. 4. From those plots, we observe that on both real-world datasets, our MP-MTL method behaves similarly at different training-data percentages and outperforms DP-MTRL and DP-AGGR, especially when ϵ is small.

Appendix I
Lemmas of Differential Privacy

- Post-Processing immunity. This property helps us safely use the output of a differentially private algorithm without additional information leaking, as long as we do not touch the dataset $\mathcal{D}$ again.

  Lemma 4 (Post-Processing immunity. Proposition 2.1 in Dwork et al. [3]). *Let algorithm $\mathcal{A}_1(\mathcal{B}_1) : \mathcal{D} \to \theta_1 \in \mathcal{C}_1$ be an (ϵ, δ) - differential privacy algorithm, and let $f : \mathcal{C}_1 \to \mathcal{C}_2$ be an arbitrary mapping. Then, algorithm $\mathcal{A}_2(\mathcal{B}_2) : \mathcal{D} \to \theta_2 \in \mathcal{C}_2$ is still (ϵ, δ) - differentially private, i.e., for any set $\mathcal{S} \subseteq \mathcal{C}_2$,*

$$\mathbb{P}(\theta_2 \in \mathcal{S} \mid \mathcal{B}_2 = \mathcal{D}) \leq e^{\epsilon}\mathbb{P}(\theta_2 \in \mathcal{S} \mid \mathcal{B}_2 = \mathcal{D}') + \delta.$$

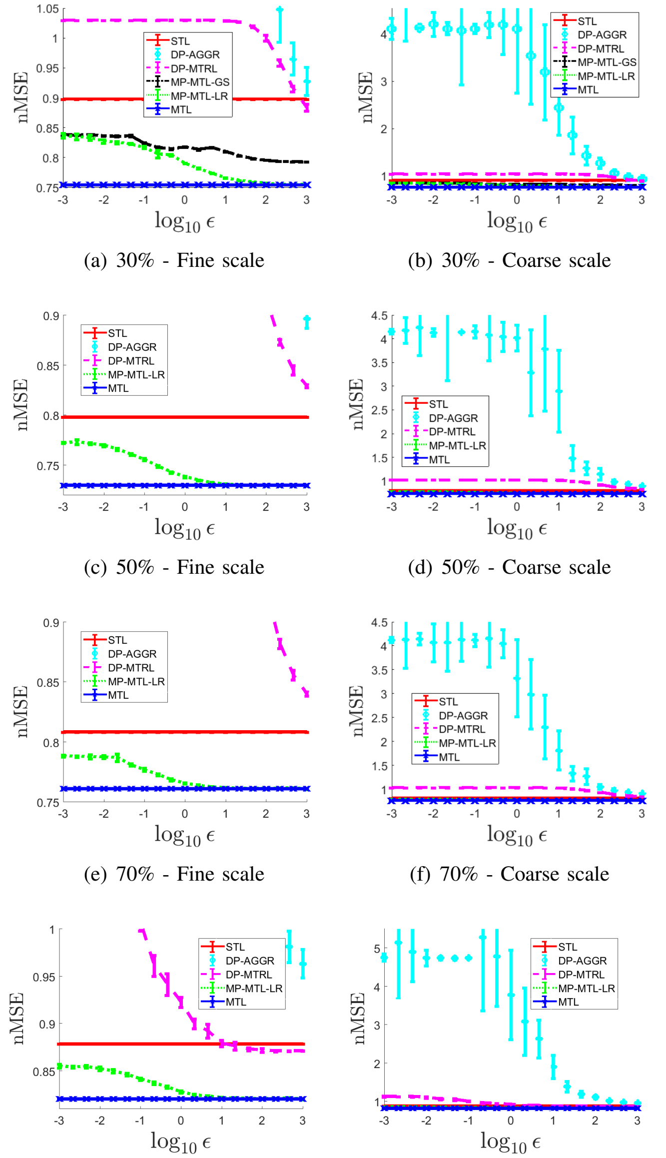
- Group privacy. This property guarantees the graceful increment of the privacy budget when more output variables need differentially private protection.

  Lemma 5 (Group privacy. Lemma 2.2 in Vadhan [12]). *Let algorithm $\mathcal{A}(\mathcal{B}) : \mathcal{D} \to \theta \in \mathcal{C}$ be an (ϵ, δ) - differential privacy algorithm. Then, considering two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in k entries, the algorithm satisfies for any set $\mathcal{S} \subseteq \mathcal{C}$*

$$\mathbb{P}(\theta \in \mathcal{S} \mid \mathcal{B} = \mathcal{D}) \leq e^{k\epsilon}\mathbb{P}(\theta \in \mathcal{S} \mid \mathcal{B} = \mathcal{D}') + ke^{k\epsilon}\delta.$$

- Combination. This property guarantees the linear incrementing of the privacy budget when the dataset $\mathcal{D}$ is repeatedly used.

  Lemma 6 (Combination. Theorem 3.16 in Dwork et al. [3]). *Let algorithm $\mathcal{A}_i : \mathcal{D} \to \theta_i \in \mathcal{C}_i$ be an (ϵ_i, δ_i) - differential privacy algorithm for all $i \in [k]$. Then, for $\mathcal{A}_{[k]} : \mathcal{D} \to$*



(a) 30% - Fine scale (b) 30% - Coarse scale

(c) 50% - Fine scale (d) 50% - Coarse scale

(e) 70% - Fine scale (f) 70% - Coarse scale

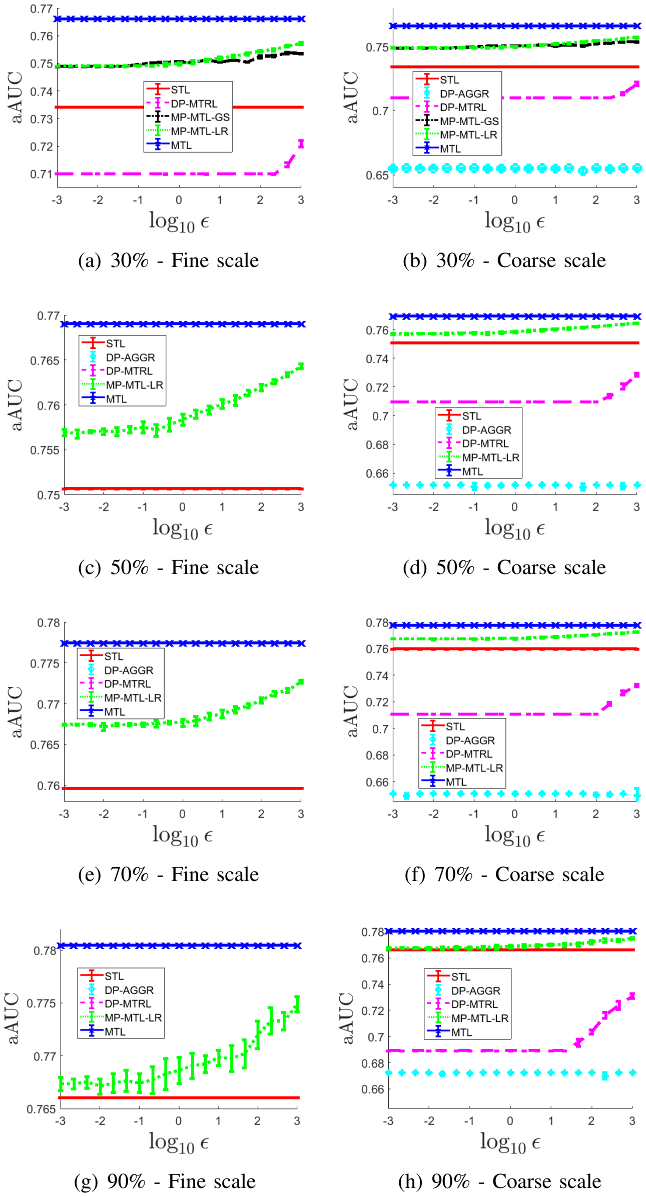
(g) 90% - Fine scale (h) 90% - Coarse scale

Figure 3. Privacy-accuracy tradeoff on School Data. (a) and (b) correspond to a training-data percentage of 30%, (c) and (d) correspond to a training-data percentage of 50%, (e) and (f) correspond to a training-data percentage of 70%, (g) and (h) correspond to a training-data percentage of 90%. (a), (c), (e) and (g) use fine scales of vertical axes to focus on the performances of our algorithms; (b), (d), (f) and (h) use coarse scales of vertical axes to focus on the baseline algorithms. In all the panels, MTL denotes the method with the best performance among the four non-private MTL methods proposed by Ji and Ye [6], Liu et al. [9], Zhang and Yeung [13] and DP-AGGR without perturbations; MP-MTL-LR denotes Algorithm 2, whereas MP-MTL-GS denotes Algorithm 3; STL denotes the method with the better performance between the ℓ_1- and ℓ_2-regularized methods.

$(\theta_1, \theta_2, \cdots, \theta_k) \in \bigotimes_{j=1}^{k} \mathcal{C}_j$ is a $(\sum_i \epsilon_i, \sum_i \delta_i)$ - differentially private algorithm.

- Adaptive composition. This property guarantees privacy when an iterative algorithm is adopted on *different* datasets that may nevertheless contain information relating to the same individual.

  Lemma 7 (Adaptive composition. Directly taken Theorem 3.5 in Kairouz et al. [8]). *Let algorithm $\mathcal{A}_1(\mathcal{B}_1) : \mathcal{D}_1 \to \theta_1$ be an (ϵ_1, δ_1) - differential privacy algorithm, and for $t = 2, \ldots, T$, let $\mathcal{A}_t(\mathcal{B}_t) : (\mathcal{D}_t, \theta_1, \theta_2, \cdots, \theta_{t-1}) \to \theta_t \in \mathcal{C}_t$ be (ϵ_t, δ_t) -*



(a) 30% - Fine scale (b) 30% - Coarse scale
(c) 50% - Fine scale (d) 50% - Coarse scale
(e) 70% - Fine scale (f) 70% - Coarse scale
(g) 90% - Fine scale (h) 90% - Coarse scale

Figure 4. Privacy-accuracy tradeoff on Patient Survey Data. (a) and (b) correspond to a training-data percentage of 30%, (c) and (d) correspond to a training-data percentage of 50%, (e) and (f) correspond to a training-data percentage of 70%, (g) and (h) correspond to a training-data percentage of 90%. (a), (c), (e) and (g) use fine scales of vertical axes to focus on the performances of our algorithms; (b), (d), (f) and (h) use coarse scales of vertical axes to focus on the baseline algorithms. In all the panels, MTL denotes the method with the best performance among the four non-private MTL methods proposed by Ji and Ye [6], Liu et al. [9], Zhang and Yeung [13] and DP-AGGR without perturbations; MP-MTL-LR denotes Algorithm 2, whereas MP-MTL-GS denotes Algorithm 3; STL denotes the method with the better performance between the ℓ_1- and ℓ_2-regularized methods.

differentially private for all given $(\theta_1, \theta_2, \cdots, \theta_{t-1}) \in \bigotimes_{t'=1}^{t-1} \mathcal{C}_{t'}$. Then, for all neighboring datasets $\mathcal{D}_t$ and $\mathcal{D}'_t$ that differ in a single entry relating to the same individual and for any set $\mathcal{S} \subseteq \bigotimes_{t=1}^{T} \mathcal{C}_t$,

$$
\begin{aligned}
&\mathbb{P}((\theta_1, \cdots, \theta_T) \in \mathcal{S} \mid \bigcap_{t=1}^{T} (\mathcal{B}_t = (\mathcal{D}_t, \boldsymbol{\theta}_{1:t-1}))) \\
&\leq e^{\epsilon} \mathbb{P}((\theta_1, \cdots, \theta_T) \in \mathcal{S} \mid \bigcap_{t=1}^{T} (\mathcal{B}_t = (\mathcal{D}'_t, \boldsymbol{\theta}_{1:t-1}))) \\
&+ 1 - (1-\delta) \prod_{t=1}^{T} (1-\delta_t),
\end{aligned}
\tag{45}
$$

where

$$
\boldsymbol{\theta}_{1:t-1} = \begin{cases} \emptyset, & t = 1 \\ \theta_1, \theta_2, \cdots, \theta_{t-1}, & t \geq 2, \end{cases}
$$

and

$$
\begin{aligned}
\epsilon = \min \Bigg\{ &\sum_{t=1}^{T} \epsilon_t, \sum_{t=1}^{T} \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^{T} 2\epsilon_t^2 \log\left(\frac{1}{\delta}\right)}, \\
&\sum_{t=1}^{T} \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^{T} 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^{T} \epsilon_t^2}}{\delta}\right)} \Bigg\}.
\end{aligned}
$$

Appendix J
Lemmas for Privacy Guarantees

The following lemma shows that STL algorithms do not increase the privacy budget when they are concatenated with an MTL algorithm.

Lemma 8. *For an (ϵ, δ) - non-iterative MP-MTL algorithm $\mathcal{A}_{mp} : (\mathbf{W} \in \mathbb{R}^{d\times m}, \mathcal{D}^m) \to \widetilde{\mathbf{W}} \in \mathbb{R}^{d\times m}$ and any STL algorithm $\mathcal{A}_{st} : (\widetilde{\mathbf{W}}, \mathcal{D}^m) \to \widehat{\mathbf{W}} \in \mathbb{R}^{d\times m}$, an algorithm $\mathcal{A}_{mp+st} : (\mathbf{W} \in \mathbb{R}^{d\times m}, \mathcal{D}^m) \to \widehat{\mathbf{W}} \in \mathbb{R}^{d\times m}$ that first uses $\mathcal{A}_{mp}$ before applying $\mathcal{A}_{st}$ is still an (ϵ, δ) - non-iterative MP-MTL algorithm. Moreover, an algorithm $\mathcal{A}_{st+mp}$ that first uses a* deterministic *STL algorithm before applying an (ϵ, δ) - non-iterative MP-MTL algorithm is also an (ϵ, δ) - non-iterative MP-MTL algorithm.*

The following result shows that adopting a series of Non-iterative MP-MTL algorithms defined in Definition 5 iteratively, we can develop an Iterative MP-MTL algorithm, as described in Algorithm 6.

Algorithm 6 Iterative MP-MTL build by Non-iterative MP-MTL

Input: Datasets $(\mathbf{X}^m, \mathbf{y}^m) = \{(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_m, \mathbf{y}_m)\}$, where $\forall i \in [m]$, $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ and $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$. Number of iterations T. Privacy loss $\{\epsilon_t, \delta_t\}_{t=1,\ldots,T}, \delta \geq 0$. Initial models of tasks $\mathbf{W}^{(0)}$.
Output: $\widehat{\mathbf{W}}^{(1:T)}$.
1: **for** $t = 1 : T$ **do**
2: $\widehat{\mathbf{W}}^{(t)} = \mathcal{A}_{\mathrm{mp}}(\mathbf{W}^{(t-1)}, \mathbf{X}^m, \mathbf{y}^m)$, where $\mathcal{A}_{\mathrm{mp}}$ denotes an (ϵ_t, δ_t)-Non-iterative MP-MTL algorithm.
3: $\mathbf{w}_i^{(t)} = \mathcal{A}_{\mathrm{st},i}(\hat{\mathbf{w}}_i^{(t)}, \mathbf{X}_i, \mathbf{y}_i)$, for $i = 1, \ldots, m$, where $\mathcal{A}_{\mathrm{st},i}$ denotes a deterministic STL algorithm for the i-th task.
4: **end for**

Lemma 9. *Use Lemmas 7 and 8. Algorithm 6 is an $(\epsilon, 1 - (1-\delta)\prod_{t=1}^{T}(1-\delta_t))$ - iterative MP-MTL algorithm, where ϵ is defined in Lemma 7.*

APPENDIX K
LEMMAS FOR UTILITY ANALYSIS

Lemma 10. *For a integer $T \geq 1$, a constant $\alpha \in \mathbb{R}$, by EulerMaclaurin formula [1], we have*

$$\sum_{t=1}^{T} t^{\alpha} = \begin{cases} O(T^{\alpha+1}/(\alpha+1)), & \alpha > -1; \\ O(1/(-\alpha-1)), & \alpha < -1. \end{cases}$$

Proof. This is the direct result of EulerMaclaurin formula [1]. □

Lemma 11. *For a integer $T \geq 1$ and a constant $Q > 0$, we have*

$$\sum_{t=1}^{T} Q^{-t} \leq \begin{cases} \frac{1}{Q-1}, & Q > 1; \\ \frac{Q^{-T}}{1-Q}, & Q < 1. \end{cases}$$

Proof. Because $\sum_{t=1}^{T} Q^{-t} = Q^{-1}\frac{1-Q^{-T}}{1-Q^{-1}}$, we complete the proof. □

Lemma 12. *For a constant $c_1, c_2 > 0$, a constant $\epsilon_0 > 0$, a integer $T \geq 1$, a mapping $s : t \in [T] \to s(t) > 0$, a mapping $S_1 : T \to S_1(T) > 0$ and a mapping $S_2 : T \to S_2(T) > 0$, then if $\sum_{t=1}^{T} \epsilon_0 s(t) \geq c_1$ and $\sum_{t=1}^{T} s(t) \leq S_1(T)$, we have*

$$1/\epsilon_0 \leq S_1(T)/c_1.$$

On the other hand, if $\sqrt{\sum_{t=1}^{T} \epsilon_0^2 s^2(t)} \geq c_2$ and $\sum_{t=1}^{T} s^2(t) \leq S_2(T)$, we have

$$1/\epsilon_0 \leq \sqrt{S_2(T)}/c_2.$$

Proof. If $\sum_{t=1}^{T} \epsilon_0 s(t) \geq c_1$, $1/\epsilon_0 \leq \sum_{t=1}^{T} s(t)/c_1 \leq S_1(T)/c_1$.

On the other hand, if $\sqrt{\sum_{t=1}^{T} \epsilon_0^2 s^2(t)} \geq c_2$, $1/\epsilon_0 \leq \sqrt{\sum_{t=1}^{T} s^2(t)}/c_2 \leq \sqrt{S_2(T)}/c_2$. □

Lemma 13. *Consider Algorithm 2. For an index $k \leq q$ that suffices the definition in Lemma 2 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LK\sqrt{m})$, set $\epsilon_t \leq 4Kk^2 d(\log d)/q^2$ for $t \in [T]$. Assume in each iteration, $\mathbf{E}$ is the defined Wishart random matrix. We have with probability at least $1 - d^{-c}$ for some constant $c > 1$ that*

$$\begin{aligned} \varepsilon_t =& \frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_* \\ &- \left\{\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_*\right\} \\ =& O\left(\frac{K^2\sqrt{m}kd\log d}{\eta\epsilon_t}\right). \end{aligned} \tag{46}$$

Proof. First, using Lemma 1 of Jiang et al. [7], we have in the t-th step, with probability at least $1 - d^{-c}$ for some constant $c > 1$,

$$\sigma_1(\mathbf{E}) = O\left(d(\log d)\sigma_1\left(\frac{\max_i s_i^{(t-1)}}{2\epsilon_t}\mathbf{I}_d\right)\right) = O(d(\log d)K/\epsilon_t).$$

We also have $\sigma_1(\mathbf{C}) \leq \|\mathbf{C}\|_F \leq \sqrt{m}\max_i \|\mathbf{C}_i\|_2 \leq K\sqrt{m}$, where $\mathbf{C}_i$ is the i-th column of $\mathbf{C}$.

As such, by Lemma 2, in the t-th iteration, for $\epsilon_t \leq 4Kk^2 d(\log d)/q^2$, where $q = \min\{d, m\}$, we have

$$\begin{aligned} \varepsilon_t =& \frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_* \\ &- \left\{\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_*\right\} \\ \leq& \frac{1}{\eta}\left(\frac{\sigma_1^2(\mathbf{C})}{\eta\lambda} + \sigma_1(\mathbf{C})\right)\left[k\frac{\sigma_1(\mathbf{E})}{2\eta\lambda}\right. \\ &\left. + (r_c - k)I(r_c > k)\sqrt{\sigma_1(\mathbf{E})} + \left(\frac{k(k-1)}{\eta\lambda} + 2k\right)\sigma_1(\mathbf{E})\right] \\ \leq& \frac{1}{\eta}\left(\frac{K^2 m}{\eta\lambda} + K\sqrt{m}\right)\left[k\frac{\sigma_1(\mathbf{E})}{2\eta\lambda}\right. \\ &\left. + q\sqrt{\sigma_1(\mathbf{E})} + \left(\frac{k(k-1)}{\eta\lambda} + 2k\right)\sigma_1(\mathbf{E})\right] \\ =& O\left(\frac{1}{\eta}\left(\frac{K^2 m}{\eta\lambda} + K\sqrt{m}\right)\left(\frac{k^2}{\eta\lambda} + 2k\right)\frac{d(\log d)K}{\epsilon_t}\right), \end{aligned}$$

where in the second inequality, the terms with $\sigma_1(\mathbf{E})$ nominate due to the condition on ϵ_t.

Further assuming $\eta = 1/L$ and $\lambda = \Theta(LK\sqrt{m})$, we complete the proof. □

Lemma 14. *Consider Algorithm 3. For an index $k \leq d$ that suffices the definition in Lemma 3 for all $t \in [T]$, $\eta = 1/L$, $\lambda = \Theta(LKd\sqrt{m})$, set $\epsilon_t \leq k^2\log(d)/4Kd(d-k)^2 m$ for $t \in [T]$. Assume in each iteration, $\mathbf{E}$ is the defined Wishart random matrix. We have with probability at least $1 - d^{-c}$ for some constant $c > 1$ that*

$$\begin{aligned} \varepsilon_t =& \frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_{2,1} \\ &- \left\{\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}\right\} \\ =& O\left(\frac{Kk\log d}{\eta\epsilon_t}\right). \end{aligned} \tag{47}$$

Proof. Similarly as in proof for Lemma 13, by Lemma 3, in the t-th iteration, we have

$$\begin{aligned} \varepsilon_t =& \frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_{2,1} \\ &- \left\{\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}\right\} \\ \leq& \frac{1}{\eta}\left[\frac{r_{c,s}}{\eta\lambda}\left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)^2 + \left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)\right] \\ &\cdot\left[\frac{k}{2\eta\lambda}\max_{j:\eta^2\lambda^2 \leq \mathbf{\Sigma}_{jj,0}}|\mathbf{E}_{jj}|\right. \\ &\left. + (r_{c,s} - k)I(r_{c,s} > k)\max_{j:\eta^2\lambda^2 > \mathbf{\Sigma}_{jj,0}}\sqrt{|\mathbf{E}_{jj}|}\right] \\ \leq& \frac{1}{\eta}\left[\frac{r_{c,s}}{\eta\lambda}\|\mathbf{C}\|_F^2 + \|\mathbf{C}\|_F\right] \\ &\cdot\left[\frac{k}{2\eta\lambda}\sigma_1(\mathbf{E}) + (r_{c,s} - k)I(r_{c,s} > k)\sqrt{\sigma_1(\mathbf{E})}\right] \\ \leq& \frac{1}{\eta}\left(\frac{dK^2 m}{\eta\lambda} + K\sqrt{m}\right)\left[k\frac{\sigma_1(\mathbf{E})}{2\eta\lambda} + (d-k)\sqrt{\sigma_1(\mathbf{E})}\right]. \end{aligned}$$

Further setting $\eta = 1/L$ and $\lambda = \Theta(LKd\sqrt{m})$, assuming $\epsilon_t \leq k^2\log(d)/4Kd(d-k)^2 m$, we have

$$\begin{aligned} \varepsilon_t =& O\left(\frac{1}{\eta}\left(\frac{dK^2 m}{\eta\lambda} + K\sqrt{m}\right)\frac{k}{\eta\lambda}\frac{d(\log d)K}{\epsilon_t}\right) \\ =& O\left(\frac{Kk\log d}{\eta\epsilon_t}\right). \end{aligned} \tag{48}$$

□

Lemma 15. *For matrices $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{W} \subset \mathbb{R}^{d\times m}$, we have*

$$\|\mathbf{W}_1 - \mathbf{W}_2\|_F = O(K\sqrt{m}).$$

Proof. Because $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{W}$, $\max_{i\in[m]} \|\mathbf{w}_{i,1}\|_2 \le K$. Therefore,

$$\|\mathbf{W}_1 - \mathbf{W}_2\|_F \le 2\|\mathbf{W}_1\|_F \le 2\sqrt{m}\max_{i\in[m]} \|\mathbf{w}_{i,1}\|_2 \le 2K\sqrt{m}.$$

□

Lemma 16. *For constants $\epsilon, \delta \ge 0$, a integer $T \ge 1$, a series constants $\epsilon_t > 0$ for $t \in [T]$, then if*

$$\epsilon = \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(\frac{1}{\delta}\right)},$$

we have

$$\sqrt{\sum_{t=1}^T \epsilon_t^2} \ge \frac{\sqrt{2}\epsilon}{2\sqrt{\log(1/\delta)+2\epsilon}}.$$

On the other hand, if

$$\epsilon = \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^T \epsilon_t^2}}{\delta}\right)},$$

we have

$$\sqrt{\sum_{t=1}^T \epsilon_t^2} \ge \max\left\{\sqrt{\frac{\epsilon}{1+\sqrt{2}/(e\delta)}}, \frac{\sqrt{2}\epsilon}{2\sqrt{\log(e+\epsilon/\sqrt{2}\delta)+2\epsilon}}\right\}.$$

Proof. If $\epsilon = \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(\frac{1}{\delta}\right)}$,

Because $(e^x-1)/(e^x+1) \le x$ for $x \ge 0$, then

$$\epsilon \le \sum_{t=1}^T \epsilon_t^2 + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(\frac{1}{\delta}\right)}.$$

Solving the inequality with respect to $\sqrt{\sum_{t=1}^T \epsilon_t^2}$, we get

$$\sqrt{\sum_{t=1}^T \epsilon_t^2} \ge \frac{\sqrt{2}\epsilon}{\sqrt{\log(1/\delta)+2\epsilon}+\sqrt{\log(1/\delta)}} \ge \frac{\sqrt{2}\epsilon}{2\sqrt{\log(1/\delta)+2\epsilon}}.$$

If we have

$$\epsilon = \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^T \epsilon_t^2}}{\delta}\right)}. \quad (49)$$

Because $(e^x-1)/(e^x+1) \le x$ for $x \ge 0$, then

$$\begin{aligned}\epsilon &\le \sum_{t=1}^T \epsilon_t^2 + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^T \epsilon_t^2}}{\delta}\right)} \\ &\le \sum_{t=1}^T \epsilon_t^2 + \frac{\sqrt{2}\sum_{t=1}^T \epsilon_t^2}{e\delta},\end{aligned}$$

where the second inequality is because $\log(e+x) \le x/e+1$ for $x \ge 0$.

As such,

$$\sqrt{\sum_{t=1}^T \epsilon_t^2} \ge \sqrt{\frac{\epsilon}{1+\sqrt{2}/(e\delta)}}.$$

On the other hand, by (49), it also holds that $\sqrt{2}\sqrt{\sum_{t=1}^T \epsilon_t^2} \le \epsilon$. Then we have

$$\begin{aligned}\epsilon &\le \sum_{t=1}^T \epsilon_t^2 + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^T \epsilon_t^2}}{\delta}\right)} \\ &\le \sum_{t=1}^T \epsilon_t^2 + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log\left(e + \frac{\epsilon}{\sqrt{2}\delta}\right)}.\end{aligned}$$

Solving the inequality with respect to $\sqrt{\sum_{t=1}^T \epsilon_t^2}$, we also get

$$\begin{aligned}\sqrt{\sum_{t=1}^T \epsilon_t^2} &\ge \frac{\sqrt{2}\epsilon}{\sqrt{\log(e+\epsilon/\sqrt{2}\delta)+2\epsilon}+\sqrt{\log(e+\epsilon/\sqrt{2}\delta)}} \\ &\ge \frac{\sqrt{2}\epsilon}{2\sqrt{\log(e+\epsilon/\sqrt{2}\delta)+2\epsilon}}.\end{aligned}$$

□

Lemma 17. *For constants $\kappa, \epsilon_0 > 0$, $c_1, c_2 > 0$, $\alpha \in \mathbb{R}$, a integer $T \ge 1$, assuming $\epsilon_t = \epsilon_0 t^\alpha$, $\varepsilon_t = O(\kappa/\epsilon_t)$ for $t \in [T]$, if $\sum_{t=1}^T \epsilon_t \ge c_1$, we have*

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1}}{c_1(\alpha/2-1)^2(\alpha+1)}}\right), & \alpha > 2; \\ O\left(\sqrt{\frac{\kappa T^3}{c_1(\alpha/2-1)^2(\alpha+1)}}\right), & -1 < \alpha < 2; \\ O\left(\sqrt{\frac{\kappa T^{2-\alpha}}{c_1(\alpha/2-1)^2(-\alpha-1)}}\right), & \alpha < -1, \end{cases}$$

and

$$\sum_{t=1}^T t\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1}}{c_1(\alpha/2-2)^2(\alpha+1)}}\right), & \alpha > 4; \\ O\left(\sqrt{\frac{\kappa T^5}{c_1(\alpha/2-2)^2(\alpha+1)}}\right), & -1 < \alpha < 4; \\ O\left(\sqrt{\frac{\kappa T^{4-\alpha}}{c_1(\alpha/2-2)^2(-\alpha-1)}}\right), & \alpha < -1. \end{cases}$$

If $\sqrt{\sum_{t=1}^T \epsilon_t^2} \ge c_2$, we have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1/2}}{c_2(\alpha/2-1)^2\sqrt{2\alpha+1}}}\right), & \alpha > 2; \\ O\left(\sqrt{\frac{\kappa T^{5/2}}{c_2(\alpha/2-1)^2\sqrt{2\alpha+1}}}\right), & -1/2 < \alpha < 2; \\ O\left(\sqrt{\frac{\kappa T^{2-\alpha}}{c_2(\alpha/2-1)^2\sqrt{-2\alpha-1}}}\right), & \alpha < -1/2, \end{cases}$$

and

$$\sum_{t=1}^T t\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1/2}}{c_2(\alpha/2-2)^2\sqrt{2\alpha+1}}}\right), & \alpha > 4; \\ O\left(\sqrt{\frac{\kappa T^{9/2}}{c_2(\alpha/2-2)^2\sqrt{2\alpha+1}}}\right), & -1/2 < \alpha < 4; \\ O\left(\sqrt{\frac{\kappa T^{4-\alpha}}{c_2(\alpha/2-2)^2\sqrt{-2\alpha-1}}}\right), & \alpha < -1/2. \end{cases}$$

Proof. If $\sum_{t=1}^T \epsilon_t = \sum_{t=1}^T \epsilon_0 t^\alpha \geq c_1$. We have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = O\left(\sum_{t=1}^T \sqrt{\kappa/\epsilon_t}\right) = O\left(\sum_{t=1}^T \sqrt{\frac{\kappa}{\epsilon_0 t^\alpha}}\right) = O\left(\sum_{t=1}^T t^{-\alpha/2}\sqrt{\frac{\kappa}{\epsilon_0}}\right).$$

Using Lemma 12, we have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = O\left(\sum_{t=1}^T t^{-\alpha/2}\sqrt{\frac{\kappa}{c_1}\sum_{t=1}^T t^\alpha}\right).$$

Then using Lemma 10, if $\alpha > 2$, i.e., $-\alpha/2 < -1$, we have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = O\left(\frac{1}{-(-\alpha/2)-1}\sqrt{\frac{\kappa}{c_1}\frac{T^{\alpha+1}}{\alpha+1}}\right) = O\left(\sqrt{\frac{\kappa}{c_1}\frac{T^{\alpha+1}}{(\alpha/2-1)^2(\alpha+1)}}\right).$$

Results under other conditions can be proved similarly. $\square$

Lemma 18. *For constants $\kappa, \epsilon_0 > 0$, $c_1, c_2 > 0$, $Q_0 \in (0,1)$, $Q > 0$, a integer $T \geq 1$, assuming $\epsilon_t = \epsilon_0 Q^{-t}$, $\varepsilon_t = O(\kappa/\epsilon_t)$ for $t \in [T]$, if $\sum_{t=1}^T \epsilon_t \geq c_1$, we have*

$$\sum_{t=1}^T Q_0^{-t}\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa Q^{-T}}{c_1(Q_0/\sqrt{Q}-1)^2(1-Q)}}\right), & 0 < Q < Q_0^2; \\ O\left(\sqrt{\frac{\kappa (Q_0^2)^{-T}}{c_1(Q_0/\sqrt{Q}-1)^2(1-Q)}}\right), & Q_0^2 < Q < 1; \\ O\left(\sqrt{\frac{\kappa (Q_0^2/Q)^{-T}}{c_1(Q_0/\sqrt{Q}-1)^2(Q-1)}}\right), & Q > 1, \end{cases}$$

and

$$\sum_{t=1}^T \sqrt{\varepsilon_t Q_0^{-t}} = \begin{cases} O\left(\sqrt{\frac{\kappa Q^{-T}}{c_1(\sqrt{Q_0}/\sqrt{Q}-1)^2(1-Q)}}\right), & 0 < Q < Q_0; \\ O\left(\sqrt{\frac{\kappa Q_0^{-T}}{c_1(\sqrt{Q_0}/\sqrt{Q}-1)^2(1-Q)}}\right), & Q_0 < Q < 1; \\ O\left(\sqrt{\frac{\kappa (Q_0/Q)^{-T}}{c_1(\sqrt{Q_0}/\sqrt{Q}-1)^2(Q-1)}}\right), & Q > 1. \end{cases}$$

If $\sqrt{\sum_{t=1}^T \epsilon_t^2} \geq c_2$, we have

$$\sum_{t=1}^T Q_0^{-t}\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa Q^{-T}}{c_2(Q_0/\sqrt{Q}-1)^2\sqrt{1-Q^2}}}\right), & 0 < Q < Q_0^2; \\ O\left(\sqrt{\frac{\kappa (Q_0^2)^{-T}}{c_2(Q_0/\sqrt{Q}-1)^2\sqrt{1-Q^2}}}\right), & Q_0^2 < Q < 1; \\ O\left(\sqrt{\frac{\kappa (Q_0^2/Q)^{-T}}{c_2(Q_0/\sqrt{Q}-1)^2\sqrt{Q^2-1}}}\right), & Q > 1, \end{cases}$$

and

$$\sum_{t=1}^T \sqrt{\varepsilon_t Q_0^{-t}} = \begin{cases} O\left(\sqrt{\frac{\kappa Q^{-T}}{c_2(\sqrt{Q_0}/\sqrt{Q}-1)^2\sqrt{1-Q^2}}}\right), & 0 < Q < Q_0; \\ O\left(\sqrt{\frac{\kappa Q_0^{-T}}{c_2(\sqrt{Q_0}/\sqrt{Q}-1)^2\sqrt{1-Q^2}}}\right), & Q_0 < Q < 1; \\ O\left(\sqrt{\frac{\kappa (Q_0/Q)^{-T}}{c_2(\sqrt{Q_0}/\sqrt{Q}-1)^2\sqrt{Q^2-1}}}\right), & Q > 1. \end{cases}$$

Proof. If $\sum_{t=1}^T \epsilon_t = \sum_{t=1}^T \epsilon_0 Q^{-t} \geq c_1$. We have

$$\sum_{t=1}^T Q_0^{-t}\sqrt{\varepsilon_t} = O\left(\sum_{t=1}^T Q_0^{-t}\sqrt{\kappa/\epsilon_t}\right) = O\left(\sum_{t=1}^T Q_0^{-t}\sqrt{\frac{\kappa}{\epsilon_0 Q^{-t}}}\right) = O\left(\sum_{t=1}^T (Q_0/\sqrt{Q})^{-t}\sqrt{\frac{\kappa}{\epsilon_0}}\right).$$

Using Lemma 12, we have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = O\left(\sum_{t=1}^T (Q_0/\sqrt{Q})^{-t}\sqrt{\frac{\kappa}{c_1}\sum_{t=1}^T Q^{-t}}\right).$$

Then using Lemma 11, if $Q < Q_0^2 < 1$, i.e., $Q_0/\sqrt{Q} > 1$, we have

$$\sum_{t=1}^T \sqrt{\varepsilon_t} = O\left(\frac{1}{Q_0/\sqrt{Q}-1}\sqrt{\frac{\kappa}{c_1}\frac{Q^{-T}}{1-Q}}\right) = O\left(\sqrt{\frac{\kappa Q^{-T}}{c_1(Q_0/\sqrt{Q}-1)^2(1-Q)}}\right).$$

Results under other conditions can be proved similarly. $\square$

Lemma 19. *For constants $L, c_3, c_4 > 0$, a integer $T \geq 1$, matrices $\widetilde{\mathbf{W}}^{(0)}, \mathbf{W}_* \in \mathcal{W} \subset \mathbb{R}^{d\times m}$, if it holds for a series of positive constants $\{\varepsilon_t\}$ that $\sum_{t=1}^T \sqrt{\varepsilon_t} = O(\sqrt{c_4 T^{c_3}})$, setting $T = \Theta((K^2 Lm/c_4)^{1/c_3})$, we have*

$$\mathcal{E} = \frac{L}{2mT}\left(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F + 2\sum_{t=1}^T \sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^T \frac{\varepsilon_t}{L}}\right)^2 = O\left(K^2 L\left[\frac{c_4}{K^2 Lm}\right]^{1/c_3}\right).$$

Proof. First, because $\varepsilon_t > 0$ for $t \in [T]$, we have

$$\sqrt{\sum_{t=1}^T \varepsilon_t} \leq \sum_{t=1}^T \sqrt{\varepsilon_t}.$$

Then combining Lemma 15, it suffices that

$$\mathcal{E} = O\left(\frac{L}{mT}\left[K\sqrt{m} + \frac{1}{\sqrt{L}}\sum_{t=1}^T \sqrt{\varepsilon_t}\right]^2\right) = O\left(\left[K\sqrt{\frac{L}{T}} + \frac{1}{\sqrt{mT}}\sum_{t=1}^T \sqrt{\varepsilon_t}\right]^2\right) = O\left(\left[K\sqrt{\frac{L}{T}} + \frac{1}{\sqrt{mT}}\sqrt{c_4 T^{c_3}}\right]^2\right).$$

Then setting $T = \Theta((K^2 Lm/c_4)^{1/c_3})$, we complete the proof. $\square$

Lemma 20. *For constants $L, c_3, c_4 > 0$, a integer $T \geq 1$, matrices $\widetilde{\mathbf{W}}^{(0)}, \mathbf{W}_* \in \mathcal{W} \subset \mathbb{R}^{d\times m}$, if it holds for a series of positive constants $\{\varepsilon_t\}$ that $\sum_{t=1}^T \sqrt{\varepsilon_t} = O(\sqrt{c_4 T^{c_3}})$, setting $T = \Theta((K^2 Lm/c_4)^{1/c_3})$, we have*

$$\mathcal{E} = \frac{2L}{m(T+1)^2}\left(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F + 2\sum_{t=1}^T t\sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^T t^2\frac{\varepsilon_t}{L}}\right)^2 = O\left(K^2 L\left[\frac{c_4}{K^2 Lm}\right]^{2/c_3}\right).$$

Proof. First, because $\varepsilon_t > 0$ for $t \in [T]$, we have

$$\sqrt{\sum_{t=1}^T t^2\varepsilon_t} \leq \sum_{t=1}^T \sqrt{t^2\varepsilon_t} = \sum_{t=1}^T t\sqrt{\varepsilon_t}.$$

Then combining Lemma 15, it suffices that

$$\begin{aligned}\mathcal{E} &= O\bigg(\frac{L}{mT^2}\bigg[K\sqrt{m}+\frac{1}{\sqrt{L}}\sum_{t=1}^{T}t\sqrt{\varepsilon_t}\bigg]^2\bigg)\\ &= O\bigg(\bigg[K\frac{\sqrt{L}}{T}+\frac{1}{\sqrt{m}T}\sum_{t=1}^{T}t\sqrt{\varepsilon_t}\bigg]^2\bigg)\\ &= O\bigg(\bigg[K\frac{\sqrt{L}}{T}+\frac{1}{\sqrt{m}T}\sqrt{c_4T^{c_3}}\bigg]^2\bigg).\end{aligned}$$

Then setting $T=\Theta((K^2Lm/c_4)^{1/c_3})$, we complete the proof. □

Lemma 21. *For constants $L, c_6 > 0$, a constant $c_5 \in (0,1)$, a constant $Q_0 \in (0,1)$, a integer $T \geq 1$, matrices $\widetilde{\mathbf{W}}^{(0)}, \mathbf{W}_* \in \mathcal{W} \subset \mathbb{R}^{d\times m}$, if it holds for a series of positive constants $\{\varepsilon_t\}$ that $\sum_{t=1}^{T}Q_0^{-t}\sqrt{\varepsilon_t} = O(\sqrt{c_6{c_5}^{-T}})$, setting $T=\Theta(\log_{1/c_5}(K^2Lm/c_6))$, we have*

$$\begin{aligned}\mathcal{E} &= \frac{Q_0^T}{\sqrt{m}}\bigg(\|\widetilde{\mathbf{W}}^{(0)}-\mathbf{W}_*\|_F+2\sum_{t=1}^{T}Q_0^{-t}\sqrt{\frac{2\varepsilon_t}{L}}\bigg)\\ &= O\bigg(K\bigg[\frac{c_6}{K^2Lm}\bigg]^{\log_{c_5}Q_0}\bigg).\end{aligned}$$

Proof. Using Lemma 15, it suffices that

$$\begin{aligned}\mathcal{E} &= O\bigg(\frac{Q_0^T}{\sqrt{m}}\bigg[K\sqrt{m}+\frac{1}{\sqrt{L}}\sum_{t=1}^{T}Q_0^{-t}\sqrt{\varepsilon_t}\bigg]\bigg)\\ &= O\bigg(Q_0^T\bigg[K+\frac{1}{\sqrt{mL}}\sum_{t=1}^{T}Q_0^{-t}\sqrt{\varepsilon_t}\bigg]\bigg)\\ &= O\bigg(Q_0^T\bigg[K+\frac{1}{\sqrt{mL}}\sqrt{c_6{c_5}^{-T}}\bigg]\bigg).\end{aligned}$$

Then setting $T=\Theta(\log_{1/c_5}(K^2Lm/c_6))$, we complete the proof. □

Lemma 22. *For constants $L, \mu, c_6 > 0$, a constant $c_5 \in (0,1)$, a constant $Q_0 \in (0,1)$, a integer $T \geq 1$, matrices $\widetilde{\mathbf{W}}^{(0)}, \mathbf{W}_* \in \mathcal{W} \subset \mathbb{R}^{d\times m}$, if it holds for a series of positive constants $\{\varepsilon_t\}$ that $\sum_{t=1}^{T}\sqrt{\varepsilon_t Q_0^{-t}} = O(\sqrt{c_6{c_5}^{-T}})$, setting $T=\Theta((K^2Lm/c_4)^{1/c_3})$, we have*

$$\begin{aligned}\mathcal{E} =&\frac{(Q_0)^T}{m}\bigg(K\sqrt{Lm}+2\sqrt{\frac{L}{\mu}}\sum_{t=1}^{T}\sqrt{\varepsilon_t(Q_0)^{-t}}\\ &+\sqrt{\sum_{t=1}^{T}\varepsilon_t(Q_0)^{-t}}\bigg)^2\\ =&O\bigg(K^2L\bigg[\frac{c_6}{K^2\mu m}\bigg]^{\log_{c_5}Q_0}\bigg).\end{aligned}$$

Proof. First, because $\varepsilon_t > 0$ for $t\in[T]$, we have

$$\sqrt{\sum_{t=1}^{T}\varepsilon_t Q_0^{-t}} \leq \sum_{t=1}^{T}\sqrt{\varepsilon_t Q_0^{-t}}.$$

Then it suffices that

$$\begin{aligned}\mathcal{E} &= O\bigg(\frac{Q_0^T}{m}\bigg[K\sqrt{Lm}+\sqrt{\frac{L}{\mu}}\sum_{t=1}^{T}\sqrt{\varepsilon_t Q_0^{-t}}\bigg]^2\bigg)\\ &= O\bigg(Q_0^T\bigg[K\sqrt{L}+\sqrt{\frac{L}{m\mu}}\sum_{t=1}^{T}\sqrt{\varepsilon_t Q_0^{-t}}\bigg]^2\bigg)\\ &= O\bigg(Q_0^T\bigg[K\sqrt{L}+\sqrt{\frac{L}{m\mu}}\sqrt{c_6{c_5}^{-T}}\bigg]^2\bigg).\end{aligned}$$

Then setting $T=\Theta(\log_{1/c_5}(K^2\mu m/c_6))$, we complete the proof. □

Appendix L
Proof of Results in the Main Text

A. Proof of Lemma 1

Proof. Under the setting of single-task learning, each task is learned independently, and thus, we have for $i=1,\ldots,m$, for any set $\mathcal{S}$,

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]},\mathbf{w}_i,\mathcal{D}_i))\\ &\quad=\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]})).\end{aligned}$$

As such, we have for $i=1,\ldots,m$,

$$\begin{aligned}&\frac{\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]},\mathbf{w}_i,\mathcal{D}_i))}{\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]},\mathbf{w}_i',\mathcal{D}_i'))}\\ =&\frac{\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]}))}{\mathbb{P}(\hat{\mathbf{w}}_{[-i]}\in\mathcal{S}\mid\mathcal{B}=(\mathbf{w}_{[-i]},\mathcal{D}_{[-i]}))}=1\leq e^0.\end{aligned}$$

□

B. Proof of Proposition 1

Proof. First, for Algorithm 2, denoting $\mathbf{\Sigma}_0=\widetilde{\mathbf{\Sigma}}^{(t)}$, the j-th diagonal element of $\mathbf{S}_{\eta\lambda}$ is

$$\begin{aligned}&\max\bigg(0,1-\frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0+\mathbf{E})}}\bigg)\\ \geq&\max\bigg(0,1-\frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)+\sigma_d(\mathbf{E})}}\bigg),\end{aligned}$$

where $\sigma_d(\mathbf{E})$ is the d-th largest singular value, i.e., the smallest singular value, of $\mathbf{E}$. As such, when $\sigma_d(\mathbf{E})=C\lambda^2$ for sufficiently large $C>0$, $\max\Big(0,1-\frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)+\sigma_d(\mathbf{E})}}\Big)\to1$.

Then $\hat{\mathbf{w}}_i^{(t-1)}=\mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}\mathbf{w}_i^{(t-1)}=\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{w}_i^{(t-1)}=\mathbf{w}_i^{(t-1)}$. Therefore, all the procedures can be decoupled to independently run for each task, thus Algorithm 2 degrades to an STL algorithm with no random perturbation.

Similarly, for Algorithm 3, for all $j\in[m]$, the j-th diagonal element of $\mathbf{S}_{\eta\lambda}$ is

$$\begin{aligned}&\max\bigg(0,1-\frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}+\mathbf{E}_{jj}|}}\bigg)\\ =&\max\bigg(0,1-\frac{\eta\lambda}{\sqrt{\mathbf{\Sigma}_{jj,0}+\mathbf{E}_{jj}}}\bigg),\end{aligned}$$

where the equality is because $\mathbf{\Sigma}_0$ is semi-positive definite and $\mathbf{E}$ is positive definite.

As such, when $\min_j\mathbf{E}_{jj}=C\lambda^2$ for sufficiently large $C>0$, $\min_j\Big[\max\Big(0,1-\frac{\eta\lambda}{\sqrt{\mathbf{\Sigma}_{jj,0}+\mathbf{E}_{jj}}}\Big)\Big]\to1$.

Then $\hat{\mathbf{w}}_i^{(t-1)}=\mathbf{S}_{\eta\lambda}\mathbf{w}_i^{(t-1)}=\mathbf{w}_i^{(t-1)}$. Therefore, all the procedures can be decoupled to independently run for each task, thus Algorithm 3 degrades to an STL algorithm with no random perturbation.

□

C. Proof of Theorem 1

Proof. For simplicity, we omit the symbol $\mathcal{B}$ used to denote the input in the conditional events in some equations.

First, we show that for all $t\in[T]$, the mapping $\mathbf{W}^{(t-1)}\to\mathbf{\Sigma}^{(t)}$ is an $(\epsilon_t,0)$-differentially private algorithm.

Case 1. For $\mathbf{\Sigma}^{(t)}=\widetilde{\mathbf{\Sigma}}^{(t)}+\mathbf{E}=\mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}}+\mathbf{E}$, we follow the proof of Theorem 4 of Jiang et al. [7].

For all $i\in[m]$, consider two adjacent parameter matrices $\mathbf{W}^{(t-1)}$ and $(\mathbf{W}')^{(t-1)}$ that differ only in the i-th column such that $\mathbf{W}^{(t-1)} = [\mathbf{w}_1^{(t-1)} \cdots \mathbf{w}_i^{(t-1)} \cdots \mathbf{w}_m^{(t-1)}]$ and $(\mathbf{W}')^{(t-1)} = [\mathbf{w}_1^{(t-1)} \cdots (\mathbf{w}_i')^{(t-1)} \cdots \mathbf{w}_m^{(t-1)}]$. Now, let

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}^{(t)} &= \mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}} = \sum_{j=1}^{m} \mathbf{w}_j^{(t-1)}(\mathbf{w}_j^{(t-1)})^{\mathrm{T}} \\ (\widetilde{\mathbf{\Sigma}}')^{(t)} &= (\mathbf{W}')^{(t-1)}((\mathbf{W}')^{(t-1)})^{\mathrm{T}} \\ &= \sum_{j\in[m],j\neq i} \mathbf{w}_j^{(t-1)}(\mathbf{w}_j^{(t-1)})^{\mathrm{T}} + (\mathbf{w}_i')^{(t-1)}((\mathbf{w}_i')^{(t-1)})^{\mathrm{T}} \\ \Delta &= \widetilde{\mathbf{\Sigma}}^{(t)} - (\widetilde{\mathbf{\Sigma}}')^{(t)} \\ &= \mathbf{w}_i^{(t-1)}(\mathbf{w}_i^{(t-1)})^{\mathrm{T}} - (\mathbf{w}_i')^{(t-1)}((\mathbf{w}_i')^{(t-1)})^{\mathrm{T}}.\end{aligned}$$

Then, we have for the conditional densities

$$\frac{p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})}{p(\mathbf{\Sigma}^{(t)} \mid (\mathbf{W}')^{(t-1)})} = \frac{p(\mathbf{\Sigma}^{(t)} = \mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}} + \mathbf{E}_1)}{p(\mathbf{\Sigma}^{(t)} = (\mathbf{W}')^{(t-1)}((\mathbf{W}')^{(t-1)})^{\mathrm{T}} + \mathbf{E}_2)}.$$

Because $\mathbf{E}_1, \mathbf{E}_2 \sim W_D(\frac{\max_i s_i^{(t-1)}}{2\epsilon_t}\mathbf{I}_D, D+1)$, letting $\mathbf{V} = \frac{\max_j s_j^{(t-1)}}{2\epsilon_t}\mathbf{I}_D$, $\alpha = \frac{\max_j s_j^{(t-1)}}{2\epsilon_t}$,

$$\begin{aligned}&\frac{p(\mathbf{\Sigma}^{(t)} = \mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}} + \mathbf{E}_1)}{p(\mathbf{\Sigma}^{(t)} = (\mathbf{W}')^{(t-1)}((\mathbf{W}')^{(t-1)})^{\mathrm{T}} + \mathbf{E}_2)} \\ &= \frac{\exp[-\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - \mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}}))/2]}{\exp[-\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - (\mathbf{W}')^{(t-1)}((\mathbf{W}')^{(t-1)})^{\mathrm{T}}))/2]} \\ &= \exp[\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - (\mathbf{W}')^{(t-1)}((\mathbf{W}')^{(t-1)})^{\mathrm{T}}))/2 \\ &- \mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - \mathbf{W}^{(t-1)}(\mathbf{W}^{(t-1)})^{\mathrm{T}}))/2] \\ &= \exp[\mathrm{tr}(\mathbf{V}^{-1}\Delta)/2] \\ &= \exp[\mathrm{tr}(\mathbf{w}_i^{(t-1)}(\mathbf{w}_i^{(t-1)})^{\mathrm{T}} - (\mathbf{w}_i')^{(t-1)}((\mathbf{w}_i')^{(t-1)})^{\mathrm{T}})/(2\alpha)] \\ &= \exp[(\mathrm{tr}(\mathbf{w}_i^{(t-1)}(\mathbf{w}_i^{(t-1)})^{\mathrm{T}}) - \mathrm{tr}((\mathbf{w}_i')^{(t-1)}((\mathbf{w}_i')^{(t-1)})^{\mathrm{T}}))/(2\alpha)] \\ &= \exp[(\mathrm{tr}((\mathbf{w}_i^{(t-1)})^{\mathrm{T}}\mathbf{w}_i^{(t-1)}) - \mathrm{tr}(((\mathbf{w}_i')^{(t-1)})^{\mathrm{T}}(\mathbf{w}_i')^{(t-1)}))/(2\alpha)] \\ &= \exp[(\|\mathbf{w}_i^{(t-1)}\|_2^2 - \|(\mathbf{w}_i')^{(t-1)}\|_2^2)/(2\alpha)] \\ &\leq \exp[s_i^{(t-1)}/(2\alpha)] = \exp\left[2\epsilon_t \frac{s_i^{(t-1)}}{2\max_j s_j^{(t-1)}}\right] \leq \exp(\epsilon_t).\end{aligned}$$

As such, we have

$$\frac{p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})}{p(\mathbf{\Sigma}^{(t)} \mid (\mathbf{W}')^{(t-1)})} \leq \exp(\epsilon_t).$$

Case 2. Consider $\mathbf{\Sigma}^{(t)} = \widetilde{\mathbf{\Sigma}}^{(t)} + \mathbf{E} = (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)} + \mathbf{E}$.

For all $i \in [m]$, consider two adjacent parameter matrices $\mathbf{W}^{(t-1)}$ and $(\mathbf{W}')^{(t-1)}$ that differ only in the i-th column such that $\mathbf{W}^{(t-1)} = [\mathbf{w}_1^{(t-1)} \cdots \mathbf{w}_i^{(t-1)} \cdots \mathbf{w}_m^{(t-1)}]$ and $(\mathbf{W}')^{(t-1)} = [\mathbf{w}_1^{(t-1)} \cdots (\mathbf{w}_i')^{(t-1)} \cdots \mathbf{w}_m^{(t-1)}]$. Let

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}^{(t)} &= (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)} \\ (\widetilde{\mathbf{\Sigma}}')^{(t)} &= ((\mathbf{W}')^{(t-1)})^{\mathrm{T}}(\mathbf{W}')^{(t-1)} \\ \Delta &= \widetilde{\mathbf{\Sigma}}^{(t)} - (\widetilde{\mathbf{\Sigma}}')^{(t)},\end{aligned}$$

where the i-th diagonal element of Δ is $\|\mathbf{w}_i^{(t-1)}\|_2^2 - \|(\mathbf{w}_i')^{(t-1)}\|_2^2$ and the other diagonal elements of Δ are zeros.

Then, we have

$$\frac{p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})}{p(\mathbf{\Sigma}^{(t)} \mid (\mathbf{W}')^{(t-1)})} = \frac{p(\mathbf{\Sigma}^{(t)} = (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)} + \mathbf{E}_1)}{p(\mathbf{\Sigma}^{(t)} = ((\mathbf{W}')^{(t-1)})^{\mathrm{T}}(\mathbf{W}')^{(t-1)} + \mathbf{E}_2)}.$$

Because $\mathbf{E}_1, \mathbf{E}_2 \sim W_m(\mathrm{diag}(\mathbf{s}^{(t-1)}/2\epsilon_t), m+1)$, letting $\mathbf{V} = \mathrm{diag}(\mathbf{s}^{(t-1)}/2\epsilon_t)$,

$$\begin{aligned}&\frac{p(\mathbf{\Sigma}^{(t)} = (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)} + \mathbf{E}_1)}{p(\mathbf{\Sigma}^{(t)} = ((\mathbf{W}')^{(t-1)})^{\mathrm{T}}(\mathbf{W}')^{(t-1)} + \mathbf{E}_2)} \\ &= \frac{\exp[-\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)}))/2]}{\exp[-\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - ((\mathbf{W}')^{(t-1)})^{\mathrm{T}}(\mathbf{W}')^{(t-1)}))/2]} \\ &= \exp[\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - ((\mathbf{W}')^{(t-1)})^{\mathrm{T}}(\mathbf{W}')^{(t-1)}))/2 \\ &\quad - \mathrm{tr}(\mathbf{V}^{-1}(\mathbf{\Sigma}^{(t)} - (\mathbf{W}^{(t-1)})^{\mathrm{T}}\mathbf{W}^{(t-1)}))/2] \\ &= \exp[\mathrm{tr}(\mathbf{V}^{-1}\Delta)/2] \\ &= \exp[(\|\mathbf{w}_i^{(t-1)}\|_2^2 - \|(\mathbf{w}_i')^{(t-1)}\|_2^2)/(2v_{ii})] \\ &\quad \leq \exp[s_i^{(t-1)}/(2v_{ii})] = \exp\left[2\epsilon_t \frac{s_i^{(t-1)}}{2s_i^{(t-1)}}\right] \leq \exp(\epsilon_t).\end{aligned}$$

As such, we also have

$$\frac{p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})}{p(\mathbf{\Sigma}^{(t)} \mid (\mathbf{W}')^{(t-1)})} \leq \exp(\epsilon_t).$$

Next, given $t \in [T]$, $\mathbf{\Sigma}^{(1:t-1)}$ (when $t=1$, $\mathbf{\Sigma}^{(1:t-1)} = \emptyset$) and the mapping $f : \mathbf{\Sigma}^{(1:t)} \to \mathbf{M}^{(t)}$, which does not touch any unperturbed sensitive information, using the *Post-Processing immunity* Lemma (Lemma 4) for the mapping $f' : \mathbf{\Sigma}^{(1:t)} \to (\mathbf{M}^{(t)}, \mathbf{\Sigma}^{(t)})$, the algorithm $(\mathbf{W}^{(t-1)}, \mathbf{\Sigma}^{(1:t-1)}) \to (\mathbf{M}^{(t)}, \mathbf{\Sigma}^{(t)})$ is still an $(\epsilon_t, 0)$-differentially private algorithm.

Then, because $\hat{\mathbf{w}}_i^{(t)} = \mathcal{A}_{\mathrm{st},i}(\mathbf{M}^{(t)}, \mathbf{w}_i^{(0:t-1)}, \mathbf{X}_i, \mathbf{y}_i)$ is an STL algorithm for the i-th task, for $i = 1, \ldots, m$, the mapping $(\mathbf{M}^{(t)}, \mathbf{w}_{[-i]}^{(0:t-1)}, \mathbf{X}_{[-i]}, \mathbf{y}_{[-i]}) \to (\hat{\mathbf{w}}_{[-i]}^{(t)})$ thus does not touch any unperturbed sensitive information for the i-th task. As such, applying the *Post-Processing immunity* Lemma again for the mapping $f'' : (\mathbf{M}^{(t)}, \mathbf{w}_{[-i]}^{(0:t-1)}, \mathbf{X}_{[-i]}, \mathbf{y}_{[-i]}, \mathbf{\Sigma}^{(1:t-1)}) \to (\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{M}^{(t)}, \mathbf{\Sigma}^{(t)})$, for the algorithm $(\mathbf{W}^{(t-1)}, \mathbf{\Sigma}^{(1:t-1)}, \mathbf{w}_{[-i]}^{(0:t-2)}, \mathbf{X}_{[-i]}, \mathbf{y}_{[-i]}) \to (\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{M}^{(t)}, \mathbf{\Sigma}^{(t)})$ (when $t=1$, $\mathbf{w}_{[-i]}^{(0:t-2)} = \emptyset$), denoting $\vartheta_{t,i} = (\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{M}^{(t)}, \mathbf{\Sigma}^{(t)}) \in \mathcal{C}_{t,i}$, we have for any set $\mathcal{S}_{t,i} \subseteq \mathcal{C}_{t,i}$

$$\begin{aligned}&\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid \mathbf{W}^{(t-1)}, \mathbf{\Sigma}^{(1:t-1)}, \mathbf{w}_{[-i]}^{(0:t-2)}, \mathcal{D}^m) \\ &\leq e^{\epsilon_t}\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid (\mathbf{W}')^{(t-1)}, \mathbf{\Sigma}^{(1:t-1)}, \mathbf{w}_{[-i]}^{(0:t-2)}, (\mathcal{D}')^m),\end{aligned}$$

where $\mathbf{W}^{(t-1)}$ and $(\mathbf{W}')^{(t-1)}$ differ only in the i-th column and $\mathcal{D}^m$ and $(\mathcal{D}')^m$ differ only in the i-th task.

Now, again, for $t = 1, \ldots, T$, we take the t-th dataset $\widetilde{\mathcal{D}}_t = \{(\mathbf{w}_1^{(t-1)}, \mathcal{D}_1), \ldots, (\mathbf{w}_m^{(t-1)}, \mathcal{D}_m)\}$. Given that $\mathbf{W}^{(t)} = \widehat{\mathbf{W}}^{(t)}$ for all $t \in [T]$, we have for any set $\mathcal{S}_{t,i} \subseteq \mathcal{C}_{t,i}$ that

$$\begin{aligned}&\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid \widetilde{\mathcal{D}}_t, \boldsymbol{\vartheta}_{1:t-1}) \\ &\leq e^{\epsilon_t}\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid \widetilde{\mathcal{D}}_t', \boldsymbol{\vartheta}_{1:t-1}),\end{aligned}$$

where $\widetilde{\mathcal{D}}_t$ and $\widetilde{\mathcal{D}}_t'$ are two adjacent datasets that differ in a single entry, the i-th data instance $(\mathbf{w}_i^{(t-1)}, \mathcal{D}_i = (\mathbf{X}_i, \mathbf{y}_i))$, and

$$\boldsymbol{\vartheta}_{1:t-1} = \begin{cases} \emptyset, & t = 1 \\ (\vartheta_{1,1}, \ldots, \vartheta_{1,m}) \ldots, (\vartheta_{t-1,1}, \ldots, \vartheta_{t-1,m}), & t \geq 2. \end{cases}$$

This renders the algorithm in the t-th iteration an $(\epsilon_t, 0)$-differentially private algorithm.

Now, again by the *Adaptive composition* Lemma (Lemma 7), for all $i \in [m]$ and for any set $\mathcal{S}' \subseteq \bigotimes_{t=1}^{T} \mathcal{C}_{t_i}$, we have

$$\begin{aligned}&\mathbb{P}((\vartheta_{1,i}, \cdots, \vartheta_{T,i}) \in \mathcal{S}' \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\widetilde{\mathcal{D}}_t, \boldsymbol{\vartheta}_{1:t-1}))) \\ &\leq e^{\tilde{\epsilon}}\mathbb{P}((\vartheta_{1,i}, \cdots, \vartheta_{T,i}) \in \mathcal{S}' \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\widetilde{\mathcal{D}}_t', \boldsymbol{\vartheta}_{1:t-1}))) \\ &\quad + \delta,\end{aligned}$$

where for all $t \in [T]$, $\mathcal{B}_t$ denotes the input for the t-th iteration.

Finally, taking $\theta_t = (\vartheta_{t,1}, \dots, \vartheta_{t,m})$ for all $t \in [T]$, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)\times T}$

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T}\mathcal{B}_t = (\mathbf{W}^{(t-1)}, \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1}))\\ \le& e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T}\mathcal{B}_t = ((\mathbf{W}')^{(t-1)}, (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}))\\ &+\delta,\end{aligned}$$

□

D. Proof of Corollary 1

Proof. For simplicity, we omit the symbol $\mathcal{B}$ used to denote the input in the conditional events in some equations.

Using Theorem 1, we only need to show that Algorithm 2 complies with our MP-MTL framework in Algorithm 1.

Consider the t-th iteration for all $t \in [T]$. Because of the norm clipping, for $i = 1, \dots, m$, the ℓ_2 norm of any parameter vector equals one. Then, we have for $i = 1, \dots, m$ that

$$\begin{aligned}s_i^{(t-1)} =& \max_{(\tilde{\mathbf{w}}_i')^{(t-1)}} |\|\tilde{\mathbf{w}}_i^{(t-1)}\|_2^2 - \|(\tilde{\mathbf{w}}_i')^{(t-1)}\|_2^2|\\ \le& \max_{(\tilde{\mathbf{w}}_i')^{(t-1)}} |\|\tilde{\mathbf{w}}_i^{(t-1)}\|_2^2| + |\|(\tilde{\mathbf{w}}_i')^{(t-1)}\|_2^2| = 2K.\end{aligned}$$

Because the norm clipping is a deterministic STL algorithm and because the mapping $\widetilde{\mathbf{W}}^{(t-1)} \to \boldsymbol{\Sigma}^{(t)}$ is an $(\epsilon_t, 0)$ - differentially algorithm, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times d}$ that

$$\begin{aligned}&\mathbb{P}(\boldsymbol{\Sigma}^{(t)} \in \mathcal{S} \mid \mathbf{w}_{[-i]}^{(t-1)}, \mathbf{w}_i^{(t-1)})\\ &=\mathbb{P}(\boldsymbol{\Sigma}^{(t)} \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}^{(t-1)}, \tilde{\mathbf{w}}_i^{(t-1)})\\ &\le e^{\epsilon_t}\mathbb{P}(\boldsymbol{\Sigma}^{(t)} \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}^{(t-1)}, (\tilde{\mathbf{w}}_i')^{(t-1)})\\ &=e^{\epsilon_t}\mathbb{P}(\boldsymbol{\Sigma}^{(t)} \in \mathcal{S} \mid \mathbf{w}_{[-i]}^{(t-1)}, (\mathbf{w}_i')^{(t-1)}),\end{aligned}$$

which renders the mapping $\mathbf{W}^{(t-1)} \to \boldsymbol{\Sigma}^{(t)}$ as an $(\epsilon_t, 0)$ - differentially algorithm as well.

Let $\mathbf{M}^{(t)} = \mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}$. As such, the 7-th step to the 9-th step can be treated as the process of first performing a mapping $f : \boldsymbol{\Sigma}^{(1:t)} \to \mathbf{M}^{(t)}$ and then applying an STL algorithm:

$$\hat{\mathbf{w}}_i^{(t)} = \mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{w}}_i^{(t-1)}, \text{ for all } i \in [m]. \tag{50}$$

Now, because (50), the 10-th step and the 11-th step are all STL algorithms, they can be treated as a complete STL algorithm performing the mapping: $(\mathbf{M}^{(t)}, \mathbf{w}_i^{(0:t-1)}, \mathbf{X}_i, \mathbf{y}_i) \to (\hat{\mathbf{w}}_i^{(t)}, \mathbf{w}_i^{(t)})$.

As such, in all the iterations, Algorithm 2 complies with Algorithm 1. Thus, the result of Theorem 1 can be applied to Algorithm 2.

Similarly, using Theorem 1, we only need to show that Algorithm 3 complies with our MP-MTL framework in Algorithm 1.

The proof for the sensitivity is the same.

Now, let $\mathbf{M}^{(t)} = \mathbf{S}_{\eta\lambda}$. As such, the 7-th step can be treated as a mapping $f : \boldsymbol{\Sigma}^{(1:t)} \to \mathbf{M}^{(t)}$.

Then, because the 8-th step, the 9-th step and the 10-th step are all STL algorithms, they can be treated as a complete STL algorithm performing the mapping: $(\mathbf{M}^{(t)}, \mathbf{w}_i^{(0:t-1)}, \mathbf{X}_i, \mathbf{y}_i) \to (\hat{\mathbf{w}}_i^{(t)}, \mathbf{w}_i^{(t)})$.

Therefore, in all the iterations, Algorithm 3 complies with Algorithm 1, and thus, the result of Theorem 1 can be applied to Algorithm 3.

□

E. Proof of Lemma 2

Proof. We invoke the results of Schmidt et al. [11] to bound the empirical optimization error.

In the t-th step, a standard proximal operator (see Ji and Ye [6]) optimizes the following problem:

$$\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_*,$$

where $\mathbf{C} = \widetilde{\mathbf{W}}^{(t-1)}$. By Theorem 3.1 of Ji and Ye [6], denote the solution of the problem by $\widehat{\mathbf{W}}_0^{(t)} = \mathbf{U}_0\mathbf{S}_{\eta\lambda,0}\mathbf{U}_0^{\mathrm{T}}\mathbf{C}$, where $\mathbf{U}_0\boldsymbol{\Lambda}_0\mathbf{U}_0^{\mathrm{T}} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$ is the SVD decomposition of $\mathbf{C}\mathbf{C}^{\mathrm{T}}$. $\mathbf{S}_{\eta\lambda,0}$ is also a diagonal matrix and $\mathbf{S}_{\eta\lambda,ii,0} = \max\{0, 1 - \eta\lambda/\sqrt{\boldsymbol{\Lambda}_{ii,0}}\}$ for $i = 1, \dots, \min\{d, m\}$.

By Algorithm 2, $\widehat{\mathbf{W}}^{(t)} = \mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}\mathbf{C}$.

Then we analyse the bound of $\frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_* - \{\frac{1}{2\eta}\|\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}_0^{(t)}\|_*\}$.

First, we have

$$\begin{aligned}&\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 - \|\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}\|_F^2\\ =&\mathrm{tr}((\widehat{\mathbf{W}}^{(t)} - \mathbf{C})^{\mathrm{T}}(\widehat{\mathbf{W}}^{(t)} - \mathbf{C})) - \mathrm{tr}((\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C})^{\mathrm{T}}(\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}))\\ =&\mathrm{tr}((\widehat{\mathbf{W}}^{(t)})^{\mathrm{T}}\widehat{\mathbf{W}}^{(t)}) - \mathrm{tr}((\widehat{\mathbf{W}}_0^{(t)})^{\mathrm{T}}\widehat{\mathbf{W}}_0^{(t)}) - 2\mathrm{tr}((\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)})^{\mathrm{T}}\mathbf{C})\\ =&\mathrm{tr}((\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)})^{\mathrm{T}}(\widehat{\mathbf{W}}^{(t)} + \widehat{\mathbf{W}}_0^{(t)})) - 2\mathrm{tr}((\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)})^{\mathrm{T}}\mathbf{C})\\ =&\mathrm{tr}((\widehat{\mathbf{W}}^{(t)} - \mathbf{C})^{\mathrm{T}}(\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}))\\ &+\mathrm{tr}((\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C})^{\mathrm{T}}(\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}))\\ \le&\sigma_1(\widehat{\mathbf{W}}^{(t)} - \mathbf{C})\|\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}\|_*\\ &+\sigma_1(\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C})\|\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}\|_*,\end{aligned} \tag{51}$$

where $\sigma_1(\cdot)$ denotes the largest singular value of the enclosed matrix.

Denote $\mathbf{T} = \mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}, \mathbf{T}_0 = \mathbf{U}_0\mathbf{S}_{\eta\lambda,0}\mathbf{U}_0^{\mathrm{T}}$. Since $\mathbf{U}$ is decomposed from a symmetric matrix, we have

$$\begin{aligned}\sigma_1(\widehat{\mathbf{W}}^{(t)} - \mathbf{C}) &= \sigma_1(\mathbf{T}\mathbf{C} - \mathbf{C}) \le \sigma_1(\mathbf{C})\sigma_1(\mathbf{T} - \mathbf{I})\\ &= \sigma_1(\mathbf{C})\sigma_1(\mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}} - \mathbf{U}\mathbf{U}^{\mathrm{T}})\\ &= \sigma_1(\mathbf{C})\sigma_1(\mathbf{U}(\mathbf{S}_{\eta\lambda} - \mathbf{I})\mathbf{U}^{\mathrm{T}}).\end{aligned}$$

Since $\mathbf{S}_{\eta\lambda} - \mathbf{I}$ is a diagonal matrix, whose i-th diagonal element is $\max\{0, 1 - \eta\lambda/\sqrt{\boldsymbol{\Lambda}_{ii}}\} - 1 \in [-1, 0)$, so $\sigma_1(\mathbf{U}(\mathbf{S}_{\eta\lambda} - \mathbf{I})\mathbf{U}^{\mathrm{T}}) \le 1$ and

$$\sigma_1(\widehat{\mathbf{W}}^{(t)} - \mathbf{C}) \le \sigma_1(\mathbf{C}). \tag{52}$$

Similarly,

$$\sigma_1(\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}) \le \sigma_1(\mathbf{C}). \tag{53}$$

On the other hand,

$$\begin{aligned}
&\|\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}\|_* = \|\mathbf{T}\mathbf{C} - \mathbf{T}_0\mathbf{C}\|_* \\
=&\left\|\sum_{j=1}^{d} \sigma_j(\mathbf{T})\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\mathbf{C} - \sum_{j=1}^{d} \sigma_j(\mathbf{T}_0)\mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}}\mathbf{C}\right\|_* \\
=&\left\|\sum_{j=1}^{d}(\sigma_j(\mathbf{T}_0) + \sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0))\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\mathbf{C}\right. \\
&\left.- \sum_{j=1}^{d} \sigma_j(\mathbf{T}_0)\mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}}\mathbf{C}\right\|_* \\
=&\left\|\sum_{j=1}^{d} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\mathbf{C}\right. \\
&\left.+ \sum_{j=1}^{d}(\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0))\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\mathbf{C}\right\|_* \\
\le&\left\|\sum_{j=1}^{d} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\mathbf{C}\right\|_* \\
&+ \left\|\sum_{j=1}^{d}(\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0))\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\mathbf{C}\right\|_*,
\end{aligned} \tag{54}$$

where $\mathbf{u}_j$ and $\mathbf{u}_{j,0}$ are the j-th column of $\mathbf{U}$ and $\mathbf{U}_0$, respectively.

Let $r_c = \mathrm{rank}(\mathbf{C}) \le \min\{d, m\}$ be the rank of $\mathbf{C}$. Then we have

$$\begin{aligned}
&\left\|\sum_{j=1}^{d}(\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0))\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\mathbf{C}\right\|_* \\
\le&\sum_{j=1}^{r_c} |\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)|\sigma_j(\mathbf{C}) \le \sigma_1(\mathbf{C}) \sum_{j=1}^{r_c} |\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)|.
\end{aligned} \tag{55}$$

Denote $\mathbf{\Sigma}_0 = \widetilde{\mathbf{\Sigma}}^{(t)} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$. Then we have for $j \in [r_c]$,

$$\begin{aligned}
&|\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)| \\
=&\left|\max\left(0, 1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0 + \mathbf{E})}}\right) - \max\left(0, 1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)}}\right)\right| \\
\le&\left|\max\left(0, 1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0) + \sigma_1(\mathbf{E})}}\right) - \max\left(0, 1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)}}\right)\right|.
\end{aligned}$$

Case 1: $\eta\lambda > \sqrt{\sigma_j(\mathbf{\Sigma}_0)}$. Then

$$\begin{aligned}
&|\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)| \\
&= \max\left(0, 1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0) + \sigma_1(\mathbf{E})}}\right) \le 1 - \frac{\eta\lambda}{\sqrt{\eta^2\lambda^2 + \sigma_1(\mathbf{E})}} \\
&\le 1 - \frac{\eta\lambda}{\eta\lambda + \sqrt{\sigma_1(\mathbf{E})}} = \frac{\sqrt{\sigma_1(\mathbf{E})}}{\eta\lambda + \sqrt{\sigma_1(\mathbf{E})}} \le \frac{\sqrt{\sigma_1(\mathbf{E})}}{\eta\lambda}
\end{aligned}$$

Case 2: $\eta\lambda \le \sqrt{\sigma_j(\mathbf{\Sigma}_0)}$. Then

$$\begin{aligned}
&|\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)| \\
=&1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0) + \sigma_1(\mathbf{E})}} - 1 + \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)}} \\
=&\eta\lambda \cdot \frac{\sqrt{\sigma_j(\mathbf{\Sigma}_0) + \sigma_1(\mathbf{E})} - \sqrt{\sigma_j(\mathbf{\Sigma}_0)}}{\sqrt{\sigma_j^2(\mathbf{\Sigma}_0) + \sigma_j(\mathbf{\Sigma}_0)\sigma_1(\mathbf{E})}} \\
=&\frac{\eta\lambda\sigma_1(\mathbf{E})}{[\sqrt{\sigma_j(\mathbf{\Sigma}_0) + \sigma_1(\mathbf{E})} + \sqrt{\sigma_j(\mathbf{\Sigma}_0)}]\sqrt{\sigma_j^2(\mathbf{\Sigma}_0) + \sigma_j(\mathbf{\Sigma}_0)\sigma_1(\mathbf{E})}} \\
\le&\frac{\eta\lambda\sigma_1(\mathbf{E})}{[\sqrt{\eta^2\lambda^2 + 0} + \sqrt{\eta^2\lambda^2}]\sqrt{\eta^4\lambda^4 + 0}} = \frac{\sigma_1(\mathbf{E})}{2\eta^2\lambda^2}.
\end{aligned}$$

Suppose that there exists an index $k \le d$ such that

$$\sigma_k^2(\mathbf{C}) = \sigma_k(\mathbf{\Sigma}_0) > \eta^2\lambda^2, \sigma_{k+1}^2(\mathbf{C}) = \sigma_{k+1}(\mathbf{\Sigma}_0) \le \eta^2\lambda^2,$$

then $\sigma_j(\mathbf{T}_0) > 0$ for $j \le k$, $k \le r_c$, and

$$\begin{aligned}
&\sum_{j=1}^{r_c} |\sigma_j(\mathbf{T}) - \sigma_j(\mathbf{T}_0)| \\
\le&k\frac{\sigma_1(\mathbf{E})}{2\eta^2\lambda^2} + (r_c - k)I(r_c > k)\frac{\sqrt{\sigma_1(\mathbf{E})}}{\eta\lambda}.
\end{aligned} \tag{56}$$

For another part of (54),

$$\begin{aligned}
&\left\|\sum_{j=1}^{d} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\mathbf{C}\right\|_* \\
=&\left\|\sum_{j=1}^{k} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\mathbf{C}\right\|_* \\
\le&\sigma_1(\mathbf{C})\left\|\sum_{j=1}^{k} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\right\|_*.
\end{aligned} \tag{57}$$

Denote $\mathbf{U}_j = \sum_{j'=1}^{j} \mathbf{u}_{j'}\mathbf{u}_{j'}, \mathbf{U}_{j,0} = \sum_{j'=1}^{j} \mathbf{u}_{j',0}\mathbf{u}_{j',0}$ for $j \in [d]$. Let $\mathbf{U}_0 = \mathbf{U}_{0,0} = \mathbf{0}$. Then $\mathbf{u}_j\mathbf{u}_j = \mathbf{U}_j - \mathbf{U}_{j-1}, \mathbf{u}_{j,0}\mathbf{u}_{j,0} = \mathbf{U}_{j,0} - \mathbf{U}_{j-1,0}$ for $j \in [d]$.

Then we have

$$\begin{aligned}
&\left\|\sum_{j=1}^{k} \sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} - \mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\right\|_* \\
=&\left\|\sum_{j=1}^{k} \sigma_j(\mathbf{T}_0)(\mathbf{U}_j - \mathbf{U}_{j,0} - (\mathbf{U}_{j-1} - \mathbf{U}_{j-1,0}))\right\|_* \\
=&\left\|\sum_{j=1}^{k-1}(\sigma_j(\mathbf{T}_0) - \sigma_{j+1}(\mathbf{T}_0))(\mathbf{U}_j - \mathbf{U}_{j,0})\right. \\
&\left.+ \sigma_k(\mathbf{T}_0)(\mathbf{U}_k - \mathbf{U}_{k,0})\right\|_* \\
\le&\sum_{j=1}^{k-1}(\sigma_j(\mathbf{T}_0) - \sigma_{j+1}(\mathbf{T}_0))\|\mathbf{U}_j - \mathbf{U}_{j,0}\|_* \\
&+ \sigma_k(\mathbf{T}_0)\|\mathbf{U}_k - \mathbf{U}_{k,0}\|_*.
\end{aligned}$$

We assume $2\sigma_1(\mathbf{E}) \le \sigma_j(\mathbf{\Sigma}_0) - \sigma_{j+1}(\mathbf{\Sigma}_0)$ for all $j \in [k]$, and apply the Theorem 6 of Jiang et al. [7]. Then for $j \in [k]$,

$$\begin{aligned}
\|\mathbf{U}_j - \mathbf{U}_{j,0}\|_* \le& \min\{2j, k\}\|\mathbf{U}_j - \mathbf{U}_{j,0}\|_2 \\
\le& \min\{2j, k\}\frac{2\sigma_1(\mathbf{E})}{\sigma_j(\mathbf{\Sigma}_0) - \sigma_{j+1}(\mathbf{\Sigma}_0)}.
\end{aligned}$$

Since $j \in [k-1]$,

$$\begin{aligned}
\sigma_j(\mathbf{T}_0) - \sigma_{j+1}(\mathbf{T}_0) =&1 - \frac{\eta\lambda}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)}} - \left(1 - \frac{\eta\lambda}{\sqrt{\sigma_{j+1}(\mathbf{\Sigma}_0)}}\right) \\
=&\eta\lambda\frac{\sqrt{\sigma_j(\mathbf{\Sigma}_0)} - \sqrt{\sigma_{j+1}(\mathbf{\Sigma}_0)}}{\sqrt{\sigma_j(\mathbf{\Sigma}_0)\sigma_{j+1}(\mathbf{\Sigma}_0)}},
\end{aligned}$$

and

$$\sigma_k(\mathbf{T}_0) =1 - \frac{\eta\lambda}{\sqrt{\sigma_k(\mathbf{\Sigma}_0)}} \le 1 - \frac{\sqrt{\sigma_{k+1}(\mathbf{\Sigma}_0)}}{\sqrt{\sigma_k(\mathbf{\Sigma}_0)}},$$

therefore,

$$
\begin{aligned}
&\left\|\sum_{j=1}^{k}\sigma_j(\mathbf{T}_0)(\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}-\mathbf{u}_{j,0}\mathbf{u}_{j,0}^{\mathrm{T}})\right\|_*\\
\le&\sum_{j=1}^{k-1}\frac{2\eta\lambda\min\{2j,k\}\sigma_1(\mathbf{E})}{(\sqrt{\sigma_j(\mathbf{\Sigma}_0)}+\sqrt{\sigma_{j+1}(\mathbf{\Sigma}_0)})\sqrt{\sigma_j(\mathbf{\Sigma}_0)\sigma_{j+1}(\mathbf{\Sigma}_0)}}\\
&+\frac{2\eta\lambda k\sigma_1(\mathbf{E})}{(\sqrt{\sigma_k(\mathbf{\Sigma}_0)}+\sqrt{\sigma_{k+1}(\mathbf{\Sigma}_0)})\sqrt{\sigma_k(\mathbf{\Sigma}_0)}} \qquad (58)\\
\le&\sum_{j=1}^{k-1}\frac{2\eta\lambda\min\{2j,k\}\sigma_1(\mathbf{E})}{(\eta\lambda+\eta\lambda)\sqrt{\eta^2\lambda^2\eta^2\lambda^2}}+\frac{2\eta\lambda k\sigma_1(\mathbf{E})}{(\eta\lambda+0)\eta\lambda}\\
\le&\left(\frac{k(k-1)}{\eta^2\lambda^2}+\frac{2k}{\eta\lambda}\right)\sigma_1(\mathbf{E}).
\end{aligned}
$$

Combining (51), (52), (53), (54), (55), (56), (57) and (58), it follows that

$$
\begin{aligned}
&\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_F^2-\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_F^2\\
\le&2\Bigg[k\frac{\sigma_1(\mathbf{E})}{2\eta^2\lambda^2}+(r_c-k)I(r_c>k)\frac{\sqrt{\sigma_1(\mathbf{E})}}{\eta\lambda} \qquad (59)\\
&+\left(\frac{k(k-1)}{\eta^2\lambda^2}+\frac{2k}{\eta\lambda}\right)\sigma_1(\mathbf{E})\Bigg]\sigma_1^2(\mathbf{C}).
\end{aligned}
$$

On the other hand,

$$\|\widehat{\mathbf{W}}^{(t)}\|_*-\|\widehat{\mathbf{W}}_0^{(t)}\|_*\le\|\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)}\|_*.$$

As such, we have

$$
\begin{aligned}
&\frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_F^2+\lambda\|\widehat{\mathbf{W}}^{(t)}\|_*\\
&-\left\{\min_{\mathbf{W}}\frac{1}{2\eta}\|\mathbf{W}-\mathbf{C}\|_F^2+\lambda\|\mathbf{W}\|_*\right\}\\
=&\frac{1}{2\eta}(\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_F^2-\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_F^2)\\
&+\lambda(\|\widehat{\mathbf{W}}^{(t)}\|_*-\|\widehat{\mathbf{W}}_0^{(t)}\|_*)\\
\le&\left(\frac{\sigma_1^2(\mathbf{C})}{\eta}+\lambda\sigma_1(\mathbf{C})\right)\Bigg[k\frac{\sigma_1(\mathbf{E})}{2\eta^2\lambda^2}\\
&+(r_c-k)I(r_c>k)\frac{\sqrt{\sigma_1(\mathbf{E})}}{\eta\lambda}+\left(\frac{k(k-1)}{\eta^2\lambda^2}+\frac{2k}{\eta\lambda}\right)\sigma_1(\mathbf{E})\Bigg]\\
=&\frac{1}{\eta}\left(\frac{\sigma_1^2(\mathbf{C})}{\eta\lambda}+\sigma_1(\mathbf{C})\right)\Bigg[k\frac{\sigma_1(\mathbf{E})}{2\eta\lambda}\\
&+\max(0,r_c-k)\sqrt{\sigma_1(\mathbf{E})}+\left(\frac{k(k-1)}{\eta\lambda}+2k\right)\sigma_1(\mathbf{E})\Bigg].
\end{aligned} \qquad (60)
$$

□

F. Proof of Lemma 3

Proof. In the t-th step, a standard proximal operator (see Liu et al. [9]) optimizes the following problem:

$$\min_{\mathbf{W}}\frac{1}{2\eta}\|\mathbf{W}-\mathbf{C}\|_F^2+\lambda\|\mathbf{W}\|_{2,1},$$

where $\mathbf{C}=\widetilde{\mathbf{W}}_i^{(t-1)}$. By Theorem 5 of Liu et al. [9], denote the solution of the problem by $\widehat{\mathbf{W}}_0^{(t)}=\mathbf{S}_{\eta\lambda,0}\mathbf{C}$, $\mathbf{\Lambda}_0$ is a diagonal matrix containing the diagonal elements of $\mathbf{C}\mathbf{C}^{\mathrm{T}}$, $\mathbf{S}_0$ is a diagonal matrix and suffices $\mathbf{S}_{ii,0}=\sqrt{\mathbf{\Lambda}_{ii,0}}$ for $i=1,\ldots,\min\{d,m\}$. $\mathbf{S}_{\eta\lambda,0}$ is also a diagonal matrix and $\mathbf{S}_{\eta\lambda,ii,0}=\max\{0,1-\eta\lambda/\mathbf{S}_{ii,0}\}$ for $i=1,\ldots,\min\{d,m\}$.

By Algorithm 2, $\widehat{\mathbf{W}}^{(t)}=\mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}\mathbf{C}$.

Then we analyse the bound of $\frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_F^2+\lambda\|\widehat{\mathbf{W}}^{(t)}\|_{2,1}$ $-\{\frac{1}{2\eta}\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_F^2+\lambda\|\widehat{\mathbf{W}}_0^{(t)}\|_{2,1}\}$.

First, similarly as in (51), we have

$$
\begin{aligned}
&\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_F^2-\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_F^2\\
=&\mathrm{tr}((\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)})(\widehat{\mathbf{W}}^{(t)}-\mathbf{C})^{\mathrm{T}})\\
&+\mathrm{tr}((\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)})(\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C})^{\mathrm{T}})\\
=&\sum_{j=1}^{d}(\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)})^j((\widehat{\mathbf{W}}^{(t)}-\mathbf{C})^j)^{\mathrm{T}} \qquad (61)\\
&+\sum_{j=1}^{d}(\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)})^j((\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C})^j)^{\mathrm{T}}\\
\le&\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_{2,1}\|\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)}\|_{2,1}\\
&+\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_{2,1}\|\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)}\|_{2,1},
\end{aligned}
$$

where $(\cdot)^j$ denotes the j-th row vector of the enclosed matrix.

Denote $\mathbf{T}=\mathbf{S}_{\eta\lambda},\mathbf{T}_0=\mathbf{S}_{\eta\lambda,0}$. Denote the indices of non-zero rows of $\mathbf{C}$ by $\mathcal{I}_c=\{j:\mathbf{C}^j\neq\mathbf{0}\}$ and let $r_{c,s}=|\mathcal{I}_c|\le d$.

We have

$$
\begin{aligned}
&\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_{2,1}=\|(\mathbf{T}-\mathbf{I})\mathbf{C}\|_{2,1}\\
&=\sum_{j=1}^{d}\sqrt{\sum_{i=1}^{m}|(\mathbf{T}-\mathbf{I})^j\mathbf{C}_i|^2}=\sum_{j\in\mathcal{I}_c}\sqrt{\sum_{i=1}^{m}|(\mathbf{T}-\mathbf{I})_{jj}\mathbf{C}_{ij}|^2}\\
&=\sum_{j\in\mathcal{I}_c}\sqrt{\sum_{i=1}^{m}|(\mathbf{T}-\mathbf{I})_{jj}|^2|\mathbf{C}_{ij}|^2}=\sum_{j\in\mathcal{I}_c}|(\mathbf{T}-\mathbf{I})_{jj}|\|\mathbf{C}^j\|_2.
\end{aligned}
$$

Since $\mathbf{S}_{\eta\lambda}-\mathbf{I}$ is a diagonal matrix, whose i-th diagonal element is $\max\{0,1-\eta\lambda/\mathbf{S}_{ii}\}-1\in[-1,0)$, so

$$\|\widehat{\mathbf{W}}^{(t)}-\mathbf{C}\|_{2,1}\le\sum_{j\in\mathcal{I}_c}\|\mathbf{C}^j\|_2\le r_{c,s}\max_{j\in[d]}\|\mathbf{C}^j\|_2. \qquad (62)$$

Similarly,

$$\|\widehat{\mathbf{W}}_0^{(t)}-\mathbf{C}\|_{2,1}\le\sum_{j\in\mathcal{I}_c}\|\mathbf{C}^j\|_2\le r_{c,s}\max_{j\in[d]}\|\mathbf{C}^j\|_2. \qquad (63)$$

On the other hand,

$$
\begin{aligned}
&\|\widehat{\mathbf{W}}^{(t)}-\widehat{\mathbf{W}}_0^{(t)}\|_{2,1}=\|\mathbf{S}_{\eta\lambda}\mathbf{C}-\mathbf{S}_{\eta\lambda,0}\mathbf{C}\|_{2,1}\\
=&\sum_{j\in\mathcal{I}_c}|\mathbf{S}_{\eta\lambda,jj}-\mathbf{S}_{\eta\lambda,jj,0}|\|\mathbf{C}^j\|_2 \qquad (64)\\
\le&\max_{j'\in[d]}\|\mathbf{C}^{j'}\|_2\sum_{j\in\mathcal{I}_c}|\mathbf{S}_{\eta\lambda,jj}-\mathbf{S}_{\eta\lambda,jj,0}|.
\end{aligned}
$$

Denote $\mathbf{\Sigma}_0=\widetilde{\mathbf{\Sigma}}^{(t)}=\mathbf{C}\mathbf{C}^{\mathrm{T}}$. Then we have for $j\in\mathcal{I}_c$,

$$
\begin{aligned}
&|\mathbf{S}_{\eta\lambda,jj}-\mathbf{S}_{\eta\lambda,jj,0}|\\
=&\left|\max\left(0,1-\frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}+\mathbf{E}_{jj}|}}\right)-\max\left(0,1-\frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}|}}\right)\right|.
\end{aligned}
$$

Case 1: $\eta\lambda>\sqrt{\mathbf{\Sigma}_{jj,0}}$. Then

$$
\begin{aligned}
&|\mathbf{S}_{\eta\lambda,jj}-\mathbf{S}_{\eta\lambda,jj,0}|\\
&=\max\left(0,1-\frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}+\mathbf{E}_{jj}|}}\right)\le1-\frac{\eta\lambda}{\sqrt{\eta^2\lambda^2+|\mathbf{E}_{jj}|}}\\
&\le1-\frac{\eta\lambda}{\eta\lambda+\sqrt{|\mathbf{E}_{jj}|}}=\frac{\sqrt{|\mathbf{E}_{jj}|}}{\eta\lambda+\sqrt{|\mathbf{E}_{jj}|}}\le\frac{\sqrt{|\mathbf{E}_{jj}|}}{\eta\lambda}
\end{aligned}
$$

Case 2: $\eta\lambda \leq \sqrt{\mathbf{\Sigma}_{jj,0}}$. Then

$$\begin{aligned}&|\mathbf{S}_{\eta\lambda,jj} - \mathbf{S}_{\eta\lambda,jj,0}| \\ \leq&1 - \frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}| + |\mathbf{E}_{jj}|}} - 1 + \frac{\eta\lambda}{\sqrt{|\mathbf{\Sigma}_{jj,0}|}} \\ =&\eta\lambda \cdot \frac{\sqrt{|\mathbf{\Sigma}_{jj,0}| + |\mathbf{E}_{jj}|} - \sqrt{|\mathbf{\Sigma}_{jj,0}|}}{\sqrt{|\mathbf{\Sigma}_{jj,0}|(|\mathbf{\Sigma}_{jj,0}| + |\mathbf{E}_{jj}|)}} \\ =&\frac{\eta\lambda|\mathbf{E}_{jj}|}{[\sqrt{|\mathbf{\Sigma}_{jj,0}| + |\mathbf{E}_{jj}|} + \sqrt{|\mathbf{\Sigma}_{jj,0}|}]\sqrt{|\mathbf{\Sigma}_{jj,0}|^2 + |\mathbf{E}_{jj}||\mathbf{\Sigma}_{jj,0}|}} \\ \leq&\frac{\eta\lambda|\mathbf{E}_{jj}|}{[\sqrt{\eta^2\lambda^2 + 0} + \sqrt{\eta^2\lambda^2}]\sqrt{\eta^4\lambda^4 + 0}} = \frac{|\mathbf{E}_{jj}|}{2\eta^2\lambda^2}.\end{aligned}$$

Suppose that there exists an integer $k \leq d$ such that

$$\sum_{j=1}^{d} I(\sqrt{\mathbf{\Sigma}_{jj,0}} \geq \eta\lambda) = k$$

then $k \leq r_{c,s}$ and

$$\begin{aligned}&\sum_{j\in\mathcal{I}_c} |\mathbf{S}_{\eta\lambda,jj} - \mathbf{S}_{\eta\lambda,jj,0}| \\ \leq&\frac{k}{2\eta^2\lambda^2} \max_{j:\eta^2\lambda^2\leq\mathbf{\Sigma}_{jj,0}} \mathbf{E}_{jj} \\ &+ \frac{(r_{c,s}-k)I(r_{c,s}>k)}{\eta\lambda} \max_{j:\eta^2\lambda^2>\mathbf{\Sigma}_{jj,0}} \sqrt{\mathbf{E}_{jj}}.\end{aligned} \tag{65}$$

Combining (61), (62), (63), (64) and (65), it follows that

$$\begin{aligned}&\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 - \|\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}\|_F^2 \\ \leq&2r_{c,s}\left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)^2\left(\frac{k}{2\eta^2\lambda^2}\max_{j:\eta^2\lambda^2\leq\mathbf{\Sigma}_{jj,0}}|\mathbf{E}_{jj}|\right. \\ &\left.+ \frac{(r_{c,s}-k)I(r_{c,s}>k)}{\eta\lambda}\max_{j:\eta^2\lambda^2>\mathbf{\Sigma}_{jj,0}}\sqrt{|\mathbf{E}_{jj}|}\right).\end{aligned} \tag{66}$$

On the other hand,

$$\|\widehat{\mathbf{W}}^{(t)}\|_{2,1} - \|\widehat{\mathbf{W}}_0^{(t)}\|_{2,1} \leq \|\widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{W}}_0^{(t)}\|_{2,1}.$$

As such, we have

$$\begin{aligned}&\frac{1}{2\eta}\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 + \lambda\|\widehat{\mathbf{W}}^{(t)}\|_{2,1} \\ &- \left\{\min_{\mathbf{W}} \frac{1}{2\eta}\|\mathbf{W} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}\right\} \\ =&\frac{1}{2\eta}(\|\widehat{\mathbf{W}}^{(t)} - \mathbf{C}\|_F^2 - \|\widehat{\mathbf{W}}_0^{(t)} - \mathbf{C}\|_F^2) \\ &+ \lambda(\|\widehat{\mathbf{W}}^{(t)}\|_{2,1} - \|\widehat{\mathbf{W}}_0^{(t)}\|_{2,1}) \\ \leq&\left[\frac{r_{c,s}}{\eta}\left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)^2 + \lambda\left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)\right] \\ &\cdot\left[\frac{k}{2\eta^2\lambda^2}\max_{j:\eta^2\lambda^2\leq\mathbf{\Sigma}_{jj,0}}|\mathbf{E}_{jj}|\right. \\ &\left.+ \frac{(r_{c,s}-k)I(r_{c,s}>k)}{\eta\lambda}\max_{j:\eta^2\lambda^2>\mathbf{\Sigma}_{jj,0}}\sqrt{|\mathbf{E}_{jj}|}\right] \\ =&\frac{1}{\eta}\left[\frac{r_{c,s}}{\eta\lambda}\left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)^2 + \left(\max_{j\in[d]}\|\mathbf{C}^j\|_2\right)\right] \\ &\cdot\left[\frac{k}{2\eta\lambda}\max_{j:\eta^2\lambda^2\leq\mathbf{\Sigma}_{jj,0}}|\mathbf{E}_{jj}|\right. \\ &\left.+ \max(0, r_{c,s}-k)\max_{j:\eta^2\lambda^2>\mathbf{\Sigma}_{jj,0}}\sqrt{|\mathbf{E}_{jj}|}\right].\end{aligned} \tag{67}$$

□

G. Proof of Theorem 2

Proof. First, consider the case with no acceleration. We first use Proposition 1 of Schmidt et al. [11] by regarding procedures from Step 5 to Step 9 as approximation for the proximal operator in (8). Note that the norm clipping only bounds the parameter space and does not affect the results of Schmidt et al. [11]. Then for ε_t defined in Lemma 13 for $t \in [T]$, we have

$$\begin{aligned}\mathcal{E} =&\frac{2L}{m(T+1)^2}\Bigg(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F \\ &+ 2\sum_{t=1}^{T} t\sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^{T} t^2\frac{\varepsilon_t}{L}}\Bigg)^2.\end{aligned}$$

Meanwhile, by Lemma 13, we have

$$\varepsilon_t = O\left(\frac{\kappa}{\epsilon_t}\right),$$

where $\kappa = \frac{K^2\sqrt{m}kd\log d}{\eta}$

On the other hand, because

$$\epsilon = \sum_{t=1}^{T} \frac{(e^{\epsilon_t}-1)\epsilon_t}{(e^{\epsilon_t}+1)} + \sqrt{\sum_{t=1}^{T} 2\epsilon_t^2 \log\left(e + \frac{\sqrt{\sum_{t=1}^{T}\epsilon_t^2}}{\delta}\right)},$$

then by Lemma 16, we have

$$\sqrt{\sum_{t=1}^{T}\epsilon_t^2} \geq \frac{\sqrt{2}\epsilon}{2\sqrt{\log(e+\epsilon/\sqrt{2}\delta) + 2\epsilon}} = c_2.$$

Then by Lemma 17, we have

$$\sum_{t=1}^{T}\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1/2}}{c_2(\alpha/2-1)^2\sqrt{2\alpha+1}}}\right), & \alpha > 2; \\ O\left(\sqrt{\frac{\kappa T^{5/2}}{c_2(\alpha/2-1)^2\sqrt{2\alpha+1}}}\right), & -1/2 < \alpha < 2; \\ O\left(\sqrt{\frac{\kappa T^{2-\alpha}}{c_2(\alpha/2-1)^2\sqrt{-2\alpha-1}}}\right), & \alpha < -1/2. \end{cases}$$

Because $\widetilde{\mathbf{W}}^{(0)}$ is the result of the norm clipping, we have $\widetilde{\mathbf{W}}^{(0)} \in \mathcal{W}$.

Finally, taking $c_3 = \phi(\alpha)$ defined in (13) and $c_4 = \frac{\kappa}{c_2(\alpha/2-1)^2\sqrt{|2\alpha+1|}}$, under the assumption that $\mathbf{W}_* \in \mathcal{W}$, using Lemma 19, we have the results for the case with no acceleration.

For the accelerated case, we use Proposition 2 of Schmidt et al. [11] to have

$$\begin{aligned}\mathcal{E} =&\frac{2L}{m(T+1)^2}\Bigg(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F \\ &+ 2\sum_{t=1}^{T} t\sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^{T} t^2\frac{\varepsilon_t}{L}}\Bigg)^2.\end{aligned}$$

Then one can prove similarly combining Lemma 13, Lemma 16, Lemma 17 and Lemma 20.

□

H. Proof of Theorem 3

Proof. First, consider the case with no acceleration. We use Proposition 1 of Schmidt et al. [11] and prove similarly as in Appendix L-G, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 19.

For the accelerated case, we use Proposition 2 of Schmidt et al. [11] and prove similarly as in Appendix L-G, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 20. □

I. Proof of Theorem 4

Proof. First, consider the case with no acceleration. We use Proposition 3 of Schmidt et al. [11] to have

$$\mathcal{E} = \frac{Q_0^T}{\sqrt{m}}\Big(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F + 2\sum_{t=1}^{T} Q_0^{-t}\sqrt{\frac{2\varepsilon_t}{L}}\Big).$$

Then one can prove similarly as in Appendix L-G, combining Lemma 13, Lemma 16, Lemma 17 and Lemma 21.

For the accelerated case, we use Proposition 4 of Schmidt et al. [11] to have

$$\mathcal{E} =\frac{(Q_0)^T}{m}\Bigg(\sqrt{2(f(\widehat{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*))} + 2\sqrt{\frac{L}{\mu}}\sum_{t=1}^{T}\sqrt{\varepsilon_t (Q_0)^{-t}} + \sqrt{\sum_{t=1}^{T}\varepsilon_t (Q_0)^{-t}}\Bigg)^2.$$

Then one can prove similarly as in Appendix L-G, using the assumption that $f(\widetilde{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*) = O(K^2 Lm)$, combining Lemma 13, Lemma 16, Lemma 17 and Lemma 22. □

J. Proof of Theorem 5

Proof. First, consider the case with no acceleration. We use Proposition 3 of Schmidt et al. [11] and prove similarly as in Appendix L-I, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 21.

For the accelerated case, we use Proposition 4 of Schmidt et al. [11] and prove similarly as in Appendix L-I, using the assumption that $f(\widetilde{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*) = O(K^2 Lm)$, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 22. □

K. Proof of Theorem 6

Proof. Consider the bound in (12), whose logarithm is

$$\phi(\alpha)\log\Big(\frac{kd\log d\sqrt{\log(e+\epsilon/\sqrt{2}\delta)+2\epsilon}}{\sqrt{m}\epsilon}\Big) - \phi(\alpha)\log((\alpha/2-1)^2\sqrt{|2\alpha+1|}) + \log(K^2 L)$$

By Assumption 1, the first term dominates. Then we should firstly maximize $\phi(\alpha)$, which results in that $\phi(\alpha) = 2/5$ and $-1/2 < \alpha < 2$. Then since $\phi(\alpha)$ is now fixed, we maximize $(\alpha/2-1)^2\sqrt{|2\alpha+1|}$, which results in $\alpha = 0$. Results under other settings can be proved similarly. □

L. Proof of Proposition 2

Proof. First, consider the method of Pathak et al. [10].

By Definition 7, an (ϵ,δ)-Iterative DP-MTL algorithm with $T = 1$ should suffice for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)}$ and all $i \in [m]$ that

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1)} \in \mathcal{S} \mid \mathbf{W}^{(0)}, \mathcal{D}^m)\\ \le& e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1)} \in \mathcal{S} \mid (\mathbf{W}')^{(0)}, (\mathcal{D}')^m) + \delta.\end{aligned}$$

On the other hand, for the ϵ given in the method of Pathak et al. [10], using Theorem 4.1 of Pathak et al. [10], taking $D = \mathcal{D}^m$ and $D' = (\mathcal{D}')^m$, we have for any set $\mathcal{S} \subseteq \mathbb{R}^d$,

$$\mathbb{P}(\hat{\mathbf{w}}^s \in \mathcal{S} \mid \mathcal{D}^m) \le e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}^s \in \mathcal{S} \mid (\mathcal{D}')^m),$$

where $\hat{\mathbf{w}}^s$ is defined in Section 3.2 of Pathak et al. [10].

Because the method of Pathak et al. [10] uses $\hat{\mathbf{w}}^s$ for all the tasks, then we have $\hat{\mathbf{w}}_i^{(1)} = \hat{\mathbf{w}}^s$ for all $i \in [m]$.

As such, denote $\mathbf{W}^{(0)}$ and $(\mathbf{W}')^{(0)}$ as the collections of models independently learned using $\mathcal{D}^m$ and $(\mathcal{D}')^m$, respectively. Then $\mathcal{D}^m$ and $(\mathcal{D}')^m$ contain all the information of $\mathbf{W}^{(0)}$ and $(\mathbf{W}')^{(0)}$, respectively. As such, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)}$, all $i \in [m]$ and $\delta = 0$ that

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1)} \in \mathcal{S} \mid \mathbf{W}^{(0)}, \mathcal{D}^m)\\ \le& e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1)} \in \mathcal{S} \mid (\mathbf{W}')^{(0)}, (\mathcal{D}')^m) + \delta,\end{aligned}$$

which shows that the method of Pathak et al. [10] is an (ϵ,δ)-Iterative DP-MTL algorithm with $T = 1$ and $\delta = 0$.

Then we consider the method of Gupta et al. [5]. Assume a constant $\delta \ge 0$ and the number of iteration T is given.

Taking $T_0 = m, t = i$ for $i \in [m]$, $\beta_t = \hat{\mathbf{w}}_i$, $\mathscr{D} = \mathcal{D}^m$, for the ϵ given in the method of Gupta et al. [5], using Theorem 1 of Gupta et al. [5], for $t \in [T]$, we have *in the t-th each iteration*, for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times m}$ and all $i \in [m]$,

$$\mathbb{P}(\hat{\mathbf{W}}^{(t)} \in \mathcal{S} \mid \mathcal{D}^m) \le e^{\epsilon}\mathbb{P}(\hat{\mathbf{W}}^{(t)} \in \mathcal{S} \mid (\mathcal{D}')^m),$$

which suggests that for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)}$ and all $i \in [m]$,

$$\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(t)} \in \mathcal{S} \mid \mathcal{D}^m) \le e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(t)} \in \mathcal{S} \mid (\mathcal{D}')^m).$$

Then for all $i \in [m]$ and for all $t \in [T]$, take the t-th output $\theta_{t,i} = \hat{\mathbf{w}}_{[-i]}^{(t)}$ and $\delta_t = 0$.

Therefore by the *Adaptive composition* Lemma (Lemma 7), for all $i \in [m]$ and for any set $\mathcal{S} \subset \mathbb{R}^{d\times(m-1)\times T}$,

$$\begin{aligned}&\mathbb{P}((\theta_{1,i},\ldots,\theta_{T,i}) \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})))\\ \le& e^{\tilde{\epsilon}}\mathbb{P}((\theta_{1,i},\cdots,\theta_{T,i}) \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = ((\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1})))\\ &+\delta,\end{aligned}$$

where for all $t \in [T]$, $\mathcal{B}_t$ denotes the input for the t-th iteration,

$$\boldsymbol{\theta}_{1:t-1} = \begin{cases}\emptyset, & t=1\\ (\theta_{1,1},\ldots,\theta_{1,m})\ldots,(\theta_{t-1,1},\ldots,\theta_{t-1,m}), & t\ge 2,\end{cases}$$

and $\tilde{\epsilon}$ is defined as follows.

$$\begin{aligned}\tilde{\epsilon} = \min\Bigg\{&\sum_{t=1}^{T}\epsilon, \sum_{t=1}^{T}\frac{(e^{\epsilon}-1)\epsilon}{(e^{\epsilon}+1)} + \sqrt{\sum_{t=1}^{T}2\epsilon^2\log\Big(\frac{1}{\delta}\Big)},\\ &\sum_{t=1}^{T}\frac{(e^{\epsilon}-1)\epsilon}{(e^{\epsilon}+1)} + \sqrt{\sum_{t=1}^{T}2\epsilon^2\log\Big(e+\frac{\sqrt{\sum_{t=1}^{T}\epsilon^2}}{\delta}\Big)}\Bigg\}.\end{aligned}$$

As such, in each t-th iteration, denote $\mathbf{W}^{(t-1)}$ and $(\mathbf{W}')^{(t-1)}$ as the collections of models independently learned using $(\mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})$ and $((\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1})$, respectively. Then $(\mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})$ and $(\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1})$ contain all the information of $\mathbf{W}^{(t-1)}$ and $(\mathbf{W}')^{(t-1)}$, respectively.

Therefore, we have for any set $\mathcal{S} \subset \mathbb{R}^{d\times(m-1)\times T}$,

$$\begin{aligned}&\mathbb{P}(\mathbf{w}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\mathbf{W}^{(t-1)}, \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})))\\ \le& e^{\tilde{\epsilon}}\mathbb{P}(\mathbf{w}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = ((\mathbf{W}')^{(t-1)}, (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1})))\\ &+\delta,\end{aligned}$$

which shows that by Definition 7, the method of Gupta et al. [5] is an $(\tilde{\epsilon},\delta)$-Iterative DP-MTL algorithm. □

M. Proof of Proposition 3

Proof. Given an (ϵ, δ) - iterative DP-MTL algorithm $\mathcal{A}(\mathcal{B})$, by Definition 7, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)\times T}$ that

$$
\begin{aligned}
&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}^{(t-1)}, \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1}))\\
&\leq \exp(\epsilon)\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = ((\mathbf{W}')^{(t-1)}, (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}))\\
&\quad + \delta.
\end{aligned}
$$

Furthermore, following the proof of the *Group privacy* Lemma (Lemma 5), shown by Vadhan [12], for protecting the entire dataset, n data instances, of the i-th task, we construct a series of datasets, $\mathcal{D}_{(0)}^m, \mathcal{D}_{(1)}^m, \ldots, \mathcal{D}_{(n)}^m$, and let $\mathcal{D}_{(0)}^m = \mathcal{D}^m, \mathcal{D}_{(n)}^m = (\mathcal{D}')^m$ such that for $j = 0, \ldots, n-1$, $\mathcal{D}_{(j)}^m$ and $\mathcal{D}_{(j+1)}^m$ are two neighboring datasets that differ in one data instance. Let a series of model matrices, $\mathbf{W}_{(0)}, \ldots, \mathbf{W}_{(n)}$, where $\mathbf{W}_{(0)} = \mathbf{W}, \mathbf{W}_{(n)} = \mathbf{W}'$, be the input model matrices in those settings. Let a series of output objects $\boldsymbol{\theta}_{1:t-1}^{(0)}, \ldots, \boldsymbol{\theta}_{1:t-1}^{(n)}$, where $\boldsymbol{\theta}_{1:t-1}^{(0)} = \boldsymbol{\theta}_{1:t-1}, \boldsymbol{\theta}_{1:t-1}^{(n)} = \mathbf{W}'$, be the output objects in those settings.

Then, we have

$$
\begin{aligned}
&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}_{(0)}^{(t-1)}, \mathcal{D}_{(0)}^m, \boldsymbol{\theta}_{1:t-1}^{(0)}))\\
&\leq \exp(\epsilon)\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}_{(1)}^{(t-1)}, \mathcal{D}_{(1)}^m, \boldsymbol{\theta}_{1:t-1}^{(1)}))\\
&\quad + \delta\\
&\quad \vdots\\
&\leq \exp(n\epsilon)\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}_{(n)}^{(t-1)}, \mathcal{D}_{(n)}^m, \boldsymbol{\theta}_{1:t-1}^{(n)}))\\
&\quad + (1 + \exp(\epsilon) + \ldots + \exp((n-1)\epsilon))\delta\\
&\leq \exp(n\epsilon)\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}_{(n)}^{(t-1)}, \mathcal{D}_{(n)}^m, \boldsymbol{\theta}_{1:t-1}^{(n)}))\\
&\quad + n\exp(n\epsilon)\delta,
\end{aligned}
$$

which renders $\mathcal{A}$ as an $(n\epsilon, n\exp(n\epsilon)\delta)$ - iterative MP-MTL algorithm. □

Appendix M
Proof of the Results in Appendix C and Appendix D

A. Proof of Corollary 2

Proof. For simplicity, we omit the symbol $\mathcal{B}$ to denote the input in the conditional events in some equations.

Use Corollary 1 and Theorem 1. Given $t \in [T]$, the algorithm $(\mathbf{P}^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}) \to (\mathbf{M}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ is an $(\epsilon_t, 0)$-differentially private algorithm, where $\mathbf{M}^{(t)} = \mathbf{U}\mathbf{S}_{\eta\lambda}\mathbf{U}^{\mathrm{T}}$.

Now, for all $i \in [m]$, applying the *Post-Processing immunity* Lemma (Lemma 4) for the mapping $f : (\mathbf{M}^{(t)}, \mathbf{p}_{[-i]}^{(t-1)}) \to \hat{\mathbf{p}}_{[-i]}^{(t-1)}$, which does not touch any unperturbed sensitive information of the i-th task, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)}$ that

$$
\begin{aligned}
&\mathbb{P}(\hat{\mathbf{p}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid \mathbf{P}^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)})\\
&\leq e^{\epsilon_t}\mathbb{P}(\hat{\mathbf{p}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid (\mathbf{P}')^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}),
\end{aligned}
$$

where $\mathbf{P}^{(t-1)}$ and $(\mathbf{P}')^{(t-1)}$ differ only in the i-th column.

Then, because in the t-th iteration the mapping $\mathbf{Q}^{(t-1)} \to \widehat{\mathbf{Q}}^{(t-1)}$ is a deterministic STL algorithm, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)}$ that

$$
\begin{aligned}
&\mathbb{P}(\hat{\mathbf{q}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid \mathbf{Q}^{(t-1)})\\
&=\mathbb{P}(\hat{\mathbf{q}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid \mathbf{q}_{[-i]}^{(t-1)}, \mathbf{q}_i^{(t-1)})\\
&=\mathbb{P}(\hat{\mathbf{q}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid \mathbf{q}_{[-i]}^{(t-1)}, (\mathbf{q}_i')^{(t-1)})\\
&=e^0\mathbb{P}(\hat{\mathbf{q}}_{[-i]}^{(t-1)} \in \mathcal{S} \mid (\mathbf{Q}')^{(t-1)}) + 0,
\end{aligned}
$$

where $\mathbf{Q}^{(t-1)}$ and $(\mathbf{Q}')^{(t-1)}$ differ only in the i-th column.

Then applying *Combination* Lemma (Lemma 6), we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$

$$
\begin{aligned}
&\mathbb{P}((\hat{\mathbf{p}}_{[-i]}^{(t-1)}, \hat{\mathbf{q}}_{[-i]}^{(t-1)}) \in \mathcal{S} \mid \mathbf{P}^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}, \mathbf{Q}^{(t-1)})\\
&\leq e^{\epsilon_t}\mathbb{P}((\hat{\mathbf{p}}_{[-i]}^{(t-1)}, \hat{\mathbf{q}}_{[-i]}^{(t-1)}) \in \mathcal{S} \mid (\mathbf{P}')^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}, (\mathbf{Q}')^{(t-1)}),
\end{aligned}
$$

Because the mapping$(\widehat{\mathbf{P}}^{(t-1)}, \widehat{\mathbf{Q}}^{(t-1)}, \mathcal{D}^m) \to (\widehat{\mathbf{P}}^{(t)}, \widehat{\mathbf{Q}}^{(t)})$ is a deterministic STL algorithm, applying Lemma 8, we further have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$ that

$$
\begin{aligned}
&\mathbb{P}((\hat{\mathbf{p}}_{[-i]}^{(t)}, \hat{\mathbf{q}}_{[-i]}^{(t)}) \in \mathcal{S} \mid \mathbf{P}^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}, \mathbf{Q}^{(t-1)}, \mathcal{D}^m)\\
&\leq e^{\epsilon_t}\mathbb{P}((\hat{\mathbf{p}}_{[-i]}^{(t)}, \hat{\mathbf{q}}_{[-i]}^{(t)}) \in \mathcal{S} \mid (\mathbf{P}')^{(t-1)}, \boldsymbol{\Sigma}^{(1:t-1)}, (\mathbf{Q}')^{(t-1)}, (\mathcal{D}')^m),
\end{aligned}
$$

where $(\mathcal{D}')^m$ differs from $\mathcal{D}^m$ in the entire dataset of the i-th task.

Now, using Theorem 1, for $t = 1, \ldots, T$, we again take the t-th dataset $\widetilde{\mathcal{D}}_t = \{(\mathbf{p}_1^{(t-1)}, \mathbf{q}_1^{(t-1)}, \mathcal{D}_1), \ldots, (\mathbf{p}_m^{(t-1)}, \mathbf{q}_m^{(t-1)}\mathcal{D}_m)\}$ and denote $\vartheta_{t,i} = (\hat{\mathbf{q}}_{[-i]}^{(t)}, \hat{\mathbf{q}}_{[-i]}^{(t)}, \mathbf{M}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \in \mathcal{C}_{t,i}$. Given the fact that $\mathbf{P}^{(t)} = \widehat{\mathbf{P}}^{(t)}$ and $\mathbf{Q}^{(t)} = \widehat{\mathbf{Q}}^{(t)}$ for all $t \in [T]$, we have for any set $\mathcal{S}_{t,i} \subseteq \mathcal{C}_{t,i}$ that

$$
\begin{aligned}
&\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid \widetilde{\mathcal{D}}_t, \boldsymbol{\vartheta}_{1:t-1})\\
&\leq e^{\epsilon_t}\mathbb{P}(\vartheta_{t,i} \in \mathcal{S}_{t,i} \mid \widetilde{\mathcal{D}}_t', \boldsymbol{\vartheta}_{1:t-1}),
\end{aligned}
$$

where $\widetilde{\mathcal{D}}_t$ and $\widetilde{\mathcal{D}}_t'$ are two adjacent datasets that differ in a single entry, the i-th "data instance" $(\mathbf{p}_i^{(t-1)}, \mathbf{q}_i^{(t-1)}, \mathcal{D}_i = (\mathbf{X}_i, \mathbf{y}_i))$, and

$$
\boldsymbol{\vartheta}_{1:t-1} = \begin{cases} \emptyset, & t = 1\\ (\vartheta_{1,1}, \ldots, \vartheta_{1,m}) \ldots, (\vartheta_{t-1,1}, \ldots, \vartheta_{t-1,m}), & t \geq 2. \end{cases}
$$

This renders the algorithm in the t-th iteration as an $(\epsilon_t, 0)$-differentially private algorithm.

Then, again by the *Adaptive composition* Lemma (Lemma 7), for all $i \in [m]$ and for any set $\mathcal{S}' \subseteq \bigotimes_{t=1}^{T} \mathcal{C}_{t_i}$, we have

$$
\begin{aligned}
&\mathbb{P}((\vartheta_{1,i}, \cdots, \vartheta_{T,i}) \in \mathcal{S}' \mid \bigcap_{t=1}^{T} (\mathcal{B}_t = (\widetilde{\mathcal{D}}_t, \boldsymbol{\vartheta}_{1:t-1})))\\
&\leq e^{\tilde{\epsilon}}\mathbb{P}((\vartheta_{1,i}, \cdots, \vartheta_{T,i}) \in \mathcal{S}' \mid \bigcap_{t=1}^{T} (\mathcal{B}_t = (\widetilde{\mathcal{D}}_t', \boldsymbol{\vartheta}_{1:t-1})))\\
&\quad + \delta,
\end{aligned}
$$

where for all $t \in [T]$, $\mathcal{B}_t$ denotes the input for the t-th iteration.

Finally, for all $t \in [T]$, taking $\theta_t = (\vartheta_{t,1}, \ldots, \vartheta_{t,m})$ and given the fact that $\widehat{\mathbf{W}}^{(t)} = \widehat{\mathbf{P}}^{(t)} + \widehat{\mathbf{Q}}^{(t)}$, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)\times T}$ that

$$
\begin{aligned}
&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = (\mathbf{W}^{(t-1)}, \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1}))\\
&\leq e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S} \mid \bigcap_{t=1}^{T} \mathcal{B}_t = ((\mathbf{W}')^{(t-1)}, (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}))\\
&\quad + \delta,
\end{aligned}
$$

where $(\mathbf{W}')^{(t-1)}$ are associated with the setting in which the i-th task has been replaced. □

B. Proof of Proposition 4

Proof. For simplicity, we omit the symbol $\mathcal{B}$ used to denote the input in the conditional events in some equations.

First, the procedure from the 4-th step to the 5-th step is a standard output perturbation of Chaudhuri et al. [2]; thus, we have for all $i \in [m]$, for all neighboring datasets $\mathcal{D}^m$ and $(\mathcal{D}')^m$ that differ in a single data instance of the i-th task, and for any set $\mathcal{S} \in \mathbb{R}^d$,

$$\begin{aligned}&\mathbb{P}(\tilde{\mathbf{w}}_i^{(t-1)} \in \mathcal{S} \mid \tilde{\mathbf{w}}_i^{(0:t-2)}, \mathcal{D}^m, \mathbf{M}^{(t-1)})\\ &\leq \exp(\epsilon_{\text{dp},t})\mathbb{P}(\tilde{\mathbf{w}}_i^{(t-1)} \in \mathcal{S} \mid \tilde{\mathbf{w}}_i^{(0:t-2)}, (\mathcal{D}')^m, \mathbf{M}^{(t-1)}),\end{aligned}$$

where $\tilde{\mathbf{w}}_i^{(0:t-2)} = \emptyset$ when $t = 1$.

Then, because the mapping $(\widetilde{\mathbf{W}}^{(t-1)}, \mathbf{\Sigma}^{(1:t-1)}) \rightarrow \theta_t = (\mathbf{\Sigma}^{(t)}, \mathbf{M}^{(t)}, \widetilde{\mathbf{W}}^{(t-1)}) \in \mathcal{C}_t$ does not touch any unperturbed sensitive information of $(\mathbf{X}_i, \mathbf{y}_i, \mathbf{w}_i^{(0:t-1)})$, the *Post-Processing immunity* Lemma (Lemma 4) can be applied such that we have for any set $\mathcal{S}' \subseteq \mathcal{C}_t$ that

$$\begin{aligned}&\mathbb{P}(\theta_t \in \mathcal{S}' \mid \widetilde{\mathbf{W}}^{(0:t-2)}, \mathcal{D}^m, \mathbf{M}^{(t-1)})\\ &\leq \exp(\epsilon_{\text{dp},t})\mathbb{P}(\theta_t \in \mathcal{S}' \mid \widetilde{\mathbf{W}}^{(0:t-2)}, (\mathcal{D}')^m, \mathbf{M}^{(t-1)}),\end{aligned}$$

which means that

$$\begin{aligned}&\mathbb{P}(\theta_t \in \mathcal{S}' \mid \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})\\ &\leq \exp(\epsilon_{\text{dp},t})\mathbb{P}(\theta_t \in \mathcal{S}' \mid (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}),\end{aligned}$$

where

$$\boldsymbol{\theta}_{1:t-1} = \begin{cases} \emptyset, & t = 1\\ \theta_1, \theta_2, \cdots, \theta_{t-1}, & t \geq 2.\end{cases}$$

Then, by the *Adaptive composition* Lemma (Lemma 7), we have for any set $\mathcal{S}'' \subseteq \bigotimes_{t=1}^T \mathcal{C}_t$ that

$$\begin{aligned}&\mathbb{P}(\theta_{1:T} \in \mathcal{S}'' \mid \bigcap_{t=1}^T (\mathcal{B}_t = (\mathcal{D}^m, \boldsymbol{\theta}_{1:t-1})))\\ &\leq \exp(\epsilon_{\text{dp}})\mathbb{P}(\theta_{1:T} \in \mathcal{S}'' \mid \bigcap_{t=1}^T (\mathcal{B}_t = ((\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1})))\\ &\quad + \delta_{\text{dp}}.\end{aligned}$$

Because the mapping $(\theta_t, \mathcal{D}_{[-i]}, \tilde{\mathbf{w}}_{[-i]}^{(0:t-2)}, \mathbf{W}^{(t-1)}) \rightarrow \hat{\mathbf{w}}_{[-i]}^{(t)}$ does not touch any unperturbed sensitive information of $(\mathbf{X}_i, \mathbf{y}_i, \mathbf{w}_i^{(0:t-1)})$ for all $t \in [T]$ ($\mathbf{W}^{(t-1)}$ is actually not used in the mapping), the *Post-Processing immunity* Lemma (Lemma 4) can be applied such that we have for any set $\mathcal{S}_0 \subseteq \mathbb{R}^{d\times(m-1)\times T}$ that

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S}_0 \mid \bigcap_{t=1}^T (\mathcal{B}_t = (\mathcal{D}^m, \boldsymbol{\theta}_{1:t-1}, \mathbf{W}^{(t-1)})))\\ &\leq e^{\epsilon_{\text{dp},t}}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S}_0 \mid \bigcap_{t=1}^T (\mathcal{B}_t = ((\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}, (\mathbf{W}')^{(t-1)})))\\ &\quad + \delta_{\text{dp}},\end{aligned}$$

where $(\mathbf{W}')^{(t-1)}$ is associated with the setting in which a single data instance of the i-th task has been replaced.

Therefore, Algorithm 5 is an $(\epsilon_{\text{dp}}, \delta_{\text{dp}})$ - iterative DP-MTL algorithm.

Next, for the conditional density of $\mathbf{\Sigma}^{(t)}$ given $\mathbf{W}^{(t-1)}$, we have

$$\begin{aligned}&p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})\\ &= \int_{\widetilde{\mathbf{W}}^{(t-1)}} p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)}, \widetilde{\mathbf{W}}^{(t-1)})\\ &\quad p(\widetilde{\mathbf{W}}^{(t-1)} \mid \mathbf{W}^{(t-1)}) d\widetilde{\mathbf{W}}^{(t-1)}\\ &= \int_{\widetilde{\mathbf{W}}^{(t-1)}} p(\mathbf{\Sigma}^{(t)} \mid \widetilde{\mathbf{W}}^{(t-1)}) p(\widetilde{\mathbf{W}}^{(t-1)} \mid \mathbf{W}^{(t-1)}) d\widetilde{\mathbf{W}}^{(t-1)}\\ &= \int_{\widetilde{\mathbf{W}}^{(t-1)}} p(\mathbf{\Sigma}^{(t)} \mid \widetilde{\mathbf{W}}^{(t-1)}) \prod_{i=1}^m p(\tilde{\mathbf{w}}_i^{(t-1)} \mid \mathbf{w}_i^{(t-1)}) d\widetilde{\mathbf{W}}^{(t-1)}.\end{aligned}$$

Because, for all $i \in [m]$ and some constant $c = \frac{\tilde{s}_i^{(t-1)}}{\epsilon_{\text{dp},t}}$, we have

$$p(\tilde{\mathbf{w}}_i^{(t-1)} \mid \mathbf{w}_i^{(t-1)}) \propto \exp\Big(-c\|\tilde{\mathbf{w}}_i^{(t-1)} - \mathbf{w}_i^{(t-1)}\|_2\Big),$$

given $(\mathbf{W}')^{(t-1)}$ such that for some $i \in [m]$, $(\mathbf{w}_i')^{(t-1)} \neq \mathbf{w}_i^{(t-1)}$, letting $(\tilde{\mathbf{w}}_i')^{(t-1)} = \tilde{\mathbf{w}}_i^{(t-1)} - \mathbf{w}_i^{(t-1)} + (\mathbf{w}_i')^{(t-1)}$, we have

$$\begin{aligned}&\|(\tilde{\mathbf{w}}_i')^{(t-1)} - (\mathbf{w}_i')^{(t-1)}\|_2 = \|\tilde{\mathbf{w}}_i^{(t-1)} - \mathbf{w}_i^{(t-1)}\|_2\\ &\Rightarrow p((\tilde{\mathbf{w}}_i')^{(t-1)} \mid \mathbf{w}_i^{(t-1)}) = p(\tilde{\mathbf{w}}_i^{(t-1)} \mid \mathbf{w}_i^{(t-1)}),\end{aligned}$$

and $d(\tilde{\mathbf{w}}_i')^{(t-1)} = d\tilde{\mathbf{w}}_i^{(t-1)}$.

Furthermore, based on the proof of Theorem 1 in Section L-C, we know that for neighboring matrices $\widetilde{\mathbf{W}}^{(t-1)}$ and $(\widetilde{\mathbf{W}}')^{(t-1)}$ that differ in the i-th column, we have

$$p(\mathbf{\Sigma}^{(t)} \mid \widetilde{\mathbf{W}}^{(t-1)}) \leq \exp(\epsilon_{\text{mp},t}) p(\mathbf{\Sigma}^{(t)} \mid (\widetilde{\mathbf{W}}')^{(t-1)}).$$

Therefore, for all $i \in [m]$, given $(\mathbf{W}')^{(t-1)}$ such that $(\mathbf{w}_i')^{(t-1)} \neq \mathbf{w}_i^{(t-1)}$, under the choice for $(\tilde{\mathbf{w}}_i')^{(t-1)}$, we have

$$\begin{aligned}&p(\mathbf{\Sigma}^{(t)} \mid \mathbf{W}^{(t-1)})\\ &= \int_{\widetilde{\mathbf{W}}^{(t-1)}} p(\mathbf{\Sigma}^{(t)} \mid \widetilde{\mathbf{W}}^{(t-1)}) \prod_{j=1}^m p(\tilde{\mathbf{w}}_j^{(t-1)} \mid \mathbf{w}_j^{(t-1)}) d\widetilde{\mathbf{W}}^{(t-1)}\\ &\leq \int_{(\widetilde{\mathbf{W}}')^{(t-1)}} e^{\epsilon_{\text{mp},t}} p(\mathbf{\Sigma}^{(t)} \mid (\widetilde{\mathbf{W}}')^{(t-1)}) p((\tilde{\mathbf{w}}_i')^{(t-1)} \mid (\mathbf{w}_i')^{(t-1)})\\ &\quad \prod_{j\in[m], j\neq i} p(\tilde{\mathbf{w}}_j^{(t-1)} \mid \mathbf{w}_j^{(t-1)}) d(\widetilde{\mathbf{W}}')^{(t-1)}.\\ &= \int_{(\widetilde{\mathbf{W}}')^{(t-1)}} \exp(\epsilon_{\text{mp},t}) p(\mathbf{\Sigma}^{(t)} \mid (\widetilde{\mathbf{W}}')^{(t-1)})\\ &\quad p((\widetilde{\mathbf{W}}')^{(t-1)} \mid (\mathbf{W}')^{(t-1)}) d(\widetilde{\mathbf{W}}')^{(t-1)}\\ &= \exp(\epsilon_{\text{mp},t}) p(\mathbf{\Sigma}^{(t)} \mid (\mathbf{W}')^{(t-1)}),\end{aligned}$$

which renders the mapping $\mathbf{W}^{(t-1)} \rightarrow \mathbf{\Sigma}^{(t)}$ as an $(\exp(\epsilon_{\text{mp},t}), 0)$ - differentially private algorithm.

Then, according to the proof of Theorem 1 in Section L-C, Algorithm 5 is an $(\epsilon_{\text{mp}}, \delta_{\text{mp}})$ - iterative MP-MTL algorithm. □

APPENDIX N
PROOF OF RESULTS IN APPENDIX E-A

A. Proof of Theorem 7

Proof. First, consider the case with no acceleration. We first use Proposition 1 of Schmidt et al. [11] by regarding procedures from Step 5 to Step 9 as approximation for the proximal operator in (8). Note that the norm clipping only bounds the parameter space and does not affect the results of Schmidt et al. [11]. Then for ε_t defined in Lemma 13 for $t \in [T]$, we have

$$\begin{aligned}\mathcal{E} =& \frac{2L}{m(T+1)^2}\Bigg(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F \\ &+ 2\sum_{t=1}^{T} t\sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^{T} t^2 \frac{\varepsilon_t}{L}}\Bigg)^2.\end{aligned}$$

Meanwhile, by Lemma 13, we have

$$\varepsilon_t = O\left(\frac{\kappa}{\epsilon_t}\right),$$

where $\kappa = \frac{K^2\sqrt{m}kd\log d}{\eta}$

On the other hand, let

$$c_1 = \epsilon = \sum_{t=1}^{T} \epsilon_t,$$

then by Lemma 17, we have

$$\sum_{t=1}^{T}\sqrt{\varepsilon_t} = \begin{cases} O\left(\sqrt{\frac{\kappa T^{\alpha+1}}{c_1(\alpha/2-1)^2(\alpha+1)}}\right), & \alpha > 2; \\ O\left(\sqrt{\frac{\kappa T^{3}}{c_1(\alpha/2-1)^2(\alpha+1)}}\right), & -1 < \alpha < 2; \\ O\left(\sqrt{\frac{\kappa T^{2-\alpha}}{c_1(\alpha/2-1)^2(-\alpha-1)}}\right), & \alpha < -1, \end{cases}$$

Because $\widetilde{\mathbf{W}}^{(0)}$ is the result of the norm clipping, we have $\widetilde{\mathbf{W}}^{(0)} \in \mathcal{W}$.

Finally, taking $c_3 = \phi(\alpha)$ defined in (13) and $c_4 = \frac{\kappa}{c_2(\alpha/2-1)^2|\alpha+1|}$, under the assumption that $\mathbf{W}_* \in \mathcal{W}$, using Lemma 19, we have the results for the case with no acceleration.

For the accelerated case, we use Proposition 2 of Schmidt et al. [11] to have

$$\begin{aligned}\mathcal{E} =& \frac{2L}{m(T+1)^2}\Bigg(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F \\ &+ 2\sum_{t=1}^{T} t\sqrt{\frac{2\varepsilon_t}{L}} + \sqrt{2\sum_{t=1}^{T} t^2 \frac{\varepsilon_t}{L}}\Bigg)^2.\end{aligned}$$

Then one can prove similarly combining Lemma 13, Lemma 16, Lemma 17 and Lemma 20. □

B. Proof of Theorem 8

Proof. First, consider the case with no acceleration. We use Proposition 1 of Schmidt et al. [11] and prove similarly as in Appendix N-A, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 19.

For the accelerated case, we use Proposition 2 of Schmidt et al. [11] and prove similarly as in Appendix N-A, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 20. □

C. Proof of Theorem 9

Proof. First, consider the case with no acceleration. We use Proposition 3 of Schmidt et al. [11] to have

$$\mathcal{E} = \frac{Q_0^T}{\sqrt{m}}\left(\|\widetilde{\mathbf{W}}^{(0)} - \mathbf{W}_*\|_F + 2\sum_{t=1}^{T} Q_0^{-t}\sqrt{\frac{2\varepsilon_t}{L}}\right).$$

Then one can prove similarly as in Appendix N-A, combining Lemma 13, Lemma 16, Lemma 17 and Lemma 21.

For the accelerated case, we use Proposition 4 of Schmidt et al. [11] to have

$$\begin{aligned}\mathcal{E} =& \frac{(Q_0)^T}{m}\Bigg(\sqrt{2(f(\widehat{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*))} + 2\sqrt{\frac{L}{\mu}}\sum_{t=1}^{T}\sqrt{\varepsilon_t(Q_0)^{-t}} \\ &+ \sqrt{\sum_{t=1}^{T}\varepsilon_t(Q_0)^{-t}}\Bigg)^2.\end{aligned}$$

Then one can prove similarly as in Appendix N-A, using the assumption that $f(\widetilde{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*) = O(K^2Lm)$, combining Lemma 13, Lemma 16, Lemma 17 and Lemma 22. □

D. Proof of Theorem 10

Proof. First, consider the case with no acceleration. We use Proposition 3 of Schmidt et al. [11] and prove similarly as in Appendix L-I, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 21.

For the accelerated case, we use Proposition 4 of Schmidt et al. [11] and prove similarly as in Appendix L-I, using the assumption that $f(\widetilde{\mathbf{W}}^{(0)}) - f(\mathbf{W}_*) = O(K^2Lm)$, combining Lemma 14, Lemma 16, Lemma 17 and Lemma 22. □

E. Proof of Theorem 11

Proof. Consider the bound in (25). First, by Assumption 1, $\mathcal{E}$ is minimized by maximizing $\phi(\alpha)$ and $(\alpha/2-1)^2|\alpha+1|$, which are maximized simultaneously when $\alpha = 0$. Results under other settings can be proved similarly. □

Appendix O
Proof of Results in Appendix E-B

Proof. Results in this settings are the corollaries of Theorem 2, Theorem 3, Theorem 4, Theorem 5 and Theorem 6, respectively, replacing the term $\sqrt{\log(e+\epsilon/\delta)}$ with the term $\sqrt{\log(1/\delta)}$ by Lemma 16. □

Appendix P
Proof of Results in Appendix J

A. Proof of Lemma 8

Proof. For simplicity, we omit the symbol $\mathcal{B}$ in the conditional events.

Because $\widetilde{\mathbf{W}} = \mathcal{A}_{\mathrm{MP}}(\mathbf{W}, \mathbf{X}^m, \mathbf{y}^m)$ is an (ϵ,δ)-Non-iterative MP-MTL algorithm, by Definition 5, we have for $i \in [m]$ and for any set $\mathcal{S}' \subseteq \mathbb{R}^{d\times(m-1)}$,

$$\begin{aligned}&\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in \mathcal{S}' \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\ &\quad\le e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in \mathcal{S}_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i', \mathcal{D}_i') + \delta.\end{aligned} \tag{68}$$

In the following, we follow the proof of Theorem B.1 of Dwork et al. [3].

For any $C_1 \subseteq \mathbb{R}^{d\times(m-1)}$, define

$$\begin{aligned}\mu(C_1) =&(\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in C_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\ &- e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in C_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i', \mathcal{D}_i'))_+\end{aligned}$$

and then, μ is a measure on C_1 and $\mu(C_1) \le \delta$ by (68). As a result, we have for all $s_1 \in C_1$,

$$\begin{aligned}&\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\ &\le e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i', \mathcal{D}_i') + \mu(ds_1).\end{aligned}$$

As such, for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$ and $\mathcal{S}_1$, which denotes the projection of $\mathcal{S}$ onto C_1:

$$\begin{aligned}
&\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, \tilde{\mathbf{w}}_{[-i]}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\quad \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\quad \left[e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \mu(ds_1)\right] \\
&= \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\quad e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \mu(S_1) \\
&= \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) \\
&\quad e^{\epsilon}\mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \mu(S_1) \\
&\leq e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, \tilde{\mathbf{w}}_{[-i]}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \delta.
\end{aligned}$$

The second equality uses the independence of the learning process for $\hat{\mathbf{w}}_{[-i]}$ given $(\tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]})$.

The procedure is similar to proving that the algorithm $\mathcal{A}_{\text{st+mp}}$ that first uses a *deterministic* STL algorithm $\mathcal{A}_{\text{st}}$ before applying $\mathcal{A}_{\text{mp}}$ is also an (ϵ, δ) - non-iterative MP-MTL algorithm.

For a *deterministic* STL algorithm $\widetilde{\mathbf{W}} = \mathcal{A}_{\text{st}}(\mathbf{W}, \mathbf{X}^m, \mathbf{y}^m)$, for $i \in [m]$, by the independence of the learning process for $\tilde{\mathbf{w}}_{[-i]}$ given $(\mathbf{w}_{[-i]}, \mathcal{D}_{[-i]})$, we have

$$\begin{aligned}
&p(\tilde{\mathbf{w}}_{[-i]} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\quad = p(\tilde{\mathbf{w}}_{[-i]} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i).
\end{aligned} \tag{69}$$

Because the STL algorithm is *deterministic*, when the input is given, it is reasonable to assume that the output is given. As such, we also have for $i \in [m]$,

$$\begin{aligned}
&p(\cdot \mid \mathbf{w}_i, \mathcal{D}_i) = p(\cdot \mid \tilde{\mathbf{w}}_i, \mathcal{D}_i) \\
&p(\cdot \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}) = p(\cdot \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]})
\end{aligned} \tag{70}$$

Then, for an (ϵ, δ)-Non-iterative MP-MTL algorithm $\widehat{\mathbf{W}} = \mathcal{A}_{\text{mp}}(\widetilde{\mathbf{W}}, \mathbf{X}^m, \mathbf{y}^m)$, by Definition 5, we have for $i \in [m]$ and for any set $\mathcal{S}' \subseteq \mathbb{R}^{d\times(m-1)}$

$$\begin{aligned}
&\mathbb{P}(\hat{\mathbf{w}}_{[-i]} \in \mathcal{S}' \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \tilde{\mathbf{w}}_i, \mathcal{D}_i) \\
&\quad \leq e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]} \in \mathcal{S}' \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \delta,
\end{aligned} \tag{71}$$

where $\tilde{\mathbf{w}}'_i$ can be replaced with $\mathbf{w}'_i$ because Definition 5 allows replacing $\tilde{\mathbf{w}}_i$ with any different model.

As such,

$$\begin{aligned}
&\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, \tilde{\mathbf{w}}_{[-i]}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\quad \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&= \int_{\mathcal{S}_1} \mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \tilde{\mathbf{w}}_i, \mathcal{D}_i) \\
&\quad \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \left[\left(e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) \right.\right. \\
&\quad \left.\left. + \delta\right) \wedge 1\right] \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \left[e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) \right. \\
&\quad \left. \wedge 1 + \delta\right] \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) \\
&\leq \int_{\mathcal{S}_1} \left[e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) \right. \\
&\quad \left. \wedge 1\right] \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}_i, \mathcal{D}_i) + \delta \\
&= \int_{\mathcal{S}_1} \left[e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, s_1) \in \mathcal{S} \mid \tilde{\mathbf{w}}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) \right. \\
&\quad \left. \wedge 1\right] \mathbb{P}(\tilde{\mathbf{w}}_{[-i]} \in ds_1 \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \delta \\
&= e^{\epsilon}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}, \tilde{\mathbf{w}}_{[-i]}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}, \mathcal{D}_{[-i]}, \mathbf{w}'_i, \mathcal{D}'_i) + \delta.
\end{aligned}$$

The second inequality uses (71), the first equality uses (70), and the second equality uses (69). ☐

B. Proof of Lemma 9

Proof. For simplicity, we omit the symbol $\mathcal{B}$ to denote the input in the conditional events in some equations.

First, because for all $t \in [T]$, $\widehat{\mathbf{W}}^{(t)} = \mathcal{A}_{\text{mp}}(\mathbf{W}^{(t-1)}, \mathbf{X}^m, \mathbf{y}^m)$ is an (ϵ_t, δ_t)-Non-iterative MP-MTL algorithm and because for all $i \in [m]$ $\mathbf{w}_i^{(t)} = \mathcal{A}_{\text{st},i}(\hat{\mathbf{w}}_i^{(t)}, \mathbf{X}_i, \mathbf{y}_i)$ is a deterministic STL algorithm for the i-th task, then by the proof of Lemma 8, we have that the mapping $(\mathbf{X}^m, \mathbf{y}^m, \mathbf{W}^{(t-1)}) \rightarrow (\widehat{\mathbf{W}}^{(t)}, \mathbf{W}^{(t)})$ is an (ϵ_t, δ_t)-Non-iterative MP-MTL algorithm for all $t \in [T]$. In other words, for all $i \in [m]$, we have for any set $\mathcal{S} \subseteq \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$ that

$$\begin{aligned}
&\mathbb{P}((\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{w}_{[-i]}^{(t)}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}^{(t-1)}, \mathcal{D}_{[-i]}, \mathbf{w}_i^{(t-1)}, \mathcal{D}_i) \\
&\quad \leq e^{\epsilon_t}\mathbb{P}((\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{w}_{[-i]}^{(t)}) \in \mathcal{S} \mid \mathbf{w}_{[-i]}^{(t-1)}, \mathcal{D}_{[-i]}, (\mathbf{w}'_i)^{(t-1)}, \mathcal{D}'_i) \\
&\qquad + \delta_t.
\end{aligned} \tag{72}$$

Then, for $t = 1, \ldots, T$, take the t-th dataset $\widetilde{\mathcal{D}}_t = \{(\mathbf{w}_1^{(t-1)}, \mathcal{D}_1 = (\mathbf{X}_1, \mathbf{y}_1)), \ldots, (\mathbf{w}_m^{(t-1)}, \mathcal{D}_m = (\mathbf{X}_m, \mathbf{y}_m))\}$, i.e., treat $(\mathbf{w}_i^{(t-1)}, \mathcal{D}_i = (\mathbf{X}_i, \mathbf{y}_i))$ as the i-th data instance of the dataset $\widetilde{\mathcal{D}}_t$ for all $i \in [m]$. For all $i \in [m]$ and for all $t \in [T]$, take the t-th output $\theta_{t,i} = (\hat{\mathbf{w}}_{[-i]}^{(t)}, \mathbf{w}_{[-i]}^{(t)})$. By (72), we have for all $t \in [T]$, for all $i \in [m]$, and for any set $\mathcal{S}_t \subseteq \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$ that

$$\mathbb{P}(\theta_{t,i} \in \mathcal{S}_t \mid \widetilde{\mathcal{D}}_t) \leq e^{\epsilon_t}\mathbb{P}(\theta_{t,i} \in \mathcal{S}_t \mid \widetilde{\mathcal{D}}'_t) + \delta_t,$$

where $\widetilde{\mathcal{D}}_t$ and $\widetilde{\mathcal{D}}'_t$ are two adjacent datasets that differ in a single entry, the i-th data instance $(\mathbf{w}_i^{(t-1)}, \mathcal{D}_i = (\mathbf{X}_i, \mathbf{y}_i))$, which renders the algorithm in the t-th iteration an (ϵ_t, δ_t)-differentially private algorithm. As such, by the *Adaptive composition* Lemma (Lemma

7), for all $i \in [m]$ and for any set $\mathcal{S} \subseteq \bigotimes_{t=1}^{T} \mathcal{C}_t$, where $\mathcal{C}_t = \mathbb{R}^{d\times(m-1)} \times \mathbb{R}^{d\times(m-1)}$, we have

$$\begin{aligned}&\mathbb{P}((\theta_{1,i},\cdots,\theta_{T,i}) \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\widetilde{\mathcal{D}}_t, \boldsymbol{\theta}_{1:t-1})))\\ \leq & e^{\epsilon}\mathbb{P}((\theta_{1,i},\cdots,\theta_{T,i}) \in \mathcal{S} \mid \bigcap_{t=1}^{T}(\mathcal{B}_t = (\widetilde{\mathcal{D}}'_t, \boldsymbol{\theta}_{1:t-1})))\\ &+ 1 - (1-\delta)\prod_{t=1}^{T}(1-\delta_t),\end{aligned}$$

where for all $t \in [T]$, $\mathcal{B}_t$ denotes the input for the t-th iteration,

$$\boldsymbol{\theta}_{1:t-1} = \begin{cases} \emptyset, & t=1\\ (\theta_{1,1},\ldots,\theta_{1,m})\ldots,(\theta_{t-1,1},\ldots,\theta_{t-1,m}), & t\geq 2,\end{cases}$$

and ϵ is defined in Lemma 7.

Then, we have for any set $\mathcal{S}' \subseteq \mathbb{R}^{d\times(m-1)\times T}$,

$$\begin{aligned}&\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S}' \mid \bigcap_{t=1}^{T}\mathcal{B}_t = (\mathbf{W}^{(t-1)}, \mathcal{D}^m, \boldsymbol{\theta}_{1:t-1}))\\ \leq & e^{\epsilon}\mathbb{P}(\hat{\mathbf{w}}_{[-i]}^{(1:T)} \in \mathcal{S}' \mid \bigcap_{t=1}^{T}\mathcal{B}_t = ((\mathbf{W}')^{(t-1)}, (\mathcal{D}')^m, \boldsymbol{\theta}_{1:t-1}))\\ &+ 1 - (1-\delta)\prod_{t=1}^{T}(1-\delta_t).\end{aligned}$$

□

References

[1] T. M. Apostol. An elementary view of euler's summation formula. *The American Mathematical Monthly*, 106(5):409–418, 1999.

[2] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, 2011.

[3] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[4] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

[5] S. K. Gupta, S. Rana, and S. Venkatesh. Differentially private multi-task learning. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 101–113. Springer, 2016.

[6] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pages 457–464. ACM, 2009.

[7] W. Jiang, C. Xie, and Z. Zhang. Wishart mechanism for differentially private principal components analysis. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[8] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 2017.

[9] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.

[10] M. Pathak, S. Rane, and B. Raj. Multiparty differential privacy via aggregation of locally trained classifiers. In *Advances in Neural Information Processing Systems*, pages 1876–1884, 2010.

[11] M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

[12] S. Vadhan. The complexity of differential privacy. *Work. Pap., Cent. Res. Comput. Soc., Harvard Univ. http://privacytools.seas. harvard. edu/publications/complexity-differential-privacy*, 2016.

[13] Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 733–742. AUAI Press, 2010.