Integrating Feature Correlation in Differential Privacy with Applications in DP-ERM

Tianyu Wang [1] **Luhao Zhang** [2] **Rachel Cummings** [1]

Abstract

Standard differential privacy treats all features uniformly, overlooking the distinction between sensitive and insensitive features in practice. We introduce a relaxed definition of differential privacy that accounts for such privacy heterogeneity, allowing certain features to be treated as insensitive even when correlated with sensitive ones. We introduce CorrDP, a correlation-aware framework that relaxes privacy for insensitive features while accounting for their correlations with sensitive features, quantified via total variation distance. We design algorithms for differentially private empirical risk minimization (DP-ERM) under the CorrDP framework, incorporating distance-dependent noise into gradients for theoretically enhanced utility guarantees. When the correlation distance is unknown, we estimate it from the dataset while maintaining comparable privacy-utility guarantees.

1. Introduction

Differential Privacy (DP, [4]), which enforces a worst-case bound on privacy loss, ensures that the output of an algorithm does not depend significantly on any single data point. While DP has become a standard for privacy-preserving algorithms, it assumes that all features are equally sensitive, leading to overly conservative privacy-utility trade-offs.

Empirical Risk Minimization (ERM) is among the fundamental and well-studied problems of all privacy-preserving machine learning problems [3]. The goal of ERM is to find the best parameter $\theta \in \mathbb{R}^m$ of a loss function to minimize empirical risk given on a dataset $\mathcal{D}$. In the standard (ϵ, δ)-DP setting for a convex loss, DP-ERM algorithms achieve a minimax optimal utility guarantee $\tilde{O}\left(\sqrt{m}/(n\epsilon)\right)$ [2] via incorporating noise into the gradient [1, 9]. However, this guarantee can be overly conservative, particularly in high-dimensional settings with few sensitive features, as it applies uniform noise to all parameter dimensions.

In many real-world applications, feature sensitivity varies, and features are often correlated. Applying a uniform privacy mechanism that disregards the sensitivity levels can lead to excessive information loss and suboptimal utility, distorting the dataset more than necessary. Recent semi-sensitive DP approaches [8, 7] address such distinctions by treating some features as private and others as public. In practice, this strategy is problematic because these public features may be correlated with private features, undermining their privacy. An ideal privacy mechanism would account for such correlations by robustly protecting the sensitive measurements while adding minimal noise to less sensitive but correlated features.

In this work, we investigate whether relaxing standard privacy definitions can improve the privacy-utility tradeoff, particularly in datasets with correlated features. Specifically, we aim to answer the following questions: (1) What is an appropriate privacy mechanism for datasets with correlated features? (2) Can a relaxed notion of differential privacy improve the privacy-utility tradeoff? If so, how much?

To address these questions, we propose CorrDP, a new correlation-aware differential privacy framework that accounts for feature dependencies while ensuring rigorous privacy and utility guarantees. This notion offers a natural way to quantify correlations between sensitive and insensitive features, and integrates probability distances like total variation distance with standard mechanisms like the Laplace mechanism. Applying CorrDP to DP-ERM, we design CorrDP-SGD algorithms that incorporate distance-dependent noise for smooth loss functions. CorrDP improves the privacy-utility tradeoff, achieving a utility improvement by a factor of $\sqrt{m/m_s}$ under mild conditions, where m_s denotes the number of sensitive features. When the correlation distance is unknown, we propose an estimation procedure that uses an upper confidence bound, mitigating estimation and sensitivity errors while maintaining the same level of privacy-utility performance.

[1]Department of Industrial Engineering and Operations Research, Columbia University, USA [2]Department of Applied Mathematics and Statistics, Johns Hopkins University, USA. Correspondence to: Tianyu Wang <tw2837@columbia.edu>.

2. CorrDP: Setup and Mechanisms

Denote $\mathcal{X}$ as the data domain where each point $X \in \mathcal{X} \subseteq \mathbb{R}^m$ is of high dimensional with $X = (X^{\mathcal{S}}, X^{\mathcal{U}})^\top$. Here, $\mathcal{S}$ denotes the indices of sensitive features that require protection, and $\mathcal{U}$ denotes the indices of insensitive features such that $\mathcal{S} \cup \mathcal{U} = [m]$ and $\mathcal{S} \cap \mathcal{U} = \emptyset$.

In the standard (global) differential privacy, all feature components of the data point in the data domain are assumed to be equally sensitive, thus the same privacy constraint is enforced on the change of any feature component between $\mathcal{D}, \mathcal{D}'$. This ignores the heterogeneity among feature changes in neighboring databases. To capture this, we present the concept of CorrDP by incorporating a "distance" metric between databases $d(\mathcal{D}, \mathcal{D}')$, which quantifies the degree of privacy loss for feature differences there.

Definition 2.1 (Neighboring Database)**.** Databases $\mathcal{D}$ and $\mathcal{D}'$ are neighboring, denoted $\|\mathcal{D} - \mathcal{D}'\|_1 = 1$, if they differ in at most one entry, e and e', which themselves differ only in their sensitive features or one insensitive feature.

Definition 2.2 (CorrDP)**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-correlated differentially private if for all potential output sets R in the output space of $\mathcal{A}$ and for all neighboring databases $\mathcal{D}, \mathcal{D}'$:

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in R) \leq e^{\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in R) + \delta.$$

We need $d(\mathcal{D}, \mathcal{D}')$ to satisfy certain properties.

Definition 2.3 (Axioms of Sensitivity Distance)**.** For neighboring databases $\mathcal{D}, \mathcal{D}'$ differing in entries e and e', the distance metric d captures the impact of changes on sensitive features and must satisfy the following: (1) $d(\mathcal{D}, \mathcal{D}') \in [0, 1]$. (2) When e and e' differ in sensitive features ($\mathcal{S}$), $d(\mathcal{D}, \mathcal{D}') = 1$. (3) When e and e' differ only in insensitive features ($\mathcal{U}$), $d(\mathcal{D}, \mathcal{D}') \in [0, 1)$. If insensitive features are independent of the sensitive features, then $d(\mathcal{D}, \mathcal{D}') = 0$.

The notion of neighboring database excludes changes in both sensitive and insensitive features because then $d(\mathcal{D}, \mathcal{D}') = 1$, and the CorrDP constraint becomes exactly DP, eliminating any benefits from the relaxed CorrDP notion. Definition 2.1 (and Assumption 3.1 in Section 3) can be relaxed to allow changes in multiple insensitive features.

Throughout the main body, we define $d(\mathcal{D}, \mathcal{D}')$ based on the Total Variation (TV) Distance, as in Definition 2.4, which naturally satisfies the desired properties above.

Definition 2.4 (Choice of d)**.** For the two neighboring databases $\mathcal{D}, \mathcal{D}'$ that differ in two entries e and e', define:

$$d(\mathcal{D}, \mathcal{D}') := \max_{\mathcal{I}:\mathcal{I}\subseteq[m+1]} TV(\mathbb{P}_{X^{\mathcal{S}}|e^{\mathcal{I}}}, \mathbb{P}_{X^{\mathcal{S}}|(e')^{\mathcal{I}}}), \quad (1)$$

where $\mathbb{P}_{X^{\mathcal{S}}|e^{\mathcal{I}}}$ is the conditional distribution of the sensitive features $X^{\mathcal{S}}$ given $X^{\mathcal{I}} = e^{\mathcal{I}}$. TV distance is defined as $TV(\mathbb{P}, \mathbb{Q}) := \sup_{A\in\mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$ for a measurable space $(\Omega, \mathcal{F})$ and probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$. Then (1) satisfies the axioms in Definition 2.3.

Next we present some concrete examples demonstrating how the CorrDP framework can be instantiated and applied.

Example 2.5. Consider the feature $X = (X^{(1)}, X^{(2)})^\top$, where $\mathcal{S} = \{1\}$, $\mathcal{U} = \{2\}$. $X^{(1)}$ and $X^{(2)}$ are partially correlated with $\mathbb{P}_{X^{(1)}|X^{(2)}=k} \sim \text{Bernoulli}(k/3), \forall k \in \{1, 2\}$ for entries in the database $\mathcal{D}, \mathcal{D}'$. Suppose two databases $\mathcal{D}$ and $\mathcal{D}'$ differ by e and e'. If $e = (1, 2)^\top$ and $e' = (0, 1)^\top$, then $d(\mathcal{D}, \mathcal{D}') = 1$ since $TV(\mathbb{P}_{X^{(1)}|X^{(1)}=1}, \mathbb{P}_{X^{(1)}|X^{(1)}=0}) = 1$.

However, if $e = (1, 2)^\top$ and $e' = (1, 1)^\top$, then $d(\mathcal{D}, \mathcal{D}') = TV(\mathbb{P}_{X^{(1)}|X^{(2)}=2}, \mathbb{P}_{X^{(1)}|X^{(2)}=1}) = 1/3 < 1$. In this case, the entrywise difference and its influence on the associated privacy constraint is not as large as in the previous case since the two entries only differ in the insensitive coordinate.

Remark 2.6. The set $d(\mathcal{D}, \mathcal{D}')$ recovers some existing DP notions. When all the features are sensitive, $d(\mathcal{D}, \mathcal{D}') = 1, \forall \mathcal{D}, \mathcal{D}'$ neighboring databases, which recovers the standard definition of differential privacy. When only some private features need to be protected and other public features are known to be independent, then $d(\mathcal{D}, \mathcal{D}') = 1$ if and only if $e^{\mathcal{S}} \neq (e')^{\mathcal{S}}$, which recovers the definition of semi-sensitive DP [8, 7].

2.1. Standard Mechanism for CorrDP

We show how to adopt the standard Laplace Mechanism to achieve privacy under CorrDP, and show the utility improvements from using CorrDP.

First, we recall the sensitivity of a function with a vector input and output:

Definition 2.7 (ℓ_1, Coordinate, and Correlated Sensitivity)**.** For any $K \geq 1$, the ℓ_1-sensitivity of function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ is: $\Delta f := \max_{\substack{\mathcal{D},\mathcal{D}'\in\mathbb{N}^{|\mathcal{X}|} \\ \|\mathcal{D}-\mathcal{D}'\|_1=1}} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$, and the sensitivity of the k-th coordinate f_k of f is: $\Delta f_k := \max_{\substack{\mathcal{D},\mathcal{D}'\in\mathbb{N}^{|\mathcal{X}|} \\ \|\mathcal{D}-\mathcal{D}'\|_1=1}} |f_k(\mathcal{D}) - f_k(\mathcal{D}')|$. The correlated sensitivity is: $\Delta_C f = \min\{\sum_{j\in\mathcal{S}} \Delta f_j + \sum_{j\in\mathcal{U}} \Delta f_j TV(j), \Delta f\}$ and $TV(j) = \max_{x_1,x_2\in\mathcal{X}} TV(\mathbb{P}_{X^{\mathcal{S}}|x_1^{(j)}}, \mathbb{P}_{X^{\mathcal{S}}|x_2^{(j)}}) \leq 1$.

We note that $\Delta f \leq \sum_{k\in[K]} \Delta f_k$, where equality holds when a common pair of neighboring databases $(\mathcal{D}^*, (\mathcal{D}^*)')$ with $\|\mathcal{D}^* - (\mathcal{D}^*)'\|_1 = 1$ simultaneously attains the maximum absolute difference Δf_k for $k \in [K]$. Correlated sensitivity accounts for the heterogeneous privacy budget of insensitive features $j \in \mathcal{U}$, incorporating their correlation with sensitive features through a weighting factor, $TV(j)$. Next, we show how modifying Laplace Mechanism with correlated sensitivity enables CorrDP while improving accuracy guarantees.

Definition 2.8 (CorrDP Laplace Mechanism)**.** For a measurable function $f(\cdot) : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ where the i-th dimension of f only applies to the i-th feature of $\mathcal{D}$, the CorrDP Laplace mechanism is defined as: $\widetilde{\mathcal{M}}_L(\mathcal{D}, f(\cdot), \epsilon) = f(\mathcal{D}) + (Y_1, \ldots, Y_K)$, where $Y_i \sim_{i.i.d.} \text{Lap}(\Delta_C f/\epsilon)$.

Theorem 2.9 (CorrDP Laplace Guarantees)**.** *The CorrDP Laplace mechanism is $(\epsilon, 0)$-CorrDP, and $\forall \beta \in (0, 1]$, $\mathbb{P}[\|f(\mathcal{D}) - \widetilde{\mathcal{M}}_L(\mathcal{D}, f(\cdot), \epsilon)\|_\infty \leq \frac{\Delta_C f \log(K/\beta)}{\epsilon}] \geq 1 - \beta$.*

The standard Laplace mechanism (e.g., Chapter 3.3 in [5]) has a high-probability accuracy guarantee that $\|f(\mathcal{D}) - \mathcal{M}_L(\mathcal{D}, f(\cdot), \epsilon)\|_\infty \leq \frac{\Delta f \log(K/\beta)}{\epsilon}$. Comparing this with Theorem 2.9, we see that the CorrDP Laplace mechanism gives a better accuracy guarantee exactly when the lack of correlation across features leads to lower sensitivity.

3. CorrDP ERM

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i\in[n]}$ and the individual loss function $\ell(\theta; (X, Y))$ where X denotes the feature and Y denotes the label, DP-ERM aims to obtain a differentially private solution $\theta^{priv} \in \mathbb{R}^m$ close to the nonprivate solution: $\hat{\theta} \in \text{argmin}_{\theta\in\Theta}\{F(\theta, \mathcal{D}) := 1/n \sum_{i=1}^n \ell(\theta; (x_i, y_i))\}$. We measure the utility loss of θ^{priv} as the additional empirical loss from adding privacy:

$$R(\theta^{priv}) := F(\theta^{priv}, \mathcal{D}) - F(\hat{\theta}, \mathcal{D}). \tag{2}$$

Existing DP-ERM models satisfying (ϵ, δ)-DP will automatically satisfy our relaxed (ϵ, δ)-CorrDP notion. Our main goal is to determine whether relaxing to CorrDP and appropriately modifying the DP-ERM algorithms can improve utility. We next give some assumptions that will be necessary for our results.

Assumption 3.1 (Neighboring Database in ERM)**.** The two elements e and e' in neighboring databases $\mathcal{D}$ and $\mathcal{D}'$ only differ in sensitive features or one insensitive feature in X, and cannot differ in the label Y.

Building on Definition 2.1, this assumption further excludes label changes of Y between $\mathcal{D}, \mathcal{D}'$. This is supported by prior work [8], which also assumed labels were public and did not require privacy protection. When labels are private while features are public, one can refer to [6]. We can relax this Assumption as we show in the full version.

For the DP-ERM problem, we consider general convex loss with bounded decision and feature domains.

Assumption 3.2 (Regularity of Loss Function)**.** The loss function ℓ is L-Lipschitz, i.e., $|\ell(\theta_1; (x, y) - \ell(\theta_2; (x, y))| \leq L\|\theta_1 - \theta_2\|_2$.

Assumption 3.3 (Boundness of Domain)**.** The decision domain $\mathcal{X}$ is bounded such that $\|x\|_2 \leq B$ with each component $|x^{(i)}| \leq B_i, i \in [m], \forall x \in \mathcal{X}$. Without loss of generality, we set $B = 1$ and each $B_i = \Theta(1/\sqrt{m})$. Furthermore, the parameter is bounded such that $\|\theta\|_2 \leq D$.

The boundness of domain and parameters is naturally imposed for theoretical analyses in DP-ERM algorithms [9]. Unbounded gradients can be handled via gradient clipping.

Finally, we link the sensitivity of features to the parameter coordinate.

Assumption 3.4 (Sensitivity of Gradient Coordinate)**.** For the i-th gradient coordinate, $i \in [m]$, $(\nabla_\theta \ell(\theta; (x_1, y)) - \nabla_\theta \ell(\theta; (x_2, y))_i \leq C_1 L|x_1^{(i)} - x_2^{(i)}| + C_2 \sum_{j\neq i} \frac{L}{m}|x_1^{(j)} - x_2^{(j)}|$ for some constants C_1, C_2.

This assumption requires the loss function to be smooth with respect to changes in x. Furthermore, the sensitivity of the i-th parameter coordinate is mainly controlled by the corresponding i-th feature component. This naturally holds when $\ell(\theta; (x, y))$ can be further represented by $\tilde{\ell}(\theta^\top x, y)$ for some $\tilde{\ell}$, i.e., **generalized linear model (GLM)**. In the full version, we demonstrate that linear and logistic regression satisfies these assumptions.

3.1. CorrDP-SGD Algorithm and Guarantees

We present CorrDP-SGD in Algorithm 1, which incorporates CorrDP based on the standard DP-SGD [2, 1]. The key part is the modified noise scale in Line 3, where each coordinate σ_i in the noise variance $\boldsymbol{\sigma}$ is set as:

$$\sigma_i^2 = \begin{cases} \frac{(\log(1/\delta)+1)L^2T}{n^2\epsilon^2}, \text{ if } i \in \mathcal{S}; \\ \frac{(\log(1/\delta)+1)L^2T \max\{TV(i), m_s^2/m^2\}}{n^2\epsilon^2}, \text{ else} \end{cases} \tag{3}$$

where $TV(j) = \max_{x_1, x_2 \in \mathcal{X}} TV(\mathbb{P}_{X^{\mathcal{S}}|x_1^{(j)}}, \mathbb{P}_{X^{\mathcal{S}}|x_2^{(j)}})$.

Compared with the standard DP-SGD mechanism [2], the only difference is the noise variance term for $i \in \mathcal{U}$, where the unit scale 1 is replaced with $\max\{TV(i), m_s^2/m^2\} \leq 1$. Despite the smaller noise imposed on the insensitive features, we show that CorrDP still guarantees privacy.

Theorem 3.5 (Privacy Guarantee of CorrDP-SGD)**.** *Under Assumptions 3.1, 3.2, 3.3 and 3.4, for $\epsilon \in (0, c_1]$ for some constant c_1 and $\delta > 0$, Algorithm 1 is (ϵ, δ)-CorrDP.*

Theorem 3.6 (Utility Guarantee of CorrDP-SGD)**.** *Under Assumptions 3.1, 3.2, 3.3 and 3.4, for Algorithm 1 with step sizes $\alpha_t = \frac{D}{\sqrt{(L^2 + \sum_{i=1}^m \sigma_i^2)t}}$ and $T = \Theta(n^2)$, if $F(\theta, \mathcal{D})$ is convex, then $R(\theta^{priv}) = \tilde{O}\left(\frac{\sqrt{(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), m_s/4\}) \log(1/\delta)}}{n\epsilon}\right)$. Furthermore, if $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$, then $R(\theta^{priv}) = \tilde{O}\left(\frac{\sqrt{m_s}}{n\epsilon}\right)$.*

This result demonstrates that if many features are insensitive and weakly correlated with sensitive ones, CorrDP-SGD

Algorithm 1 CorrDP Stochastic Gradient Descent (CorrDP-SGD)

Require: Parameter Domain Θ, iterations number T, and step sizes α_t; Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i\in[n]}$; Subsampling size n_q with $1 \le n_q \le n$; CorrDP parameters (ϵ, δ).

1: Initialize θ_1 and set diagonal entries of the noise variance $\{\sigma_i^2\}_{i\in[m]}$ according to (3).
2: **for** $t = 1, \dots, T$ **do**
3: Randomly sample n_q datapoints $\{(x_{(i)}, y_{(i)})\}_{i\in[n_q]}$ from the whole dataset $\mathcal{D}$.
4: Generate the noise $b \sim N(0, \text{diag}(\boldsymbol{\sigma}^2))$ and update:

$$\theta_{t+1} = \Pi_\Theta\left(\theta_t - \alpha_t\left(\frac{1}{n_q}\sum_{i=1}^{n_q}\nabla_\theta \ell(\theta_t; (x_{(i)}, y_{(i)})) + b\right)\right).$$

5: **end for**

Ensure: $\theta^{priv} = \theta_{T+1}$.

significantly improves utility over the standard minimax result of DP-ERM bound $\tilde{O}\left(\frac{\sqrt{m}}{n\epsilon}\right)$ when $m_s = o(m)$. These utility gains extend to strongly convex and general nonconvex losses and other first-order algorithms, including SVRG [9] or Adam.

3.2. Lower Bound

We provide a near-matching lower bound in terms of n, ϵ and problem dimension.

Theorem 3.7 (Lower bound for (ϵ, δ)-CorrDP algorithm)**.** *Let $n, m \in \mathbb{N}, \epsilon > 0$ and $\delta = o(1/n)$. Denote $\{TV^{(i)}\}_{i\in\mathcal{U}}$ by $\{TV(i)\}_{i\in\mathcal{U}}$ sorted in descending order. For every (ϵ, δ)-CorrDP algorithm outputting θ^{priv}, there exists a $\mathcal{D}$ such that with probability at least 1/3,*

$$R(\theta^{priv}) = \Omega\left(\min\left\{1, \frac{\sqrt{m_s + \max_{k\in[m_u]}\{k(TV^{(k)})^2\}}}{n\epsilon}\right\}\right).$$

Comparing this lower bound against the utility upper bound of CorrDP-SGD in Theorem 3.6, we see that $\max_{k\in[m_u]}\{k(TV^{(k)})^2\} \le \sum_{i\in[m_u]} TV(i)$.

3.3. Estimating of TV distance in CorrDP-SGD

In this section, we address the challenge of handling an unknown TV distance in Algorithm 1. The noise terms σ_i (Equation (3)) in CorrDP-SGD depend on the maximum of the conditional total variation distance $TV(i)$, which is generally unknown. A naive approach would be to empirically estimate $\{TV(i)\}_{i\in\mathcal{U}}$ from the dataset $\mathcal{D}$ as $\widehat{TV}_{\mathcal{D}}(i)$ and substitute them directly. However, this leads to privacy leakage since the estimation procedure itself depends on $\mathcal{D}$. We show that in many scenarios, the estimation error is negligible after proper processing. For example, if domain knowledge provides an exact or near-exact upper bound: $U_i = TV(i)(1 + o(1))$, replacing $TV(i)$ with $U_i, i \in \mathcal{U}$ ensures that utility and privacy guarantees remain unaffected.

We focus on the more general case when no prior estimate of $TV(i)$ is available, but its estimation is regular and smooth. Our goal will be to find adjusted expressions for the noise terms in Equation (3) to preserve similar privacy and utility guarantees as if $TV(i)$ were known. In the following, we adjust the estimation of $\{TV(i)\}_{i\in\mathcal{U}}$ in the noise terms σ_i. To proceed, we require two assumptions as follows:

Assumption 3.8 (Bounded Estimation Error)**.** Within $\mathcal{D}$, each entry is i.i.d. sampled from $\mathbb{P}^*$, and with probability at least $1 - \beta$, for each $i \in [m]$, $|\widehat{TV}_{\mathcal{D}}(i) - TV(i)| \le c_2\frac{\sqrt{\log(1/\beta)}}{n^\gamma}$ for some constants $c_2 \in (0, \infty)$ and $\gamma \in (0, \frac{1}{2}]$.

Assumption 3.9 (Bounded Sensitivity)**.** Given two neighboring databases $\mathcal{D}$ and $\mathcal{D}'$, $\forall i \in \mathcal{U}$, $|\widehat{TV}_{\mathcal{D}'}(i) - \widehat{TV}_{\mathcal{D}}(i)| \le \frac{c_3}{n}$ for some constants $c_3 < \infty$.

Sensitivity is $O(1/n)$ in many empirical estimators since each sample contributes $1/n$ to the empirical distribution. Changing one sample influences the probability mass by at most $1/n$, resulting in a proportional effect on TV distance estimation. In the full version, we give empirical estimators $\widehat{TV}_{\mathcal{D}}(i)$ that satisfy Assumptions 3.8 and 3.9.

We impose the following modified noise for $TV(i)_{i\in\mathcal{U}}$:

Definition 3.10 (In-Sample TV Estimation)**.** When $\{TV(i)\}_{i\in\mathcal{U}}$ is unknown, replace $TV(i)$ with $\widetilde{TV}(i) = \widehat{TV}_{\mathcal{D}}(i) + 2c_2\frac{\sqrt{\log(m_u/\delta)}}{n^\gamma}$ in the expression for σ_i^2 in (3).

The additional term in the estimator accounts for inflation due to the estimation error. The sensitivity error does not explicitly appear in Definition 3.10, because its magnitude $O(1/n)$, is dominated by the estimation error from Assumption 3.8. However, Assumption 3.9 remains necessary to control variance differences in the privacy analysis.

Theorem 3.11 (Guarantees of CorrDP with In-Sample TV Estimation)**.** *When the estimator in Definition 3.10 is used for the noise terms σ_i, when $n = \Omega(\log(1/\delta))$ and under Assumptions 3.8 and 3.9, Algorithm 1 is $(\epsilon, 2\delta)$-CorrDP and achieves the same utility guarantee as Theorem 3.6.*

The full version of this paper also considers general loss functions including neural networks, demonstrates that other distance that can be incorporated in this framework in place of $TV(\cdot, \cdot)$, and evaluates the empirical performance of CorrDP-SGD.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[2] Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

[3] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[4] Dwork, C. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

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

[6] Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., and Zhang, C. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

[7] Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Meka, R., and Zhang, C. On convex optimization with semi-sensitive features. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1916–1938. PMLR, 2024.

[8] Shen, Z., Krishnaswamy, A., Kulkarni, J., and Munagala, K. Classification with partially private features. *arXiv preprint arXiv:2312.07583*, 2023.

[9] Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.