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

Tianyu Wang
Columbia University

Luhao Zhang
Johns Hopkins University

Rachel Cummings
Columbia University

Abstract

Standard differential privacy imposes uniform privacy constraints across all features, overlooking the inherent distinction between sensitive and insensitive features in practice. In this paper, we introduce a relaxed definition of differential privacy that accounts for such heterogeneity, allowing certain features to be treated as insensitive even when correlated with sensitive ones. We propose a correlation-aware framework, CorrDP, which relaxes privacy for insensitive features while accounting for their correlations with sensitive features, with the correlations quantified using 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 improved theoretical utility guarantees. When the correlation distance is unknown, we estimate it from the dataset and show that it achieves a comparable privacy-utility guarantee. We perform experiments on synthetic and real-world datasets and show that CorrDP-based DP-ERM algorithms consistently outperform the standard DP framework in the presence of insensitive features.

1 INTRODUCTION

The growing reliance on personal data in domains such as economics and healthcare has amplified the need for privacy-preserving algorithms. Differential Privacy (DP, Dwork (2006)), 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. As a result, DP has become a standard methodology for safeguarding privacy during analysis of sensitive data. However, traditional DP assumes that all features of an individual's data are equally sensitive, often leading to overly conservative privacy-utility trade-offs in practice.

In many real-world applications, feature sensitivity can vary, where some are highly sensitive (e.g., health status) and other are less so (e.g., age group). However, these features are often correlated. Consider a healthcare dataset containing both a patient's blood pressure readings and age. Blood pressure measurements, which can reveal medical conditions, are more sensitive, while age may be less sensitive if presented in broad categories (e.g., age groups). However, these features are often correlated—age influences typical blood pressure ranges. Applying a uniform privacy mechanism that disregards these distinctions can lead to excessive information loss and suboptimal utility. Adding equal noise to both features also overlooks their correlation, distorting the dataset more than necessary. Recent semi-sensitive DP approaches (Shen et al., 2023; Ghazi et al., 2024) address such distinctions by treating some features as private (e.g., blood pressure) and others as public (e.g., age). However, this approach risks privacy violations because revealing age can expose the conditional distribution of blood pressure, undermining its privacy. An ideal privacy mechanism would therefore account for such correlations by robustly protecting the sensitive measurements while adding minimal noise to less sensitive but correlated features.

Empirical Risk Minimization (ERM) is among the most fundamental and well-studied problems in privacy-preserving machine learning (Chaudhuri et al., 2011). The goal of ERM is to find the best parameter $\theta \in \mathbb{R}^m$ of a (convex) loss function to minimize empirical risk given a dataset $\mathcal{D}$ of size n; privacy of the output parameter is often achieved via incorporating noise into the gradient when conducting gradient descent at each step, i.e., DP-SGD (Abadi et al., 2016; Wang et al., 2017). In the standard (ϵ, δ)-DP setting, DP-ERM algorithms achieve a minimax optimal utility guarantee $\tilde{O}\left(\sqrt{m}/(n\epsilon)\right)$. While this utility guarantee

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

is theoretically minimax optimal (Bassily et al., 2014), this guarantee can be overly conservative, particularly in high-dimensional settings with few sensitive features, as DP-SGD applies uniform noise to all dimensions.

In this work, we investigate whether relaxing the DP definition can improve the privacy-utility tradeoff, particularly in datasets with features that are known to less sensitive. Specifically, we aim to answer: (1) What is an appropriate privacy mechanism for datasets with insensitive features that may be correlated with sensitive features? and (2) How much can a relaxed notion of differential privacy improve the privacy-utility tradeoff? To address these questions, we design a privacy framework and mechanisms that consider the correlation among features while ensuring rigorous privacy and utility guarantees.

Contributions. Motivated by the limitations of standard DP and practical applications where not all features are equally sensitive, we propose a new correlation-aware differential privacy framework (CorrDP). Our framework leverages feature correlations to achieve improved utility guarantees with modified algorithms, including extensions of both the standard Laplace mechanism and DP-ERM. In particular, our contributions are as follows:

1. **New Frameworks Incorporating Feature Correlation in DP:** We introduce CorrDP, a new DP notion incorporating feature correlations in the privacy constraint between neighboring databases. This notion offers a natural way to quantify correlations between sensitive and insensitive features, and collapses to the standard DP definition when all features are sensitive. It seamlessly integrates probability distances such as total variation distance, and standard mechanisms like the Laplace mechanism can be adapted to fit within this framework.

2. **Improved Privacy-Utility Tradeoff in DP-ERM:** We apply CorrDP into DP-ERM and design corresponding CorrDP-SGD algorithms that incorporate distance-dependent noise for loss functions satisfying certain smoothness properties (see Assumption 3.4), which includes generalized linear models. Our theoretical results show that CorrDP offers an improved privacy-utility tradeoff, providing a utility improvement by a factor of $\sqrt{m/m_s}$ under mild conditions, when m_s features are sensitive.

3. **Distance Estimation:** When the correlation distance is unknown at the time of noise addition, we propose an estimation procedure that uses an upper confidence bound for the empirical estimate. This procedure, integrated within the CorrDP framework, eliminates the effects of both estimation and sensitivity errors while maintaining the same level of privacy-utility performance.

4. **Empirical Evaluation**: To demonstrate the practical utility of distance estimation and the CorrDP framework, we conduct experiments using both synthetic and real-world datasets. Our experiments confirm that using CorrDP achieves better accuracy for the same fixed privacy budget.

1.1 Related work

Relaxations of DP. Recent studies consider relaxations of DP to achieve better privacy-utility tradeoffs, especially for ERM. When relaxing the DP notion at the feature level, Ghazi et al. (2021) considered label DP, which protects only the labels rather than the entire feature set. This is a special case of the notion called *semi-sensitive DP*, where the feature domain includes both sensitive and insensitive features (Chua et al., 2024). Under such a setup, Shen et al. (2023) designed a specific boosting algorithm for linear classification, while Ghazi et al. (2024) investigated the general DP-ERM problem when the sensitive domain size is finite. Empirically these relaxations have been successfully applied to vision (Shi et al., 2021) and language (Schneider et al., 2024) tasks without utility guarantees provided. In general, all these methods often rely on strong assumptions about the independence between insensitive and sensitive features. On the other hand, metric DP (Andrés et al., 2013; Zhao & Chen, 2022; Imola et al., 2022) relaxes the standard DP notion under a general metric space. However, no prior work in the metric DP framework studied the possibility of capturing correlations between features for improved utility guarantees for DP-ERM. (See Appendix B.1 for more discussion.) In contrast, our CorrDP notion specifies this by incorporating feature correlations with different sensitivity levels for general ERM losses, while allowing infinitely sensitive domains. Besides feature-level relaxations, other work incorporates the heterogeneity of elements within the data domain to relax privacy constraints. These relaxations include distance in entry pairs (Acharya et al., 2020), task-aware latent representation (Cheng et al., 2022), and graphs of data generating distributions (Geumlek & Chaudhuri, 2019). All these relaxations operate under *local DP* without a trusted data collector, compared to our setup under *global DP* (Dwork, 2006).

Utility Improvements in DP-ERM. In the context of DP-ERM, other utility improvements arise from public data or better gradient geometry. When the loss function in DP-ERM is convex with the feature dimension m, the minimax optimal utility guarantee

$\tilde{O}\left(\frac{\sqrt{m}}{n\epsilon}\right)$ can often be improved by relaxing certain conditions, typically through adaptive gradients and/or the use of public data. When the gradient lies in a low-dimensional subspace, the utility bound can be improved to $\tilde{O}\left(\frac{1}{n\epsilon}\right)$ (Zhou et al., 2020). In more general settings where better gradient geometry is available, the private adaptive gradient method can further enhance utility. For example, Asi et al. (2021) introduced non-isotropic noise into the gradient, while Li et al. (2022) utilized side information as a preconditioner to adapt to the gradient geometry. These approaches both achieve a better utility rate $\tilde{O}(\frac{c}{n\epsilon})$ with $c \ll \sqrt{m}$. When incorporating public data into training, Alon et al. (2019); Lowy et al. (2024) gave information theoretic utility lower bounds depending on the number of public data point. As highlighted in Section 4, our utility improvements are achieved by leveraging a refined noise scale that accounts for feature correlation with a reasonable privacy condition. This approach does not fundamentally rely on the availability of public data or assumptions about gradient geometry, and can be further improved when some favorable conditions, such as low-dimensional subspaces or improved gradient geometry, are present.

Correlated Data. Correlation between entries in the database may cause unexpected privacy issues. As discussed in Kifer & Machanavajjhala (2011), one of the challenges when releasing correlated datasets is tuning noise to balance the privacy-utility tradeoff. Song et al. (2017) proposed *Pufferfish privacy*, which is a generalization of DP and handles correlated entries. Zhang et al. (2022a) provides a discussion on the correlated data in DP. To the best of our knowledge, all these works focus on correlation between individuals rather than features. Zhang et al. (2022b) introduced attribute privacy for sensitive features, extending Pufferfish to allow feature correlations. Their approach protects privacy at the dataset- or distribution-level, whereas our framework focuses on the individual-level. (See Appendix B.1 for a more detailed discussion.) Chaudhuri & Courtade (2025) proposed *Add-remove Heterogeneous DP*, where privacy levels depend on each user's data; in contrast, our framework relaxes privacy in a dataset-independent way, based on statistical relation of insensitive attributes to sensitive ones. Aliakbarpour et al. (2025) proposed *Bayesian Coordinate DP (BCDP)* in a local DP setting without a trusted curator, which typically yields worse utility than central models.

2 CORRDP: SETUP AND MECHANISMS

Let $\mathcal{X}$ be the data domain where each point $X \in \mathcal{X} \subseteq \mathbb{R}^m$ can be partitioned as $X = (X^{\mathcal{S}}, X^{\mathcal{U}})^\top$, where $\mathcal{S}$ is the indices of sensitive features that require protection, and $\mathcal{U}$ is the indices of insensitive features, with $\mathcal{S} \cup \mathcal{U} = [m]$ and $\mathcal{S} \cap \mathcal{U} = \emptyset$. Let m_s be the number of sensitive features: $m_s = |\mathcal{S}|$. While the main body of the paper adopts a binary framework distinguishing sensitive and insensitive features, one can naturally extend CorrDP by assigning a sensitivity parameter to each feature. This yields a relaxed CorrDP notion without compromising theoretical guarantees. Appendix C.3 provides a detailed discussion and illustrates applications to ERM.

In the standard (global) differential privacy (Definition B.1 in Appendix B), all feature components of a data point are assumed to be equally sensitive, and thus the same privacy constraint is enforced on the change of any feature component between neighboring databases $\mathcal{D}, \mathcal{D}'$. To capture heterogeneity among feature changes, we present the concept of CorrDP that incorporates 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 if they differ in at most one entry, e and e', which themselves differ only in their sensitive features or insensitive features. See Remark 2.4 for commentary on this restriction.

Definition 2.2 (CorrDP)**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-correlated differentially private for a distance metric d if for all subsets $R \subseteq Range(\mathcal{A})$ 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 require the distance metric $d(\mathcal{D}, \mathcal{D}')$ to satisfy certain properties.

Definition 2.3 (Axioms of Sensitivity Distance)**.** For two neighboring databases $\mathcal{D}, \mathcal{D}'$, the distance metric d captures how much the changes between D and D' depend on the sensitive component, and must satisfy the following: (1) $d(\mathcal{D}, \mathcal{D}') \in [0, 1]$, (2) when e and e' differ in sensitive features (i.e., in $\mathcal{S}$), $d(\mathcal{D}, \mathcal{D}') = 1$, (3) when e and e' differ only in their insensitive features (i.e., in $\mathcal{U}$), $d(\mathcal{D}, \mathcal{D}') \in [0, 1)$, and (4) when all differing insensitive features are independent of the sensitive features, then $d(\mathcal{D}, \mathcal{D}') = 0$.

Definition 2.3 ensures that sensitivity appropriately captures correlations between sensitive and insensitive features. When the insensitive features are only weakly

correlated, the privacy constraint is relaxed; if neighboring databases differ in sensitive features, then CorrDP coincides with standard differential privacy.

Remark 2.4. The notion of neighbors excludes changes in both sensitive and insensitive features because then $d(\mathcal{D}, \mathcal{D}') = 1$, and the CorrDP constraint becomes exactly DP. Although CorrDP can be defined to include this case, no additional benefits can be gained from using this relaxed notion.

In the main body, we define $d(\mathcal{D}, \mathcal{D}')$ based on the Total Variation (TV) Distance, which naturally satisfies the desired properties above. In Appendix B.2, we describe other properties preserved by CorrDP, and in Appendix B.3, we discuss other distances that can be used.

Definition 2.5 (Choice of d)**.** For neighboring databases $\mathcal{D}, \mathcal{D}'$ that differ in two entries e and e',

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

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 used.

Example 2.6. Consider $X = (X^{(1)}, X^{(2)})^\top$, where $\mathcal{S} = \{1\}, \mathcal{U} = \{2\}$. $X^{(1)}$ and $X^{(2)}$ are partially correlated with $\mathbb{P}_{X^{(1)}|X^{(2)}=k} \sim \text{Bernoulli}(k/3), \forall k \in \{1, 2\}$. Suppose databases $\mathcal{D}, \mathcal{D}'$ differ in entries e, 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 influence of this change on the privacy constraint is smaller since the two entries only differ in the insensitive coordinate.

Example 2.7. Consider an economic dataset that includes a sensitive income feature $X^{(1)}$, and an insensitive geography feature $X^{(2)}$. If $\mathcal{D}, \mathcal{D}'$ differ only in the geographical information of one individual, this geographical information could still inadvertently reveal information about the individual's income level, as the income distribution is correlated with location. Then this geography feature also requires protection, albeit with a reduced budget of $\epsilon/d(\mathcal{D}, \mathcal{D}')$. In the extreme case where income depends solely on the geography, a change in location directly reveals the exact income level, and the privacy budget for this feature reduces to ϵ, since the geography information alone is sufficient to determine the individual's income.

2.1 Laplace Mechanism with CorrDP

We show how to adopt the Laplace Mechanism to achieve privacy under CorrDP, and show the utility improvements. First, we introduce sensitivity under CorrDP.

Definition 2.8 (ℓ_1, Coordinate, and Correlated Sensitivity)**.** The ℓ_1-sensitivity of function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ is: $\Delta f := \max_{\substack{\mathcal{D},\mathcal{D}' \\ \text{neighbors}}} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$, and the sensitivity of the k-th coordinate f_k of f is: $\Delta f_k := \max_{\substack{\mathcal{D},\mathcal{D}' \\ \text{neighbors}}} |f_k(\mathcal{D}) - f_k(\mathcal{D}')|$. 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 $TV(j) = \max_{x_1,x_2\in\mathcal{X}} TV(\mathbb{P}_{X^{\mathcal{S}}|x_1^{(j)}}, \mathbb{P}_{X^{\mathcal{S}}|x_2^{(j)}})$.

Note that $\Delta f \le \sum_{k\in[K]} \Delta f_k$, where equality holds when there exists a pair of neighboring databases that attains the maximum coordinate sensitivity Δf_k for all $k \in [K]$. The definition of correlated sensitivity accounts for the heterogeneous privacy constraints for insensitive features $j \in \mathcal{U}$, incorporating their correlation with sensitive features through a weighting factor, $TV(j)$.

Next we show that the Laplace Mechanism can be modified to satisfy CorrDP by using correlated sensitivity, and can lead to improved accuracy guarantees.

Definition 2.9 (CorrDP Laplace Mechanism)**.** For a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ where the i-th dimension of f only applies to the i-th feature of $\mathcal{D}$, the CorrDP Laplace mechanism is defined as: $\widetilde{\mathcal{M}}_L(\mathcal{D}, f, \epsilon) = f(\mathcal{D}) + (Y_1, \ldots, Y_K)$, where $Y_i \sim_{i.i.d.} \text{Lap}(\Delta_C f/\epsilon)$.

Theorem 2.10 (CorrDP Laplace Guarantees)**.** *The CorrDP Laplace mechanism is* $(\epsilon, 0)$*-CorrDP, and* $\forall\beta \in (0, 1]$, $\mathbb{P}[\|f(\mathcal{D}) - \widetilde{\mathcal{M}}_L(\mathcal{D}, f(\cdot), \epsilon)\|_\infty \le \frac{\Delta_C f \log(K/\beta)}{\epsilon}] \ge 1 - \beta$.

The proof is given in Appendix A.1. The standard Laplace mechanism $\mathcal{M}_L$ (Dwork, 2006) has a high-probability accuracy guarantee of $\|f(\mathcal{D}) - \mathcal{M}_L(\mathcal{D}, f, \epsilon)\|_\infty \le \frac{\Delta f \log(K/\beta)}{\epsilon}$. Comparing this with Theorem 2.10, we see that the CorrDP Laplace mechanism yields better accuracy exactly when the lack of correlation across features leads to lower sensitivity, thereby requiring less noise to be added.

3 CORRDP ERM

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i\in[n]}$ and an individual loss function $\ell(\theta; (X, Y))$ where X denotes the features and Y denotes the label, DP-ERM aims to obtain a solution $\theta^{priv} \in \mathbb{R}^m$ close to the nonprivate solution: $\hat{\theta} \in \text{argmin}_{\theta\in\Theta}\{F(\theta, \mathcal{D}) := 1/n \sum_{i=1}^n \ell(\theta; (x_i, y_i))\}$ with the guarantee of being differentially private. Across each privacy mechanism and setting, we measure the

utility gap 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 that satisfy (ϵ, δ)-DP will automatically satisfy our relaxed (ϵ, δ)-CorrDP notion, so the existence of CorrDP algorithms is a priori known. Our main goal is to see whether relaxing to CorrDP and appropriately modifying the DP-ERM algorithms will lead to utility improvements. We next give some assumptions that will be necessary for our results.

Assumption 3.1 (Neighboring Database in ERM)**.** The entry that differs across neighboring databases $\mathcal{D}, \mathcal{D}'$ only differs in sensitive features or insensitive features in X, and cannot differ in the label Y.

This assumption refines Definition 2.1 by excluding label changes in Y. This follows prior work (Shen et al., 2023), which treated labels as public and not requiring privacy protection. Since changing labels may change all features (due to the relationship between features and label), then CorrDP in this setting will collapse to standard DP, and no additional improvements can be seen. When labels are private and features are public, one can use Label DP (Ghazi et al., 2021). See Appendix C.3 for a detailed comparison between CorrDP and Label DP.

For the ERM problem, we consider general smooth convex losses 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 $\forall x \in \mathcal{X}$, $\|x\|_2 \leq B$, and the parameter is bounded: $\|\theta\|_2 \leq D$.

The boundness of domain and parameters is naturally imposed for theoretical analyses in DP-ERM algorithms (Wang et al., 2017). In practice, unbounded gradients can be handled via gradient clipping. In Appendix C.3, we relax Assumption 3.3 and show that our main results still hold.

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

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

This assumption requires the loss function to be smooth with respect to changes in x, and the sensitivity of the i-th parameter coordinate to be mainly controlled by the i-th feature component.

We note that Assumptions 3.3 and 3.4 for CorrDP-SGD are slightly stronger than the standard assumptions than DP-SGD, but Assumption 3.3 holds when all feature coordinates have relatively similar ranges (which can be done from normalization in model fitting) and Assumption 3.4 holds when $\ell(\theta; (x, y))$ can be represented by $\tilde{\ell}(\theta^\top x, y)$ for some $\tilde{\ell}$, i.e., *generalized linear model (GLM)*, such as OLS or logistic regression. We provide details in Appendix C.2.

3.1 CorrDP-SGD Algorithm and Guarantees

We now present CorrDP-SGD in Algorithm 1, which incorporates CorrDP with DP-SGD (Bassily et al., 2014; Abadi et al., 2016). The key change is the modified noise scale in Line 1, where the noise terms σ_i are set. Compared with the noise added in the standard DP-SGD mechanism (Bassily et al., 2014), in the noise variance term σ_i^2 for $i \in \mathcal{U}$, 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 the privacy guarantee of CorrDP still holds.

Algorithm 1 CorrDP Stochastic Gradient Descent (CorrDP-SGD)

input Parameter domain Θ, number of iterations T, step sizes α_t, dataset $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$ with sensitive features $\mathcal{S}$ and insensitive features $\mathcal{U}$, sample size n_q with $1 \leq n_q \leq n$, CorrDP parameters (ϵ, δ)

1: Initialize θ_1, set $m_s = |\mathcal{S}|$ and $TV(i) = \max_{x_1, x_2 \in \mathcal{X}} TV(\mathbb{P}_{X^{\mathcal{S}}|x_1^{\mathcal{U}}}, \mathbb{P}_{X^{\mathcal{S}}|x_2^{\mathcal{U}}})$ $\forall i \in \mathcal{U}$, and set diagonal entries of the noise variance $\{\sigma_i^2\}_{i \in [m]}$

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

2: **for** $t = 1, \ldots, T$ **do**

3: Randomly sample n_q datapoints $\{(x_{(i)}, y_{(i)})\}_{i \in [n_q]}$ from the dataset $\mathcal{D}$.

4: Generate noise $b \sim N(0, \text{diag}(\boldsymbol{\sigma}^2))$ and update:

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

5: **end for**

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

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

The full proof of this main result is given in Appendix A.2. We give a proof sketch here, starting with the result for the full-batch gradient descent ($n_q = n$). The analysis is similar to the standard moment accountant argument (Abadi et al., 2016), replacing the parame-

ter ϵ in the DP constraint with $\epsilon/d(\mathcal{D},\mathcal{D}')$, which is dataset-dependent. This modified moment accountant argument requires a refined analysis of the distributional distance between $\mathcal{D}$ and $\mathcal{D}'$ for the sensitive and insensitive components. When $\mathcal{D}$ and $\mathcal{D}'$ differ only in sensitive features $\mathcal{S}$, the noise imposed on the sensitive features $\mathcal{S}$ is the same as the standard noise to preserve the privacy for the sensitive features, and the noise term m_s^2/m^2 imposed on the insensitive features $\mathcal{U}$ ensures that privacy for the insensitive features is preserved from Assumption 3.4. Conversely, when $\mathcal{D}$ and $\mathcal{D}'$ only differ in insensitive features, the noise term $TV(i)$ imposed on the insensitive feature $i \in \mathcal{U}$ ensures privacy for that i-th feature. For the general version of the stochastic gradient descent where $n_q < n$, the full batch result can be directly extended to SGD using privacy amplification via sampling (Balle et al., 2018).

Remark 3.6. When Assumption 3.1 is replaced with a stronger assumption that e and e' differ in sensitive features or in one insensitive feature in X, we can simplify $TV(i)$ as $\max_{x_1,x_2\in\mathcal{X}} TV(\mathbb{P}_{X^{\mathcal{S}}|x_1^{(i)}}, \mathbb{P}_{X^{\mathcal{S}}|x_2^{(i)}})$ for the same privacy guarantee of Theorem 3.5.

Theorem 3.7 (Utility Guarantee of CorrDP-SGD)**.** *Under Assumptions 3.1, 3.2, 3.3 and 3.4, for Algorithm 1 with step sizes* $\alpha_t = D/\sqrt{(L^2+\sum_{i=1}^m \sigma_i^2)t}$ *and* $T = \Theta(n^2)$*, if* $F(\theta,\mathcal{D})$ *is convex, then* $R(\theta^{priv}) = \tilde{O}\left(\sqrt{(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), m_s/4\})\log(1/\delta)}/(n\epsilon)\right)$.

The stepsize choice follows by verifying the square norm of the gradient at each step is $\mathbb{E}[\|F(\theta_t,\mathcal{D})\|_2^2] \le L^2 + \sum_{i=1}^m \sigma_i^2$ and applying Theorem 2 of Shamir & Zhang (2013). The full proof of Theorem 3.7 is given in Appendix A.3.

Corollary 3.8 (Improved Utility Guarantee)**.** *Under the same conditions as Theorem 3.7, if* $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$*, then* $R(\theta^{priv}) = \tilde{O}\left(\sqrt{m_s}/(n\epsilon)\right)$.

This result demonstrates that if there are many insensitive features that are each not very correlated with sensitive features, i.e., $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$, then compared with the DP-ERM utility of $\tilde{O}\left(\sqrt{m}/(n\epsilon)\right)$ Bassily et al. (2014); Wang et al. (2018), CorrDP-SGD significantly improves utility guarantee when $m_s = o(m)$. Such improved utility guarantees under heterogeneous noise scales in Algorithm 1 can be applied to improve utility performance under strongly convex and general nonconvex losses (e.g., Corollary C.1) and other privacy-preserving first-order algorithms including SVRG (Wang et al., 2017) or Adam, while reducing the gradient complexity.

Extension to Neural Networks. Beyond GLMs, CorrDP can also be applied to more general loss functions of the form $\ell(f(\theta;x),y)$ that do not satisfy Assumption 3.4. Consider general loss functions of the form $\ell(f(\theta;x),y)$, where $f(\theta;x) : \mathbb{R}^{m_\theta} \times \mathbb{R}^{m_x} \to \mathbb{R}$. Here, $\theta \in \mathbb{R}^{m_\theta}$ and $x \in \mathbb{R}^{m_x}$ have different dimensions ($m_\theta \neq m_x$). The most common application of such general loss functions is neural networks (NN) and CorrDP-SGD can be run to differentially privately train neural networks while adding reduced noise to the first layer, which will also result in improved performance relative to DP-SGD (under mild conditions).

We consider a fully connected NN, with the following assumptions similar to Assumption 3.4:

Assumption 3.9 (Sensitivity of Fully Connected NN)**.** Consider $f(\theta;x) = g(\theta^{(K)}\cdots g((\theta^{(0)})^\top x))$, where $g(\cdot)$ denotes the activation function of the NN parameterized by $\theta = (\theta^{(0)},\ldots,\theta^{(K)})$, with the total dimension $m_\theta = \sum_{i=0}^K m_{\theta_i}$. Then for parameter $\theta^{(0)} \in \mathbb{R}^{m\times d_0}$, where d_0 is the number of neurons in the first hidden layer: $(\nabla_\theta \ell(f(\theta^{(0)};x_1),y) - \nabla_\theta \ell(f(\theta^{(0)};x_2),y))_i \le C_1 L|x_1^{(i)} - x_2^{(i)}| + C_2 \sum_{j\neq i} \frac{L}{m}|x_1^{(j)} - x_2^{(j)}|, i \in [m_{\theta_0}]$.

When adapting Algorithm 1 in the context of neural networks, we adjust the noise $b = (b^{(0)},\ldots,b^{(K)})$, with $b^{(0)} \in \mathbb{R}^{m\times d_0}$, and the total dimension is m_θ. Each noise term is sampled independently: for $b_{i,j}^{(k)} \sim N(0,\sigma_{i,j}^{(k)}), k \in \{0,1,\ldots,K\}$, the noise scale $\forall i \in [m], j \in [d_0]$ is set as:

$$(\sigma_{i,j}^{(0)})^2 = \begin{cases} \frac{(\log(1/\delta)+1)L^2T}{n^2\epsilon^2}, \text{ if } i \in \mathcal{S}; \\ \frac{(\log(1/\delta)+1)L^2T\max\{TV(i), m_s^2/m^2\}}{n^2\epsilon^2}, \text{ else} \end{cases} \tag{4}$$

For the noise applied to the subsequent layers $\forall k \ge 1$, we use $(\sigma_{i,j}^{(k)})^2 = \frac{(\log(1/\delta)+1)L^2T}{n^2\epsilon^2}$. Under Assumption 3.9, running CorrDP-SGD (Algorithm 1) to add noise to i-th row of $\theta^{(0)}$ for $i \in \mathcal{U}$ according to Equation (4), satisfies (ϵ,δ)-CorrDP.

When $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$, in NNs, CorrDP improves the utility guarantees from $\tilde{O}\left(\frac{\sqrt{m_\theta}}{n\epsilon}\right)$ under regular DP (Wang et al., 2017) to $\tilde{O}\left(\frac{\sqrt{m_\theta - (m-m_s)d_0}}{n\epsilon}\right)$ by instantiating Theorem 3.7. If the number of neurons in the first hidden layer d_0 is large compared with those of subsequent layers, this provides significant utility improvements. In specialized NNs like RNN modules, insensitive features may be processed in the first several layers without sensitive feature modules, and so there may be opportunities for more significant utility improvements via CorrDP.

3.2 Lower Bound

We provide a near-matching lower bound in terms of n, ϵ, and problem dimension. The key is to build the equivalence between CorrDP and standard DP definitions so that we can translate known lower bounds

from DP (Bassily et al., 2014) to CorrDP. A full proof is in Appendix A.4.

Theorem 3.10 (Lower bound for (ϵ, δ)-CorrDP algorithm)**.** *Let $n, m \in \mathbb{N}, \epsilon > 0$ and $\delta = o(1/n)$. Consider $\{TV(i)\}_{i\in\mathcal{U}}$ sorted in descending order, denoted $\{TV^{(i)}\}_{i\in\mathcal{U}}$. For every (ϵ, δ)-CorrDP algorithm that outputs θ^{priv}, there exists a $\mathcal{D}$ such that with probability at least 1/3,*

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

Comparing this lower bound to the utility upper bound of CorrDP-SGD in Theorem 3.7, observe that $(TV^{(1)})^2 \leq \max_{k\in[(m-m_s)]}\{k(TV^{(k)})^2\} \leq \sum_{i\in[(m-m_s)]} TV(i)$, so this is near-matching in terms of the problem dimension.

4 ESTIMATING TV DISTANCE IN CORRDP-SGD

In this section, we show how to extend CorrDP-SGD to handle the fact that the TV distance may not be known. The expression for the noise terms σ_i (Equation (3)) in CorrDP-SGD depends on the maximum of the conditional total variation distance $TV(i)$. However, this term may be unknown in general, and can leak information when estimated from $\mathcal{D}$. In this section, we demonstrate that in many scenarios, the error from this additional estimation step is insignificant after proper processing. As a simple example, imagine there is domain knowledge of the exact value or a near-exact upper bound: $U_i = TV(i)(1 + o(1))$. It is easy to see that the utility and privacy guarantees are not affected in this case by simply replacing each unknown $TV(i)$ with $U_i, i \in \mathcal{U}$.

We will mainly focus on the more general case when such knowledge is not available, but the estimation of TV distance is regular and smooth. Our goal will be to find adjusted expressions for the noise terms in Equation (3) while still obtaining similar privacy and utility guarantees to the case where $TV(i)$ are known.

In Equation (3), when $\{TV(i)\}_{i\in\mathcal{U}}$ are unknown, the straightforward approach is to empirically estimate them from the dataset $\mathcal{D}$ as $\widehat{TV}_{\mathcal{D}}(i)$ and use these in place of $\{TV(i)\}_{i\in\mathcal{U}}$. However, this ignores that the estimation procedure itself can leak information about the database, and thus this procedure alone will not lead to correct privacy guarantees. Instead, we adjust the estimation of $\{TV(i)\}_{i\in\mathcal{U}}$ in the noise terms σ_i. We require two assumptions: that this estimator has bounded error and bounded sensitivity.

Assumption 4.1. With probability at least $1 - \beta$, for each $i \in [m]$, $|\widehat{TV}_{\mathcal{D}}(i) - TV(i)| \leq c_2\sqrt{\log(1/\beta)}/n^{\gamma}$ for some $c_2 \in (0, \infty)$ and $\gamma \in (0, \frac{1}{2}]$.

Assumption 4.2. For all neighbors $\mathcal{D}, \mathcal{D}'$, $\forall i \in \mathcal{U}, |\widehat{TV}_{\mathcal{D}'}(i) - \widehat{TV}_{\mathcal{D}}(i)| \leq \frac{c_3}{n}$ for some $c_3 < \infty$.

Assumption 4.1 ensures the estimation error of the TV distance decreases with the sample size. Assumption 4.2 ensures the stability of the TV distance estimate, requiring that changing one sample leads to a bounded impact on the estimator. This is reasonable since one sample only influences the probability mass by at most $1/n$. In Appendix D, we provide examples of the empirical estimator $\widehat{TV}_{\mathcal{D}}(i)$ that satisfy Assumptions 4.1 and 4.2, and their corresponding value of γ. For example, in Examples D.1 and D.3 where $X^S|X^U = x$ follows a Gaussian distribution and X follows a joint Gaussian distribution, then $\gamma = \frac{1}{2}$ because mean estimation in this setting has a convergence rate of exactly $O_p(n^{-1/2})$. For a more general case of $X^S|X^U = x$, the corresponding γ is strictly less than $\frac{1}{2}$ when using nonparametric estimation methods (Tsybakov, 2008), such as kernel or histogram estimate (see Example D.2) to estimate the conditional distribution.

In this general case, we use the following estimation procedure for $TV(i)$ in the noise term σ_i^2.

Definition 4.3 (In-Sample TV Estimation)**.** When $\{TV(i)\}_{i\in\mathcal{U}}$ is unknown, replace $TV(i)$ with $\widetilde{TV}(i) = \widehat{TV}_{\mathcal{D}}(i) + 2c_2\sqrt{\log((m-m_s)/\delta)}/n^{\gamma}$ in the expression for σ_i^2 in Equation (3).

The additional term in the estimator ensures sufficient noise is added, since the empirical distance $\widehat{TV}(i)$ may underestimate $TV(i)$. The error of the sensitivity does not explicitly appear in Definition 4.3 because its magnitude is $O(1/n)$ by Assumption 4.2, and thus is dominated by the estimation error as a consequence of Assumptions 4.1, and is explicitly taken into account by Definition 4.3 when n is large.

Theorem 4.4 (Guarantees of CorrDP with In-Sample TV Estimation)**.** *When the estimator in Definition 4.3 is used for the noise terms σ_i, Assumptions 4.1 and 4.2 hold, and $n = \Omega(\log(1/\delta))$, then Algorithm 1 is $(\epsilon, 2\delta)$-CorrDP and achieves the same utility guarantee as Theorem 3.7.*

The proof of Theorem 4.4 is more involved than the standard analyses in Theorem 3.5 due to the dependence of $\widetilde{TV}$ on the database $\mathcal{D}$, and can be found in Appendix A.5. First, since the empirical estimator $\widetilde{TV}$ can underestimate the true TV distance, the analysis introduces a high-probability upper bound to ensure that $\widetilde{TV}(i) \geq TV(i), \forall i \in \mathcal{U}$. This guarantees conservative noise calibration and enables control of the Renyi divergence, following the proof technique for Theorem 3.5. Second, the noise variances for $\mathcal{D}$ and $\mathcal{D}'$ may be different, which introduces an additional difference that must be accounted for in the privacy

analysis. We show that the resulting difference can be bounded by $\Theta(1/n^2)$ using Assumption 4.2, and thus still yields valid privacy guarantees.

Remark 4.5 (Publicly available data). Suppose there is a public database $\tilde{\mathcal{D}}$ of size $\tilde{n}$ that shares the same covariate distribution with our database $\mathcal{D}$. Then setting $\widetilde{TV}(i) := \widehat{TV}_{\tilde{\mathcal{D}}}(i) + 2c_2\sqrt{\log((m-m_s)/\delta)}/\tilde{n}^{\gamma}$ from the empirical estimate $\widehat{TV}_{\tilde{\mathcal{D}}}(i)$ and plugging this estimate into the expression for σ_i in Equation (3), will preserve the guarantees of Theorem 4.4.

5 NUMERICAL EXPERIMENTS

In this section, we empirically demonstrate that CorrDP achieves a better privacy-utility tradeoff than standard DP for empirical risk minimization. We use CorrDP-SGD (Algorithm 1) as the foundational algorithm across different experimental setups, and compare the utility of CorrDP against three baseline methods: (i) *Semi*: adds noise only on the feature component corresponding to the sensitive feature; (ii) *Standard*: standard DP-SGD that adds uniform noise to all features (Abadi et al., 2016); (iii) *Partial*: discards the sensitive features and runs non-private algorithms only on insensitive features. The Semi baseline should provide an upper bound on utility for CorrDP since Semi provides a weaker privacy guarantee that doesn't account for correlations across sensitive and insensitive features. Our goal is to show that CorrDP achieves better utility than the *Standard* and *Partial* baselines, and performs comparably to *Semi*. Our loss setup encompasses strongly convex losses (e.g., linear and logistic regression) and non-convex losses (e.g., neural networks). For privacy parameters, we set $\delta = 10^{-4}$ in synthetic datasets and $\delta = 10^{-5}$ in real datasets.

We specify the noise magnitude used by each method in our problem setting. Across all methods, the default configuration (*Standard*) injects additive noise $b \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\sigma}^2))$ at each round, with $\sigma_i^2 = C \cdot \frac{\log(1/\delta)+1}{\epsilon^2}$ for some constant $C > 0$.

- For *Semi*, when using OLS or logistic regression, we change $\sigma_i^2 = 0$ for indices $i \in U$; when using neural networks, we change $(\sigma^{(0)}_{(i,j)})^2 = 0$ for $i \in U$ at the first layer, and apply the standard noise scale to all subsequent layers.
- For CorrDP, when using OLS or logistic regression, we change $\sigma_i^2 = C \cdot \frac{\log(1/\delta)+1}{\epsilon^2} TV(i)$ for indices $i \in [m]$; when using neural networks, we change $(\sigma^{(0)}_{(i,j)})^2 = C \cdot \frac{\log(1/\delta)+1}{\epsilon^2} TV(i)$ at the first layer, and apply the standard noise scale to all subsequent layers. In both cases, the unknown $TV(i)$ is estimated by the upper bound estimate of the empirical counterpart $\widetilde{TV}(i)$ using finite samples, as described in Section 4.

Full details of the experimental setup, along with additional results, are provided in Appendix E.

5.1 Synthetic dataset and results

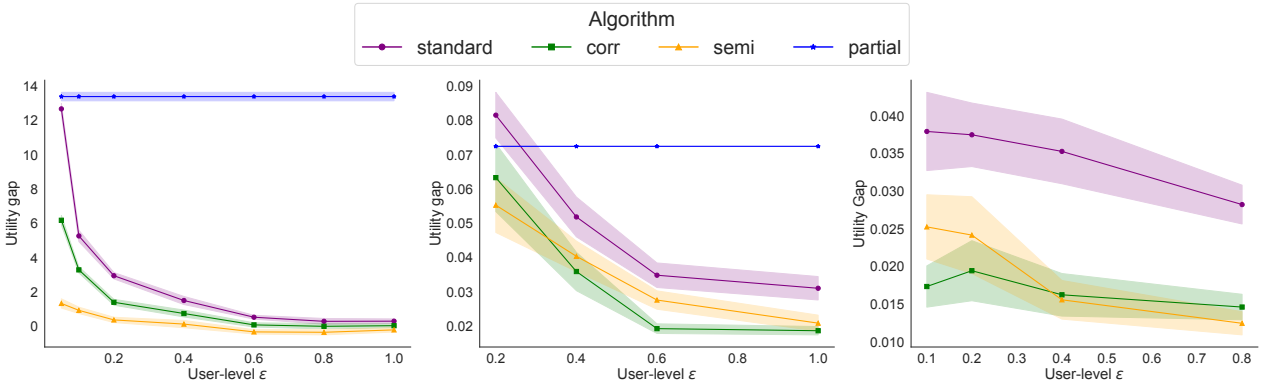
We first consider synthetic data, and run a (ridge) linear regression model with loss $\ell(\theta; (x, y)) = (y - \theta^\top x)^2 + \frac{1}{n}\|\theta\|_2^2$. We generate the synthetic training data according to a multivariate Gaussian, $X = (X^{\mathcal{S}}, X^{\mathcal{U}})^\top \sim N(\boldsymbol{\mu}, \Sigma)$, $Y = \theta^\top X + \eta$ with $\eta \sim N(0, 5^2)$, $m = 100$ features, sensitive features $\mathcal{S} = \{1, 2, \ldots, 10\}$, and $n = 2000$ samples. Other parameter values including $\boldsymbol{\mu}$ and Σ are specified in Appendix E.1. Therefore, the conditional distribution of sensitive features can be directly calculated and estimated from the posterior of the Gaussian distribution, which is still Gaussian.

Figure 1a presents empirical findings on this dataset, which plots the utility gap (Equation (2)) as a function of ϵ for all four algorithms, where lower utility gap indicates better performance. We observe that CorrDP always outperforms Standard and Partial, especially so in the high privacy regime. The performance of CorrDP is comparable to that of Semi, especially for more reasonable values of ϵ, indicating that substantial performance is not lost with the stronger privacy guarantees. All three methods that incorporate the sensitive features (CorrDP, Semi, and Standard) outperform Partial, which ignores these features. In Appendix E.1, we include additional results that vary the number of sensitive features m_s.

5.2 Real-world datasets and results

We also empirically evaluate the performance of CorrDP on four real-world classification or regression datasets: Adult (adu), Sepsis (sep), Credit Card (cre) and Medical Cost (med). These datasets include both discrete and continuous features, where all discrete features are processed with one-hot encoding. For the three classification datasets (Adult, Sepsis, Credit Card), we use the same categorization as Shen et al. (2023) of these features into sensitive (private) and insensitive (public). In the Medical Cost dataset (regression task), we set sex, smoke, region as insensitive features and all others as sensitive.

Adult dataset. We filter the dataset with 45,175 samples and 9 features, and employ a logistic regression model for binary classification. Additional details are presented in Appendix E.2. In Figure 1b, the baseline accuracy for the non-private algorithm is 83%, and we calculate the accuracy gap of private algorithms relative to this. We observe that CorrDP even outperforms Semi,



(a) Synthetic Data (OLS) (b) Adult (Log. Reg) (c) Sepsis (2-NN)

Figure 1: Privacy-utility trade-offs for least-square regression, logistic regression, and neural network training. The shaded area around each curve shows the standard deviation error band. We do not run the Partial algorithm on the Sepsis dataset since it is not comparable to use the same architecture only with partial features.

and suffers less than a 4% accuracy drop compared with the non-private baseline, even in a relatively high privacy regime ($0.2 < \epsilon < 0.5$).

Sepsis dataset. We randomly sample 10,000 points from the dataset with 3 features and utilize a two-layer neural network, where the activation function and the number of the hidden layer are 'Softplus' and 5 respectively, and the activation function of the final layer is the 'softmax'. Based on the extension of CorrDP to NNs presented in Section 3.1, for CorrDP and Semi, less noise can be added than Standard by only adding noise in the corresponding input component of the first layer. Additional experimental details are presented in Appendix E.3. In Figure 1c, the baseline accuracy for non-private algorithm is 69%, and we observe that CorrDP has similar performance to Semi and consistently outperforms Standard, despite the stochasticity of the neural network initiation and batch gradient descent. All these results indicate CorrDP provides strong privacy and utility guarantees while relaxing the scale of noise added.

Credit Card and Medical Cost dataset. We also apply logistic regression on the Credit Card dataset, and linear regression and two-layer neural network on the Medical Cost dataset (Figure 3); our empirical results there continue to show that CorrDP outperforms Standard and performs comparably to Semi. Due to space constraints, full discussion of these datasets and results are deferred to Appendix E.4.

6 DISCUSSION

In this work, we studied CorrDP, a relaxed notion of differential privacy that incorporates feature correlation across partitioned sensitive and insensitive features. We showed how this notion can be applied to the DP-ERM and DP-SGD algorithms. Our theoretical results show that CorrDP leads to utility improvements for the same fixed ϵ value, relative to standard DP. The essence of these improvements comes from adding less noise to features that are independent from, or only weakly correlated with, the sensitive variables. We verified these findings empirically across real and synthetic datasets. We also gave extensions of these results, showing that our algorithmic framework could be applied (i) to neural networks with more general loss functions, (ii) when there was uncertainty about the level of correlation, and (iii) using general notions of distance across features.

Our work has several limitations that point to promising directions for future research. First, the framework is most effective when the distinction between sensitive and insensitive features is clearly defined. In settings where sensitivity lies on a spectrum and the boundary is less clear, we introduce relaxations in Appendix C.3 that still yield improved utility guarantees. Second, ensuring privacy while maintaining strong utility under CorrDP-SGD requires additional assumptions on the feature space and the structure of the fitted model. Third, the noise calibration in CorrDP-SGD relies on an (upper bound) estimate of the TV distance. While we provide a method for computing such an estimate, this step can be computationally demanding in high-dimensional problems, with some potential remedies discussed at the end of Appendix D. Advances in distance estimation methods would directly enhance the efficiency of this step in CorrDP. Furthermore, extending the theoretical guarantees of CorrDP to other statistical tasks, such as hypothesis testing and related inferential problems, would be worth investigating.

Acknowledgements

R.C. was supported in part by NSF grant CNS-2138834 (CAREER), and by the Center for Smart Streetscapes, an NSF Engineering Research Center, under grant

agreement EEC-2133516.

References

Adult income dataset. https://www.kaggle.com/datasets/wenruliu/adult-income-dataset. Accessed: Jan 2025.

Default of credit card clients. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients. Accessed: Mar 2025.

Medical cost dataset. https://www.openml.org/search?type=data&status=active&id=46289. Accessed: Mar 2025.

Sepsis survival dataset. https://www.kaggle.com/datasets/joebeachcapital/sepsis-survival-minimal-clinical-records/data. Accessed: Jan 2025.

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.

Acharya, J., Bonawitz, K., Kairouz, P., Ramage, D., and Sun, Z. Context aware local differential privacy. In *International Conference on Machine Learning*, pp. 52–62. PMLR, 2020.

Aliakbarpour, M., Chaudhuri, S., Courtade, T., Fallah, A., and Jordan, M. Enhancing feature-specific data protection via bayesian coordinate differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2025.

Alon, N., Bassily, R., and Moran, S. Limits of private learning with access to public data. *Advances in neural information processing systems*, 32, 2019.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 901–914, 2013.

Asi, H., Duchi, J., Fallah, A., Javidbakht, O., and Talwar, K. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021.

Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

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.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 1–10, 2014.

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

Chaudhuri, S. and Courtade, T. A. Managing correlations in data and privacy demand. *arXiv preprint arXiv:2509.02856*, 2025.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Cheng, J., Tang, A., and Chinchali, S. Task-aware privacy preservation for multi-dimensional data. In *International Conference on Machine Learning*, pp. 3835–3851. PMLR, 2022.

Chua, L., Cui, Q., Ghazi, B., Harrison, C., Kamath, P., Krichene, W., Kumar, R., Manurangsi, P., Narra, K. G., Sinha, A., et al. Training differentially private ad prediction models with semi-sensitive features. *arXiv preprint arXiv:2401.15246*, 2024.

Devroye, L., Mehrabian, A., and Reddad, T. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

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

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.

Geumlek, J. and Chaudhuri, K. Profile-based privacy for locally private computations. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 537–541. IEEE, 2019.

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.

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.

Hayfield, T. and Racine, J. S. Nonparametric econometrics: The np package. *Journal of statistical software*, 27:1–32, 2008.

Imola, J., Kasiviswanathan, S., White, S., Aggarwal, A., and Teissier, N. Balancing utility and scalability in metric differential privacy. In *Uncertainty in Artificial Intelligence*, pp. 885–894. PMLR, 2022.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Kifer, D. and Machanavajjhala, A. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204, 2011.

LemaÃŽtre, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5, 2017.

Li, T., Zaheer, M., Reddi, S., and Smith, V. Private adaptive optimization with side information. In *International Conference on Machine Learning*, pp. 13086–13105. PMLR, 2022.

Li, W., Wang, H., Huang, Z., and Pang, Y. Private wasserstein distance with random noises. *arXiv preprint arXiv:2404.06787*, 2024.

Liu, H., Xu, D., Tian, Y., Peng, C., Wu, Z., and Wang, Z. Wasserstein generative adversarial networks based differential privacy metaverse data sharing. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Lowy, A., Li, Z., Huang, T., and Razaviyayn, M. Optimal differentially private model training with public data. In *Forty-first International Conference on Machine Learning*, 2024.

Rakotomamonjy, A. and Liva, R. Differentially private sliced wasserstein distance. In *International Conference on Machine Learning*, pp. 8810–8820. PMLR, 2021.

Rohde, A. and Steinberger, L. Geometrizing rates of convergence under local differential privacy constraints. *The Annals of Statistics*, 48(5):2646–2670, 2020.

Sart, M. Density estimation under local differential privacy and hellinger loss. *Bernoulli*, 29(3):2318–2341, 2023.

Schneider, D., Sajadmanesh, S., Sehwag, V., Sarfraz, S., Stiefelhagen, R., Lyu, L., and Sharma, V. Masked differential privacy. *arXiv preprint arXiv:2410.17098*, 2024.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

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

Shi, W., Cui, A., Li, E., Jia, R., and Yu, Z. Selective differential privacy for language modeling. *arXiv preprint arXiv:2108.12944*, 2021.

Song, S., Wang, Y., and Chaudhuri, K. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1291–1306, 2017.

Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

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.

Wang, D., Gaboardi, M., and Xu, J. Empirical risk minimization in non-interactive local differential privacy: Efficiency and high dimensional case. *arXiv preprint arXiv:1802.04085*, 2018.

Zhang, T., Zhu, T., Liu, R., and Zhou, W. Correlated data in differential privacy: definition and analysis. *Concurrency and Computation: Practice and Experience*, 34(16):e6015, 2022a.

Zhang, W., Ohrimenko, O., and Cummings, R. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 757–766, 2022b.

Zhao, Y. and Chen, J. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28, 2022.

Zhou, Y., Wu, Z. S., and Banerjee, A. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We provide them in Sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We describe the sample complexity for the ideal case when we know the TV distance in Section 3 and the sample complexity when TV distances need to be incorporated in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] We provide the source code in the supplementary files.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] We provide all the assumptions for the results in the main body in Section 3 and additional assumption relaxation in Appendix C.3.
   (b) Complete proofs of all theoretical results. [Yes] We provide proofs of all the theoretical results in Appendice A.
   (c) Clear explanations of any assumptions. [Yes] We explain the interpretation of each assumption in the main body along in Section 2.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] We provide the detailed instructions of reproducing main results in the supplemental files.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We provide the details of experimental results in Section 5 and Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We provide the error bar across random seed in the description of Figure 1.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We cite the dataset resources in References, with URLs.
   (b) The license information of the assets, if applicable. [Yes] Same as above.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix to "Integrating Feature Correlation in Differential Privacy with Applications in DP-ERM"

A Omitted Proofs

A.1 Proof of Theorem 2.10

Theorem 2.10 (CorrDP Laplace Guarantees)**.** *The CorrDP Laplace mechanism is* $(\epsilon, 0)$*-CorrDP, and* $\forall \beta \in (0, 1]$, $\mathbb{P}[\|f(\mathcal{D}) - \widetilde{\mathcal{M}}_L(\mathcal{D}, f(\cdot), \epsilon)\|_\infty \leq \frac{\Delta_C f \log(K/\beta)}{\epsilon}] \geq 1 - \beta$.

Proof. Consider two neighboring databases $\mathcal{D}, \mathcal{D}' \in \mathbb{N}^{|\mathcal{X}|}$, and let $f(\cdot)$ be some function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$. Let $p_{\mathcal{D}}$ denote the probability density function of $\widetilde{M}_L(\mathcal{D}, f, \epsilon)$ and $p_{\mathcal{D}'}$ denote the probability density function of $\widetilde{M}_L(\mathcal{D}', f, \epsilon)$ with $\mathcal{D}_{:,i}$ and $\mathcal{D}'_{:,i}$ as their i-th column (feature) respectively. Then for any arbitrary point $z \in \mathbb{R}^K$:

$$\begin{aligned}
\frac{p_{\mathcal{D}}(z)}{p_{\mathcal{D}'}(z)} &= \prod_{i=1}^{K} \left(\frac{\exp(-\frac{\epsilon|f(\mathcal{D}_{:,i}) - z_i)|}{\Delta_C f})}{\exp(-\frac{\epsilon|f(\mathcal{D}_{:,i}) - z_i)|}{\Delta_C f})} \right) \\
&= \prod_{i=1}^{K} \exp\left(\frac{\epsilon(|f(\mathcal{D}'_{:,i}) - z_i| - |f(\mathcal{D}_{:,i}) - z_i|)}{\Delta_C f} \right) \\
&\leq \prod_{i=1}^{K} \exp\left(\frac{\epsilon|f(\mathcal{D}_{:,i}) - f(D'_{:,i})|\mathbf{1}_{D_i \neq \mathcal{D}'_{:,i}}}{\Delta_C f} \right) \\
&\leq \exp\left(\frac{\epsilon \sum_{k \in [K]} \Delta f_k \mathbf{1}_{D_{:,k} \neq D'_{:,k}}}{\Delta_C f} \right) \\
&\leq \exp\left(\frac{\epsilon}{d(\mathcal{D}, \mathcal{D}')} \right) \mathbf{1}_{\mathcal{D}_{:,\mathcal{S}} = \mathcal{D}'_{:,\mathcal{S}}} + \exp(\epsilon) \mathbf{1}_{\mathcal{D}_{:,\mathcal{S}} \neq \mathcal{D}'_{:,\mathcal{S}}},
\end{aligned}$$

where the second inequality comes from the fact that $\Delta f \leq \sum_{k \in [K]} \Delta f_k$, and the third inequality follows from considering the two cases when $\mathcal{D}_{:,\mathcal{S}} = \mathcal{D}_{:,\mathcal{S}}$ and $\mathcal{D}_{:,\mathcal{S}} \neq \mathcal{D}'_{:,\mathcal{S}}$ (since only one insensitive feature can change by Definition 2.1).

For the utility guarantee, we have:

$$\begin{aligned}
\mathbb{P}\left[\|f(\mathcal{D}) - \widetilde{\mathcal{M}}_L(\mathcal{D}, f(\cdot), \epsilon)\|_\infty \geq \frac{\Delta_C f}{\epsilon} \log(K/\beta) \right] &= \mathbb{P}\left[\max_{i \in [K]} |Y_i| \geq \frac{\Delta_C f}{\epsilon} \log(K/\beta) \right] \\
&\leq K \mathbb{P}\left[|Y_i| \geq \frac{\Delta_C f}{\epsilon} \log(K/\beta) \right] \\
&= K \cdot \frac{\beta}{K} \\
&= \beta,
\end{aligned}$$

where the first equality comes from the definition of the mechanism, the second equality comes from a union bound over $i \in [K]$, and the second step comes from the fact each $Y_i \sim \text{Lap}(\Delta_C f/\epsilon)$ and tail bounds on the Laplace distribution. We note that this analysis is nearly identical to the utility guarantee of the Laplace Mechanism in Dwork (2006), but is included for completeness. □

A.2 Proof of Theorem 3.5

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

A.2.1 Helper Lemmas

We first introduce and prove the following lemmas, which will be used in the proof of Theorem 3.5.

Lemma A.1 (High-Probability Composition of CorrDP)**.** *Suppose with probability $1 - \delta_2$ (taken over the randomness in the dataset and the algorithmic mechanism), the algorithm $\mathcal{A}$ satisfies (ϵ, δ_1)-CorrDP. Then $\mathcal{A}$ is $(\epsilon, \delta_1 + \delta_2)$-CorrDP.*

Proof. Consider the good event E that $\mathcal{A}$ is (ϵ, δ_1)-CorrDP with $\mathbb{P}(E) = 1 - \delta_2$. Then:

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}|E) \leq e^{\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in \mathcal{S}|E) + \delta_1.$$

On the contrary, under E, we have: $\mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}|E) \leq 1$.

Combining these facts:

$$\begin{aligned}
\mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}) &= (1-\delta_2)\mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}|E) + \delta_2 \mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}|\bar{E}) \\
&\leq (1-\delta_2)(e^{\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}} \mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}|E) + \delta_1) + \delta_2 \cdot 1 \\
&= (1-\delta_2) e^{\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in \mathcal{S}|E) + \delta_1 + \delta_2 - \delta_1\delta_2 \\
&\leq e^{\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in \mathcal{S}) + \delta_1 + \delta_2
\end{aligned}$$

where the second inequality is because $\mathbb{P}(\mathcal{A}(\mathcal{D}') \in \mathcal{S}) \geq (1-\delta_2)\mathbb{P}(\mathcal{A}(\mathcal{D}') \in \mathcal{S}|E)$ from the total probability decomposition. □

Definition A.2 (Renyi Divergence)**.** The α-Renyi divergence between distributions P and Q is defined as:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim P}\left[\left(\frac{dP(x)}{dQ(x)}\right)^{\alpha-1}\right],$$

For special values $\alpha = 1, \infty$, the α-Renyi divergence is defined by taking the corresponding limit of the right-hand side.

Lemma A.3 (α-Renyi Divergence between two Multivariate Gaussians)**.** *For two m-dimensional distributions $P = N(\mu_1, \Sigma_1)$, $Q = N(\mu_2, \Sigma_2)$,*

$$D_\alpha(P\|Q) = \begin{cases} \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - m\right), & \text{if } \alpha = 1 \\ \frac{1}{\alpha-1}\log\left(\frac{|\Sigma_2|^{\alpha/2}|\Sigma_1|^{(1-\alpha)/2}}{|(1-\alpha)\Sigma_1+\alpha\Sigma_2|^{1/2}}\right) + \frac{\alpha}{2}(\mu_1-\mu_2)^\top((1-\alpha)\Sigma_1+\alpha\Sigma_2)^{-1}(\mu_1-\mu_2), & \text{else} \end{cases}.$$

Proof. We prove the case $\alpha \neq 1$ by direct calculation from the definition, and then obtain the $\alpha = 1$ case by taking a limit. First, denote the densities of P and Q as $p(x)$ and $q(x)$ respectively:

$$p(x) = \frac{1}{(2\pi)^{m/2}|\Sigma_1|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\Big), \quad q(x) = \frac{1}{(2\pi)^{m/2}|\Sigma_2|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2)\Big).$$

For $\alpha \in (0,1)\cup(1,\infty)$, we have

$$\begin{aligned}
p(x)^\alpha q(x)^{1-\alpha} &= (2\pi)^{-m/2}|\Sigma_1|^{-\alpha/2}|\Sigma_2|^{-(1-\alpha)/2} \\
&\quad \times \exp\Big(-\frac{1}{2}\big[\alpha(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + (1-\alpha)(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2)\big]\Big).
\end{aligned}$$

In the exponent, we expand the quadratic form in x:

$$\begin{aligned}
&\alpha(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + (1-\alpha)(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2) \\
&= x^\top \underbrace{\left(\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1}\right)}_{=:A} x - 2\underbrace{\left(\alpha\Sigma_1^{-1}\mu_1 + (1-\alpha)\Sigma_2^{-1}\mu_2\right)}_{=:b}^\top x + \alpha\mu_1^\top\Sigma_1^{-1}\mu_1 + (1-\alpha)\mu_2^\top\Sigma_2^{-1}\mu_2.
\end{aligned}$$

Let

$$A := \alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1}, \qquad b := \alpha\Sigma_1^{-1}\mu_1 + (1-\alpha)\Sigma_2^{-1}\mu_2.$$

Then

$$x^\top Ax - 2b^\top x = (x - A^{-1}b)^\top A(x - A^{-1}b) - b^\top A^{-1}b.$$

Hence

$$p(x)^\alpha q(x)^{1-\alpha} = (2\pi)^{-m/2}|\Sigma_1|^{-\alpha/2}|\Sigma_2|^{-(1-\alpha)/2}\exp\Big(-\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b)\Big)\exp\Big(-\frac{1}{2}\big[c_0 - b^\top A^{-1}b\big]\Big),$$

where we set

$$c_0 := \alpha\mu_1^\top\Sigma_1^{-1}\mu_1 + (1-\alpha)\mu_2^\top\Sigma_2^{-1}\mu_2.$$

Note that the integral of the centered Gaussian is standard:

$$\int_{\mathbb{R}^m}(2\pi)^{-m/2}\exp\Big(-\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b)\Big)\,dx = |A|^{-1/2}.$$

Therefore

$$\int p(x)^\alpha q(x)^{1-\alpha}\,dx = |\Sigma_1|^{-\alpha/2}|\Sigma_2|^{-(1-\alpha)/2}|A|^{-1/2}\exp\Big(-\frac{1}{2}\big[c_0 - b^\top A^{-1}b\big]\Big).$$

Define $\Sigma_\alpha := (1-\alpha)\Sigma_1 + \alpha\Sigma_2$. Two standard (but checkable) matrix identities hold for positive definite Σ_1, Σ_2 and scalar α:

$$|A| = |\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1}| = |\Sigma_1|^{-\alpha}|\Sigma_2|^{-(1-\alpha)}|\Sigma_\alpha|^{-1}, \tag{5}$$

$$c_0 - b^\top A^{-1}b = \alpha(1-\alpha)(\mu_1 - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_1 - \mu_2). \tag{6}$$

These can be verified by expanding both sides or by diagonalizing in a basis where the matrices commute; they are standard in formulas for Gaussian Chernoff/Renyi divergences.

Substituting (5) and (6) into the integral, we get

$$\begin{aligned}\int p(x)^\alpha q(x)^{1-\alpha}\,dx &= |\Sigma_1|^{-\alpha/2}|\Sigma_2|^{-(1-\alpha)/2}\big(|\Sigma_1|^{-\alpha}|\Sigma_2|^{-(1-\alpha)}|\Sigma_\alpha|^{-1}\big)^{-1/2}\\ &\quad\times\exp\Big(-\frac{1}{2}\alpha(1-\alpha)(\mu_1 - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_1 - \mu_2)\Big)\\ &= \frac{|\Sigma_2|^{\alpha/2}|\Sigma_1|^{(1-\alpha)/2}}{|\Sigma_\alpha|^{1/2}}\exp\Big(-\frac{\alpha(1-\alpha)}{2}(\mu_1 - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_1 - \mu_2)\Big).\end{aligned}$$

Finally, plug into the definition, by Definition A.2,

$$\begin{aligned}D_\alpha(P\|Q) &= \frac{1}{\alpha-1}\log\int p(x)^\alpha q(x)^{1-\alpha}\,dx\\ &= \frac{1}{\alpha-1}\log\left(\frac{|\Sigma_2|^{\alpha/2}|\Sigma_1|^{(1-\alpha)/2}}{|\Sigma_\alpha|^{1/2}}\right) + \frac{1}{\alpha-1}\log\exp\Big(-\frac{\alpha(1-\alpha)}{2}(\mu_1 - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_1 - \mu_2)\Big)\\ &= \frac{1}{\alpha-1}\log\left(\frac{|\Sigma_2|^{\alpha/2}|\Sigma_1|^{(1-\alpha)/2}}{|\Sigma_\alpha|^{1/2}}\right) + \frac{\alpha}{2}(\mu_1 - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_1 - \mu_2),\end{aligned}$$

which is exactly the expression stated in the lemma for $\alpha \neq 1$.

Specially, when $\alpha = 1$, the Kullback-Leibler divergence between two Gaussians is known to be

$$\mathrm{KL}\big(\mathcal{N}(\mu_1,\Sigma_1)\,\|\,\mathcal{N}(\mu_2,\Sigma_2)\big) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top\Sigma_2^{-1}(\mu_2 - \mu_1) - m\right).$$

Since $D_\alpha(P\|Q)$ is continuous in α and Renyi divergence converges to KL divergence as $\alpha \to 1$, taking the limit $\alpha \to 1$ in the above expression gives the $\alpha = 1$ line in the statement. This completes the proof. □

Lemma A.4 (Property of CorrDP Privacy Loss). *Define* $\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \log \mathbb{E}_{\theta \sim \mathcal{M}(\mathcal{D})} \left(\frac{\mathbb{P}[\mathcal{M}(\mathcal{D})=\theta]}{\mathbb{P}[\mathcal{M}(\mathcal{D}')=\theta]} \right)^{\lambda}$. *Then:*

- ***Composability****: Suppose* $\mathcal{M}$ *consists of a sequence of* K *adaptive mechanisms* $\mathcal{M}_1, \ldots, \mathcal{M}_K$. *Then:*

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{i=1}^{K} \alpha_{\mathcal{M}_i}(\lambda).$$

- ***Tail bound****: For any* $\epsilon > 0$, *the mechanism* $\mathcal{M}$ *is* $(\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}, \delta)$*-differentially private for,*

$$\delta \geq \exp\left(\alpha_{\mathcal{M}}(\lambda) - \frac{\lambda}{d(\mathcal{D}, \mathcal{D}')} \epsilon \right),$$

for some $\lambda \geq 0$.

Lemma A.4 follows immediately from Theorem 2 in Abadi et al. (2016) by replacing ϵ there with $\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}$.

Lemma A.5. *Consider any algorithm* $\mathcal{A}$ *that is* (ϵ, δ)*-CorrDP on a database* $\mathcal{D}$ *with* n *entries. Then if it is executed in a subsampled dataset* $\mathcal{D}_{n_q} \subseteq \mathcal{D}$ *by selecting each data point independently with probability* $\zeta \in (0, 1]$, *i.e., the sampling fraction* ζ *and run* $\mathcal{A}$. *This subsampled version of algorithm* $\mathcal{A}$ *is* $(\zeta\epsilon, \zeta\delta)$*-CorrDP.*

Lemma A.5 follows immediately from Theorem 9 in Balle et al. (2018) by replacing ϵ there with $\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}$.

A.2.2 Proving Theorem 3.5

We now return to the proof of Theorem 3.5.

Proof. We start the case of the full-batch gradient descent with $n_q = n$.

When $n_q = n$, we prove through the moment accountant argument in Abadi et al. (2016). Consider the overall gradient mechanism $\mathcal{M}$ consisting of the privacy guarantee at each iteration $\mathcal{M}_1 \times \ldots \mathcal{M}_i \ldots \times \mathcal{M}_T$.

First, we define the privacy loss at an outcome $\theta \in \Theta$ as:

$$c(\theta; \mathcal{M}, \mathcal{D}, \mathcal{D}') := \log \frac{\mathbb{P}[\mathcal{M}(\mathcal{D}) = \theta]}{\mathbb{P}[\mathcal{M}(\mathcal{D}') = \theta]}. \tag{7}$$

For the given mechanism $\mathcal{M}$, we define the λ-th moment $\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}')$ as the log of the moment generating function evaluated at the value λ:

$$\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') := \log \mathbb{E}_{\theta \sim \mathcal{M}(\mathcal{D})}[\exp(\lambda c(\theta; \mathcal{M}, \mathcal{D}, \mathcal{D}'))]. \tag{8}$$

Due to the composability and the tail bound properties of the moment accountant (Lemma A.4), we only need to show:

$$\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') \leq \frac{\lambda\epsilon}{2Td(\mathcal{D}, \mathcal{D}')}. \tag{9}$$

Then we set λ such that $\delta \geq \exp(-\frac{\lambda\epsilon}{2d(\mathcal{D},\mathcal{D}')})$ (i.e., $\lambda = \frac{2\log(1/\delta)}{\epsilon} \geq \frac{2d(\mathcal{D},\mathcal{D}')\log(1/\delta)}{\epsilon} = \Theta(\log(1/\delta))$ for a constant ϵ) and satisfy the privacy condition.

Consider the two distributions with respect to input $\mathcal{D}$ and $\mathcal{D}'$ at each time step t. That is:

$$P := \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(\theta_t, (x_i, y_i)) + b$$

$$Q := \frac{1}{n} \sum_{i=1}^{n-1} \nabla_\theta \ell(\theta_t, (x_i, y_i)) + \frac{1}{n} \nabla_\theta \ell(\theta_t, (x'_n, y'_n)) + b.$$

We can represent P and Q by $N(\mu_{\mathcal{D}}, \text{diag}(\boldsymbol{\sigma}^2))$ and $N(\mu_{\mathcal{D}'}, \text{diag}(\boldsymbol{\sigma}^2))$, where:

$$\mu_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_z \ell(z, y_i)|_{z=\theta_t^\top x_i} x_i,\ \mu_{\mathcal{D}'} = \frac{1}{n}\sum_{i=1}^{n-1} \nabla_z \ell(z, y_i)|_{z=\theta_t^\top x_i} x_i + \frac{1}{n}\nabla_z \ell(z, y_n)|_{z=\theta_t^\top x'_n} x'_n.$$

Recall the definition of α-Renyi divergence in Definition A.2. Then $\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \lambda D_{\lambda+1}(P\|Q)$. Plugging Lemma A.3 in, we have:

$$\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \frac{\lambda(\lambda+1)}{2} \sum_{i\in[m]} \left(\frac{(\mu_{\mathcal{D}} - \mu_{\mathcal{D}'})_i}{\sigma_i}\right)^2.$$

And our goal is to let $\frac{\lambda(\lambda+1)}{2}\sum_{i\in[m]}\left(\frac{(\mu_{\mathcal{D}}-\mu_{\mathcal{D}'})_i}{\sigma_i}\right)^2 \le \frac{\lambda\epsilon}{2Td(\mathcal{D},\mathcal{D}')}$. Specifically, we allocate the privacy cost of each coordinate equally, i.e., $(\frac{(\mu_{\mathcal{D}}-\mu_{\mathcal{D}'})_i}{\sigma_i})^2 \le \frac{\epsilon}{(\lambda+1)mTd(\mathcal{D},\mathcal{D}')}, \forall i \in [m]$. This is equivalent to saying $\sigma_i^2 \ge \frac{(\lambda+1)mTd(\mathcal{D},\mathcal{D}')(\mu_{\mathcal{D}}-\mu_{\mathcal{D}'})_i^2}{\epsilon^2}$. Based on Assumptions 3.3 and 3.4, we have:

$$|\mu_{\mathcal{D}} - \mu_{\mathcal{D}'}|_i \le \frac{1}{n}(C_1 L B_i \mathbf{1}_{\{x_n^{(i)} \neq x_n'^{(i)}\}} + C_2 \frac{L}{m}\sum_{j\neq i} B_j \mathbf{1}_{\{x_n^{(j)} \neq x_n'^{(j)}\}}). \tag{10}$$

From Assumption 3.3, we have: $mB_i^2 = \Theta(B^2)$. Following the definition of neighboring database in Assumption 3.1, we consider the following cases:

- When $\mathcal{D}, \mathcal{D}'$ differ in the subsets of sensitive feature: $d(\mathcal{D}, \mathcal{D}') = 1$. For the sensitive feature $i \in \mathcal{S}$, $\sigma_i^2 \ge \frac{2(\lambda+1)L^2 m B_i^2 T}{n^2\epsilon^2}$; For the insensitive feature $i \in \mathcal{U}$, $|\mu_{\mathcal{D}} - \mu_{\mathcal{D}'}|_i \le \frac{C_2 L}{nm}\sum_{j\in\mathcal{S}} B_j$, plugging it back, we have: $\sigma_i^2 \ge \frac{2L^2T(\lambda+1)B^2}{n^2\epsilon^2}(\frac{m_s}{m})^2$.
- When $\mathcal{D}, \mathcal{D}'$ differ in the subsets of insensitive feature: $d(\mathcal{D}, \mathcal{D}') \in (0, 1]$. The noise scale of the sensitive feature matches the one in standard DP in the previous case. Therefore, we only need to consider insensitive features $i \in \mathcal{U}$, $|\mu_{\mathcal{D}} - \mu_{\mathcal{D}'}|_i \le \frac{C_2 L}{mn}\sum_{j\in\mathcal{U}} B_j + \frac{C_1 B_i L}{n} \le \frac{2C_1 LB}{n}$. Plugging it back, we have: $\sigma_i^2 \ge \frac{2L^2T(\lambda+1)B^2TV(i)}{n^2\epsilon^2}$.

Therefore, the following noise choice satisfies the privacy guarantee by setting $\lambda^* = \Theta(\log(1/\delta))$ as discussed above.

$$\sigma_i^2 = \begin{cases} \frac{(\lambda^*+1)B^2L^2T}{n^2\epsilon^2}, \text{ if } i \in \mathcal{S}; \\ \frac{(\lambda^*+1)B^2L^2T\max\{TV(i), m_s^2/m^2\}}{n^2\epsilon^2}, \text{ else} \end{cases}. \tag{11}$$

Then we show that our results hold for the case when $n_q < n$. At each step, the current update with respect to the full batch in Algorithm 1 is equivalent to:

$$\theta_{t+1} = \Pi_\Theta\left(\theta_t - \frac{\alpha_t}{n}\sum_{i=1}^{n}(\nabla_\theta \ell(\theta_t; (x_i, y_i)) + b_i)\right),\ \text{where } b_i \sim N(0, n\text{diag}(\boldsymbol{\sigma}^2)).$$

If we apply this algorithm to the SGD with subsampling ratio ζ with $n_q = \zeta n$ with a randomly selected subset $\{(x_{(i)}, y_{(i)}\}_{i\in[n_q]}$, the corresponding gradient descent each time becomes:

$$\begin{aligned} &\theta_{t+1} = \Pi_\Theta\left(\theta_t - \frac{\alpha_t}{n_q}\sum_{i=1}^{n_q}(\nabla_\theta \ell(\theta_t; (x_i, y_i)) + b_i)\right),\ \text{where } b_i \sim N(0, n\text{diag}(\boldsymbol{\sigma}^2)). \\ \iff\quad &\theta_{t+1} = \Pi_\Theta\left(\theta_t - \alpha_t\left(\frac{1}{n_q}\sum_{i=1}^{n_q}(\nabla_\theta \ell(\theta_t; (x_{(i)}, y_{(i)}))) + b\right)\right),\ \text{where } b \sim N(0, \frac{1}{\zeta}\text{diag}(\boldsymbol{\sigma}^2)). \end{aligned}$$

Applying Lemma A.5, this gives $(\zeta\epsilon, \zeta\delta)$-CorrDP. From the noise terms defined in Equation (3), by changing the noise scale from $b \sim N(0, \frac{1}{\zeta}\text{diag}(\boldsymbol{\sigma}^2)$ to $b \sim N(0, \text{diag}(\boldsymbol{\sigma}^2))$ gives $(\epsilon, \zeta\delta)$-CorrDP. □

A.3 Proof of Theorem 3.7

Theorem 3.7 (Utility Guarantee of CorrDP-SGD). *Under Assumptions 3.1, 3.2, 3.3 and 3.4, for Algorithm 1 with step sizes* $\alpha_t = D/\sqrt{(L^2 + \sum_{i=1}^{m} \sigma_i^2)t}$ *and* $T = \Theta(n^2)$*, if* $F(\theta, \mathcal{D})$ *is convex, then* $R(\theta^{priv}) = \tilde{O}\left(\sqrt{(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), m_s/4\})\log(1/\delta)}/(n\epsilon)\right)$.

Before proving Theorem 3.7, we first state the following lemma, which follows immediately from Theorem 2 in Shamir & Zhang (2013).

Lemma A.6 ((Shamir & Zhang, 2013)). *Recall the stochastic gradient descent* $\theta_{t+1} = \prod_{\Theta}(\theta_t - \alpha_t G(\theta_t))$*, where* $\mathbb{E}[G(\theta_t)] = \nabla F(\theta_t, \mathcal{D})$ *and* $\mathbb{E}[\|G(\theta_t)\|_2^2] \le G^2$ *in Algorithm 1. For any* $T > 1$*, if* $\alpha_t = \frac{D}{G\sqrt{t}}$*, then:*

$$\mathbb{E}[F(\theta_T, \mathcal{D}) - F(\hat{\theta}, \mathcal{D})] \le O\left(\frac{DG\log T}{\sqrt{T}}\right) = \tilde{O}\left(\frac{DG}{\sqrt{T}}\right). \tag{12}$$

We now return to the proof of Theorem 3.7.

Proof. Lemma A.6 gives a bound on $R(\theta^{priv})$, but in order to apply the lemma, we need to compute the upper bound of $\mathbb{E}[\|G(\theta_t)\|_2^2]$ in Algorithm 1. We begin as follows:

$$\mathbb{E}[\|G(\theta_t)\|_2^2] = \mathbb{E}[\|\nabla F(\theta_t, \mathcal{D}) + b\|_2^2] = \mathbb{E}[\|\nabla F(\theta_t, \mathcal{D})\|_2^2] + \mathbb{E}[\|b\|_2^2] \le L^2 + \sum_{i\in[m]} \sigma_i^2. \tag{13}$$

Recall the definition of $\{\sigma_i\}_{i\in[m]}$ in (11), we can further bound the second term of Equation (13) by:

$$\begin{aligned}
\sum_{i\in[m]} \sigma_i^2 &\le \frac{B^2D^2T\log(1/\delta)}{n^2\epsilon^2}(m_s + \sum_{i\in\mathcal{U}} \min\{TV(i), \frac{m_s^2}{m^2}\}) \\
&\le \frac{B^2D^2T\log(1/\delta)}{n^2\epsilon^2}\left(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), \frac{(m-m_s)m_s^2}{m^2}\})\right) \\
&\le \frac{B^2D^2\log(1/\delta)}{\epsilon^2}\left(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), \frac{m_s}{4}\}\right).
\end{aligned} \tag{14}$$

Above, the second inequality follows from the fact there are $m - m_s$ insensitive features in all. The third inequality follows from $\frac{m_s(m-m_s)}{m^2} \le \frac{1}{4}$, and $T = \Theta(n^2)$.

Plugging this upper bound back into Equation (13) gives,

$$\mathbb{E}[\|G(\theta_t)\|_2^2] \le L^2 + \frac{B^2D^2\log(1/\delta)}{\epsilon^2}\left(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), \frac{m_s}{4}\}\right).$$

We can now apply Lemma A.6 with:

$$G = \sqrt{L^2 + \frac{B^2D^2\log(1/\delta)}{\epsilon^2}\left(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), \frac{m_s}{4}\}\right)} \le L + \frac{BD\sqrt{\log(1/\delta)}}{\epsilon}\sqrt{m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), \frac{m_s}{4}\}}$$

and eliminate constant B, D, C to obtain the bound in Theorem 3.7. □

A.4 Proof of Theorem 3.10

Theorem 3.10 (Lower bound for (ϵ, δ)-CorrDP algorithm). *Let* $n, m \in \mathbb{N}, \epsilon > 0$ *and* $\delta = o(1/n)$*. Consider* $\{TV(i)\}_{i\in\mathcal{U}}$ *sorted in descending order, denoted* $\{TV^{(i)}\}_{i\in\mathcal{U}}$*. For every* (ϵ, δ)*-CorrDP algorithm that outputs* θ^{priv}*, there exists a* $\mathcal{D}$ *such that with probability at least 1/3,*

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

We first introduce the following lemma.

Lemma A.7 (Lower bound for 1-way marginals). *Let $m, n \in \mathbb{N}, \epsilon > 0$ and $\delta = o(1/n)$. There is a number $M = \Omega(\min\{n, \frac{\sqrt{m}}{\epsilon}\})$ such that for every (ϵ, δ)-differential private algorithm $\mathcal{A}$, there is a dataset $\mathcal{D} = \{d_1, \ldots, d_n\} \subseteq \{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}^m$ with $\|\sum_{i=1}^n d_i\|_2 \in [M-1, M+1]$ such that, with probability at least 2/3, we have: $\|\mathcal{A}(\mathcal{D}) - q(\mathcal{D})\|_2 = \Omega(\min\{1, \frac{\sqrt{m}}{n\epsilon}\})$, where $q(\mathcal{D}) = \frac{1}{n}\sum_{i\in[n]} d_i$.*

Lemma A.7 is nearly identical to Lemma 5.1 (Part 2) in Bassily et al. (2014). The only modification is that we invoke the following Lemma A.8 to obtain the required lower bound with probability 2/3. The remainder of the proof is identical.

Lemma A.8 (Modified Corollary 3.8 of Bun et al. (2014)). *There exists $n^* = \Omega\left(\frac{\sqrt{m}}{\epsilon}\right)$ such that for every $n \le n^*$, there exists a dataset $\mathcal{D} = \{d_1, \ldots, d_n\} \subseteq \{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}^m$ such that, with probability at least 2/3, $\|\mathcal{A}(\mathcal{D}) - q(\mathcal{D})\|_2 > \frac{2}{27}$.*

We now return to the proof of Theorem 3.10.

Proof. We consider sensitive and insensitive components separately, for any (ϵ, δ)-CorrDP algorithm.

First, consider the sensitive components. Then (ϵ, δ)-CorrDP algorithm is the same as (ϵ, δ)-DP algorithm since $d(\mathcal{D}, \mathcal{D}') = 1$. For this, we construct $D_{\mathcal{S}} = \{e_1, \ldots, e_n\} \subseteq \{-\frac{1}{\sqrt{m_s}}, \frac{1}{\sqrt{m_s}}\}^{m_s}$. Applying Lemma A.7, we know there exists such $D_{\mathcal{S}}$ with $\|\sum_{i=1}^n e_i\| \in [M_1 - 1, M_1 + 1]$ and $M_1 = \Omega(\min\{n, \frac{\sqrt{m_s}}{\epsilon}\})$, such that $\|\mathcal{A}(\mathcal{D})_{\mathcal{S}} - q(\mathcal{D})_{\mathcal{S}}\|_2 = \Omega(\min\{1, \frac{\sqrt{m_s}}{n\epsilon}\})$ with probability at least 2/3.

Then, consider the insensitive components. Denote $k^* \in \operatorname{argmax}_{k\in[(m-m_s)]}\{k(TV^{(k)})^2\}$, note that for the change of the insensitive components, if we only change the insensitive component where the indice corresponding to $\{TV^{(k)}\}_{k\in[k^*]}$, the algorithm becomes $(\frac{\epsilon}{TV^{(k)}}, \delta)$-DP. For this, we construct $D_{\mathcal{U}} = \{e'_1, \ldots, e'_n\} \subseteq \{-\frac{1}{\sqrt{k^*}}, \frac{1}{\sqrt{k^*}}\}^{k^*} \cup \mathbf{0}^{m_s - k^*}$. Applying Lemma A.7, we know there exists $D_{\mathcal{U}}$ with $\|\sum_{i=1}^n e_i\| \in [M_2 - 1, M_2 + 1]$ and $M_2 = \Omega(\min\{n, \frac{\sqrt{k^*(TV^{(k^*)})^2}}{\epsilon}\})$, such that $\|\mathcal{A}(\mathcal{D})_{\mathcal{U}} - q(\mathcal{D})_{\mathcal{U}}\|_2 = \Omega(\min\{1, \frac{\sqrt{k^*(TV^{(k^*)})^2}}{n\epsilon}\})$ with probability at least 2/3.

If we set $d_i = e_i \oplus e'_i$, we have: $\|\sum_{i=1}^n d_i\|_2 \in [\sqrt{M_1^2 + M_2^2 - 2}, \sqrt{M_1^2 + M_2^2 + 2}]$. Combining the discussion of the two paragraphs above, this happens at least probability 1/3.

Set $d = (x, y)$ and consider the following convex loss function as:

$$\ell(\theta; d) = -\theta^\top d, \text{ s.t. } \|\theta\|_2 \le 1.$$

Then with probability 1/3,

$$\begin{aligned}
F(\theta^{priv}, \mathcal{D}) - F(\hat{\theta}, \mathcal{D}) &= \|\sum_{i=1}^n d_i\|_2 (1 - (\theta^{priv})^\top \hat{\theta}) \\
&\ge \frac{1}{2n}\|\sum_{i=1}^n d_i\|_2 \|\theta^{priv} - \hat{\theta}\|_2^2 \\
&= \Omega\left(\frac{\sqrt{M_1^2 + M_2^2}}{n}\right) \\
&= \Omega\left(\min\left\{1, \frac{\sqrt{m_s + \max_{k\in[(m-m_s)]}\{k(TV^{(k)})^2\}}}{n\epsilon}\right\}\right).
\end{aligned}$$

□

A.5 Proof of Theorem 4.4

Theorem 4.4 (Guarantees of CorrDP with In-Sample TV Estimation). *When the estimator in Definition 4.3 is used for the noise terms σ_i, Assumptions 4.1 and 4.2 hold, and $n = \Omega(\log(1/\delta))$, then Algorithm 1 is $(\epsilon, 2\delta)$-CorrDP and achieves the same utility guarantee as Theorem 3.7.*

Proof. Without loss of generality, we consider each $TV(i) > \frac{m_s^2}{m^2}$ for $i \in \mathcal{U}$. Otherwise, the estimation error procedure does not introduce any privacy loss since the noise scale formula is the same. Following the same notation as in the proof of Theorem 3.5, we still compare distributions P and Q. However, P and Q become $N(\mu_{\mathcal{D}}, \text{diag}(\boldsymbol{\sigma}_P^2))$ and $N(\mu_{\mathcal{D}'}, \text{diag}(\boldsymbol{\sigma}_Q^2))$ with different variance terms depending on the estimation of $\widetilde{TV}$, which depends on $\mathcal{D}$ and $\mathcal{D}'$ further. Setting $C = (\log(1/\delta) + 1)L^2$ and $T = \Theta(n^2)$ in (3), then

$$\sigma_i^2 = \frac{C T \widetilde{TV}_{\mathcal{D}}(i)}{n^2\epsilon^2} = \frac{C\widetilde{TV}_{\mathcal{D}}(i)}{\epsilon^2} = \frac{C}{\epsilon^2}(\widehat{TV}_{\mathcal{D}}(i) + 2c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}).$$

We compute the moment privacy loss:

$$\begin{aligned}\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \sum_{i=1}^{m}[\frac{\lambda+1}{2}\log\sigma_{Q,i}^2 - \frac{\lambda}{2}\log\sigma_{P,i}^2 - \frac{1}{2}\log((\lambda+1)\sigma_{Q,i}^2 - \lambda\sigma_{P,i}^2)] \\ + \frac{\lambda(\lambda+1)}{2}\sum_{i\in[m]}\frac{[(\mu_{\mathcal{D}} - \mu_{\mathcal{D}'})_i]^2}{(\lambda+1)\sigma_{Q,i}^2 - \lambda\sigma_{P,i}^2}.\end{aligned} \tag{15}$$

For the privacy guarantee, following Lemma A.1, it is equivalent to showing $\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') \le \frac{\lambda\epsilon}{Td(\mathcal{D},\mathcal{D}')}$ with probability $1-\delta$ in this case. Compared with the proof in Theorem 3.5, we only focus on the change in the noise term from $TV(i)$ to $\widehat{TV}_{\mathcal{D}}(i) + 2c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}$.

Based on Assumption 4.1, we have: $TV(i) \le \widetilde{TV}(i), \forall i \in \mathcal{U}$ with probability $1-\delta$. We then want to bound the right-hand side in the first line and second lines of (15).

We can bound the first part (i.e., the first line in (15)), through an application of Lemma A.9, stated below.

Lemma A.9 (Bounds on Variability Loss)**.** *Let λ and u be any constants. Suppose (n, v) satisfies $n \ge \frac{10C\max\{\lambda,1\}}{u}$ and $|u - v| \le \frac{C}{n}$ for some constant $C > 0$. Then*

$$(\lambda+1)\log u - \lambda\log v - \log[(\lambda+1)u - \lambda v] \le \frac{2\lambda(\lambda+1)C^2}{3n^2u^2}. \tag{16}$$

Specifically, the sensitivity of the estimand $|\widetilde{TV}_{\mathcal{D}}(i) - \widetilde{TV}_{\mathcal{D}'}(i)| = |\widehat{TV}_{\mathcal{D}}(i) - \widehat{TV}_{\mathcal{D}'}(i)| \le \frac{C_2}{n}$. For each component i, we then bound it by $\frac{2\lambda(\lambda+1)C_2^2}{3n^2\widetilde{TV}_{\mathcal{D}}^2(i)} \le \frac{2\lambda(\lambda+1)C_2^2}{3n^2TV_{\mathbb{P}^*}^2(i)} \le \frac{\lambda\epsilon}{4T}$, here the last inequality holds as long as $T \le \frac{n^2\epsilon^2 TV_{\mathbb{P}^*}(i)}{(\lambda+1)C_2^2}$, which can be satisfied when $T = \Theta(n^2)$.

For the second part (i.e., the second line in (15)), we only need to focus on the indices $i \in \mathcal{U}$. More specifically, for $|\lambda| \le C_2$, keeping other terms except TV in (3) as constant, for each component $i \in [m]$, we have:

$$\begin{aligned}&(\lambda+1)\sigma_Q^2(i) - \lambda\sigma_P^2(i) \\ &= (\lambda+1)(\widehat{TV}_{\mathcal{D}'}(i) + 2c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}) - \lambda(\widehat{TV}_{\mathcal{D}}(i) + 2c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}) \\ &\ge \widehat{TV}_{\mathcal{D}}(i) + c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma} + \lambda(\widehat{TV}_{\mathcal{D}'}(i) - \widehat{TV}_{\mathcal{D}}(i) + \frac{c_3}{n}) \\ &\ge TV(i).\end{aligned}$$

where the first inequality follows from the fact $n^{2(1-\gamma)} \ge \frac{\lambda^2 c_3^2}{c_2^2\log((m-m_s)/\delta)} = \Theta(\log(1/\delta))$. Then we can show $\frac{\lambda(\lambda+1)}{2}\sum_{i\in[m]}\frac{[(\mu_{\mathcal{D}}-\mu_{\mathcal{D}'})_i]^2}{(\lambda+1)\sigma_{Q,i}^2 - \lambda\sigma_{P,i}^2} \le \frac{\lambda(\lambda+1)}{2}\sum_{i\in[m]}\frac{[(\mu_{\mathcal{D}}-\mu_{\mathcal{D}'})_i]^2}{TV(i)^2} \le \frac{\lambda\epsilon}{2Td(\mathcal{D},\mathcal{D}')}$, where the latter inequality follows from the proof in Theorem 3.7.

Note that we do not need to consider the case $TV_{\mathbb{P}^*}(i) = 0$ for some feature $i \in \mathcal{U}$ that are independent with the sensitive features, which can be done via a preprocessing check. That is, if we find $\widehat{TV}_{\mathcal{D}}(i) \le c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}$, then with probability $1-\delta$, $TV(i) \le c_2\frac{\sqrt{\log((m-m_s)/\delta)}}{n^\gamma}, \forall i \in \mathcal{U}$.

The utility guarantees can be derived with the same rate with respect to n as in Theorem 3.7. This is because, $\frac{\sum_{i\in[m]}\sigma_i^2}{\sum_{i\in[m]}\sigma_i'^2} = \Theta(1)$ comparing the noise σ without in-sample estimation and the noise σ' with in-sample estimation. Following the convexity condition there, we can still derive the rate $\tilde{O}\left(\frac{\sqrt{m_s}}{n\epsilon}\right)$. □

A.5.1 Proof of Lemma A.9

Proof. Define $\Delta = v - u$. Therefore,

$$\begin{aligned}(\lambda+1)\log u - \lambda\log v - \log\big((\lambda+1)u - \lambda v\big) &= (\lambda+1)\log u - \lambda\log(u+\Delta) - \log(u-\lambda\Delta)\\ &= -\lambda\log\left(1+\frac{\Delta}{u}\right) - \log\left(1-\frac{\lambda\Delta}{u}\right).\end{aligned}$$

For $|x| \le 1/10$, the following inequality holds:

$$-\log(1+x) \le -x + \frac{2}{3}x^2. \tag{17}$$

Applying this bound to $x = \Delta/u$ and $x = -\lambda\Delta/u$, we obtain

$$\begin{aligned}-\lambda\log\left(1+\frac{\Delta}{u}\right) &\le -\lambda\left(\frac{\Delta}{u} - \frac{2}{3}\frac{\Delta^2}{u^2}\right) = -\lambda\frac{\Delta}{u} + \frac{2\lambda}{3}\frac{\Delta^2}{u^2},\\ -\log\left(1-\frac{\lambda\Delta}{u}\right) &\le \lambda\frac{\Delta}{u} + \frac{2\lambda^2}{3}\frac{\Delta^2}{u^2}.\end{aligned}$$

Summing the two inequalities cancels the linear terms and yields

$$-\lambda\log\left(1+\frac{\Delta}{u}\right) - \log\left(1-\frac{\lambda\Delta}{u}\right) \le \frac{2\lambda(\lambda+1)}{3}\frac{\Delta^2}{u^2}.$$

By assumption $|\Delta| \le C/n$, hence

$$\frac{\max\{\lambda,1\}|\Delta|}{u} \le \frac{\max\{\lambda,1\}C}{nu}.$$

If $n \ge \frac{10C\max\{\lambda,1\}}{u}$, then

$$\frac{|\Delta|}{u} \le \frac{1}{10}, \qquad \frac{\lambda|\Delta|}{u} \le \frac{1}{10},$$

so the inequality (17) applies. Finally, using $\Delta^2 \le C^2/n^2$ gives

$$-\lambda\log\left(1+\frac{\Delta}{u}\right) - \log\left(1-\frac{\lambda\Delta}{u}\right) \le \frac{2\lambda(\lambda+1)\Delta^2}{3u^2} \le \frac{2\lambda(\lambda+1)C^2}{3n^2u^2}.$$

□

B Additional details from Section 2

B.1 Definitions and variants of differential privacy

We provide the definitions of (global) differential privacy, semi-differential privacy and metric-differential privacy for completeness and further comparison.

Definition B.1 ((Global) Differential Privacy (Dwork, 2006))**.** A randomized algorithm $\mathcal{A}$ is (ϵ,δ)-differentially private if for all neighboring databases $\mathcal{D}, \mathcal{D}'$ and for all potential output parameters R in the output space of $\mathcal{A}$, we have:

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

Definition B.2 (Semi-Differential Privacy (Shi et al., 2021)). A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-semidifferentially private if for all neighboring databases $\mathcal{D}, \mathcal{D}'$ (that differ in some *selected features of a single example*) and for all potential output parameters R in the output space of $\mathcal{A}$, we have:

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

Definition B.3 (Metric-Differential Privacy (Andrés et al., 2013)). A randomized algorithm $\mathcal{A}$ is (ϵ, δ) metric-differentially private if for all neighboring databases $\mathcal{D}, \mathcal{D}'$, and for all potential output parameters R in the output space of $\mathcal{A}$, for some distance metric $\tilde{d}(\cdot, \cdot)$, we have:

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

Comparison between CorrDP and metric DP. Although both CorrDP and metric DP provide privacy guarantees that account for variations in data similarity, there are key differences worth highlighting. In metric DP, common distance types set in $\tilde{d}$ include Euclidean distance, Wasserstein distance, angular distance, and temporal distance, which all impose stronger privacy constraints if two elements are close naturally. However, existing literature on metric DP has not explored formalizations to capture feature correlation as CorrDP does. Under metric DP, it is possible to capture feature correlation by setting $\tilde{d}(\mathcal{D}, \mathcal{D}') = \mathbf{1}_{\{e^{\mathcal{S}} \neq (e')^{\mathcal{S}}\}} + \kappa\mathbf{1}_{\{e^{\mathcal{S}}=(e')^{\mathcal{S}}, e^{\mathcal{U}} \neq (e')^{\mathcal{U}}\}}$ for some $\kappa > 1$. However, this definition would still not account for dependencies across features without a proper value of κ. In contrast, the definition in CorrDP naturally captures feature correlation.

Comparison between CorrDP and Attribute Privacy (Zhang et al., 2022b). The attribute privacy notion in Zhang et al. (2022b) is designed to protect an entire column (attribute) across the whole database, e.g., prevent an attacker from learning the distribution of a sensitive attribute in the population. However, CorrDP is a row-level privacy notion that protects each individual's record via correlations between sensitive and insensitive features. Additionally, our work and Zhang et al. (2022b) consider different adversarial models. In the attribute privacy of Zhang et al. (2022b), leaking one individual's information is allowed as long as the column statistics remain hidden. However, in CorrDP, the influence of any single individual needs to be explicitly controlled.

Comparison between CorrDP and Label DP. As can be seen from Definition 2.2, LabelDP is a special case of CorrDP, where the labels Y are treated as sensitive features and the inputs X are insensitive components, provided X and Y are independent. However, this inclusion does not hold in the DP-ERM setting. Since ERM aims to predict Y from X, the independence assumption between X and Y does not hold. Under CorrDP, the change in insensitive input features leads to changes in the sensitive components (i.e., labels), which must be protected. In contrast, LabelDP assumes that all input features are public and do not contribute to privacy leakage, focusing solely on protecting privacy of labels.

B.2 Properties of CorrDP

Many properties of standard differential privacy are preserved under CorrDP, including post-processing, composition, and (under appropriate conditional) subsampling.

Proposition B.4 (Basic Properties of CorrDP). *Let $\mathcal{M}$ be a randomized mechanism that satisfies (ϵ, δ)-CorrDP with respect to a protected feature set $\mathcal{S}$ and a dataset-dependent correlation metric $d(\cdot, \cdot)$. Then the following properties hold.*

1. ***Post-processing.*** *For any (possibly randomized) measurable mapping f, the composed mechanism $f \circ \mathcal{M}$ also satisfies (ϵ, δ)-CorrDP with respect to the same $\mathcal{S}$ and d.*
2. ***Composition.*** *Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are independent mechanisms that satisfy (ϵ_1, δ_1)-CorrDP and (ϵ_2, δ_2)-CorrDP, respectively, with respect to the same protected feature set $\mathcal{S}$ and the same correlation metric d. Then their sequential composition $(\mathcal{M}_1, \mathcal{M}_2)$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-CorrDP.*
3. ***Subsampling.*** *Let* Sub *be a (possibly randomized) subsampling procedure, and consider the subsampled mechanism $\widetilde{\mathcal{M}}(\mathcal{D}) := \mathcal{M}(\mathsf{Sub}(\mathcal{D}))$. Suppose $\mathcal{M}$ satisfies (ϵ, δ)-CorrDP with respect to a correlation metric $d_{\text{sub}}(\cdot, \cdot)$ defined on subsampled datasets,*

$$d_{\text{sub}}(\mathsf{Sub}(\mathcal{D}), \mathsf{Sub}(\mathcal{D}')) = d(\mathcal{D}, \mathcal{D}')$$

almost surely with respect to the randomness of Sub. *Then* $\widetilde{\mathcal{M}}$ *satisfies* (ϵ, δ)-*CorrDP with respect to the same protected feature set* $\mathcal{S}$ *and the same correlation metric* d.

Proof. We prove each property directly from the definition of CorrDP.

Post-processing. Fix any measurable set E in the output space of $f \circ \mathcal{M}$. Then there exists a measurable set F in the output space of $\mathcal{M}$ such that

$$\Pr[f(\mathcal{M}(\mathcal{D})) \in E] = \Pr[\mathcal{M}(\mathcal{D}) \in F].$$

Therefore, for any pair of datasets $\mathcal{D}, \mathcal{D}'$,

$$\Pr[f(\mathcal{M}(\mathcal{D})) \in E] = \Pr[\mathcal{M}(\mathcal{D}) \in F] \leq e^{\epsilon/d(\mathcal{D},\mathcal{D}')} \Pr[\mathcal{M}(\mathcal{D}') \in F] + \delta = e^{\epsilon/d(\mathcal{D},\mathcal{D}')} \Pr[f(\mathcal{M}(\mathcal{D}')) \in E] + \delta.$$

Hence $f \circ \mathcal{M}$ satisfies (ϵ, δ)-CorrDP.

Composition. This result follows immediately from Theorem B.1 in Dwork et al. (2014) by replacing ϵ there with $\frac{\epsilon}{d(\mathcal{D},\mathcal{D}')}$.

Subsampling under correlation stability. Fix any measurable set A. Conditioning on the randomness of Sub and using that $\mathcal{M}$ is (ϵ, δ)-CorrDP with respect to d_{sub},

$$\begin{aligned}\Pr[\widetilde{\mathcal{M}}(\mathcal{D}) \in E] &= \mathbb{E}\Big[\Pr[\mathcal{M}(\mathsf{Sub}(\mathcal{D})) \in E \mid \mathsf{Sub}]\Big] \\ &\leq \mathbb{E}\Big[e^{\epsilon/d_{\text{sub}}(\mathsf{Sub}(\mathcal{D}),\mathsf{Sub}(\mathcal{D}'))} \Pr[\mathcal{M}(\mathsf{Sub}(\mathcal{D}')) \in E \mid \mathsf{Sub}] + \delta\Big].\end{aligned}$$

By the assumed stability condition, we have:

$$\begin{aligned}\Pr[\widetilde{\mathcal{M}}(\mathcal{D}) \in E] &\leq e^{\epsilon/d(\mathcal{D},\mathcal{D}')}\mathbb{E}\Big[\Pr[\mathcal{M}(\mathsf{Sub}(\mathcal{D}')) \in E \mid \mathsf{Sub}]\Big] + \delta \\ &= e^{\epsilon/d(\mathcal{D},\mathcal{D}')} \Pr[\widetilde{\mathcal{M}}(\mathcal{D}') \in E] + \delta.\end{aligned}$$

Therefore $\widetilde{\mathcal{M}}$ satisfies (ϵ, δ)-CorrDP. □

B.3 Other distance measures in CorrDP

Here we consider two other potential distance measures that could be used used in the CorrDP framework beyond TV distance, to demonstrate that other distance measures can be used as well. Any choice of distance measure should simply satisfy the axioms of Definition 2.3 and satisfy Assumptions 4.1 and 4.2.

Hellinger distance. Hellinger distance is defined as:

$$H(\mathbb{P}, \mathbb{Q}) = \frac{1}{\sqrt{2}}\sqrt{\int (\sqrt{\mathbb{P}(x)} - \sqrt{\mathbb{Q}(x)})^2 dx}.$$

Like TV distance, Hellinger distance is bounded between 0 and 1, and thus $H(\cdot, \cdot)$ can be used in place of $TV(\cdot, \cdot)$ in Definition 2.5. Furthermore, in terms of empirical estimation, similar to Assumptions 4.1 and 4.2, Hellinger distance has well-understood concentration properties, which could be utilized for convergence results (Rohde & Steinberger, 2020; Sart, 2023), and it exhibits low sensitivity to changes in individual data points, which would enable similar results to Theorem 4.4.

Wasserstein Distance. Wasserstein Distance measures the minimal cost of transforming one distribution into another, and is defined as:

$$W(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|],$$

where $\Pi(P, Q)$ is the set of joint distributions with marginals P and Q. With advantageous analytical properties such as computationally tractability and computability from finite samples, it has been used in text document similarity measurement (Li et al., 2024), domain adaptation (Rakotomamonjy & Liva, 2021), and generative adversarial networks (Liu et al., 2023). Under certain regularity conditions (e.g., bounded support), Wasserstein distance satisfies the required sensitivity and concentration assumptions.

C Additional details from Section 3

C.1 Utility guarantee of CorrDP-SGD of other losses

Under other conditions of $F(\theta, \mathcal{D})$, we can strengthen the accuracy guarantee as follows:

Corollary C.1. *Under Assumptions 3.1, 3.2, 3.3 and 3.4 (as well as $F(\theta, \mathcal{D})$ is G-smooth with respect to θ), if $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$ and we apply Algorithm 1 with $\alpha_t = \Theta(1/G)$:*

- *If $F(\theta, \mathcal{D})$ satisfies Polyak-Lojasiewicz condition with respect to θ (Karimi et al., 2016), i.e., $\|\nabla_\theta F(\theta, \mathcal{D})\|_2^2 \geq 2\mu(F(\theta, \mathcal{D}) - F(\hat{\theta}, \mathcal{D}))$ for some $\mu > 0$ for any θ, then $R(\theta^{priv}) = O\left(\frac{m_s}{n^2\epsilon^2}\right)$ if we set $\theta^{priv} = \theta_{T+1}$;*
- *For general $F(\cdot, \cdot)$, $\mathbb{E}[\|\nabla_\theta F(\theta^{priv}, \mathcal{D})\|_2^2] = \tilde{O}\left(\frac{\sqrt{m_s}}{n\epsilon}\right)$ if we set $\theta^{priv} = \theta_m$ where m is uniformly sampled from $\{1, \ldots, T\}$.*

Proof. In both cases, from Inequality (14) and $\sum_{i\in\mathcal{U}} TV(i) = \Theta(m_s)$, we have:

$$\sum_{i\in[m]} \sigma_i^2 \leq \frac{B^2D^2T\log(1/\delta)}{n^2\epsilon^2}\left(m_s + \min\{\sum_{i\in\mathcal{U}} TV(i), m_s/4\}\right) = \Theta\left(\frac{m_sT\log(1/\delta)}{n^2\epsilon^2}\right). \tag{18}$$

First if $F(\theta, \mathcal{D})$ satisfies Polyak-Lojasiewicz condition and is G-smooth, we have:

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1}, \mathcal{D}) - F(\theta_t, \mathcal{D})] &\leq \mathbb{E}\left[-\frac{1}{G}\nabla_\theta F(\theta_t, \mathcal{D})^\top(\nabla_\theta F(\theta_t, \mathcal{D}) + b) + \frac{1}{2G}\|\nabla_\theta F(\theta_t, \mathcal{D}) + b\|^2\right] \\
&= -\frac{1}{2G}\|\nabla_\theta F(\theta_t, \mathcal{D})\|^2 + \frac{1}{2G}\mathbb{E}[\|b\|^2] \\
&\leq -\frac{\mu}{G}(F(\theta_t, \mathcal{D}) - F(\hat{\theta}, \mathcal{D})) + \frac{\sum_{i\in[m]}\sigma_i^2}{2G}
\end{aligned}$$

where the first inequality follows by $F(\theta, \mathcal{D})$ is G-smooth and the stepsize $\alpha_t = \Theta(1/G)$. The second inequality follows from the Polyak-Lojasiewicz condition. Rearanging the terms and summing over $t = 1, \ldots, T$, recall $\theta^{priv} = \theta_{T+1}$, the definition of $R(\theta)$ and the bound in Inequality (14), we have:

$$R(\theta^{priv}) \leq \left(1 - \frac{\mu}{G}\right)^T R(\theta_1) + \frac{T\sum_{i\in[m]}\sigma_i^2}{2G} \leq \left(1 - \frac{\mu}{G}\right)^T R(\theta_1) + O\left(\frac{m_sT\log(1/\delta)}{n^2\epsilon^2}\right).$$

Then, setting $T = O\left(\log\left(\frac{n^2\epsilon^2}{m_s\log(1/\delta)}\right)\right)$, we have the utility guarantee:

$$R(\theta^{priv}) \leq O\left(\log^2(n)\frac{m_s\log(1/\delta)}{n^2\epsilon^2}\right) = \tilde{O}\left(\frac{m_s}{n^2\epsilon^2}\right).$$

Second, for the general case of $F(\theta, \mathcal{D})$, denote the noise imposed at the step t as b_t, we have:

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1}, \mathcal{D}) - F(\theta_t, \mathcal{D})] &\leq \mathbb{E}\left[-\frac{1}{G}\nabla_\theta F(\theta_t, \mathcal{D})^\top(\nabla_\theta F(\theta_t, \mathcal{D}) + b_t) + \frac{1}{2G}\|\nabla_\theta F(\theta_t, \mathcal{D}) + b_t\|^2\right] \\
&= -\frac{1}{2G}\|\nabla_\theta F(\theta_t, \mathcal{D})\|^2 + \frac{1}{2G}\mathbb{E}[\|b_t\|^2].
\end{aligned}$$

From this, we have:

$$\frac{1}{2G}\|\nabla_\theta F(\theta_t, \mathcal{D})\|_2^2 \leq \mathbb{E}[F(\theta_t) - F(\theta_{t+1})] + \frac{\sum_{i\in[m]}\sigma_i^2}{2G}.$$

Then $\mathbb{E}_{m,\{b_t\}_{t\in[T]}}[\|\nabla_\theta F(\theta_m, \mathcal{D})\|^2] = \frac{1}{T}\sum_{i\in[T]}\mathbb{E}_{b_i}[\|\nabla_\theta F(\theta_t, \mathcal{D})\|^2]$. Rearanging the above terms and summing over $t = 1, \ldots, T$, we have:

$$\begin{aligned}
\mathbb{E}[\|\nabla_\theta F(\theta^{priv}, \mathcal{D})\|_2^2] = \mathbb{E}_m[\|\nabla F(\theta_m, \mathcal{D})\|_2^2] &\leq \frac{2G(F(\theta_1, \mathcal{D}) - \mathbb{E}[F(\theta_{T+1}, \mathcal{D})])}{T} + \sum_{i\in[m]}\sigma_i^2 \\
&\leq \frac{2GR(\theta_1)}{T} + \sum_{i\in[m]}\sigma_i^2 \leq \frac{2GR(\theta_1)}{T} + O\left(\frac{m_sT\log(1/\delta)}{n^2\epsilon^2}\right),
\end{aligned}$$

where the second inequality follows by the fact $\mathbb{E}[F(\theta_{T+1}, \mathcal{D})] \geq F(\hat{\theta}, \mathcal{D})$, and the third inequality follows by the bound in Inequality (14). Then, setting $T = O\left(\frac{n\epsilon}{\sqrt{m_s \log(1/\delta)}}\right)$, we have the utility guarantee:

$$\mathbb{E}[\|\nabla_\theta F(\theta^{priv}, \mathcal{D})\|_2^2] = \tilde{O}\left(\frac{\sqrt{m_s \log(1/\delta)}}{n\epsilon}\right) = \tilde{O}\left(\frac{\sqrt{m_s}}{n\epsilon}\right).$$

□

C.2 Examples that Satisfy Assumptions

We demonstrate that both linear regression with Ordinary Least Square (OLS) and logistic regression satisfy these assumptions.

Example C.2 (Linear Regression, OLS)**.** Consider the squared loss $\ell(\theta; (x, y)) = (\theta^\top x - y)^2$ with bounded domain as in Assumption 3.3. If y is bounded, then Assumption 3.2 is satisfied with $L = 2\sup_x |\theta^\top x - y|\|x\|_2 \leq 2B(D + \max_y |y|) < \infty$. The sensitivity of the i-th coordinate $(\nabla_\theta \ell(\theta; (x_1, y)) - \nabla_\theta \ell(\theta; (x_2, y)))_i$ can be written as $2(y - \theta^\top x_1 + \theta_i x_2^{(i)})(x_1^{(i)} - x_2^{(i)}) + 2\sum_{j \neq i} \theta_j x_2^{(i)}(x_1^{(j)} - x_2^{(j)})$. If the boundary parameter satisfies $B|\theta_i| \leq \frac{CD}{m}, \forall i \in [m]$ for some constant $C > 0$, then Assumption 3.4 is satisfied with $C_1 = 4B + 2CD/m$ and $C_2 = 2CD$.

Example C.3 (Logistic Regression)**.** Consider the logistic loss $\ell(\theta; (x, y)) = \log(1 + e^{\theta^\top x}) - y\theta^\top x$ for $y \in \{0, 1\}$ with bounded domain as in Assumption 3.3. Define $\sigma(t) = (1 + e^{-t})^{-1}$. The sensitivity of the i-th coordinate $(\nabla_\theta \ell(\theta; (x_1, y)) - \nabla_\theta \ell(\theta; (x_2, y)))_i$ can be written as:

$$(\sigma(\theta^\top x_1) - y)(x_1^{(i)} - x_2^{(i)}) + x_2^{(i)}(\sigma(\theta^\top x_1) - \sigma(\theta^\top x_2)). \tag{19}$$

Since $0 \leq \sigma(\cdot) \leq 1$, the first term in (19) satisfies $|(\sigma(\theta^\top x_1) - y)(x_1^{(i)} - x_2^{(i)})| \leq |x_1^{(i)} - x_2^{(i)}|$; since $\sigma(\cdot)$ is 1/4-Lipschitz, the second term in (19) satisfies $x_2^{(i)}|\sigma(\theta^\top x_1) - \sigma(\theta^\top x_2)| \leq \frac{B}{4}|\theta^\top(x_1 - x_2)| \leq \frac{B}{4}\sum_{j \in [m]} |\theta_j||x_1^{(j)} - x_2^{(j)}|$. If the boundary parameter satisfies $B|\theta_i| \leq \frac{CD}{m}, \forall i \in [m]$ for some constant $C > 0$, then Assumption 3.4 is satisfied with $C_1 = 2 + CD/(4m)$ and $C_2 = CD/4$.

C.3 Additional discussions and relaxations of assumptions

In this subsection, we provide additional discussions and examples of relaxing the various modeling assumptions. For the relaxing assumptions, we consider cases where (i) the gradient and domain assumptions are violated; (ii) sensitive and insensitive features are not distinguished clearly; and (iii) label privacy.

Relaxations of Gradient and Domain Assumptions. We can relax Assumptions 3.3 and 3.4 with the following alternative:

Assumption C.4 (Relaxation of Bounded Domain and Gradients)**.** There exists some constant C, such that:

$$(\nabla_\theta \ell(\theta; (x_1, y)) - \ell(\theta; (x_2, y)))_i \leq CL\left(1_{x_1^{(i)} \neq x_2^{(i)}} + 1/m \sum_{j \neq i} 1_{x_1^{(j)} \neq x_2^{(j)}}\right), \forall i$$

This assumption does not require that both gradients and domains are bounded. Instead, it only requires that the gradient can possibly change in a bounded range. Therefore, Assumption C.4 is a variant of the bound derived in the proof of Theorem 3.5 to ensure the validity of our results.

Corollary C.5. *Under Assumptions 3.1, 3.2, C.4, for $\epsilon, \delta > 0$, Algorithm 1 is (ϵ, δ)-CorrDP.*

This corollary follows the same proof as Theorem 3.5, where the only difference is to replace Equation (10), which comes from Assumptions 3.3 and 3.4, with Assumption C.4.

Relaxations of Distinction between Sensitive and Insensitive Features. Building on our basic binary categorization for sensitive and insensitive features, it is possible to incorporate a generalization to continuous

sensitivity scores to allow different levels of protections across features. We can instead consider a sensitivity score $s_i \in [0,1]$ for each feature, with the special case of binary distinctions captured as $s_i = 0$ for each sensitive feature and $s_i = 1$ for each insensitive feature. The distance d in Definition 2.5 can be modified by using these scores as a weight vector $\{s_j\}_{j\in[m]}$ for each feature:

$$d(\mathcal{D},\mathcal{D}') = \max_{j\in[m],e^j\neq(e')^j}(1-s_j)\max_{i:i\subseteq[m]} TV(P_{X^{\mathcal{S}}|e^i}, P_{X^{\mathcal{S}}|(e')^i}) + s_j.$$

To demonstrate the key properties of this augmented definition, note that if e and e' differ in some completely sensitive feature with $s_j = 1$, then $d(D,D') = 1$, which reduces the original DP definition. On the other hand, if e and e' only differ in some completely insensitive feature with $s_j = 0$, then $d(D,D') = \max_{i:i\subseteq[m]} TV(P_{X^{\mathcal{S}}|e^i}, P_{X^{\mathcal{S}}|(e')^i})$. Otherwise, if e and e' differ in some semi-sensitive feature $s_j > 0$, the proposed d could be, e.g., a weighted combination of 1 and $\max_{i:i\subseteq[m]} TV(P_{X^{\mathcal{S}}|e^i}, P_{X^{\mathcal{S}}|(e')^i})$, where the latter is the original value of distance without the continuous relaxation. This definition could be directly incorporated into the noise terms in Algorithm 1 by changing $TV(i)$ when $i \notin \mathcal{S}$ to:

$$\sigma_i = (1-s_i)\max_{x_1,x_2\in X} TV(P_{X^{\mathcal{S}}|x_1^{\mathcal{U}}}, P_{X^{\mathcal{S}}|x_2^{\mathcal{U}}}) + s_i.$$

This change would, however, lead to an increase of the noise scale added in Algorithm 1, which would in turn weaken the utility guarantee in Theorem 3.7. However, as long as there exists some $s_i < 1$ for insensitive features, the noise would be smaller than that under standard DP-SGD and would still lead to an improved utility compared to standard DP-SGD.

Relaxations of Label Privacy. Under our current binary framework, labels must be treated as fully sensitive if we want to protect them completely. To allow a partial level of label protection, we can introduce a label-specific scale $s_l \in [0,1]$ into the distance metric d. Concretely, compared with the original d, one could use distance:

$$\tilde{d}(\mathcal{D},\mathcal{D}') = \max\{s_l \mathbb{1}_{\{\text{labels change}\}}, d(D,D')\} \in [0,1].$$

Here $s_l = 1$ recovers full DP protection on the label (so $\tilde{d} = 1$ whenever it changes), while $s_l = 0$ treats it as fully public. With this $\tilde{d}$ in hand, the CorrDP-SGD noise terms in Equation (3) simply change from $\max\{TV(i), m_s^2/m^2\}$ to $\max\{TV(i), m_s^2/m^2, s_l\}$ and the proof of Theorem 3.5 goes through identically. Note that we incorporate s_l into the noise scale of all insensitive features since changing labels affects all gradient coordinates. Then as long as s_l is smaller than 1, CorrDP still provides meaningful utility improvements compared with classical DP due to smaller noise scales.

D Additional Details in Section 4

In this appendix, we present concrete examples with two features that satisfy Assumptions 4.1 and 4.2, along with proofs that they satisfy these two assumptions. These examples are presented in Appendix D.1. In Appendix D.2, we extend to scenarios involving multiple sensitive features.

D.1 Examples satisfying Assumptions 4.1 and 4.2

In this subsection, we consider the two-feature problem with one sensitive and one insensitive feature, denoted as X^S and X^U respectively, in separate cases for when these features are discrete or continuous. We show how assumptions are satisfied under the dataset $\mathcal{D} = \{(x_k^S, x_k^U)\}_{k=1}^n$.

X^U is discrete. We discuss two cases depending on whether X^S is discrete or not.

When X^S is discrete, we apply the following empirical estimate.

Example D.1 (Estimation and Sensitivity Guarantees for Discrete Features)**.** Suppose $X^S \in \{a_i \mid i \in [K_S]\}$ and $X^U \in \{b_j \mid j \in [K_U]\}$ with joint distribution $\mathbb{P}(X^S = a_i, X^U = b_j) = p_{i,j}$. For any $j \in [K_U]$, define the conditional distribution

$$\mathbb{P}_j(i) := \mathbb{P}(X^S = a_i \mid X^U = b_j) = \frac{p_{i,j}}{\sum_{k\in[K_S]} p_{k,j}}.$$

Define the empirical counts $n_{i,j} := \sum_{k=1}^{n} \mathbb{1}_{\{x_k^S = a_i,\, x_k^U = b_j\}}$, then the plug-in estimator of $\mathbb{P}_j$ is

$$\widehat{\mathbb{P}}_j(i) := \frac{n_{i,j}}{\sum_{k\in[K_S]} n_{k,j}}, \forall i \in [K_S],$$

and the corresponding empirical estimator of the total variation distance $\forall j_1, j_2 \in [K_U]$ is

$$\widehat{TV}_{\mathcal{D}}(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) := \frac{1}{2} \sum_{i\in[K_S]} \left| \widehat{\mathbb{P}}_{j_1}(i) - \widehat{\mathbb{P}}_{j_2}(i) \right|. \tag{20}$$

Then, under $\min_{i,j} p_{i,j} > 0$, the conditional total variation distance estimator

$$\widehat{TV}_{\mathcal{D}} = \max_{j_1, j_2 \in [K_U]} \widehat{TV}_{\mathcal{D}}(\mathbb{P}_{j_1}, \mathbb{P}_{j_2})$$

satisfies Assumptions 4.1 and 4.2 with parameters $c_2 = \sqrt{\frac{\log(K_S K_U)}{\min_{i,j} p_{i,j}}}, c_3 = \frac{1}{\min_{i,j} p_{i,j}}, \gamma = \frac{1}{2}$.

Proof. First, applying the multinomial concentration inequality and a union bound over (i, j), for any $\beta \in (0, 1)$, with probability at least $1 - \beta$,

$$\left| \frac{n_{i,j}}{n} - p_{i,j} \right| \le \sqrt{\frac{\log(K_S K_U/\beta)}{n}} \quad \text{uniformly over } (i, j).$$

Since $\min_{i,j} p_{i,j} > 0$, we have $n_j = \sum_i n_{i,j} = \Theta(n)$ uniformly over j. For any $j_1, j_2 \in [K_U]$, consider $TV(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) := \frac{1}{2} \sum_{i\in[K_S]} |\mathbb{P}_{j_1}(i) - \mathbb{P}_{j_2}(i)|$. Then using the identity $\left|\frac{a}{b} - \frac{c}{d}\right| \le \frac{|a-c|}{b} + \frac{|c|}{bd}|b - d|$, and applying concentration bounds to $n_{i,j}$ and n_j, we obtain

$$\left| \widehat{\mathbb{P}}_j(i) - \mathbb{P}_j(i) \right| \le C \sqrt{\frac{\log(K_S K_U/\beta)}{n \min_{i,j} p_{i,j}}}, \forall i \in [K_S],$$

for some constant $C > 0$. Summing over i and taking a union bound over (j_1, j_2),

$$\left| \widehat{TV}_{\mathcal{D}}(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) - TV(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) \right| \le C \sqrt{\frac{\log(K_S K_U/\beta)}{n \min_{i,j} p_{i,j}}}.$$

Therefore, Assumption 4.1 is satisfied with $c_2 = \sqrt{\frac{\log(K_S K_U)}{\min_{i,j} p_{i,j}}}$ and $\gamma = \frac{1}{2}$.

For sensitivity, changing one sample affects at most one $n_{i,j}$ and one n_j by 1. Hence, $\left| \widehat{\mathbb{P}}_j(i) - \widehat{\mathbb{P}}'_j(i) \right| \le \frac{1}{n_j} \le \frac{1}{n \min_{i,j} p_{i,j}}$. Therefore, for any (j_1, j_2), $\left| \widehat{TV}_{\mathcal{D}}(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) - \widehat{TV}_{\mathcal{D}'}(\mathbb{P}_{j_1}, \mathbb{P}_{j_2}) \right| \le \frac{1}{n \min_{i,j} p_{i,j}}$.

Taking the maximum over (j_1, j_2) yields

$$\sup_{\mathcal{D}, \mathcal{D}'} \left| \widehat{TV}_{\mathcal{D}} - \widehat{TV}_{\mathcal{D}'} \right| \le \frac{1}{n \min_{i,j} p_{i,j}} = \Theta(n^{-1}).$$

Therefore, Assumption 4.2 is satisfied with $c_3 = \frac{1}{\min_{i,j} p_{i,j}}$. □

When X^S is continuous, we apply the following histogram estimate.

Example D.2 (Estimation and Sensitivity Guarantees for Histogram Smoothing)**.** Suppose $X^U \in \{b_j \mid j \in [K_U]\}$ with $\mathbb{P}(X^U = b_j) = p_j$, and $X^S \in \mathbb{R}$ satisfies $|X^S| \le B$. For each $j \in [K_U]$, let $p_j(\cdot)$ denote the conditional density of X^S given $X^U = b_j$.

We partition $[-B, B]$ into K_S bins $\{B_i\}_{i\in[K_S]}$ of equal width h. Define the empirical bin counts $n_{i,j} := \sum_{k=1}^{n} \mathbb{1}_{\{x_k^S \in B_i,\, x_k^U = b_j\}}, n_{\cdot,j} := \sum_{i\in[K_S]} n_{i,j}$. Then the histogram estimator of the conditional density is $\widehat{p}_j(x) := \frac{n_{i,j}}{n_{\cdot,j} h}, \forall x \in B_i$. The corresponding plug-in estimator of the total variation distance is

$$\widehat{TV}_{\mathcal{D}}(p_{j_1}, p_{j_2}) := \frac{1}{2} \int_{-B}^{B} |\widehat{p}_{j_1}(x) - \widehat{p}_{j_2}(x)|\, dx = \frac{1}{2} \sum_{i\in[K_S]} \left| \frac{n_{i,j_1}}{n_{\cdot,j_1}} - \frac{n_{i,j_2}}{n_{\cdot,j_2}} \right|.$$

Setting the histogram bandwidth $h = \Theta(n^{-1/3})$, then the conditional total variation estimator

$$\widehat{TV}_{\mathcal{D}} = \max_{j_1, j_2 \in [K_U]} \widehat{TV}_{\mathcal{D}}(p_{j_1}, p_{j_2})$$

satisfies Assumptions 4.1 and 4.2 with parameters $c_2 = B\sqrt{\frac{\log K_S K_U}{\min_j p_j}}, c_3 = \frac{B}{\min_j p_j}, \gamma = \frac{1}{3}$.

Proof. For any $j_1, j_2 \in [K_U]$, we decompose

$$\left|\widehat{TV}_{\mathcal{D}}(p_{j_1}, p_{j_2}) - TV(p_{j_1}, p_{j_2})\right| \le TV(p_{j_1}, \widehat{p}_{j_1}) + TV(p_{j_2}, \widehat{p}_{j_2}).$$

We bound each term using the standard bias–variance decomposition for histogram density estimators (see, e.g., as a special case of kernel density estimators in Chapter 1.2 of Tsybakov (2008)). For any $j \in [K_U]$,

$$TV(p_j, \widehat{p}_j) = \frac{1}{2}\|\widehat{p}_j - p_j\|_1 \le B\Big(h + \sqrt{\frac{\log(K_S/\beta)}{n_j h}}\Big),$$

with probability at least $1 - \beta$.

Since $\min_j p_j > 0$, we have $n_j = \Theta(n)$ uniformly over j with high probability. Taking $h = \Theta(n^{-1/3})$ yields

$$TV(p_j, \widehat{p}_j) \le B\sqrt{\frac{\log(K_S/\beta)}{n^{2/3} \min_j p_j}}.$$

Applying this bound to both j_1 and j_2, and taking a union bound over all (j_1, j_2), we obtain

$$\left|\widehat{TV}_{\mathcal{D}}(p_{j_1}, p_{j_2}) - TV(p_{j_1}, p_{j_2})\right| \le B\sqrt{\frac{\log(K_S K_U/\beta)}{n^{2/3} \min_j p_j}}.$$

Therefore, Assumption 4.1 is satisfied with $c_2 = B\sqrt{\frac{\log K_S K_U}{\min_j p_j}}, \gamma = \frac{1}{3}$.

For sensitivity, changing one sample affects at most one bin count $n_{i,j}$ by ± 1, and hence changes $\widehat{p}_j$ by at most $O(1/(n_j h))$ on a single bin. Therefore, $\|\widehat{p}_j - \widehat{p}'_j\|_1 \le \frac{C}{n_j}$, which implies

$$\left|TV(p_{j_1}, \widehat{p}_{j_1}) - TV(p_{j_1}, \widehat{p}'_{j_1})\right| \le \frac{C}{n_j} \le \frac{C}{n \min_j p_j}.$$

Applying the same argument to j_2 and taking the maximum over (j_1, j_2), we obtain

$$\sup_{\mathcal{D}, \mathcal{D}'} \left|\widehat{TV}_{\mathcal{D}} - \widehat{TV}_{\mathcal{D}'}\right| \le \frac{C}{n \min_j p_j} = \Theta(n^{-1}).$$

Therefore, Assumption 4.2 is satisfied with $c_3 = \frac{B}{\min_j p_j}$. □

X^U is continuous. When X^S is continuous, we present the example of estimation if X is jointly Gaussian distributed:

Example D.3 (Estimation and Sensitivity Guarantees for Gaussian Marginal Distributions)**.** Suppose $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Define the estimators

$$\hat{\phi} := \frac{\sum_{i=1}^n (x_i^U - \bar{x}^U)(x_i^S - \bar{x}^S)}{\sum_{i=1}^n (x_i^U - \bar{x}^U)^2}, \quad \hat{\psi} := \bar{x}^S - \hat{\phi}\bar{x}^U, \quad \hat{\eta}^2 := \frac{1}{n}\sum_{i=1}^n (x_i^S - \hat{\phi}x_i^U - \hat{\psi})^2. \tag{21}$$

The empirical conditional distribution estimator is $\widehat{P}_u := N(\hat{\phi}u + \hat{\psi}, \hat{\eta}^2)$, and the plug-in estimator of the conditional total variation distance is

$$\widehat{TV}_{\mathcal{D}}(u_1, u_2) := TV(\widehat{P}_{u_1}, \widehat{P}_{u_2}).$$

Then the conditional total variation estimator

$$\widehat{TV}_{\mathcal{D}} := \max_{1\le i,j\le n} \widehat{TV}_{\mathcal{D}}(x_i^U, x_j^U).$$

satisfies Assumptions 4.1 and 4.2 with parameters $c_2 = C\frac{\sqrt{\log n}}{\eta}, c_3 = \frac{C}{\eta}, \gamma = \frac{1}{2}$ for some constant C.

Proof. First, for any u, the conditional distribution satisfies $X^S \mid X^U = u \sim N(\phi u + \psi, \eta^2)$ for some $\phi, \psi, \eta \in \mathbb{R}$ determined by $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. For each $X^{(i)} = x_i, i \in \mathcal{U}$, given $\begin{pmatrix} X^{\mathcal{S}} \\ X^{(i)} \end{pmatrix} \sim \begin{pmatrix} \Sigma_{\mathcal{S}\mathcal{S}} & \Sigma_{\mathcal{S}i} \\ \Sigma_{i\mathcal{S}} & \Sigma_{ii} \end{pmatrix}$, then the posterior distribution $X^{\mathcal{S}} \sim N(\hat{\mu}_{\mathcal{S}}(x_i), \hat{\Sigma}_{\mathcal{S}}(x_i))$ with:

$$\hat{\mu}_{\mathcal{S}}(x_i) := \boldsymbol{\mu}_{\mathcal{S}|x_i} = \boldsymbol{\mu}_{\mathcal{S}} + \Sigma_{\mathcal{S}i}\,\Sigma_{ii}^{-1}\,(x_i - \mu_i), \quad \hat{\Sigma}_{\mathcal{S}}(x_i) := \Sigma_{\mathcal{S}\mathcal{S}} - \Sigma_{\mathcal{S}i}\,\Sigma_{ii}^{-1}\,\Sigma_{i\mathcal{S}}. \tag{22}$$

For any $u_1, u_2 \in \mathbb{R}$, we have

$$P_{u_k} = N(\phi u_k + \psi, \eta^2), \qquad \widehat{P}_{u_k} = N(\hat{\phi} u_k + \hat{\psi}, \hat{\eta}^2), \qquad k = 1, 2.$$

Since the two Gaussian distributions have the same variance within each pair,

$$TV(p_{u_1}, p_{u_2}) = 2\Phi\left(\frac{|\phi|\,|u_1 - u_2|}{2\eta}\right) - 1, \quad \widehat{TV}_{\mathcal{D}}(u_1, u_2) = 2\Phi\left(\frac{|\hat{\phi}|\,|u_1 - u_2|}{2\hat{\eta}}\right) - 1,$$

with $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. Therefore,

$$\begin{aligned}\left|\widehat{TV}_{\mathcal{D}}(u_1, u_2) - TV(p_{u_1}, p_{u_2})\right| &\le 2\sup_{t\in\mathbb{R}} \Phi'(t) \left|\frac{|\hat{\phi}|\,|u_1 - u_2|}{2\hat{\eta}} - \frac{|\phi|\,|u_1 - u_2|}{2\eta}\right| \\ &\le \frac{1}{\sqrt{2\pi}} |u_1 - u_2| \left|\frac{|\hat{\phi}|}{\hat{\eta}} - \frac{|\phi|}{\eta}\right|\end{aligned}$$

Next,

$$\left|\frac{|\hat{\phi}|}{\hat{\eta}} - \frac{|\phi|}{\eta}\right| \le \frac{1}{\hat{\eta}}\,|\hat{\phi} - \phi| + |\phi| \left|\frac{1}{\hat{\eta}} - \frac{1}{\eta}\right|.$$

By standard concentration for least-squares estimators (21), with probability at least $1 - \beta$,

$$|\hat{\phi} - \phi| \le C\sqrt{\frac{\log(1/\beta)}{n}}, \qquad |\hat{\eta} - \eta| \le C\sqrt{\frac{\log(1/\beta)}{n}}.$$

Moreover, since $x \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\max_{1\le i\le n} |x_i^U| \le C\sqrt{\log(n/\beta)}$ with probability at least $1 - \beta$, so that for $u_1, u_2 \in \{x_1^U, \ldots, x_n^U\}$,

$$|u_1 - u_2| \le C\sqrt{\log(n/\beta)}.$$

Combining these bounds yields $\left|\widehat{TV}_{\mathcal{D}}(u_1, u_2) - TV(p_{u_1}, p_{u_2})\right| \le C\frac{\sqrt{\log(n/\beta)}}{\eta\sqrt{n}}$. Therefore, Assumption 4.1 is satisfied with $\gamma = \frac{1}{2}$ and $c_2 = C\frac{\sqrt{\log n}}{\eta}$ for some constant C.

For sensitivity, changing one sample perturbs $\hat{\phi}$ and $\hat{\eta}$ by at most $O(n^{-1})$. Since Φ is Lipschitz and $\widehat{TV}_{\mathcal{D}}(u_1, u_2)$ depends smoothly on $\hat{\phi}/\hat{\eta}$, it follows that $\left|\widehat{TV}_{\mathcal{D}} - \widehat{TV}_{\mathcal{D}'}\right| \le \frac{C}{\eta} \cdot \frac{1}{n} = \Theta(n^{-1})$. Therefore, Assumption 4.2 is satisfied with $c_3 = \frac{1}{\eta}$. □

When X^S is discrete, we can run a multi-class logistic regression to estimate the likelihood $P(X^S|X_U = u)$ and calculate the TV distance based on the definition. When the parametric distribution assumption does not hold and X^S is discrete, the kernel density estimates satisfy the same rate as Example D.2.

D.2 Extensions to Multiple Sensitive Features

Our previous examples can also be extended to include multiple sensitive features. In particular, the key arguments carry over with modified dimensional dependence:

1. When X^S and X^U are both discrete, let $X^S \in \prod_{k\in[m_s]}\{a_{1,k},\dots,a_{K_{S,k},k}\}$ denote a vector of m_s categorical sensitive features. Then the joint distribution can be written as $P\big(X^S=(a_{i_1,1},\dots,a_{i_{m_s},m_s}),\, X^U=b_j\big)=p_{i_1,\dots,i_{m_s},j}$. Applying the same uniform concentration arguments as in Example D.1 over all joint categories, we obtain $c_2=\sqrt{\frac{\log\big((\prod_{k\in[m_s]}K_{S,k})\cdot K_U\big)}{\min_{i_1,\dots,i_{m_s},j}p_{i_1,\dots,i_{m_s},j}}}, \gamma=\frac{1}{2}$.

2. When X^S can be continuous and X^U is discrete, we partition the space of the entire X^S into bins $\{B_j\}$. The empirical density estimator becomes $\hat{p}_j(x)=\frac{\sum_{k\in[n]}\mathbf{1}\{x_k^S\in B_j,\,x_k^U=b_j\}}{n_j h^{m_s}}$. Then using the same bias–variance decomposition as in Example D.2, the estimation error scales as $O(h)+O\Big(\sqrt{\frac{1}{nh^{m_s}}}\Big)$. Then balancing the two terms in the estimation error with $h=\Theta(n^{-1/(m_s+2)})$ yields $\gamma=\frac{1}{m_s+2}$.

3. When (X^S,X^U) follows a joint Gaussian distribution, the conditional distribution $X^S \mid X^U$ remains Gaussian with parameters given by (22). Therefore, the same total variation bounds as in Example D.3 apply componentwise, yielding $\gamma=\frac{1}{2}$ by Proposition 2.1 of Devroye et al. (2018).

General Guideline for Multivariate Mixed Features. We can apply the kernel estimate to estimate distance when X^S is high-dimensional with both discrete and continuous features via a product kernel, i.e., a Gaussian kernel for continuous features and a discrete kernel for discrete features, which is automated in standard statistical packages (e.g., Hayfield & Racine (2008)).

In general, to estimate $TV(\mathbb{P}_{X^S|X^U=x_1},\mathbb{P}_{X^S|X^U=x_2})$ from $\mathcal{D}$ when X^U is continuous and X^S is continuous, the following methods can be used:

- Nonparametric methods: Use a kernel in X^U-space and define weights to construct weighted empirical distributions. Compute (or approximate) the TV distance of these two estimated distributions. For each $X^U=x$,
$$w_i(x)=\frac{K_h(x_i^U-x)}{\sum_{j=1}^n K_h(x_j^U-x)},$$
and estimate $\hat{p}_x(s)=\sum_{i=1}^n w_i(x_1)K_h(s-x_i^S)$. Then we can estimate the TV distance via kernel density or histogram.

- Parametric methods: Fit a parametric form for $P(X^S|X^U=x)$, e.g., Gaussian distribution in Example D.3. This way, we can compute TV distance in a closed form or a simple integral or binned manner.

High-Dimensional Estimation. When an analytical solution is unavailable or in a very high dimension, TV distance can be estimated using Monte Carlo sampling. For problems with very high dimensions, one can leverage generative modeling techniques (e.g., f-GANs and variational representations) to approximate TV distance since $TV(\mathbb{P},\mathbb{Q})=\sup_{\|f\|_\infty\le 1}\mathbb{E}_{\mathbb{P}}[f(X)]-\mathbb{E}_{\mathbb{Q}}[f(Y)]$, and f can be approximated with some parametrized f_θ as a discriminated network with clipping for bounded output. More importantly, exact estimation of TV distance is not necessary – our framework only requires an upper bound that is not overly loose, as discussed in Section 4.

E Additional details on experiments

All experiments were run on a PC laptop with Processor 8 Core(s), Apple M1 with 16GB RAM. It took around 40 hours to run all the experiments. All classification and regression problems are implemented through `scikit-learn` and `PyTorch`. We restrict the neighboring databases to allowing changes in sensitive features or one of the insensitive features.

E.1 Synthetic data

Setup. We simulate the data-generating process according $X = (X^{\mathcal{S}}, X^{\mathcal{U}})^\top \sim N(\boldsymbol{\mu}, \Sigma)$, $Y = \theta^\top X + \eta$. Here, $\eta \sim N(0, 5^2)$. Among all features in X, we set the indices of sensitive features $\mathcal{S} = [m_s]$ and the indices of insensitive features $\mathcal{U} = [m] \backslash [m_s]$. For the parameters $\boldsymbol{\mu}$ and Σ in the marginal distribution, we set

$$\mu_i = (-1)^{i+1}, i \in [m] \text{ and } \Sigma_{i,j} = \begin{cases} 1 & \text{if } i = j \in \mathcal{S} \\ 2 & \text{if } i = j \in \mathcal{U} \\ 0.5 & \text{if } i, j \in \mathcal{S}, i \neq j \\ 0.5 & \text{if } i, j \in \mathcal{U}, i \neq j \\ 0.1 & \text{if } i \in \mathcal{S}, j \in \mathcal{U} \end{cases}.$$

For each component of the true parameter $i \in [m]$, $\theta_i = (-1)^{i+1}\sqrt{i+1}$. We set $m = 100, n = 2000, m_s = 10, (m - m_s) = 90$.

Gradient Descent. We consider the ridge regression loss with:

$$\ell(\theta; (x, y)) = (y - \theta^\top x)^2 + \frac{1}{n}\|\theta\|_2^2.$$

For each private or non-private algorithm, at each step, we apply the full gradient descent to the whole dataset with a fixed stepsize $\alpha = 0.001$. We set the total iteration number $T = 4000$.

TV distance estimation. For two neighboring databases, we only allow the changes of one insensitive feature. Given the marginal distribution of both sensitive and insensitive features $X \sim N(\boldsymbol{\mu}, \Sigma)$, there are two different cases:

- For those features with indices $i\%2 = 0$, i.e., the prior mean $\mu_i = -1$;
- For those features with indices $i\%2 = 1$, i.e, the prior mean $\mu_i = 1$.

Following (22) and recall $\sigma_i^2 = C\frac{\log(1/\delta)+1}{\epsilon^2}TV(i)$. To ensure the (ϵ, δ)-CorrDP guarantee, we set $x_i = \mu_i - 2\sqrt{\log(2m_s/\delta)}, x_i' = \mu_i + 2\sqrt{\log(2m_s/\delta)}$ such that the i-th component is bounded within $[x_i, x_i']$ with probability $1 - \frac{\delta}{2}, \forall i \in [m_s]$. For the empirical estimate, use the empirical mean and variance from the dataset.

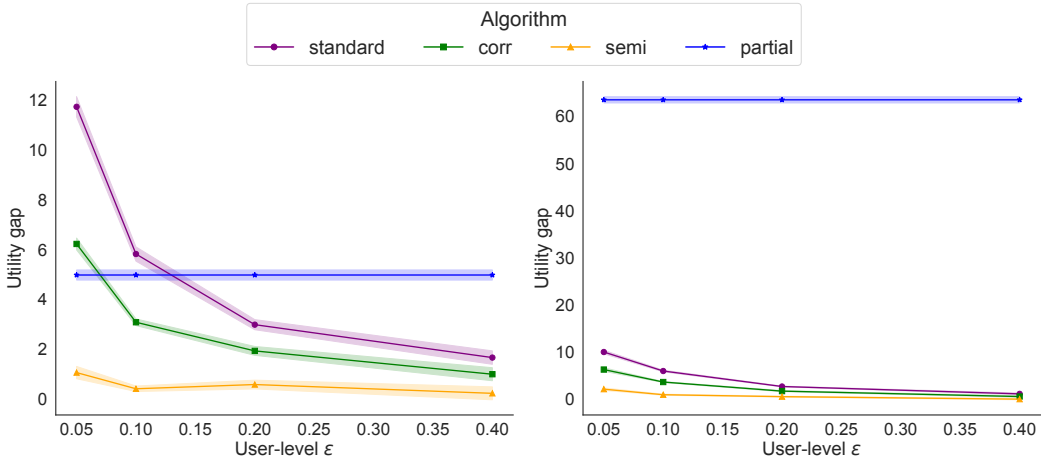
Additional Results. To show the robustness of our improvements against Standard, we vary the number of sensitive features m_s to 5, 20, 50, 80 (i.e., $0.05m, 0.2m, 0.5m, 0.8m$) respectively, with the results presented in Figure 2. We observe qualitatively similar results as in the main body (compare to Figure 1a), suggesting that the number of sensitive features does not impact the overall findings.

E.2 Adult dataset (adu)

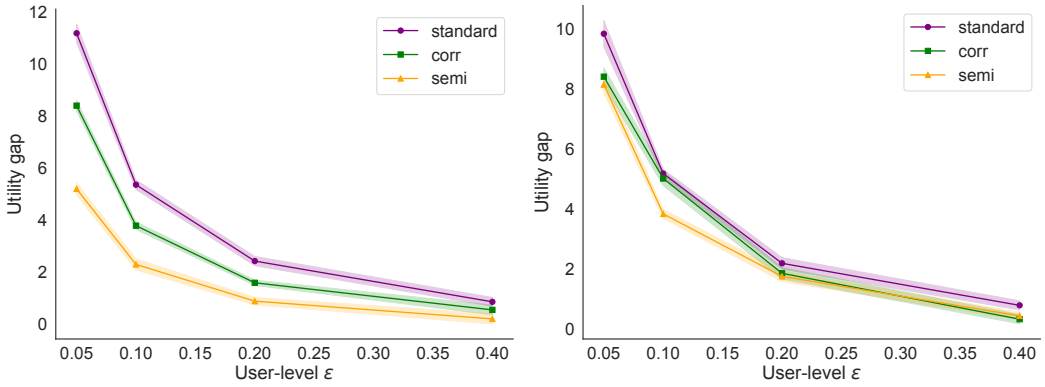
Setup. The dataset contains around 48,800 individuals in US and the standard task on this dataset is to predict whether the annual income of an individual reaches \$50k. The dataset contains 14 features. We filter the dataset and obtain 45,175 sample points with 9 features. We set race, gender, workclass, fnlwgt as insensitive features and age, educational-num, hours-per-week, martial-status, relationship as sensitive features. We categorize all the features we use as follows:

Table 1: Feature Categorization of Adult Dataset

	Sensitive	Insensitive
Continuous	age, educational-num, hours-per-week	fnlwgt
Discrete	martial-status, relationship	race, gender, workclass



(a) $m_s = 0.05m$ (b) $m_s = 0.2m$



(c) $m_s = 0.5m$ (d) $m_s = 0.8m$

Figure 2: Privacy-utility trade-offs for least-square regression under (ϵ, δ)-DP or (ϵ, δ)-CorrDP on synthetic data.

Gradient descent. We use one-hot encoding for all the discrete features and obtain 31 feature dimensions in total. We consider the standard logistic regression loss $\ell(\theta; (x, y)) = -(y \log \sigma(\theta^\top x) + (1-y)\log(1-\sigma(\theta^\top x)))$ for binary $y \in \{0, 1\}$, where $\sigma(u) = 1/(1+\exp(-u))$. For each private or nonprivate algorithm, at each step, we apply the full gradient descent to the whole dataset with a fixed stepsize $\alpha = 0.05$. We set the total iteration number $T = 5000$.

TV distance estimation. In this case for two neighboring databases, we only allow changes of one insensitive feature. This gives rise to the addition of two effects, one for continuous feature, the other for discrete features. Within the dataset, for one particular insensitive feature x_u, x'_u, we take a rough upper bound for calculating the sum of the conditional distance with respect to the sensitive continuous features $X^{\mathcal{S},c}$ and discrete features $X^{\mathcal{S},d}$:

$$\max_{x_u, x'_u} TV(\mathbb{P}_{X^{\mathcal{S}}|x_u}, \mathbb{P}_{X^{\mathcal{S}}|x'_u}) \le \max\{\max_{x_u, x'_u} TV(\mathbb{P}_{X^{\mathcal{S},c}|x_u}, \mathbb{P}_{X^{\mathcal{S},c}|x'_u}) + \max_{x_u, x'_u} TV(\mathbb{P}_{X^{\mathcal{S},d}|x_u}, \mathbb{P}_{X^{\mathcal{S},d}|x'_u}), 1\}$$

- For discrete insensitive features $u \in \{\text{race, gender, workclass}\}$:
  - To obtain the discrete distribution $\mathbb{P}_{X^{\mathcal{S},d}|x_u}$, we apply the multiclass logistic regression ('multinomial') and estimate the probability of all the value combinations of all the discrete variables. Then we calculate the TV distance of these two discrete distributions by its formula and find the maximum x_u, x'_u to obtain $TV(u)$.
  - To obtain the continuous distribution $\mathbb{P}_{X^{\mathcal{S},c}|x_u}$, we use the empirical distribution with data points with $X_u = x_u$. Then we calculate the TV distance by histogram approximation as Example D.2.
- For continuous insensitive features $u \in \{\text{fnlwgt}\}$:

– To obtain the discrete distribution $\mathbb{P}_{X^{\mathcal{S},d}|x_u}$, we apply the multiclass logistic regression ('multinomial') and estimate the probability of all the value combinations of all the discrete variables. Then we calculate the TV distance by its formula and find the maximum pair x_u, x'_u to obtain $TV(u)$.
– To obtain the continuous distribution $\mathbb{P}_{X^{\mathcal{S},c}|x_u}$, we model them as Gaussian and apply Equation (22) to estimate the posterior, with $x_u = 0, 1$ (since we do MinMaxScaler and each resulting feature lies in [0, 1]). Then we calculate the TV distance by integral following Example D.3.

Furthermore, we set the same noise for the one-hot encoding columns if they correspond to the same raw feature. In this way, we calculate the privacy scale of each insensitive feature in Table 2.

Table 2: Relative privacy noise scale for insensitive features in Adult dataset (Note that 1 means the same relative privacy noise scale as sensitive features)

Feature Name	race	gender	workclass	fnlwgt
Scale	0.23	0.90	1	0.22

Notably, workclass is highly correlated with the features educational-num, hours-per-week. Thus although the feature workclass does not belong to the class of sensitive features, it still requires strong levels of privacy protection due to it's correlation with the sensitive features.

E.3 Sepsis dataset (sep)

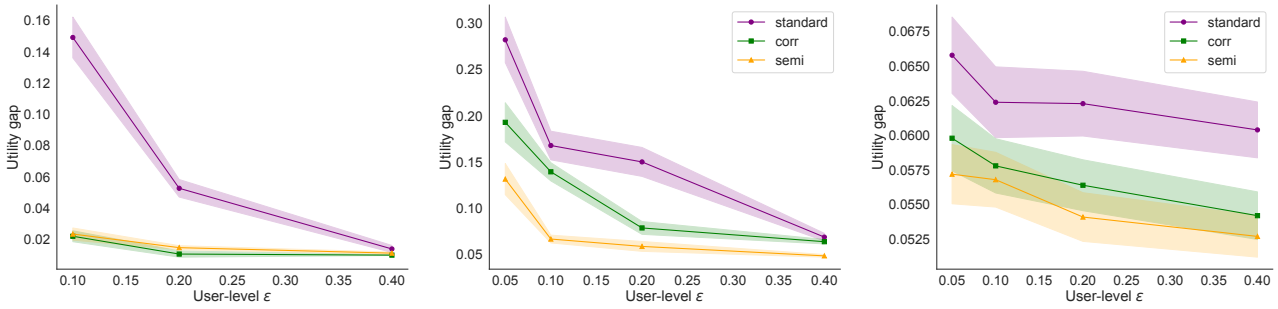
Setup. The dataset contains 110,204 patients in Norway who were diagnosed with infections. The standard goal on this dataset is to predict whether a patient survived an additional 9 days. The dataset contains 3 features, i.e., sex, episode number and age. We set sex and episode number as insensitive features and age as the sensitive feature. The original sepsis dataset is imbalanced with 92.6% data from one class ('Yes') and the remaining 7.4% from the other class ('No'). To handle the imbalance problem, we use `SMOTE` algorithms (Chawla et al., 2002) using the `imblearn` (imbalanced-learn) (LemaÃŽtre et al., 2017) Python package, which combines SMOTE-based oversampling with Edited Nearest Neighbours (ENN) cleaning. Specifically, SMOTENC generates synthetic minority samples by interpolating numerical features while assigning categorical features via nearest-neighbor voting (for features indexed by 1, 2, and 3), and ENN subsequently removes samples that are inconsistent with their local neighborhood to reduce noise. Furthermore, we randomly sample 10,000 data points in the training set for each random seed.

Gradient descent. We consider a two-layer neural network, where (1) the activation function and the dimension of the hidden layer are 'Softplus' and 5, respectively; (2) the activation function of the final layer is 'Softmax'. We set the batch size as 1024 with a constant step size of 0.005. We set the total iteration number $T = 200$, and in each iteration, we clip the gradient norm to 5. The private scale only applies to the insensitive components $\mathcal{U}$ of weight and the bias in the first layer. Denote $\hat{y} = \sigma_2(w_2^\top \sigma_1(w_1 x + b_1) + b_2)$. Then for $w_1 \in \mathbb{R}^{5\times 3}, b \in \mathbb{R}^5$, and the privacy scale only applies to the parameter estimate of $(w_1)_{i,j}, b_i, \forall i \in [5], j \in \mathcal{U}$.

TV distance estimation. For two neighboring databases, we allow changes of both insensitive features. Here both sex and episode number are category features, with 2 and 5 categories respectively. Therefore, we calculate the max TV distance between the conditional distribution of age on all 10 joint parameter combinations of sex and episode number. The calculated distance $TV_{\max} = 0.3, \forall i \in \{\text{sex}, \text{episode number}\}$ through histogram approximation.

E.4 Credit Card dataset (cre) and Medical Cost dataset (med)

Credit Card dataset. We randomly sample 30,000 points from the dataset and use logistic regression. We set sex, education, marriage, age as insensitive features and use 19 others as sensitive features, including pay, bill due, and realized. We follow TV histogram and compute $TV(\text{sex}) = 0.103, TV(\text{education}) = 0.8113, TV(\text{marriage}) = 0.8295, TV(\text{age}) = 1$. We use one-hot encoding for all the discrete features. At each step, we apply the full gradient descent to the whole dataset with a fixed stepsize $\alpha = 0.005$ and total iteration number $T = 2000$.



(a) Credit Card (Logistic Regression) (b) Medical Cost (Linear Regression) (c) Medical Cost (2-NN)

Figure 3: Privacy-utility tradeoff for logistic regression in the Credit Card dataset and linear regression / 2-NN in the Medical Cost dataset.

Medical Cost dataset. We randomly sample 1,000 points from the dataset and use linear regression or a two-layer neural network (2-NN). 2-NN is set with a 10 hidden layer with the activation function of "Softplus". We set age, sex, region as insensitive features, and bmi, children, smoker as sensitive features. After one-hot encoding, we have 8 features in total and compute $TV_{\max} = 0.36$. In the neural network, we set the clipping norm to be 10.

Empirical results on both datasets are presented in Figure 3.