Learning with Locally Private Examples by Inverse Weierstrass Private Stochastic Gradient Descent

Jean Dufraiche [1] **Paul Mangold** [2] **Michael Perrot** [1] **Marc Tommasi** [1]

Abstract

Releasing data once and for all under noninteractive Local Differential Privacy (LDP) enables complete data reusability, but the resulting noise may create bias in subsequent analyses. In this work, we leverage the Weierstrass transform to characterize this bias in binary classification. We prove that inverting this transform leads to a bias-correction method to compute unbiased estimates of nonlinear functions on examples released under LDP. We then build a novel stochastic gradient descent algorithm called Inverse Weierstrass Private SGD (IWP-SGD). It converges to the true population risk minimizer at a rate of $\mathcal{O}(1/n)$, with n the number of examples. We empirically validate IWP-SGD on binary classification tasks using synthetic and real-world datasets.

1. Introduction

Machine Learning (ML) models are increasingly deployed in domains involving sensitive data, such as healthcare, speech recognition, prediction, and forecasting. These models are vulnerable to inference attacks that allow adversaries to extract information about individual training examples (Hu et al., 2022). This has motivated the use of Differential Privacy (DP) (Dwork & Roth, 2014) as a rigorous standard to assess privacy in ML. To achieve meaningful guarantees, DP typically requires data to be centralized by a trusted curator, in charge of enforcing privacy. Unfortunately, this raises several risks: the trusted authority may fall victim to attacks that lead to major data breaches (Primoff & Kess, 2017; Lu, 2019), and data may be misappropriated by untrustworthy third parties that do not prioritize privacy.

Local Differential Privacy (LDP) (Kasiviswanathan et al., 2008; Duchi et al., 2018) addresses this challenge by requiring each data holder to privatize their data locally before release, effectively ensuring privacy without relying on a trusted curator. While this provides strong privacy guarantees, applying it in ML requires adapting the downstream learning process. Existing methods can be categorized into interactive and noninteractive approaches. In *interactive* methods, the learner adaptively queries data holders over multiple rounds, incurring a communication cost (Smith et al., 2017). In contrast, *noninteractive* methods require each user to release one or several privatized versions of their data in a single shot, eliminating the need for further interaction during learning (Zheng et al., 2017).

In practice, designing LDP mechanisms involves two considerations: whether downstream learning tasks are known, and whether data release can be adapted during learning. In some scenarios, the *downstream learning problem is known*, and task-specific algorithms can be designed to correct for LDP noise; however, this limits the potential for the data to be reused for other purposes. In contrast, many real-world scenarios *involve unknown downstream tasks* or require that data remain reusable in the long run. This motivates the use of task-agnostic, noninteractive LDP methods.

In *task-agnostic LDP*, each data holder publishes a one-time privatized representation of their data, without prior knowledge of the downstream learning task. Such a mechanism allows institutions or users to safely publish privatized datasets that remain usable for future analyses, for example, hospitals sharing medical records for research purposes. Yet, despite its generality, noninteractive and task-agnostic LDP raises a significant challenge. Indeed, learning from noisy (private) data may bias the process, as previously identified in supervised learning with noisy features (Bishop, 1995) and labels (van Rooyen & Williamson, 2018). Naively applying standard ML frameworks to privatized data may yield suboptimal models: new algorithms tailored for noninteractive and task-agnostic LDP are thus needed.

Contributions. In this paper, we develop a principled view of learning under noninteractive and task-agnostic LDP, and design new algorithms for locally private ML. We show that standard LDP mechanisms can be viewed, in expectation, as functional transforms: the Gaussian mechanism corresponds to the Weierstrass transform, while Randomized Response induces what we call the Bernoulli transform.

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France [2]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France. Correspondence to: Jean Dufraiche <jean.dufraiche@inria.fr>.

Preprint. February 19, 2026.

This perspective allows us to fully characterize the bias induced by LDP on data-dependent computations.

Crucially, *inverting these transforms* allows the design of algorithms that provably mitigate privacy-induced bias, yielding unbiased estimators for the underlying data-dependent quantities. In learning contexts, we leverage the *inverse* of these transforms to construct unbiased gradient estimators for loss functions. Applying this principle to first-order optimization, we introduce *Inverse Weierstrass Private Stochastic Gradient Descent (IWP-SGD)*. We show that IWP-SGD asymptotically recovers, in expectation over the noise, the population risk minimizer of the original, non-private problem, as the number of samples n grows to infinity. Interestingly, the convergence rate of IWP-SGD scales as $\mathcal{O}(1/n)$ similarly to classic interactive LDP approaches (Smith et al., 2017). Finally, we empirically validate IWP-SGD on binary classification tasks using synthetic and real-world datasets. To the best of our knowledge, this is the first method that asymptotically recovers the non-private population risk minimizer in a fully task-agnostic LDP setting using a single privatized release per data point.

Our contributions can be summarized as follows:

- We formalize the processing of data released under the Gaussian and Randomized Response mechanisms as transform operators of the intended computations, enabling a unified analysis of their induced bias (Section 3). This view allows us to fully characterize the bias induced by the Gaussian and Randomized Response in standard risk minimization for binary classification (Section 4).
- We construct an unbiased gradient estimator by inverting the transform associated with gradient computation on LDP records. Using this estimator, we propose Inverse Weierstrass Private SGD (IWP-SGD). We formally analyze IWP-SGD, showing that it recovers, in expectation, the solution to the original problem (Section 5).
- We empirically evaluate IWP-SGD on synthetic and real-world binary classification tasks, showing that it successfully removes the bias induced by LDP (Section 6).

1.1. Related Work

Interactive LDP Methods. Interactive methods permit adaptive communication between learners and data owners, with each owner potentially answering multiple sequential queries. A famous example is distributed SGD, where each data owner shares a noisy version of the gradient computed on their local data (Smith et al., 2017; Duchi et al., 2018). It is also the root of private SGD-based algorithms in Wang & Xu (2019) to perform sparse linear regression. As the learner explicitly queries gradient evaluations at successive model updates, these approaches do not permit data reusability and suffer from a large communication cost.

Noninteractive and Task-Specific LDP Methods. Noninteractive methods, by definition, prohibit adaptive communication: data owners release one or several privatized statistics only once. For example, Wang et al. (2019) and Zheng et al. (2017) tackle the estimation of generalized linear models under LDP using Chebychev and Bernstein polynomial approximations of the loss gradients. These approaches use multiple Gaussian-perturbed versions of each data point to construct a biased gradient estimator for which the bias shrinks with the number of noisy data releases. In Wang et al. (2018), data owners compute noisy loss evaluations over a grid of model values to estimate the population risk, which is optimized later. As a grid-based method, it suffers from an exponential dimension dependency, and loss evaluations are not reusable for other ML problems involving different losses. In contrast, our method only requires one noisy release to construct an unbiased estimator, directly reduces noise variance, while allowing reusing the data for a large class of downstream tasks.

Learning with Noisy Data. Beyond LDP, several notable works have considered settings where data points are subject to local randomization, although these studies are not directly concerned with privacy. Bishop (1995) approximates the bias induced by Gaussian noise addition in the features as an implicit regularization, while van Rooyen & Williamson (2018) models label corruption identically to the way we model the Randomized Response mechanism in Section 3.2: our bias characterization encompasses both as special cases. Scaman et al. (2024) identifies a learning bias when training and test data distributions differ in a worst-case scenario. Prior works on deconvolution methods (Fan, 1991) aim to recover the density of data from repeated noisy observations. In our work, we do not aim to estimate the noiseless distribution, but directly tackle computations performed on noisy inputs.

Noninteractive and Task-Agnostic LDP. Several instances of noninteractive and task-agnostic methods exist in the literature. Zheng et al. (2017) study a debiasing method for sparse linear regression, while Wang & Xu (2019) consider the case where only labels are private. Duchi et al. (2018) study the optimal noninteractive and task-agnostic LDP methods for mean and median estimation. Our method proposes a principled solution that generalizes these results to a broader class of learning problems under Gaussian and Randomized Response LDP mechanisms.

2. Privacy Setting and Notations

Notations. We consider a supervised learning setting with a bounded feature space $\mathcal{X} \subset \mathbb{R}^p$ and a binary label space $\mathcal{Y} = \{-1, 1\}$. Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \mathcal{Y}$, and let (x, y) be an example drawn from $\mathcal{D}$. We denote by $\|\cdot\|$

the euclidean norm and for any subset $\mathcal{Z} \subset \mathbb{R}^d$, we write $\|\mathcal{Z}\| = \sup_{z\in\mathcal{Z}} \|z\|$. The Laplacian of a twice differentiable function f is $\Delta[f] = \sum_i \partial^2_{x_i}[f]$ and its composition k times is denoted $\Delta^k[f] = (\Delta \circ \cdots \circ \Delta)[f]$.

Privacy. For the remainder of the paper, we consider a privacy setting in which data is released once and for all, without adapting the mechanism to any specific downstream task. We call it the *task-agnostic* setting. To this end, we leverage Local Differential Privacy, defined as follows.

Definition 2.1 (Local Differential Privacy (LDP) (Kasiviswanathan et al., 2008))**.** Let $\mathcal{M} : E \to F$ be a randomized algorithm. Let $\epsilon, \delta > 0$, the mechanism $\mathcal{M}$ satisfies (ϵ, δ)-LDP if, for any $z, z' \in E$ and any subset $\mathcal{C} \subseteq F$,

$$\mathbb{P}(\mathcal{M}(z) \in \mathcal{C}) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(z') \in \mathcal{C}) + \delta.$$

If $\delta = 0$, we say that $\mathcal{M}$ satisfies ϵ-LDP.

To enforce LDP, one can use the Gaussian mechanism (Dwork & Roth, 2014) for continuous variables.

Proposition 2.2 (Gaussian Mechanism)**.** *Assume a bounded subset $\mathcal{X} \subset \mathbb{R}^p$. By the Gaussian mechanism, the release of*

$$\mathcal{G}_{\epsilon,\delta}(x) = x + w, \quad w \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}_p\right),$$

with $\sigma^2 = 8\log(1.25/\delta)\left\|\mathcal{X}\right\|^2/\epsilon^2$ is (ϵ, δ)-LDP.

Similarly, one can use the Randomized Response (Kairouz et al., 2014) to enforce LDP for binary variables.

Proposition 2.3 (Randomized Response (RR))**.** *Assume $\mathcal{Y} = \{-1, 1\}$. By the Randomized Response, the release of*

$$\mathcal{B}_{\epsilon}(y) = \begin{cases} y & \text{with probability } S(\epsilon) \\ -y & \text{with probability } 1 - S(\epsilon) \end{cases},$$

where $S(\epsilon) = 1/(1 + e^{-\epsilon})$, is ϵ-LDP.

Throughout the paper, we consider the following task-agnostic (ϵ, δ)-LDP mechanism, which releases continuous features with the Gaussian mechanism, and a binary label with RR. Formally, for an example (x, y) drawn from $\mathcal{D}$,

$$(\tilde{x},\ \tilde{y}) = (\mathcal{G}_{\epsilon_x,\delta}(x),\ \mathcal{B}_{\epsilon_y}(y)) \ . \tag{1}$$

The total privacy guarantee of this mechanism is $\epsilon = \epsilon_x + \epsilon_y$, combining budgets over features and labels.

3. Privacy as a Transform

When learning from the LDP release defined in Equation (1), any data-dependent quantity, such as a loss or a gradient, can be viewed as a function h evaluated on a randomized version of the data. Ideally, one would like these quantities to be unbiased, in the sense that their expectation with respect to the privacy noise coincides with the value of h evaluated on the original data (x, y). However, this is generally not the case. Instead, local randomization of data induces a systematic transformation of the function h. We formalize this with a transform as follows:

$$\mathbb{T}_{\epsilon,\delta}[h] :\ (x,\ y) \mapsto \mathbb{E}_{(\tilde{x},\tilde{y})}[h(\tilde{x},\ \tilde{y})].$$

This operator maps the original function h to its average evaluation on noisy releases $(\tilde{x}, \tilde{y})$ of a given data point (x, y). This perspective fully captures the effect of local randomization. As a consequence, bias analysis and correction can be carried out by the study of $\mathbb{T}_{\epsilon,\delta}$ without any assumptions about the data distribution. In this section, we first study the transforms associated with the Gaussian and Randomized Response mechanisms in isolation. Then, we present how to characterize the joint effect of both mechanisms through the composition of their respective transforms.

3.1. Weierstrass Transform: a Tool for Gaussian Noise

First, we remark that applying a Gaussian noise to the inputs of an arbitrary function and considering the expectation induces a Gaussian smoothing operator known as the Generalized Weierstrass transform (Bilodeau, 1962).

Definition 3.1 (Generalized Weierstrass transform)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$. The Weierstrass transform is the function $\mathbb{W}_{\sigma^2}[f]$ defined for any $x \in \mathbb{R}^p$ and $\sigma > 0$ as

$$\mathbb{W}_{\sigma^2}[f](x) = \mathbb{E}_{w\sim\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[f(x + w)\right].$$

The alternative parameterization $\sigma^2 = 2t$ is commonly used in the literature on the Weierstrass transform. We focus on a class of sufficiently regular functions for which the Weierstrass transform admits a well-defined series representation (Bilodeau, 1962; Eddington, 1913).

Definition 3.2 (Class of Gaussian growing and slowly growing iterated Laplacians function)**.** For constants $M, a > 0$, let $\Phi_{M,a}(\mathbb{R}^p)$ denote the set of infinitely continuously differentiable functions f from $\mathbb{R}^p$ to $\mathbb{R}$ such that for any $x \in \mathbb{R}^p$

$$|f(x)| \leq M\exp(a||x||^2), \tag{2}$$
$$|\Delta^k f(x)| \leq A_x \cdot (4a)^k k!,\ \forall k \in \mathbb{N}, \tag{3}$$

for some $A_x > 0$ that depends only on x.

Note that (2) requires f to grow slower than the exponential of a quadratic function, which is a fairly mild condition. The condition (3) is met, for example, for finite linear combinations of exponentials, sines, cosines, polynomials, and band-limited functions. According to a known result in the study of the heat equation (Bilodeau, 1962; Eddington, 1913), the Weierstrass transform of functions in $\Phi_{M,a}(\mathbb{R}^p)$ admits the following series expression.

Theorem 3.3 (Series expression of $\mathbb{W}_{\sigma^2}$)**.** *Let $f \in \Phi_{M,a}(\mathbb{R}^p)$. Then, for any $\sigma^2 < 1/2a$, the generalized Weierstrass transform $\mathbb{W}_{\sigma^2}[f]$ admits the following expression*

$$\mathbb{W}_{\sigma^2}[f] = \sum_{k=0}^{\infty} \frac{\sigma^{2k}}{2^k k!} \Delta^k[f] \ .$$

Sketch of proof. The proof is given in Appendix B. First, using the analyticity of the heat equation solution, we remark that $\mathbb{W}_{\sigma^2}[f]$ is such an analytic solution. With the parameterization $\sigma^2 = 2t$, we use the Taylor expression of $t \mapsto \mathbb{W}_{2t}[f]$ around a positive $t_0 > 0$. We then take the limit as t_0 goes to zero to obtain a formal expression of $\mathbb{W}_{2t}[f]$. Given that f is in $\Phi_{M,a}(\mathbb{R}^p)$, the resulting series converges and we can identify it to $\mathbb{W}_{\sigma^2}[f]$. □

Remark 3.4. The condition $\sigma^2 < 1/2a$ is not a strong limitation since, even for fast-increasing functions, a can be very small. For example, the exponential loss $f(x) = \exp(-\theta^\top xy)$ is in $\Phi_{M_a,a}(\mathbb{R}^p)$, with $M_a = \exp(\|\Theta\|^2/4a)$, for any $(\theta, y) \in \Theta \times \mathcal{Y}$ and any arbitrary small $a > 0$ (see Appendix B.2 for more details).

3.2. Bernoulli Transform: a Tool for Binary Label Noise

We also define the analogous transform associated with the Randomized Response (RR) for an arbitrary real-valued function of binary inputs. The same transform is studied in van Rooyen & Williamson (2018). For clarity, we call it the *Bernoulli* transform, referencing the random draw of a Bernoulli variable in the RR mechanism.

Definition 3.5 (Bernoulli transform)**.** Let $g : \{-1, 1\} \to \mathbb{R}$, we define for any $\epsilon > 0$ and any $y \in \{-1, 1\}$,

$$\begin{aligned}\mathbb{B}_\epsilon[g](y) &= \mathbb{E}_{\mathcal{B}_\epsilon}\left[g(\mathcal{B}_\epsilon(y))\right] \\ &= S(\epsilon)g(y) + (1 - S(\epsilon))g(-y).\end{aligned}$$

The Bernoulli transform is the expected value of a function of a binary data point y under the Randomized Response mechanism with a given privacy budget ϵ.

3.3. Combining Weierstrass and Bernoulli Transforms

When computing a function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ of continuous and binary variables, we can compose the transforms defined for both the Gaussian and RR mechanisms as follows:

$$\begin{aligned}\mathbb{T}_{\epsilon,\delta}[h](x, y) &= \mathbb{E}_{(\tilde{x},\tilde{y})}\left[h\left(\tilde{x}, \tilde{y}\right)\right] \\ &= \mathbb{B}_{\epsilon_y}\left[z \mapsto \mathbb{W}_{\sigma^2}[h(\cdot, z)](x)\right](y).\end{aligned} \tag{4}$$

This transform accounts for the simultaneous effect of Gaussian noise on continuous variables and sign flipping on labels. It will allow to analyze the combined effects of Gaussian and Randomized Response mechanisms on the population risk in binary classification.

4. Bias in Risk Minimization

We now turn to the learning problem and study how task-agnostic LDP affects the outcome of risk minimization. In particular, we focus on the bias incurred by LDP when the goal is to learn the minimizer of the true population risk directly. This is essential as such bias is intrinsic to the problem at hand, and cannot be compensated for by increasing the number of records used for training. Depending on the settings, LDP noise may or may not change the population risk minimizer. When it does not, the original population risk minimizer can be recovered, provided that enough samples are available. In some other problems, however, the injected noise modifies the expected loss, resulting in a biased minimizer that might be far from the original model. When this occurs, increasing the sample size is not sufficient to eliminate the discrepancy between the learned solution and the true population risk minimizer.

In this section, we study the population risk obtained when losses are evaluated on task-agnostic LDP releases generated by the Gaussian and RR mechanisms. By expressing the expected noisy loss as the composition of the Weierstrass and Bernoulli transforms, we explicitly characterize when and how these mechanisms shift the population risk.

Risk Minimization. Let $\Theta \subset \mathbb{R}^k$ be a bounded convex set of model parameters. Let $\ell : \Theta \times \mathbb{R}^p \times \mathcal{Y} \to \mathbb{R}$ be a real-valued function defined for any tuple $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$ as the loss incurred when predicting an example with features x using a model θ, given that the true label is y. We can then evaluate the quality of any model $\theta \in \Theta$ with the population risk $\mathcal{R}$, defined as follows:

$$\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta, x, y)].$$

For the remainder of the paper, we consider loss functions satisfying the following regularity assumption.

Assumption 4.1 (Loss regularity)**.** The functions $x \mapsto \ell(\theta, x, y)$ and $x \mapsto \partial_{\theta_j}\ell(\theta, x, y)$ for $j \in \{1, \ldots, k\}$ are in $\Phi_{M,a}(\mathbb{R}^p)$ (see Definition 3.2) for any $(\theta, y) \in \Theta \times \mathcal{Y}$.

Bias in Noisy Risk Minimization. Let $\epsilon, \delta > 0$, for any model $\theta \in \Theta$, we define the expected population risk when the loss is evaluated on the (ϵ, δ)-LDP release of (x, y) defined in Equation (1) as follows:

$$\tilde{\mathcal{R}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\ell(\theta, \tilde{x}, \tilde{y})\right].$$

We first analyze the pointwise loss $\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\ell(\theta, \tilde{x}, \tilde{y})\right]$ for a given pair (x, y) and model θ. Recalling Equation (4) with $h(\cdot, \cdot) = \ell(\theta, \cdot, \cdot)$, we have

$$\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\ell\left(\theta,\ \tilde{x},\ \tilde{y}\right)\right] = \mathbb{T}_{\epsilon,\delta}[\ell(\theta, \cdot, \cdot)](x, y). \tag{5}$$

Developing the expression of the composed transform $\mathbb{T}_{\epsilon,\delta}$ and averaging over $(x, y) \sim \mathcal{D}$ yields a relation between the

noisy population risk $\tilde{\mathcal{R}}$ and the original population risk $\mathcal{R}$ in the following theorem (proven in Appendix C).

Theorem 4.2 (Bias induced by the Gaussian and Randomized Response mechanisms in binary classification)**.** *Let Δ_x denote the Laplacian with respect to the variable x and assume that ℓ satisfies Assumption 4.1 with $a < 1/2\sigma^2$. Recall $S(\epsilon_y) = 1/(1+e^{-\epsilon_y})$. For any $\theta \in \Theta$,*

$$\tilde{\mathcal{R}}(\theta) - \mathcal{R}(\theta) = \underbrace{(1 - S(\epsilon_y))\left(\mathbb{E}_{x,y}\left[\ell(\theta, x, -y)\right] - \mathcal{R}(\theta)\right)}_{\textit{label noise contribution}}$$
$$+ \underbrace{S(\epsilon_y) \sum_{k=1}^{\infty} \frac{\sigma^{2k}}{2^k k!} \mathbb{E}_{x,y}\left[\Delta_x^k \ell(\theta, x, y)\right]}_{\textit{feature noise contribution}}$$
$$+ \underbrace{(1 - S(\epsilon_y)) \sum_{k=1}^{\infty} \frac{\sigma^{2k}}{2^k k!} \mathbb{E}_{x,y}\left[\Delta_x^k \ell(\theta, x, -y)\right]}_{\textit{interactions of feature and label noise}}.$$

Theorem 4.2 shows that Randomized Response on labels induces a mixture between the risks associated with true and corrupted labels, while Gaussian feature noise induces a systematic smoothing of the loss through iterated Laplacians. For non-private labels ($\epsilon_y \to \infty$), our result admits the work of Bishop (1995), proposed in the more restrictive low noise regime, as a particular case. Indeed, the loss functions they consider admit a second-order Taylor approximation with respect to the features and they derive an approximate expression of the expected risk on noisy data that matches exactly ours truncated at $k = 1$. For non-private features ($\epsilon_x \to \infty$), we recover the corruption of labels in van Rooyen & Williamson (2018).

For some specific loss functions, the bias term in Theorem 4.2 has a closed-form expression. For instance, if the derivatives of ℓ with respect to the features vanish after a certain order, then $\tilde{\mathcal{R}}$ reduces to a finite sum. Even for infinitely differentiable loss functions ℓ with non-zero derivatives, we can sometimes derive a closed-form expression of the bias. This is, for example, the case for the exponential loss.

Example 1 (Exponential loss). Consider $\ell(\theta, x, y) = \exp(-\theta^\top xy)$, we have for any $\theta \in \Theta$,

$$\tilde{\mathcal{R}}(\theta) = e^{\sigma^2\|\theta\|^2/2}\left(S(\epsilon_y)\mathcal{R}(\theta) + (1 - S(\epsilon_y))\mathcal{R}(-\theta)\right).$$

Define $\tilde{\theta}^* \in \arg\min_\theta \tilde{\mathcal{R}}(\theta)$ and $\theta^* \in \arg\min_\theta \mathcal{R}(\theta)$, applying the logarithm preserves the minimum so $\tilde{\theta}^*$ also minimizes

$$\log\left(S(\epsilon_y)\mathcal{R}\left(\theta\right) + (1 - S(\epsilon_y))\mathcal{R}\left(-\theta\right)\right) + \frac{\sigma^2}{2}\left\|\theta\right\|^2.$$

The term in $\|\theta\|^2$ can be seen as further regularization induced by the feature noise, while the term $\mathcal{R}(-\theta)$ promotes predicting the wrong label in some cases due to label contamination. Both of these effects steer the optimal solution away from θ^*, creating a gap between θ^* and $\tilde{\theta}^*$. We exhibit this gap empirically in Figure 1 and 2 in Section 6.

Note that, considering $\mathcal{D}$ as an empirical distribution over a dataset of n examples leads to the same bias characterization in empirical risk minimization. Having characterized the population risk bias induced when learning from (ϵ, δ)-LDP published examples from Gaussian and RR mechanisms under the lens of the Weierstrass and Bernoulli transforms, we now turn to the natural next step of correcting it.

5. Bias Correction

In this section, we leverage the framework of privacy seen as a transform introduced in Section 3 to design a practical method that corrects the bias we identified in the previous section. To this end, we start by defining the inverse of the aforementioned transforms.

Theorem 5.1 (Inverse of $\mathbb{B}_\epsilon$ and $\mathbb{W}_{\sigma^2}$)**.** *Define $\tilde{S}(\epsilon) = 1/(1 - e^{-\epsilon})$. Let $g : \mathcal{Y} \to \mathbb{R}$ and $\epsilon > 0$, for any $\tilde{y} \in \mathcal{Y}$,*

(i) $\mathbb{B}_\epsilon^{-1}[g](\tilde{y}) = \tilde{S}(\epsilon)g(\tilde{y}) + \big(1 - \tilde{S}(\epsilon)\big)g(-\tilde{y})$.

Let $f \in \Phi_{M,a}(\mathbb{R}^p)$, for any $\sigma^2 < 1/4a$ and $\tilde{x} \in \mathbb{R}^p$

(ii) $\mathbb{W}_{\sigma^2}^{-1}[f](\tilde{x}) = \sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!} \Delta^k[f](\tilde{x})$.

Proofs of the two inverse transforms can be found in Appendix D.1. Remark that σ^2 is lower than $1/4a$ instead of $1/2a$. That is because to prove that $\mathbb{W}_{\sigma^2}^{-1}$ is the inverse of $\mathbb{W}_{\sigma^2}$, we apply a composition of their two series expression, resulting in a stronger constraint on σ^2. Computing $\mathbb{B}^{-1}[g]$ requires two evaluations of g and matches the noisy label correction in van Rooyen & Williamson (2018, Theorem 5). $\mathbb{W}^{-1}[f]$ can be computed by deriving a closed-form expression (see Section 5.2) or approximated by truncating the sum (see Appendix D.7).

The invertibility of these transforms plays a fundamental role in our study: when computing the inverse transform on noisy data, we obtain an unbiased estimate of the original function. Indeed, taking the expectation of $\mathbb{W}_{\sigma^2}^{-1}[f](\tilde{x})$ (resp. $\mathbb{B}_\epsilon^{-1}[g](\tilde{y})$) amounts to computing the Weierstrass (resp. Bernoulli) transform, which recovers the function on the original data record x (resp. y). Next, we will show that these two transforms can be applied and inverted sequentially, which will allow us to build novel, unbiased estimators of gradients of binary classification losses.

Composition of $\mathbb{B}^{-1}$ and $\mathbb{W}^{-1}$. For any function $h : \mathbb{R}^p \times \mathcal{Y} \to \mathbb{R}$, the inverse of $\mathbb{T}_{\epsilon,\delta}$ is

$$\mathbb{T}_{\epsilon,\delta}^{-1}[h](\tilde{x}, \tilde{y}) = \mathbb{B}_{\epsilon_y}^{-1}\left[z \mapsto \mathbb{W}_{\sigma^2}^{-1}[h(\cdot, z)](\tilde{x})\right](\tilde{y}).$$

Algorithm 1 Inverse Weierstrass Private SGD (IWP-SGD)

Input: Dataset $\tilde{D}_n = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$ of (ϵ, δ)-LDP released data $(x_i, y_i) \sim \mathcal{D}$ according to the mechanism of Equation (1). Initial model $\theta_0 \in \Theta$ and step size $\gamma > 0$. Loss function $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and projection Π_Θ on the bounded convex set Θ.
for $t \in \{1, \dots, n\}$ **do**
  Compute the IWP gradient $\nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}_t, \tilde{y}_t)$ (Equation (7)).
  Update $\theta_t = \Pi_\Theta(\theta_{t-1} - \gamma \nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}_t, \tilde{y}_t))$.
end for
Output: Model after the last update θ_n.

It provides a basis for the definition of the following unbiased loss estimator from (ϵ, δ)-LDP releases via Gaussian and Randomized Response mechanisms, we call the Inverse Weierstrass Private (IWP) loss estimator:

$$\tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}, \tilde{y}) = \mathbb{T}_{\epsilon,\delta}^{-1}[\ell(\theta, \cdot, \cdot)](\tilde{x}, \tilde{y}). \tag{6}$$

We differentiate it to obtain the IWP gradient estimator

$$\nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}, \tilde{y}) = \mathbb{T}_{\epsilon,\delta}^{-1}[\nabla_\theta \ell(\theta, \cdot, \cdot)](\tilde{x}, \tilde{y}), \tag{7}$$

with the convention that $\mathbb{T}$ and $\mathbb{T}^{-1}$ act component-wise on vector-valued functions such as the gradient. We show in Appendix D.2 that the IWP gradient estimator is indeed the gradient of the IWP loss estimator.

The following theorem, proven in Appendix D.3, states the unbiasedness guarantees of both $\tilde{\ell}_{\epsilon,\delta}$ and $\nabla_\theta \tilde{\ell}_{\epsilon,\delta}$.

Theorem 5.2 (Unbiasedness of IWP loss and gradient estimators)**.** *Assume ℓ satisfies Assumption 4.1 with $a < 1/4\sigma^2$. Let $\epsilon, \delta > 0$, for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\theta \in \Theta$, define $(\tilde{x}, \tilde{y})$ as a (ϵ, δ)-LDP release defined in Equation (1), the IWP loss estimator defined in Equation (6) satisfies:*

(i) $\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}, \tilde{y})\right] = \ell(\theta, x, y)$,

(ii) $\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}, \tilde{y})\right] = \nabla_\theta \ell(\theta, x, y)$.

Using the IWP gradient estimator, we introduce IWP-SGD in Algorithm 1. It is a single-pass projected SGD over a dataset $\tilde{D}_n = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$ consisting of (ϵ, δ) releases of samples (x_i, y_i) drawn i.i.d. from $\mathcal{D}$. It relies on the update $\theta_t = \Pi_\Theta\left(\theta_{t-1} + \gamma \nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta_t, \tilde{x}_t, \tilde{y}_t)\right)$ for each t in $\{1, \dots, n\}$, with $\gamma > 0$, $\theta_0 \in \Theta$ and Π_Θ the projection on Θ. Following standard convergence analyses of SGD with unbiased stochastic gradient, we bound the variance of the IWP gradient estimator in the following theorem.

Theorem 5.3 (Variance of the IWP gradient estimator)**.** *Let $\epsilon, \delta > 0$, an original feature-label pair $x, y \in \mathcal{X} \times \mathcal{Y}$ and its corresponding (ϵ, δ)-LDP release $(\tilde{x}, \tilde{y})$ defined in Equation (1). Let ℓ satisfy Assumption 4.1 with $a < 1/4\sigma^2$. Given that $\mathcal{X}$ and Θ are bounded sets, denoting*

$$C = \sup_{(\theta,x,y) \in \Theta \times \mathcal{X} \times \mathcal{Y}, s < \sigma^2} \max\Big\{ \|\nabla_\theta \ell(\theta, x, y)\|, \; \left\|\mathbb{W}_s^{-1}\left[\nabla_\theta \nabla_x \ell(\theta, \cdot, y)\right](x)\right\| \Big\}, \tag{8}$$

the variance of the IWP gradient estimator admits the following upper bound

$$\mathbb{E}\left\|\nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta, \tilde{x}, \tilde{y}) - \nabla_\theta \ell(\theta, x, y)\right\|^2 \leq C^2\left(\sigma^2 + 4\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y) - 1\right)(1 + \sigma^2)\right) . \tag{9}$$

Sketch of proof. The proof is given in Appendix D.4. We prove the result using a combination of the marginal variances from both $\tilde{x}$ and $\tilde{y}$ together with total variance law. It yields an exact variance expression that we can bound using Equation (8), relying on the increasing property of $\mathbb{W}_{\sigma^2}$ proven in Appendix B.2. □

The resulting variance bound in Theorem 5.3 distinguishes the three contributions of $\mathbb{W}^{-1}$, $\mathbb{B}^{-1}$, and the compound contribution of both in the resulting variance. It shows dependency in the feature noise variance and labels privacy budget ϵ_y. A dependency on the loss can be further quantified via the constant C. Indeed, the growth rate of C with $\|\mathcal{X}\|$ and $\|\Theta\|$ is affected by the regularity of the loss. For example, for the quadratic loss with linear models, C is of order $\mathcal{O}(p\,\|\mathcal{X}\|\,\|\Theta\|)$. Similarly, for the exponential loss with linear models, C is of order $\mathcal{O}\big(\exp\left(\epsilon_x^2\right)\left(p + \|\mathcal{X}\|\,\|\Theta\| + \sigma^2\,\|\Theta\|^2\right)\big)$ (see proofs in Appendix D.6). Now that we have bounded the variance of the IWP gradient estimator, we can analyze the IWP-SGD convergence guarantees using known results on SGD with unbiased stochastic gradients.

5.1. Convergence guarantees of IWP-SGD

We give convergence guarantees of IWP-SGD under the strong convexity and smoothness assumptions (see Appendix A for proper definitions).

Assumption 5.4 (Strong convexity and smoothness)**.** $\mathcal{R}$ is μ-strongly convex and $\mathcal{K}$-smooth, with $\mathcal{K} > 0$ and $\mu > 0$.

These assumptions are common in the analysis of SGD's convergence (Moulines & Bach, 2011; Stich, 2019) and are used solely for this purpose in this paper.

Theorem 5.5 (Convergence guarantees of IWP-SGD)**.** *Let ℓ satisfy Assumption 4.1 and be such that $\mathcal{R}$ satisfies Assumption 5.4. Let the privacy budget be $\epsilon = \epsilon_x + \epsilon_y$, $\delta > 0$ such that $\sigma^2 < 1/4a$. Denote $\theta^* = \arg\min_\theta \mathcal{R}(\theta)$. Assume $\mathcal{X}$ and Θ are bounded convex sets and let C be as*

defined in Equation (8). For any $n \in \mathbb{N}$ the number of training samples, initial model $\theta_0 \in \Theta$ and step-size $\gamma \leq \frac{1}{2\mathcal{K}}$, Algorithm 1 is (ϵ, δ)-LDP and its output θ_n satisfies

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leq (1-\gamma\mu)^n \|\theta_0 - \theta^*\|^2 + \mathcal{O}\left(\frac{\gamma C^2}{\mu}\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\frac{\log(1.25/\delta)}{\epsilon_x^2}\right).$$

In addition, for an appropriate step size $\gamma = \mathcal{O}(\log(n)/n)$,

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leq \tilde{\mathcal{O}}\left(\|\theta_0 - \theta^*\|^2 \exp\left(-\frac{\mu n}{2\mathcal{K}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{C^2}{\mu^2 n}\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\frac{\log(1.25/\delta)}{\epsilon_x^2}\right),$$

where $\tilde{\mathcal{O}}$ hides logarithmic terms in n.

Sketch of proof. The proof, given in Appendix D.8, is a direct application of Stich (2019) with our unbiased noisy gradient estimator and its variance bound given in Theorem 5.3. We also use their derivation for the appropriate step-size $\gamma = \mathcal{O}(\log(n)/n)$. □

Theorem 5.5 shows that the last iterate of IWP-SGD is converging to the population risk minimizer when the number of examples grows to infinity. The convergence rate of IWP-SGD matches the dependency on n of locally private SGD in Smith et al. (2017, Theorem 20, item 4). However, depending on the choice of loss, we may, through C, suffer from a higher dependency on the dimensions of $\mathcal{X}$ and Θ than Smith et al. (2017), which, for 1-Lipschitz, smooth and strongly-convex losses, show a linear dependency on the dimension of the model space Θ. In practice, one can obtain estimators with lower variance by considering batches of examples at each iteration. However, it does not change the fact that each example can only be used once in the optimization process, and, thus, has limited impact on the total number of records required to approximate the solution.

The general framework of IWP-SGD can be instantiated in various settings. In particular, when applied to generalized linear models, the method admits a more tractable form, which we develop in the following subsection.

5.2. Application to Generalized Linear Models

Let $\ell(\theta, x, y) = f\left(\theta^\top xy\right)$ be a loss that satisfies the generalized linear loss assumption. In this case, by Equation (6), the IWP loss estimators becomes

$$\tilde{\ell}_{\epsilon,\delta}(\theta,\ \tilde{x},\ \tilde{y}) = \tilde{S}(\epsilon_y)\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) + \left(1-\tilde{S}(\epsilon_y)\right)\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(-\theta^\top\tilde{x}\tilde{y}\right),$$

and the IWP gradient estimator can be expressed as

$$\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = \tilde{S}(\epsilon_y)\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) + \left(1-\tilde{S}(\epsilon_y)\right)\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(-\theta^\top\tilde{x}\tilde{y}\right).$$

This form involves the Weierstrass transform of the scalar valued function f instead of $\ell(\theta, \cdot, y)$, which simplifies the IWP gradient estimator expression. It further yields a closed-form expression for some losses f such as the quadratic or exponential losses.

Example 2 (Quadratic loss). Let $f\left(\theta^\top xy\right) = \frac{1}{2}(\theta^\top xy - 1)^2$ for any $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$, the IWP loss is

$$\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = \tilde{S}(\epsilon_y)f(\theta^\top\tilde{x}\tilde{y}) + \left(1-\tilde{S}(\epsilon_y)\right)f(-\theta^\top\tilde{x}\tilde{y}) - \frac{\sigma^2}{2}\|\theta\|^2,$$

with the corresponding IWP gradient estimator

$$\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = \tilde{S}(\epsilon_y)\nabla_\theta f(\theta^\top\tilde{x}\tilde{y}) + \left(1-\tilde{S}(\epsilon_y)\right)\nabla_\theta f(-\theta^\top\tilde{x}\tilde{y}) - \sigma^2\theta.$$

In that case, the IWP gradient estimator is acting like ℓ_2 regularization with the negative constant $-\sigma^2$ and requires two gradient evaluations.

Example 3 (Exponential loss). Let $f\left(\theta^\top xy\right) = e^{-\theta^\top xy}$ for any $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$, the IWP loss is

$$\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = e^{-\sigma^2\|\theta\|^2/2}\tilde{S}(\epsilon_y)f(\theta^\top\tilde{x}\tilde{y}) + e^{-\sigma^2\|\theta\|^2/2}\left(1-\tilde{S}(\epsilon_y)\right)f(-\theta^\top\tilde{x}\tilde{y}),$$

with the corresponding IWP gradient estimator

$$\begin{aligned}\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = {} & e^{-\sigma^2\|\theta\|^2/2}\tilde{S}(\epsilon_y)\nabla_\theta f(\theta^\top\tilde{x}\tilde{y}) \\ & + e^{-\sigma^2\|\theta\|^2/2}\left(1-\tilde{S}(\epsilon_y)\right)\nabla_\theta f(-\theta^\top\tilde{x}\tilde{y}) \\ & - \sigma^2 e^{-\sigma^2\|\theta\|^2/2}\tilde{S}(\epsilon_y)f(\theta^\top\tilde{x}\tilde{y})\theta \\ & - \sigma^2 e^{-\sigma^2\|\theta\|^2/2}\left(1-\tilde{S}(\epsilon_y)\right)f(-\theta^\top\tilde{x}\tilde{y})\theta.\end{aligned}$$

There, the IWP gradient estimator requires two gradient evaluations and two loss evaluations. As in the previous example, there is a similar term to ℓ_2 regularization with a negative constant of order $-\sigma^2\exp(\epsilon_y + \epsilon_x^2)$.

6. Experiments

This section empirically validates the claimed guarantees of convergence and absence of bias of Section 5. We compare three different SGD approaches: (i) SGD - real data: on the original dataset without noise, (ii) SGD - noisy data: on the (ϵ, δ)-LDP released dataset via the mechanism of

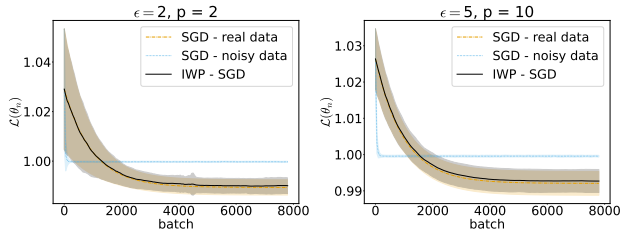


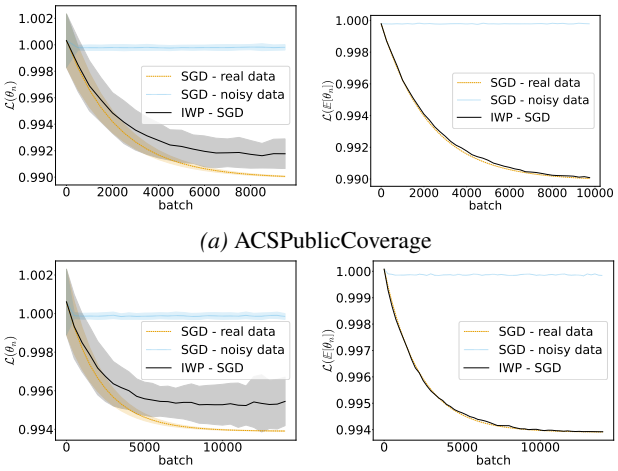
Figure 1. Comparison of SGD convergence of the exponential loss under $(2, 10^{-5})$-LDP for the 2-dimensional synthetic data and $(5, 10^{-5})$-LDP for the 10-dimensional synthetic data.

Equation (1) and (iii) IWP-SGD: Algorithm 1 on the same (ϵ, δ)-LDP released dataset. All the experiments are binary classification problems with linear models minimizing the exponential loss with ℓ_2 regularization. It is thus a strongly convex problem having a unique minimizer. We average 100 random draws of data and noise of the LDP mechanism for the synthetic data, and noise only for the real data. Across all three methods, we report the empirical risk on a test dataset defined as $\mathcal{L}(\theta) = \frac{1}{n}\sum_i \ell(\theta, x_i, y_i)$. Additional details on the experimental setup are provided in Appendix E.

Synthetic Data. We study two synthetic binary classification problems in dimension $p = 2$ and $p = 10$ generated with the `make_classification` routine of `scikit-learn` (Pedregosa et al., 2011) having features within $[-1, 1]^p$. We conduct the experiments on $n = 10^6$ samples for two privacy guarantees : $(2, 10^{-5})$-LDP for $p = 2$ and $(5, 10^{-5})$-LDP for $p = 10$.

The constant loss over the batches in Figure 1 shows models fitted via SGD on noisy data converge to a different model than models fitted via SGD on the real data. Whereas the loss of models fitted via IWP-SGD follow the one of the models fitted on the real data. It illustrates the absence of bias for IWP-SGD and its presence for SGD on noisy data.

Real Data. We study the ACSIncome and ACSPublicCoverage problems of the Folktables dataset (Ding et al., 2021). ACSIncome consists of predicting whether an individual's income is above \$50 000 and ACSPublicCoverage consists of predicting individual coverage from health insurance. For both problems, we select the three variables *AGEP* (age in years), *SEX* and *SCHL* (educational attainment). For ACSIncome we add *WKHP* (usual hours worked per week over the past year) and for ACSPublicCoverage we add *PINCP* (total annual income). We employ the Gaussian mechanism, which is suitable for continuous and ordinal variables. Although suboptimal for binary variables, we also apply it to the *SEX* attribute for consistency and practicality. We merge the data of the five largest states yielding datasets of respectively 668 859 rows and 883 984 rows for ACSIncome and ACSPublicCoverage. The data is then randomly split into training (80%) and test (20%) sets.



(a) ACSPublicCoverage

(b) ACSIncome

Figure 2. Comparison of SGD convergence of the exponential loss under $(2, 10^{-5})$-LDP on ACSPublicCoverage and ACSIncome. Left hand plots show the averaged loss of fitted model (θ_n) while right hand plots are showing the loss of the averaged model $(\mathbb{E}\theta_n)$ over random draws.

Figure 2 shows the average loss over fitted models (θ_n) and also the loss of the averaged model $(\mathbb{E}\theta_n)$ over the random draws. It allows us to distinguish the remaining excess risk resulting from bias or variance of the IWP-SGD outputs. As with synthetic data, we remark that the constant loss of models fitted via SGD on noisy data with the number of batches illustrates the presence of bias in this setting. Whereas the IWP-SGD outputs are showing a decrease to a remaining low loss close to the one on real data. We can interpret this remaining gap as a consequence of the variance of IWP-SGD. Indeed, on the second plot, the averaged output model $(\mathbb{E}\theta_n)$ for IWP-SGD is showing a null difference with the loss of the optimal model obtained through SGD.

7. Conclusion

In this paper, we characterized the bias that occurs when learning from an LDP-released dataset using Gaussian and Randomized Response mechanisms. Linking these mechanisms with transform operators, we derived an expression of the bias on the population risk under these LDP mechanisms. This view of privacy as a transform yielded the construction of a theoretically-grounded debiasing technique, which takes the form of a variant of SGD called IWP-SGD.

Our results show, theoretically and empirically, that the bias induced by the use of private noisy examples for SGD can be avoided at the cost of a higher variance of the noisy gradient estimator, illustrating a bias-variance tradeoff. This opens a pathway to study LDP through the lens of transform

operators. Extending the framework of this paper to other locally private mechanisms presents promising avenues for future exploration.

Impact Statement

This paper presents work that can help in the design of Machine Learning projects using locally private examples. It can help argue against the idea that directly learning from noisy examples in Differential Privacy is a problem that is too hard to be solved using practical algorithms. Future societal consequences might be the publication of locally private data for later use in fields where no public dataset exist.

Acknowledgments

We acknowledge the support of the French National Research Agency (ANR) through the grant ANR-23-CE23-0011-01 (Project FaCTor) and ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR.

References

Ahlfors, L. V. *Complex Analysis*. McGraw-Hill Book Company, 3 edition, 1979.

Bilodeau, G. G. The Weierstrass transform and Hermite polynomials. *Duke Mathematical Journal*, 29(2):293 – 308, 1962. doi: 10.1215/S0012-7094-62-02929-0. URL https://doi.org/10.1215/S0012-7094-62-02929-0.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Boyd, J. P. The devil's invention: Asymptotic, superasymptotic and hyperasymptotic series. *Acta Applicandae Mathematica*, 56:1–98, 1999. URL https://api.semanticscholar.org/CorpusID:3091422.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. doi: 10.1080/01621459.2017.1389735. URL https://doi.org/10.1080/01621459.2017.1389735.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Eddington, A. S. On a formula for correcting statistics for the effects of a known probable error of observation. *Monthly Notices of the Royal Astronomical Society*, 73(5):359–360, 03 1913. ISSN 0035-8711. doi: 10.1093/mnras/73.5.359. URL https://doi.org/10.1093/mnras/73.5.359.

Fan, J. On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *The Annals of Statistics*, 19(3):1257 – 1272, 1991. doi: 10.1214/aos/1176348248. URL https://doi.org/10.1214/aos/1176348248.

Fritz, J. *Partial Differential Equations*. Applied Mathematical Sciences. Springer New York, 1991. ISBN 9780387906096. URL https://books.google.fr/books?id=cBib_bsGGLYC.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), September 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL https://doi.org/10.1145/3523273.

Kairouz, P., Oh, S., and Viswanath, P. Extremal mechanisms for local differential privacy. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/c16cf23dd72c445d3050d0fcd3f28728-Paper.pdf.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 531–540, 2008. doi: 10.1109/FOCS.2008.27.

Komatsu, H. A characterization of real analytic functions. *Proceedings of the Japan Academy*, 36(3):90 – 93, 1960. doi: 10.3792/pja/1195524081. URL https://doi.org/10.3792/pja/1195524081.

Lu, J. Assessing the cost, legal fallout of capital one data breach. *Law360 Expert Analysis*, 08 2019.

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Primoff, W. and Kess, S. The equifax data breach: What cpas and firms need to know now: Certified public accountant, 12 2017. Nom - New York Times Co; eWeek; Federal Trade Commission–FTC; American Institute of Certified Public Accountants; Copyright - Copyright New York State Society of Certified Public Accountants Dec 2017.

Rudin, W. *Principles of mathematical analysis*. McGraw-Hill,, United States, 3rd edition, 1976.

Scaman, K., Even, M., Le Bars, B., and Massoulie, L. Minimax excess risk of first-order methods for statistical learning with data-dependent oracles. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3709–3717. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/scaman24a.html.

Smith, A., Thakurta, A., and Upadhyay, J. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77, 2017. doi: 10.1109/SP.2017.35.

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method, 2019. URL https://arxiv.org/abs/1907.04232.

van Rooyen, B. and Williamson, R. C. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018. URL http://jmlr.org/papers/v18/16-315.html.

Wang, D. and Xu, J. On sparse linear regression in the local differential privacy model. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6628–6637. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wang19m.html.

Wang, D., Gaboardi, M., and Xu, J. Empirical risk minimization in non-interactive local differential privacy revisited. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/13f320e7b5ead1024ac95c3b208610db-Paper.pdf.

Wang, D., Smith, A., and Xu, J. Noninteractive locally private learning of linear models via polynomial approximations. In Garivier, A. and Kale, S. (eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 898–903. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/wang19c.html.

Zheng, K., Mou, W., and Wang, L. Collect at once, use effectively: Making non-interactive locally private learning possible. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4130–4139. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/zheng17c.html.

A. Definitions and Theorems

Notations. We recall the main notations of the paper:

- Data lies in $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{Y} = \{-1, 1\}$, parameters in $\Theta \subset \mathbb{R}^k$, and $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \mathcal{Y}$.
- The loss function is $\ell : \Theta \times \mathbb{R}^p \times \mathcal{Y} \to \mathbb{R}$ and the associated risk $\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta, x, y)]$.
- We denote $\|\cdot\|$ the euclidean norm, and for any subset $\mathcal{Z} \subset \mathbb{R}^d$, we write $\|\mathcal{Z}\| = \sup_{z\in\mathcal{Z}} \|z\|$.
- For $a, b : \mathbb{N} \to \mathbb{R}^+$, we write $a = \mathcal{O}(b)$ if there exists $C > 0$ such that for all $n \in \mathbb{N}$, $a(n) \leq Cb(n)$.
- For a complex number $z = \alpha + i\beta$ with $(\alpha, \beta) \in \mathbb{R}^2$ and $i^2 = -1$, we denote $\Re(z) = \alpha$ its real part, $\Im(z) = \beta$ its imaginary part, and $|z| = \sqrt{\alpha^2 + \beta^2}$ its modulus.

Subsets of $\mathbb{R}^d$. We also recall some definitions on subsets of $\mathbb{R}^d$ for completeness.

Definition A.1 (Convex subset of $\mathbb{R}^d$)**.** A subset $\mathcal{Z} \subset \mathbb{R}^d$ is convex if for any $z, z' \in \mathcal{Z}$ and $t \in [0, 1]$, $tz + (1 - t)z' \in \mathcal{Z}$.

Definition A.2 (Open and closed subsets of $\mathbb{R}^d$)**.** A subset $\mathcal{Z} \subset \mathbb{R}^d$ is open if for any $z \in \mathcal{Z}$, there exist $\delta > 0$ such that for any $\tilde{z} \in \mathbb{R}^d$ such that $\|z - \tilde{z}\| < \delta$, $\tilde{z} \in \mathcal{Z}$. The set $\mathcal{Z}$ is said to be closed if $\{z \in \mathbb{R}^d \mid z \notin \mathcal{Z}\}$ is open.

Definition A.3 (Compact subset of $\mathbb{R}^d$)**.** A subset $\mathcal{Z} \subset \mathbb{R}^d$ is said to be compact if it is closed and bounded ($\|\mathcal{Z}\| < \infty$).

Regularity of functions. We now define classical regularity definitions for real-valued functions.

Definition A.4 (Convexity and strong convexity)**.** A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex (where $\mu > 0$) if $\forall x, y \in \mathbb{R}^p$, we have $f(x) - f(y) \geq \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2$, and convex if this holds for $\mu = 0$.

Definition A.5 (Smoothness)**.** A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is $\mathcal{K}$-smooth if $\forall x, x' \in \mathbb{R}^p$, we have $\|\nabla f(x) - \nabla f(x')\| \leq \mathcal{K}\|x - x'\|$.

Definition A.6 (Laplacian)**.** The Laplacian of a twice differentiable function f is $\Delta[f] = \sum_i \partial^2_{x_i}[f]$ and its composition k times is denoted $\Delta^k[f] = (\Delta \circ \cdots \circ \Delta)[f]$.

Analytic functions. We define real and complex analytic functions as follows.

Definition A.7 (Real analytic functions (Komatsu, 1960))**.** A function $f : \mathbb{R} \to \mathbb{R}$ is real analytic on a subset $\Omega \subset \mathbb{R}$ if it is infinitely continuously differentiable on Ω and for any compact $K \subset \Omega$, there exist $A > 0$ such that for any $k \in \mathbb{N}^*$,

$$\sup_{x\in K} \left| \frac{d^k f}{dx^k}(x) \right| \leq A^{k+1} k! \ .$$

Definition A.8 (Complex analytic function (Ahlfors, 1979))**.** A function $f : \mathbb{C} \to \mathbb{C}$ is complex analytic on a subset $K \subset \mathbb{C}$ if for any $x_0 \in \Omega$, it admits a convergent power series in a neighborhood of x_0.

Two key results from complex analysis are Morera's theorem (Ahlfors, 1979, Page 122) and Cauchy's integral theorem (Ahlfors, 1979, Theorem 2, Page 109). These will allow us to give explicit expressions of the Weierstrass transform.

Theorem A.9 (Morera's theorem)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ be a complex-valued function that is continuous on $K \subset \mathbb{C}$ an open set in the complex plane. If*

$$\oint_\Gamma f(x)dx = 0$$

for every closed piecewise smooth contour Γ in K, then f is complex analytic.

Theorem A.10 (Cauchy's integral theorem)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ be a complex analytic function on $K \subset \mathbb{C}$ an open set in the complex plane. Then for any closed piecewise smooth contour Γ in K*

$$\oint_\Gamma f(x)dx = 0.$$

In particular, a corollary of these two theorems is that for a continuous complex function $f : \mathbb{C} \to \mathbb{C}$, it holds that

$$f \text{ is analytic on } K \subset \mathbb{C} \textit{ if and only if } \oint_\Gamma f(x)dx = 0, \text{ for any closed piecewise smooth contour } \Gamma \subset K \ .$$

B. The Weierstrass Transform

B.1. Expression of the Weierstrass Transform – Proof of Theorem 3.3

We provide a proof of the following theorem, which gives an expression of the Weierstrass transform.

Theorem 3.3 (Series expression of $\mathbb{W}_{\sigma^2}$)**.** *Let $f \in \Phi_{M,a}(\mathbb{R}^p)$. Then, for any $\sigma^2 < 1/2a$, the generalized Weierstrass transform $\mathbb{W}_{\sigma^2}[f]$ admits the following expression*

$$\mathbb{W}_{\sigma^2}[f] = \sum_{k=0}^{\infty} \frac{\sigma^{2k}}{2^k k!} \Delta^k[f] \ .$$

To prove this theorem, we use the following lemma from Fritz (1991, Chapter 7, Problem 3). No proof is given for this problem in the initial textbook, we thus provide one in the following.

Lemma B.1 ($t \mapsto \mathbb{W}_{2t}$ is analytic for continuous Gaussian growing functions)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ a continuous function satisfying Equation (2) we refer to as Gaussian growth and recall: $|f(x)| \leq M \exp(a \|x\|^2)$, for any $x \in \mathbb{R}^p$. Then for any $x \in \mathbb{R}^p$, $t \mapsto \mathbb{W}_{2t}[f](x)$ is real analytic on $]0, 1/4a[$.*

Proof. Let f satisfying $|f(x)| \leq M \exp(a \|x\|^2)$ on $\mathbb{R}^p$ and a fixed $x \in \mathbb{R}^p$ throughout the proof. The overall goal is to show the analyticity of $t \mapsto \mathbb{W}_{2t}[f](x)$ on the larger complex domain $\Omega = \{t \in \mathbb{C} \ s.t. \ \Re(1/4t) > a\}$ which contains the real interval $]0, 1/4a[$. Indeed, the complex analyticity on a larger open set implies the real analyticity on the contained real open interval $]0, 1/4a[$. The analyticity of $t \mapsto \mathbb{W}_{2t}[f](x)$ on Ω is shown by verifying it is analytic on any compact subset $K \subset \Omega$. This latter objective is done using Morera's theorem (see Theorem A.9). We must then verify two properties:

(i) $t \mapsto \mathbb{W}_{2t}[f](x)$ is continuous on K,

(ii) $\oint_\Gamma \mathbb{W}_{2t}[f](x)dt = 0$ for any closed piecewise smooth contour Γ in K.

By Morera's theorem, if these conditions are met, the function $t \mapsto \mathbb{W}_{2t}[f](x)$ is analytic on K.

(i) Continuity of $t \mapsto \mathbb{W}_{2t}[f](x)$ on K. We prove this using continuity under the integral. We show that $\mathbb{W}_{2t}[f](x)$ can be written as the integral of a continuous dominated function $F(t, x, y)$ with respect to its integration variable y. Under these conditions, $\mathbb{W}_{2t}[f](x)$ is then continuous.

Denote $\psi_{2t}(w) = \frac{1}{\sqrt{4\pi t}^p} \exp\big(-\frac{\|w\|^2}{4t}\big)$, the probability density function of a centered isotropic Gaussian distribution of variance $2t\mathrm{I}_p$. We first define $F(t, x, y) = f(y)\psi_{2t}(x - y)$ for any $(t, y) \in]0, 1/4a[\times\mathbb{R}^p$. By definition of $\mathbb{W}_{2t}$,

$$\mathbb{W}_{2t}[f](x) = \mathbb{E}_{w\sim\mathcal{N}(0,2t\mathrm{I}_p)}\left[f(x + w)\right] = \int_{\mathbb{R}^p} f(x + w)\psi_{2t}(w)dw = \int_{\mathbb{R}^p} F(t, x, w)dw.$$

We use the dominated convergence of $t \mapsto F(t, x, w)$ on K for a fixed $x \in \mathbb{R}^p$ by a function of w to show that $t \mapsto \mathbb{W}_{2t}[f](x)$ is continuous. Let $w \in \mathbb{R}^p$ and $t \in K$,

$$\begin{aligned}
|F(t, x, w)| &\leq M \exp(a \|w\|^2)|4\pi t|^{-p/2} \exp(-\Re(\|x - w\|^2 /4t)) \\
&\quad \downarrow \text{ Denote } C_K = \sup_{t\in K} |4\pi t|^{-p/2}. \\
&\leq MC_K \exp(a \|w\|^2 - \Re(\|x - w\|^2 /4t)) \ .
\end{aligned}$$

Expanding the squared norm $\|x - w\|^2$ gives $-\Re(\|x - w\|^2 /4t) \leq \big(-\|x\|^2 + 2\|x\| \|w\| - \|w\|^2\big)\Re(1/4t)$. Denoting $\delta = \inf_{t\in K}\{\Re(1/4t) - a\}$, which is positive by definition of Ω, we obtain

$$\begin{aligned}
|F(t, x, w)| &\leq MC_K \exp(a \|w\|^2 - (a + \delta)(\|w\|^2 - 2\|x\| \|w\| + \|x\|^2)) \\
&= MC_K \exp(-\delta \|w\|^2) \exp(2(a + \delta)\|x\| \|w\| - (a + \delta)\|x\|^2) \\
&\quad \downarrow \text{ Denoting } D_{K,x,a} = MC_K \exp\left((a + \delta)\|x\|^2 \frac{2a + \delta}{\delta}\right) \\
&\leq D_{K,x,a} \exp(-(\delta/2)\|w\|^2) \ ,
\end{aligned}$$

which is a scaled Gaussian function, and is thus integrable. Then, for the given $x \in \mathbb{R}^p$, $w \mapsto F(t,x,w)$ is uniformly dominated by an integrable function for any $t \in K$. So $t \mapsto \mathbb{W}_{2t}[f]$ is continuous on K.

(ii) Morera's criterion. Let Γ be a closed piecewise smooth contour in K. We write $\oint_\Gamma \mathbb{W}_{2t}[f](x)dt$ as the double integral $\oint_\Gamma \left(\int_{\mathbb{R}^p} F(t,x,w)dw\right) dt$. Since $w \mapsto |F(t,x,w)|$ is dominated by an integrable function, Fubini's theorem gives

$$\oint_\Gamma \mathbb{W}_{2t}[f](x)dt = \oint_\Gamma \left(\int_{\mathbb{R}^p} F(t,x,w)dw\right) dt = \int_{\mathbb{R}^p} \left(\oint_\Gamma F(t,x,w)dt\right) dw.$$

Remark that, for any $w \in \mathbb{R}^p$, $t \mapsto F(t,x,w)$ is analytic on K because it is the product of $t \mapsto \psi_{2t}(x-y)$ that is analytic on K for any $x, y \in \mathbb{R}^p$ and the function f that does not depend on t. Then by Cauchy's integral theorem, $\oint_\Gamma F(t,x,w)dt = 0$. Consequently, it holds that

$$\oint_\Gamma \mathbb{W}_{2t}[f](x)dt = \int_{\mathbb{R}^p} \left(\oint_\Gamma F(t,x,w)dt\right) dw = 0 \ .$$

Then, by Morera's theorem $\mathbb{W}_{2t}[f](x)$ is analytic on any arbitrary $K \subset \Omega$. Finally, $t \mapsto \mathbb{W}_{2t}[f](x)$ is analytic on Ω (and in particular on the real interval $]0, 1/4a[$). □

Using this lemma, we now prove Theorem 3.3.

Proof of Theorem 3.3. In this proof, we use the parameterization $\sigma^2 = 2t$. Then, we work on $\mathbb{W}_{2t}$ and we finally re-inject σ^2 to obtain the desired result. By Fritz (1991, Chapter 7, Equation 1.11), if f satisfies $|f(x)| \leq M \exp(a\|x\|^2)$ on $\mathbb{R}^p$, then $u(x,t) = \mathbb{W}_{2t}[f]$ is an infinitely continuously differentiable solution of the following Heat equation:

$$\partial_t u(x,t) = \Delta_x u(x,t), \quad u(x,0) = f(x), \quad (x,t) \in \mathbb{R}^p \times]0, 1/4a[, \tag{10}$$

where $\Delta_x u(x,t)$ is the Laplacian of the function $x \mapsto u(x,t)$, and the constraint on $u(x,0)$ follows from $u(x,0) = \lim_{t\to 0} \mathbb{W}_{2t}[f](x)$. Furthermore, by Lemma B.1, the function $t \mapsto u(x,t)$ is analytic on $]0, 1/4a[$. Then, for $x \in \mathbb{R}^p$ and $t \in]0, 1/4a[$, we have the following Taylor expansion around $t_0 \in]0, 1/4a[$, assuming that the series converge,

$$\mathbb{W}_{2t}[f](x) = u(x,t) = \sum_{k=0}^{\infty} \frac{\partial_t^k u(x,t_0)}{k!}(t-t_0)^k = \sum_{k=0}^{\infty} \frac{\Delta^k u(x,t_0)}{k!}(t-t_0)^k \ ,$$

where the second equality comes from the fact that u is solution of the Heat equation (10). Taking the limit $t_0 \to 0$, we obtain the following Taylor expansion around 0,

$$\mathbb{W}_{2t}[f](x) = \sum_{k=0}^{\infty} \frac{\Delta^k u(x,0)}{k!} t^k = \sum_{k=0}^{\infty} \frac{\Delta^k f(x)}{k!} t^k \ . \tag{11}$$

It remains to check if this series converges for a given $x \in \mathbb{R}^p$. As f is in $\Phi_{M,a}(\mathbb{R}^p)$, we have $|\Delta^k[f](x)| \leq A_x (4a)^k k!$. Consequently, the *root test* condition

$$\limsup_k \left|\frac{\Delta^k[f](x)}{k!}\right|^{1/k} < \infty \ .$$

is met, and the series converges for $|t| < \frac{1}{r}$, where $r = \lim_k \sup \left|\frac{\Delta^k[f](x)}{k!}\right|^{1/k}$, with the convention that $1/0 = \infty$. Since f is in $\Phi_{M,a}(\mathbb{R}^p)$, $r \leq 4a$ and the series converges for $|t| \leq \frac{1}{4a} \leq \frac{1}{r}$. The result follows from plugging $t = \sigma^2/2$ in (11). □

B.2. Weierstrass Transform Properties

We now give two useful properties of the Weierstrass transform.

Proposition B.2 (Linearity of $\mathbb{W}_{\sigma^2}$)**.** *Let $\sigma > 0$, the Weierstrass transform is linear with respect to the function it applies to. Let f and g be two functions from $\mathbb{R}^p$ to $\mathbb{R}$ and let $\alpha \in \mathbb{R}$,*

$$\mathbb{W}_{\sigma^2}[\alpha f + g] = \alpha \mathbb{W}_{\sigma^2}[f] + \mathbb{W}_{\sigma^2}[g].$$

Proof. By linearity of the expectation, for a given $x \in \mathbb{R}^p$,

$$\begin{aligned}\mathbb{W}_{\sigma^2}[\alpha f + g](x) &= \mathbb{E}_{w\in\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[\alpha f(x+w) + g(x+w)\right] \\ &= \alpha\mathbb{E}_{w\in\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[f(x+w)\right] + \mathbb{E}_{w\in\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[g(x+w)\right] \\ &= \alpha\mathbb{W}_{\sigma^2}[f](x) + \mathbb{W}_{\sigma^2}[g](x)\ ,\end{aligned}$$

which is the result. □

Proposition B.3 ($\mathbb{W}_{\sigma^2}$ is increasing)**.** *Let $\sigma > 0$, the Weierstrass transform is increasing with respect to the function it applies to. Let f and g be two functions from $\mathbb{R}^p$ to $\mathbb{R}$ such that for any $x \in \mathbb{R}^p$, $f(x) \leq g(x)$, then for any $x \in \mathbb{R}^p$,*

$$\mathbb{W}_{\sigma^2}[f](x) \leq \mathbb{W}_{\sigma^2}[g](x).$$

Proof. By increasing property of the expectation, for a given $x \in \mathbb{R}^p$,

$$\mathbb{W}_{\sigma^2}[f](x) = \mathbb{E}_{w\in\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[f(x+w)\right] \leq \mathbb{E}_{w\in\mathcal{N}(0,\sigma^2\mathrm{I}_p)}\left[g(x+w)\right] = \mathbb{W}_{\sigma^2}[g](x),$$

and the lemma follows. □

We state and prove Remark 3.4.

Proposition B.4 (The exponential loss is in $\Phi_{M_a,a}(\mathbb{R}^p)$)**.** *Let $f(x) = \exp(-\theta^\top xy)$ for a given pair $(\theta, y) \in \Theta \times \mathcal{Y}$. Define $M_a = \exp\left(\frac{\|\Theta\|^2}{4a}\right)$ for any $a > 0$. The function f is in $\Phi_{M_a,a}(\mathbb{R}^p)$ for any $a > 0$.*

Proof. We first verify the property of Equation (3). Let an arbitrary $a > 0$. Let $k \in \mathbb{N}$ and $x \in \mathbb{R}^p$, we have the following Laplacian identity

$$\Delta_x^k f(x) = |\Delta_x^k f(x)| = \|\theta\|^{2k} f(x).$$

Factorials are increasing faster than any power of a positive number. Then, denoting k_0 such that for any $k \in \mathbb{N}$,

$$k > k_0 \implies k! > (\|\theta\|/4a)^{2k},$$

we have

$$|\Delta_x^k f(x)| \leq A_x (4a)^k k!\ ,$$

with $A_x = f(x)(\|\theta\|^2/4a)^{k_0}$. We now verify that the property of Equation (2) holds with $M_a = \exp\left(\frac{\|\Theta\|^2}{4a}\right)$. Let $x \in \mathbb{R}^p$,

$$f(x) = \exp(-\theta^\top xy) \leq \exp(\|\theta\|\|x\|) \leq \exp(\|\Theta\|\|x\|).$$

To prove that that $\exp(\|\Theta\|\|x\|) \leq M_a \exp(a\|x\|^2)$, we need that

$$a\|x\|^2 - \|\Theta\|\|x\| + \log(M_a) \geq 0.$$

It forms a second degree polynomial in $\|x\|$ with a positive quadratic constant. The inequality is then true for any $\|x\| > 0$ if the discriminant $\|\Theta\|^2 - 4a\log(M_a)$ is positive. In other words, if

$$M_a \leq \exp\left(\frac{\|\Theta\|^2}{4a}\right).$$

Thus, for any $a > 0$, Equation (2) holds with $M_a = \exp\left(\frac{\|\Theta\|^2}{4a}\right)$, and f is in $\Phi_{M_a,a}(\mathbb{R}^p)$ for any $a > 0$. □

C. Bias Characterization – Proof of Theorem 4.2

We provide a proof of the following theorem.

Theorem 4.2 (Bias induced by the Gaussian and Randomized Response mechanisms in binary classification)**.** *Let Δ_x denote the Laplacian with respect to the variable x and assume that ℓ satisfies Assumption 4.1 with $a < 1/2\sigma^2$. Recall $S(\epsilon_y) = 1/(1+e^{-\epsilon_y})$. For any $\theta \in \Theta$,*

$$\begin{aligned}\tilde{\mathcal{R}}(\theta) - \mathcal{R}(\theta) = &\underbrace{(1-S(\epsilon_y))\left(\mathbb{E}_{x,y}\left[\ell(\theta,x,-y)\right] - \mathcal{R}(\theta)\right)}_{\textit{label noise contribution}} \\ &+ \underbrace{S(\epsilon_y)\sum_{k=1}^{\infty}\frac{\sigma^{2k}}{2^k k!}\mathbb{E}_{x,y}\left[\Delta_x^k \ell(\theta,x,y)\right]}_{\textit{feature noise contribution}} \\ &+ \underbrace{(1-S(\epsilon_y))\sum_{k=1}^{\infty}\frac{\sigma^{2k}}{2^k k!}\mathbb{E}_{x,y}\left[\Delta_x^k \ell(\theta,x,-y)\right]}_{\textit{interactions of feature and label noise}}.\end{aligned}$$

Proof. For a given feature-label pair $(x,y) \in \mathcal{X}\times\mathcal{Y}$, with LDP release $(\tilde{x},\tilde{y})$ defined in (1), and a model $\theta\in\Theta$, we have

$$\begin{aligned}\mathbb{E}_{\tilde{x},\tilde{y}}\left[\ell\left(\theta,\ \tilde{x},\ \tilde{y}\right)\right] &= \mathbb{B}_{\epsilon_y}\left[z \mapsto \mathbb{W}_{\sigma^2}[\ell(\theta,\cdot,z)](x)\right](y) \\ &= S(\epsilon_y)\mathbb{W}_{\sigma^2}[\ell(\theta,\cdot,y)](x) + (1-S(\epsilon_y))\mathbb{W}_{\sigma^2}[\ell(\theta,\cdot,-y)](x).\end{aligned}$$

Taking the expectation with respect to $(x,y)\sim\mathcal{D}$ yields

$$\begin{aligned}\tilde{\mathcal{R}}(\theta) &= \mathbb{E}_{x,y}\left[S(\epsilon_y)\mathbb{W}_{\sigma^2}[\ell(\theta,\cdot,y)](x) + (1-S(\epsilon_y))\mathbb{W}_{\sigma^2}[\ell(\theta,\cdot,-y)](x)\right] \\ &\quad\downarrow \text{ Using the Theorem 3.3.} \\ &= \sum_{k=0}^{\infty}\frac{\sigma^{2k}}{2^k k!}\left\{S(\epsilon_y)\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,y)\right] + (1-S(\epsilon_y))\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,-y)\right]\right\} \\ &\quad\downarrow \text{ Isolating the term } k=0. \\ &= S(\epsilon_y)\mathbb{E}_{x,y}\ell(\theta,x,y) + (1-S(\epsilon_y))\mathbb{E}_{x,y}\ell(\theta,x,-y) \\ &\quad + \sum_{k=1}^{\infty}\frac{\sigma^{2k}}{2^k k!}\left\{S(\epsilon_y)\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,y)\right] + (1-S(\epsilon_y))\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,-y)\right]\right\}.\end{aligned}$$

Based on this, we have

$$\begin{aligned}\tilde{\mathcal{R}}(\theta) &= \mathbb{E}_{x,y}\ell(\theta,x,y) + (S(\epsilon_y)-1)\mathbb{E}_{x,y}\ell(\theta,x,y) + (1-S(\epsilon_y))\mathbb{E}_{x,y}\ell(\theta,x,-y) \\ &\quad + \sum_{k=1}^{\infty}\frac{\sigma^{2k}}{2^k k!}\left\{S(\epsilon_y)\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,y)\right] + (1-S(\epsilon_y))\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,-y)\right]\right\} \\ &= \mathbb{E}_{x,y}\ell(\theta,x,y) + (1-S(\epsilon_y))\left(\mathbb{E}_{x,y}\ell(\theta,x,-y) - \mathbb{E}_{x,y}\ell(\theta,x,y)\right) \\ &\quad + \sum_{k=1}^{\infty}\frac{\sigma^{2k}}{2^k k!}\left\{S(\epsilon_y)\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,y)\right] + (1-S(\epsilon_y))\mathbb{E}_{x,y}\left[\Delta_x^k\ell(\theta,x,-y)\right]\right\},\end{aligned}$$

and the result follows by identifying $\mathcal{R}(\theta) = \mathbb{E}_{x,y}\ell(\theta,x,y)$. □

D. Bias Correction Proofs

D.1. Inverse of Transforms – Proof of Theorem 5.1

We provide a proof of the following theorem.

Theorem 5.1 (Inverse of $\mathbb{B}_\epsilon$ and $\mathbb{W}_{\sigma^2}$)**.** *Define $\tilde{S}(\epsilon) = 1/(1-e^{-\epsilon})$. Let $g:\mathcal{Y}\to\mathbb{R}$ and $\epsilon>0$, for any $\tilde{y}\in\mathcal{Y}$,*

(i) $\mathbb{B}_\epsilon^{-1}[g](\tilde{y}) = \tilde{S}(\epsilon)g(\tilde{y}) + \big(1 - \tilde{S}(\epsilon)\big)g(-\tilde{y})$.

Let $f \in \Phi_{M,a}(\mathbb{R}^p)$*, for any* $\sigma^2 < 1/4a$ *and* $\tilde{x} \in \mathbb{R}^p$

(ii) $\mathbb{W}_{\sigma^2}^{-1}[f](\tilde{x}) = \sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!} \Delta^k[f](\tilde{x})$.

Proof. We first prove (i). Let $\tilde{y} \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{B}_\epsilon^{-1}[\mathbb{B}_\epsilon[g]](\tilde{y}) &= \mathbb{B}_\epsilon^{-1}\left[S(\epsilon)g(\cdot) + (1 - S(\epsilon))g(-\cdot)\right](\tilde{y}) \\
&= \tilde{S}(\epsilon)\left\{S(\epsilon)g(\tilde{y}) + (1 - S(\epsilon))g(-\tilde{y})\right\} + (1 - \tilde{S}(\epsilon))\left\{S(\epsilon)g(-\tilde{y}) + (1 - S(\epsilon))g(\tilde{y})\right\} \\
&\downarrow \text{ Group } g(\tilde{y}) \text{ and } g(-\tilde{y}) \text{ terms.} \\
&= \left\{\tilde{S}(\epsilon)S(\epsilon) + (1 - \tilde{S}(\epsilon))(1 - S(\epsilon))\right\} g(\tilde{y}) + \left\{(1 - \tilde{S}(\epsilon))S(\epsilon) + (1 - S(\epsilon))\tilde{S}(\epsilon)\right\} g(-\tilde{y}) \\
&= \left\{1 + 2\tilde{S}(\epsilon)S(\epsilon) - \tilde{S}(\epsilon) - S(\epsilon)\right\} g(\tilde{y}) + \left\{\tilde{S}(\epsilon) + S(\epsilon) - 2\tilde{S}(\epsilon)S(\epsilon)\right\} g(-\tilde{y}) \\
&\downarrow \text{ Developing } S(\cdot) \text{ and } \tilde{S}(\cdot). \\
&= \{1\}g(\tilde{y}) + \{0\}g(-\tilde{y}) = g(\tilde{y}).
\end{aligned}$$

Now we prove (ii). We use the parameterization $\sigma^2 = 2t$, it remains to reinject σ^2 to finish the proof. By Theorem 3.3 proof, the series $t \mapsto \sum_k \frac{\Delta_x^k f(x)}{k!} t^k$ converges absolutely for t in $]0, 1/4a[$ for any $x \in \mathbb{R}^p$. Since considering the series for $-t$ leads to the same coefficients in absolute value, then the series $t \mapsto \sum_k \frac{(-1)^k \Delta_x^k f(x)}{k!} t^k$ also converges for $0 < t < 1/4a$.

We thus consider this series as a candidate for the inverse of the Weierstrass transform. For f in $\Phi_{M,a}(\mathbb{R}^p)$, $x \in \mathbb{R}^p$ and $t \in]0, 1/8a[$, using the expression of the Weierstrass transform and its candidate inverse,

$$(\mathbb{W}_{2t} \circ \mathbb{W}_{2t}^{-1})[f](x) = \mathbb{W}_{2t}\left[\sum_{j=0}^{\infty} \frac{(-t)^j}{j!} \Delta^j[f](\cdot)\right](x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \Delta^j \left[\sum_{k=0}^{\infty} \frac{(-t)^j}{j!} \Delta^j[f](\cdot)\right](x).$$

We want to reorder the series and swap $\sum_{k=0}^{\infty}$ with Δ^j to obtain

$$(\mathbb{W}_{2t} \circ \mathbb{W}_{2t}^{-1})[f](x) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{(-t)^j t^k}{j!k!} \Delta^{k+j}[f](x).$$

That is valid if the resulting series converges absolutely. This is indeed the case since

$$\begin{aligned}
\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{t^{j+k}}{j!k!} \left|\Delta^{k+j}[f](x)\right| &\leq \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{t^{j+k}}{j!k!} A_x (4a)^{j+k} (j+k)! \\
&\downarrow \text{ Reordering (valid by positivity) with } n = k + j. \\
&= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \frac{t^n}{(n-j)!j!} A_x (4a)^n n! \\
&= A_x \sum_{n=0}^{\infty} (4at)^n \sum_{j=0}^{n} \binom{n}{j} \\
&\downarrow \text{ As } \sum_{j=0}^{n} \binom{n}{j} = 2^n. \\
&= A_x \sum_{n=0}^{\infty} (8at)^n \\
&\downarrow \text{ As } 8at < 1. \\
&= \frac{A_x}{1 - 8at} < \infty.
\end{aligned}$$

Then the reordering and swap of derivative and series are valid. We can thus write

$$
\begin{aligned}
(\mathbb{W}_{2t} \circ \mathbb{W}_{2t}^{-1})[f](x) &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{(-t)^j t^k}{j!k!}\Delta^{k+j}[f](x) \\
&\quad \downarrow \text{ Same reordering with } n = k + j. \\
&= \sum_{n=0}^{\infty}\sum_{j=0}^{n} \frac{(-1)^j t^n}{(n-j)!j!}\Delta^{n}[f](x) \\
&= \sum_{n=0}^{\infty} \frac{t^n \Delta^n[f](x)}{n!} \sum_{j=0}^{n} \binom{n}{j} (-1)^j 1^{n-j} = \sum_{n=0}^{\infty} \frac{t^n \Delta^n[f](x)}{n!}(1-1)^n \\
&= \frac{0^n \Delta^0[f](x)}{0!} \\
&\quad \downarrow \text{ With the convention that } 0^0 = 1. \\
&= f(x).
\end{aligned}
$$

We reparameterize with $\sigma^2 = 2t$, then for $a < 1/4\sigma^2$, $(\mathbb{W}_{\sigma^2} \circ \mathbb{W}_{\sigma^2}^{-1})[f] = f$. □

D.2. Commutativity of the gradient operator with the transforms

Proposition D.1. *Let $\epsilon, \delta > 0$, and let ℓ be a loss which satisfies 4.1 with $a < 1/2\sigma^2$. Then $\mathbb{T}_{\epsilon,\delta}$ and $\mathbb{T}_{\epsilon,\delta}^{-1}$ applied to $\theta \mapsto \ell(\theta, \cdot, \cdot)$, commute with ∇_θ.*

Proof. Let a tuple $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$, we express $\mathbb{T}^{-1}$ with $\mathbb{B}^{-1}$ and $\mathbb{W}^{-1}$:

$$
\begin{aligned}
\nabla_\theta \mathbb{T}_{\epsilon,\delta}^{-1}[\ell(\theta,\cdot,\cdot)](x,y) &= \nabla_\theta \left[\mathbb{B}_{\epsilon_y}^{-1}\left[z \mapsto \mathbb{W}_{\sigma^2}^{-1}[\ell(\theta,\cdot,z)](x)\right](y)\right] \\
&\quad \downarrow \text{ Replacing } \mathbb{W}^{-1} \text{ with its expression.} \\
&= \nabla_\theta \left[\mathbb{B}_{\epsilon_y}^{-1}\left[z \mapsto \sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,z)\right](y)\right] \\
&\quad \downarrow \text{ Replacing } \mathbb{B}^{-1} \text{ with its expression.} \\
&= \nabla_\theta \left[\tilde{S}(\epsilon_y)\sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,y) + \left(1-\tilde{S}(\epsilon_y)\right)\sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,-y)\right] \\
&\quad \downarrow \text{ By linearity of gradients with the finite sum.} \\
&= \tilde{S}(\epsilon_y)\nabla_\theta \left[\sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,y)\right] + \left(1-\tilde{S}(\epsilon_y)\right)\nabla_\theta \left[\sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,-y)\right]. \qquad (12)
\end{aligned}
$$

We now check for y and $-y$ if we can swap the series and gradients. The reasoning is the same for both and we then present it only for y. By series differentiation (Rudin, 1976., Theorem 7.17), if the series $\sum_{m=1}^{\infty} \frac{\sigma^{2m}}{2^m m!}|\partial_{\theta_j}\Delta_x^m \ell(\theta,x,y)|$ converges for each component $j \in \{1,\ldots,k\}$, then

$$
\nabla_\theta \left[\sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\Delta_x^k \ell(\theta,x,y)\right] = \sum_{k=0}^{\infty} \frac{(-1)^k \sigma^{2k}}{2^k k!}\nabla_\theta \Delta_x^k \ell(\theta,x,y). \qquad (13)
$$

As Δ^k is a finite sum of iterated derivatives, it commutes with ∇_θ. We then need to check that for $j \in \{1,\ldots,k\}$,

$$
\sum_{m=0}^{\infty} \frac{\sigma^{2m}}{2^m m!}\Delta_x^m \left[\partial_{\theta_j}\ell(\theta,x,y)\right]
$$

is a convergent series. By hypothesis on the loss, we have $\partial_{\theta_j}\ell(\theta,x,y) \in \Phi_{M,a}(\mathbb{R}^p)$, then

$$
\sum_{m=0}^{\infty} \frac{\sigma^{2m}}{2^m m!}\left|\Delta_x^m \left[\partial_{\theta_j}\ell(\theta,x,y)\right]\right| \leq \sum_{m=0}^{\infty} \frac{\sigma^{2m}}{2^m m!} A_x (4a)^m = A_x \sum_{m=0}^{\infty} (2a\sigma^2)^m = \frac{A_x}{1-2a\sigma^2} < \infty,
$$

where the last inequality follows from $2a\sigma^2 < 1$. Injecting Equation 13 in Equation (12) and swapping Δ^k and ∇_θ yields

$$\begin{aligned}\nabla_\theta \mathbb{T}^{-1}_{\epsilon,\delta}[\ell(\theta,\cdot,\cdot)](x,y) &= \tilde{S}(\epsilon_y)\sum_{k=0}^{\infty}\frac{(-1)^k\sigma^{2k}}{2^k k!}\nabla_\theta\Delta_x^k\ell(\theta,x,y) + \left(1-\tilde{S}(\epsilon_y)\right)\sum_{k=0}^{\infty}\frac{(-1)^k\sigma^{2k}}{2^k k!}\nabla_\theta\Delta_x^k\ell(\theta,x,-y)\\ &= \mathbb{B}^{-1}_{\epsilon_y}\left[z\mapsto \mathbb{W}^{-1}_{\sigma^2}[\nabla_\theta\ell(\theta,\cdot,z)](x)\right](y).\end{aligned}$$

The exact same reasoning can be carried out with $\mathbb{T}$ replacing $\tilde{S}$ by S and the terms $(-1)^k$ by 1 which does not affect the convergence of series involved. □

D.3. Unbiasedness of $\tilde{\ell}_{\epsilon,\delta}$ and $\nabla_\theta\tilde{\ell}_{\epsilon,\delta}$ – Proof of Theorem 5.2

We prove the following theorem.

Theorem 5.2 (Unbiasedness of IWP loss and gradient estimators)**.** *Assume ℓ satisfies Assumption 4.1 with $a < 1/4\sigma^2$. Let $\epsilon,\delta > 0$, for any pair $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\theta\in\Theta$, define $(\tilde{x},\tilde{y})$ as a (ϵ,δ)-LDP release defined in Equation (1), the IWP loss estimator defined in Equation (6) satisfies:*

(i) $\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right] = \ell(\theta,x,y),$

(ii) $\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right] = \nabla_\theta\ell(\theta,x,y).$

Proof. We first prove that for any $h:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ such that for any $y\in\mathcal{Y}$, $x\mapsto h(x,y)\in\Phi_{M,a}(\mathbb{R}^p)$,

$$\mathbb{T}_{\epsilon,\delta}[\mathbb{T}^{-1}_{\epsilon,\delta}[h(\cdot,\cdot)]](x,y) = h(x,y). \tag{14}$$

We can write

$$\begin{aligned}\mathbb{T}_{\epsilon,\delta}[\mathbb{T}^{-1}_{\epsilon,\delta}[h(\cdot,\cdot)]](x,y) &= \mathbb{B}_{\epsilon_y}\left[\mathbb{W}_{\sigma^2}\left[\mathbb{B}^{-1}_{\epsilon,\delta}\left[\mathbb{W}^{-1}_{\sigma^2}[h(\cdot,\cdot)]\right]\right]\right](x,y)\\ &\downarrow \text{ As } \mathbb{B}^{-1}_{\epsilon_y} \text{ forms the sum of two functions in } \Phi_{M,a}(\mathbb{R}^p), \text{ it then commutes with } \mathbb{W}^{-1}_{\sigma^2}.\\ &= \mathbb{B}_{\epsilon_y}\left[\mathbb{W}_{\sigma^2}\left[\mathbb{W}^{-1}_{\sigma^2}\left[\mathbb{B}^{-1}_{\epsilon,\delta}[h(\cdot,\cdot)]\right]\right]\right](x,y)\\ &\downarrow \text{ Simplifying } \mathbb{W}\circ\mathbb{W}^{-1}.\\ &= \mathbb{B}_{\epsilon_y}\left[\mathbb{B}^{-1}_{\epsilon,\delta}[h(\cdot,\cdot)]\right](x,y)\\ &\downarrow \text{ Simplifying } \mathbb{B}\circ\mathbb{B}^{-1}.\\ &= h(x,y).\end{aligned}$$

Considering Equation (14) with $h:(x,y)\mapsto\ell(\theta,x,y)$ for a given $\theta\in\Theta$ yields

$$\begin{aligned}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right] &= \mathbb{T}_{\epsilon,\delta}[\mathbb{T}^{-1}_{\epsilon,\delta}[\ell(\theta,\cdot,\cdot)]](x,y)\\ &\downarrow \text{ By Equation (14).}\\ &= \ell(\theta,x,y).\end{aligned}$$

Now, for the gradient, we also have the following.

$$\begin{aligned}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right] &= \mathbb{T}_{\epsilon,\delta}\left[\mathbb{T}^{-1}_{\epsilon,\delta}[\nabla_\theta\ell(\theta,\cdot,\cdot)\right](x,y)\\ &\downarrow \text{ By commutativity of } \nabla_\theta \text{ and } \mathbb{T}^{-1}_{\epsilon,\delta}.\\ &= \mathbb{T}_{\epsilon,\delta}\left[\nabla_\theta\mathbb{T}^{-1}_{\epsilon,\delta}[\ell(\theta,\cdot,\cdot)\right](x,y)\\ &\downarrow \text{ By commutativity of } \nabla_\theta \text{ and } \mathbb{T}_{\epsilon,\delta}.\\ &= \nabla_\theta\mathbb{T}_{\epsilon,\delta}\left[\mathbb{T}^{-1}_{\epsilon,\delta}[\ell(\theta,\cdot,\cdot)\right](x,y)\\ &\downarrow \text{ By Equation (14).}\\ &= \nabla_\theta\ell(\theta,x,y).\end{aligned}$$

□

D.4. Variance of $\nabla\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})$ – Proof of Theorem 5.3

We provide a proof of the following result.

Lemma D.2 (Variance of $\mathbb{W}^{-1}_{\sigma^2}[f](x+w)$)**.** *Let $f \in \Phi_{M,a}(\mathbb{R}^p)$ such that $\left(\mathbb{W}^{-1}_{\sigma^2}[f]\right)^2$ is in $\Phi_{M,a}(\mathbb{R}^p)$. Let $\sigma^2 \in]0, 1/4a[$, $x \in \mathbb{R}^p$ and $w \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_p)$. The variance of $\mathbb{W}^{-1}_{\sigma^2}[f](x+w)$ is*

$$\begin{aligned}\mathbb{V}_w\left(\mathbb{W}^{-1}_{\sigma^2}[f](x+w)\right) &= \mathbb{W}_{\sigma^2}\left[\left(\mathbb{W}^{-1}_{\sigma^2}[f]\right)^2\right](x) - f^2(x)\\ &= 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\left\|\mathbb{W}^{-1}_{2s}[\nabla f](\cdot)\right\|^2(x)ds.\end{aligned}$$

Furthermore, if there exists $C \geq 0$ such that $\sup_{x\in\mathcal{X}}\sup_{0<s<t}\left\|\mathbb{W}^{-1}_{2s}[\nabla f](x)\right\| \leq C$, it holds that

$$\mathbb{V}_w\left(\mathbb{W}^{-1}_{\sigma^2}[f](x+w)\right) \leq C^2\sigma^2.$$

Proof. For the proof, we use the parameterization $2t = \sigma^2$. Denote $g_t(x) = \mathbb{W}^{-1}_{2t}[f](x)$ and $v(x,t) = \mathbb{E}_w[g_t^2(x+w)] = \mathbb{W}_{2t}[g_t^2](x)$. We first have that $\mathbb{E}_w\left(g_t(x+w) - f(x)\right)^2 = v(x,t) - f^2(x)$. So we focus on the term $v(x,t)$ that we will compute in the following integral form by integrating over t:

$$v(x,t) = v(x,0) + \int_0^t \partial_s v(x,s)ds = f^2(x) + \int_0^t \partial_s v(x,s)ds. \tag{15}$$

We develop $\partial_s v(x,s)$:

$$\begin{aligned}\partial_s v(x,s) &= \partial_s \mathbb{W}_{2s}[g_s^2](x)\\ &\quad\downarrow \text{ By hypothesis, } g_s^2 = (\mathbb{W}_{2s}[f])^2 \text{ is in } \Phi_{M,a}(\mathbb{R}^p)\\ &= \partial_s\left[\sum_{k=0}^{\infty}\frac{s^k}{k!}\Delta^k[g_s^2](x)\right]\\ &= \sum_{k=0}^{\infty}\frac{s^{k-1}}{(k-1)!}\Delta^k[g_s^2](x) + \sum_{k=0}^{\infty}\frac{s^k}{k!}\Delta^k[\partial_s g_s^2](x)\\ &= \sum_{k=0}^{\infty}\frac{s^k}{k!}\Delta^{k+1}[g_s^2](x) + \sum_{k=0}^{\infty}\frac{s^k}{k!}\Delta^k[\partial_s g_s^2](x)\\ &= \mathbb{W}_{2s}[\Delta g_s^2](x) + \mathbb{W}_{2s}[\partial_s g_s^2](x) = \mathbb{W}_{2s}[\Delta g_s^2 + \partial_s g_s^2](x).\end{aligned}$$

Now we develop the derivative $\partial_s g_s^2 = 2g_s\partial_s g_s = -2g_s\Delta g_s$ and $\Delta g_s^2 = 2g_s\Delta g_s + 2\|\nabla g_s\|^2$ to reinject it in the previous derivation:

$$\partial_s v(x,s) = \mathbb{W}_{2s}[2g\Delta g_s + 2\|\nabla g_s\|^2 - 2g_s\Delta g_s](x) = 2\mathbb{W}_{2s}[\|\nabla g_s\|^2](x).$$

Since $\nabla g_s = \nabla\mathbb{W}^{-1}_{2s}[f] = \mathbb{W}^{-1}_{2s}[\nabla f]$, we further simplify the expression of $\partial_s v(x,s)$ in the integral representation (15), which yields the desired result:

$$v(x,t) = f^2(x) + \int_0^t 2\mathbb{W}_{2s}[\|\mathbb{W}^{-1}_{2s}[\nabla f](\cdot)\|^2](x)ds.$$

Now, assume $\sup_{x\in\mathcal{X}}\sup_{0<s<t}\|\mathbb{W}^{-1}_{2s}[\nabla f](x)\| \leq C$, denoting $\psi_{2s}(w) = \frac{1}{\sqrt{4\pi s}}\exp\left(-\frac{\|w\|^2}{4s}\right)$ the probability density function of a centered isotropic Gaussian distribution of variance $2s\mathrm{I}_p$, we have

$$\mathbb{W}_{2s}[\|\mathbb{W}^{-1}_{2s}[\nabla f](\cdot)\|^2](x) = \int \|\mathbb{W}^{-1}_{2s}[\nabla f](x-w)\|^2\psi_{2s}(w)dw \leq \sup_{x\in\mathcal{X}}\sup_{0<s<t}\|\mathbb{W}^{-1}_{2s}[\nabla f](x)\|^2\int\psi_{2s}(w)dw = C^2.$$

And the result follows from $v(x,t) \leq f^2(x) + 2\int_0^t C^2 ds = f^2(x) + 2tC^2$. □

We now have an unbiased estimator of any function $f \in \Phi_{M,a}(\mathbb{R}^p)$ at any point $x \in \mathbb{R}^p$ from a Gaussian-perturbed release of the point $x + w$ for which we can compute the variance exactly.

The variance can be derived in closed-form for known functions like:

- if $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$, then $\mathbb{V}_w(\mathbb{W}_{2t}^{-1}[f](x+w)) = f^2(x) + 2t\,\|\Sigma x + b\|^2 + 2t^2\,\mathrm{Tr}(\Sigma^2)$ with $\Sigma = (A + A^\top)/2$,
- if $f(x) = \exp(\alpha^\top x),\ \alpha \in \mathbb{R}^p$, then $\mathbb{V}_w(\mathbb{W}_{2t}^{-1}[f](x+w)) = \exp(2a^\top x + 2t\|a\|^2)$.

Particular cases of the examples are respectively the mean squared error and the exponential loss (see subsection 5.2).

Similarly, the existence of an inverse $\mathbb{B}_\epsilon^{-1}$, means that for any function $g : \{-1, 1\} \to \mathbb{R}$, any $y \in \{-1, 1\}$ and $\epsilon > 0$,

$$\mathbb{B}_\epsilon^{-1}[g](\tilde{y}), \quad \tilde{y} \sim \mathcal{B}_\epsilon(y)$$

is an unbiased estimator of $g(y)$. We give an exact expression for its variance in the following theorem.

Lemma D.3 (Variance of $\mathbb{B}_\epsilon^{-1}[g](\mathcal{B}_\epsilon(y))$)**.** *Let $g\colon \{-1, 1\} \to \mathbb{R}$, $\epsilon > 0$ and $y \in \{-1, 1\}$. The variance of $\mathbb{B}_\epsilon^{-1}[g](\tilde{y})$ with $\tilde{y} = \mathcal{B}_\epsilon(y)$ is*

$$\mathbb{V}_{\mathcal{B}_\epsilon}\left(\mathbb{B}_\epsilon^{-1}[g](\tilde{y})\right) = \tilde{S}(\epsilon)(\tilde{S}(\epsilon) - 1)(g(1) - g(-1))^2 \ .$$

Proof. For clarity, we denote $S \equiv S(\epsilon)$ and $\tilde{S} \equiv \tilde{S}(\epsilon)$.

$$\begin{aligned}
\mathbb{V}_{\mathcal{B}_\epsilon}\left(\mathbb{B}_\epsilon^{-1}[g](\tilde{y})\right) &= \mathbb{E}_{\mathcal{B}_\epsilon}\left[\left(\mathbb{B}_\epsilon^{-1}[g](\tilde{y}) - g(y)\right)^2\right] \\
&= S\left(\mathbb{B}_\epsilon^{-1}[g](y) - g(y)\right)^2 + (1 - S)\left(\mathbb{B}_\epsilon^{-1}[g](-y) - g(y)\right)^2 \\
&= S\left(\tilde{S}g(y) + (1 - \tilde{S})g(-y) - g(y)\right)^2 + (1 - S)\left(\tilde{S}g(-y) + (1 - \tilde{S})g(y) - g(y)\right)^2 \\
&= S\left((\tilde{S} - 1)g(y) + (1 - \tilde{S})g(-y)\right)^2 + (1 - S)\left(\tilde{S}g(-y) - \tilde{S}g(y)\right)^2 \\
&= S(1 - \tilde{S})^2\left(g(y) - g(-y)\right)^2 + (1 - S)\tilde{S}^2\left(g(-y) - g(y)\right)^2 \\
&\downarrow \text{ } S \text{ and } \tilde{S} \text{ are replaced by their expressions and } (g(y) - g(-y))^2 = (g(1) - g(-1))^2 \text{ for any } y \in \{1, -1\}. \\
&= \left\{S(1 - \tilde{S})^2 + (1 - S)\tilde{S}^2\right\}(g(1) - g(-1))^2 \\
&= \left\{\frac{e^\epsilon}{e^\epsilon + 1}\frac{1}{(e^\epsilon - 1)^2} + \frac{1}{e^\epsilon + 1}\frac{e^{2\epsilon}}{(e^\epsilon - 1)^2}\right\}(g(1) - g(-1))^2 \\
&= \frac{e^\epsilon}{(e^\epsilon - 1)^2}\left\{\frac{1}{e^\epsilon + 1} + \frac{e^\epsilon}{e^\epsilon + 1}\right\}(g(1) - g(-1))^2 \\
&\downarrow \text{ As } \frac{e^\epsilon}{(e^\epsilon - 1)^2} = \tilde{S}(\epsilon)(\tilde{S}(\epsilon) - 1). \\
&= \tilde{S}(\epsilon)(\tilde{S}(\epsilon) - 1)(g(1) - g(-1))^2 \ .
\end{aligned}$$

□

Let us now restate the variance of the IWP gradient estimator and provide the proof.

Theorem 5.3 (Variance of the IWP gradient estimator)**.** *Let $\epsilon, \delta > 0$, an original feature-label pair $x, y \in \mathcal{X} \times \mathcal{Y}$ and its corresponding (ϵ, δ)-LDP release $(\tilde{x}, \tilde{y})$ defined in Equation (1). Let ℓ satisfy Assumption 4.1 with $a < 1/4\sigma^2$. Given that $\mathcal{X}$ and Θ are bounded sets, denoting*

$$C = \sup_{(\theta,x,y)\in\Theta\times\mathcal{X}\times\mathcal{Y}, s<\sigma^2} \max\Big\{\|\nabla_\theta \ell(\theta, x, y)\|, \ \left\|\mathbb{W}_s^{-1}\left[\nabla_\theta \nabla_x \ell(\theta, \cdot, y)\right](x)\right\|\Big\}, \tag{8}$$

the variance of the IWP gradient estimator admits the following upper bound

$$\mathbb{E}\left\|\nabla_\theta \tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) - \nabla_\theta \ell(\theta,x,y)\right\|^2 \leq C^2\left(\sigma^2 + 4\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)(1+\sigma^2)\right) \quad . \tag{9}$$

Proof. We first decompose the variance for each component of $\nabla\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})$.

$$\begin{aligned}\mathbb{E}\|\nabla_\theta\tilde{\ell}(\theta,\tilde{x},\tilde{y})\|^2 - \|\nabla\ell(\theta,x,y)\|^2 &= \sum_{j=1}^k\left(\mathbb{E}\left|\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right|^2 - \left|\partial_{\theta_j}\ell(\theta,x,y)\right|^2\right)\\ &= \sum_{j=1}^k \underbrace{\mathbb{V}\left(\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\right)}_{(A_j)} \quad .\end{aligned}$$

We are interested in computing (A_j) for any $j\in\{1,\dots,k\}$. We can decompose it using total variance law:

$$(A_j) = \underbrace{\mathbb{E}_{\tilde{y}}\left[\mathbb{V}_{\tilde{x}}\left(\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\,\middle|\,\tilde{y}\right)\right]}_{(a_j)} + \underbrace{\mathbb{V}_{\tilde{y}}\left(\mathbb{E}_{\tilde{x}}\left[\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})\,\middle|\,\tilde{y}\right]\right)}_{(b_j)} \quad .$$

We start with the term (a_j). For that, we need the following expression of $\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y})$:

$$\begin{aligned}\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) &= \partial_{\theta_j}\left[\mathbb{W}_{\sigma^2}^{-1}\left[z\mapsto\mathbb{B}_{\epsilon_y}^{-1}[\ell(\theta,z,\cdot)](\tilde{y})\right](\tilde{x})\right]\\ &= \mathbb{W}_{\sigma^2}^{-1}\left[z\mapsto\mathbb{B}_{\epsilon_y}^{-1}[\partial_{\theta_j}\ell(\theta,z,\cdot)](\tilde{y})\right](\tilde{x}).\end{aligned}$$

Then, using Lemma D.2 with $f(x) = \mathbb{B}_{\epsilon_y}^{-1}\left[\partial_{\theta_j}\ell(\theta,x,\cdot)\right](\tilde{y})$ and $t=\sigma^2/2$, we have

$$\begin{aligned}\mathbb{V}_{\tilde{x}}\left(\partial_{\theta_j}\ell(\theta,\tilde{x},\tilde{y})\mid\tilde{y}\right) &= \mathbb{W}_{\sigma^2}\left[\left(\mathbb{W}_{\sigma^2}^{-1}\left[z\mapsto\mathbb{B}_{\epsilon_y}^{-1}[\partial_{\theta_j}\ell(\theta,z,\cdot)](\tilde{y})\right](\tilde{x})\right)^2\right] - \left(\mathbb{B}_{\epsilon_y}^{-1}\left[\partial_{\theta_j}\ell(\theta,x,\cdot)\right](\tilde{y})\right)^2\\ &= 2\int_0^{\sigma^2/2}\mathbb{W}_{2s}[\|\mathbb{W}_{2s}^{-1}[\nabla_x\mathbb{B}_{\epsilon_y}^{-1}(\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y}))]\|^2](x)ds.\end{aligned}$$

We now inject this variance term in (a_j), which gives

$$\begin{aligned}(a_j) &= \mathbb{E}_{\tilde{y}}\left[2\int_0^{\sigma^2/2}\mathbb{W}_{2s}[\|\mathbb{W}_{2s}^{-1}[\nabla_x\mathbb{B}_{\epsilon_y}^{-1}(\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y}))]\|^2](x)ds\right]\\ &\quad\downarrow \text{ Developping the gradient norm } \|\nabla_x h(x)\|^2 = \textstyle\sum_{i=1}^p(\partial_{x_i}h(x))^2, \text{ and swapping } \partial_{x_i} \text{ and } \mathbb{B}^{-1}.\\ &= \sum_{i=1}^p\mathbb{E}_{\tilde{y}}\left[2\int_0^{\sigma^2/2}\mathbb{W}_{2s}[(\mathbb{W}_{2s}^{-1}[\mathbb{B}_{\epsilon_y}^{-1}(\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y}))])^2](x)ds\right]\\ &\quad\downarrow \text{ Swapping } \mathbb{W} \text{ and } \mathbb{E}_{\tilde{y}}.\\ &= \sum_{i=1}^p 2\int_0^{\sigma^2/2}\mathbb{W}_{2s}\left[\mathbb{E}_{\tilde{y}}\left[\left(\mathbb{W}_{2s}^{-1}[\mathbb{B}_{\epsilon_y}^{-1}(\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y}))]\right)^2\right]\right](x)ds\\ &= \sum_{i=1}^p 2\int_0^{\sigma^2/2}\mathbb{W}_{2s}\left[\mathbb{E}_{\tilde{y}}\left[\left(\mathbb{B}_{\epsilon_y}^{-1}[\mathbb{W}_{2s}^{-1}(\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y}))]\right)^2\right]\right](x)ds \quad .\end{aligned}$$

Now we can use the formula of $\mathbb{V}(\mathbb{B}^{-1}(g(\tilde{y}))$ from Lemma D.3 with $g(y) = \mathbb{W}_{2s}^{-1}(\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,y))$ and the fact that $\mathbb{E}_{\tilde{y}}[\mathbb{B}_{\epsilon_y}^{-1}[\mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y})]]] = \mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y})]$, which gives

$$
\begin{aligned}
(a_j) &= \sum_{i=1}^{p} 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\left[\mathbb{V}_{\tilde{y}}(\mathbb{B}_{\epsilon_y}^{-1}[\mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y})]]) + (\mathbb{E}_{\tilde{y}}[\mathbb{B}_{\epsilon_y}^{-1}[\mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,\tilde{y})]]])^2\right](x)ds \\
&= \sum_{i=1}^{p} 2\int_0^{\sigma^2/2} \tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\mathbb{W}_{2s}\left[\left(\mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,1)] - \mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,-1)]\right)^2\right](x)ds \\
&\quad + \sum_{i=1}^{p} 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\left[\left(\mathbb{W}_{2s}^{-1}[\partial_{x_i}\partial_{\theta_j}\ell(\theta,\cdot,y)]\right)^2\right](x)ds.
\end{aligned}
$$

We rearrange the terms to make the squared norm appear:

$$
\begin{aligned}
(a_j) &= 2\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\int_0^{\sigma^2/2} \mathbb{W}_{2s}\left[\|\mathbb{W}_{2s}^{-1}(\nabla_x\partial_{\theta_j}\ell(\theta,\cdot,1) - \mathbb{W}_{2s}^{-1}(\nabla_x\partial_{\theta_j}\ell(\theta,\cdot,-1)\|^2\right](x)ds \\
&\quad + 2\int_0^{\sigma^2} \mathbb{W}_{2s}\left[\|\mathbb{W}_{2s}^{-1}[\nabla_x\partial_{\theta_j}\ell(\theta,\cdot,y)]\|^2\right](x)ds.
\end{aligned}
$$

Finally, remarking that $\mathbb{E}_{\tilde{x}}[\partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) \mid \tilde{y}] = \partial_{\theta_j}\tilde{\ell}_{\epsilon,\delta}(\theta,x,\tilde{y})$ and using Lemma D.3 again gives

$$
(b_j) = \mathbb{V}_{\tilde{y}}\left(\mathbb{B}_{\epsilon_y}^{-1}\left[\partial_{\theta_j}\ell(\theta,x,\cdot)\right](\tilde{y})\right) = \tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\left(\partial_{\theta_j}\ell(\theta,x,1) - \partial_{\theta_j}\ell(\theta,x,-1)\right)^2.
$$

Plugging the results of (a_j) and (b_j) in the formula of (A_j) and summing over $j \in \{1\ldots,k\}$ yields the following expression of the variance of the IWP gradient estimator:

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) - \nabla_\theta\ell(\theta,x,y)\right\|^2 \\
&= 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta,\cdot,y)]\|^2(x)ds \\
&\quad + \tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\|\nabla_\theta\ell(\theta,x,1) - \nabla_\theta\ell(\theta,x,-1)\|^2 \\
&\quad + 2\tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y)-1\right)\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta,\cdot,1) - \nabla_x\nabla_\theta\ell(\theta,\cdot,-1)]\|^2(x)ds \ ,
\end{aligned}
$$

where the matrix norm is the Frobenius norm. Recalling the definition

$$
C = \sup_{(\theta,x,y)\in\Theta\times\mathcal{X}\times\mathcal{Y}, s<\sigma^2/2} \max\left\{\|\nabla_\theta\ell(\theta,x,y)\|, \left\|\mathbb{W}_{2s}^{-1}\left[\nabla_\theta\nabla_x\ell(\theta,\cdot,y)\right](x)\right\|\right\},
$$

we can bound the integrands with C and use the increasing property of $\mathbb{W}$ to bound the variance and get the result. □

D.5. Application to Generalized Linear Models (GLM)

We provide proofs of derivations for the case of GLM (subsection 5.2).

Loss. Recall that in GLM the loss is $\ell(\theta,\tilde{x},\tilde{y}) = f(\theta^\top\tilde{x}\tilde{y})$. It simplifies the expression of iterated Laplacians:

$$
\Delta_x^k f(\theta^\top\tilde{x}\tilde{y}) = \|\theta\|^{2k} f^{(2k)}(\theta^\top\tilde{x}\tilde{y}) \ .
$$

Then, the Weierstrass inverse has the following expression.

$$
\begin{aligned}
\mathbb{W}_{2t}^{-1}[\ell(\theta,\cdot,\tilde{y})](\tilde{x}) &= \sum_{k=0}^{\infty}\frac{\Delta_x^k\ell(\theta,\tilde{x},\tilde{y})}{k!}(-t)^k = \sum_{k=0}^{\infty}\frac{\Delta_x^k f(\theta^\top\tilde{x}\tilde{y})}{k!}(-t)^k \\
&= \sum_{k=0}^{\infty}\frac{\|\theta\|^{2k} f^{(2k)}(\theta^\top\tilde{x}\tilde{y})}{k!}(-t)^k = \sum_{k=0}^{\infty}\frac{(-t\|\theta\|^2)^k}{k!}f^{(2k)}(\theta^\top\tilde{x}\tilde{y}) \\
&= \mathbb{W}_{2t\|\theta\|^2}^{-1}[f](\theta^\top\tilde{x}\tilde{y}).
\end{aligned}
$$

Recall Equation (6):

$$\begin{aligned}\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) &= \mathbb{T}^{-1}_{\epsilon,\delta}[\ell(\theta,\cdot,\cdot)](\tilde{x},\tilde{y}) = \mathbb{W}^{-1}_{\sigma^2}\left[\mathbb{B}_{\epsilon_y}[\ell(\theta,\cdot,\cdot)](\cdot,\tilde{y})\right](\tilde{x}) \\ &\quad \downarrow \text{ Applying } \mathbb{B}^{-1}_{\epsilon_y} \text{ first.} \\ &= \mathbb{W}^{-1}_{\sigma^2}\left[\tilde{S}(\epsilon_y)\ell(\theta,\cdot,\tilde{y}) + \left(1-\tilde{S}(\epsilon_y)\right)\ell(\theta,\cdot,-\tilde{y})\right](\tilde{x}) \\ &\quad \downarrow \text{ By linearity of } \mathbb{W}^{-1}_{2t}. \\ &= \tilde{S}(\epsilon_y)\mathbb{W}^{-1}_{\sigma^2}[\ell(\theta,\cdot,\tilde{y})](\tilde{x}) + \left(1-\tilde{S}(\epsilon_y)\right)\mathbb{W}^{-1}_{\sigma^2}[\ell(\theta,\cdot,-\tilde{y})](\tilde{x}) \ .\end{aligned}$$

Replacing $\mathbb{W}^{-1}_{2t}[\ell(\theta,\cdot,\tilde{y})](\tilde{x}) = \mathbb{W}^{-1}_{2t\|\theta\|^2}[f](\theta^\top\tilde{x}\tilde{y})$ with $2t=\sigma^2$ yields

$$\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = \tilde{S}(\epsilon_y)\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) + \left(1-\tilde{S}(\epsilon_y)\right)\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(-\theta^\top\tilde{x}\tilde{y}\right).$$

Gradient. We can't directly differentiate Equation (6) with respect to θ by swaping Weierstrass transform and gradient ($\nabla_\theta\mathbb{W}^{-1}_{2t}[\ell(\theta,\cdot,\tilde{y})] = \mathbb{W}^{-1}_{2t}[\nabla_\theta\ell(\theta,\cdot,\tilde{y})]$) because here t depends on θ. We thus write the derivative explicitly,

$$\nabla_\theta\tilde{\ell}_{\epsilon,\delta}(\theta,\tilde{x},\tilde{y}) = \tilde{S}(\epsilon_y)\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) + \left(1-\tilde{S}(\epsilon_y)\right)\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(-\theta^\top\tilde{x}\tilde{y}\right).$$

It boils down to computing $\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right)$ for a given $\tilde{y}\in\{-1,1\}$. Let $j\in\{1,\dots,p\}$,

$$\begin{aligned}\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) &= \nabla_\theta\left[\sum_{k=0}^{\infty}\frac{(-\tau\|\theta\|^2)^k}{k!}f^{(2k)}(\theta^\top\tilde{x}\tilde{y})\right], \quad \text{Denoting } \tau=\sigma^2/2. \\ &= \sum_{k=0}^{\infty}\frac{(-\tau)^k}{k!}\nabla_\theta\left[\|\theta\|^{2k}f^{(2k)}(\theta^\top\tilde{x}\tilde{y})\right] \\ &= \sum_{k=0}^{\infty}\frac{(-\tau)^k}{k!}\left\{2k\|\theta\|^{2k-2}\theta f^{(2k)}(\theta^\top\tilde{x}\tilde{y}) + \|\theta\|^{2k}\tilde{x}\tilde{y}f^{(2k+1)}(\theta^\top\tilde{x}\tilde{y})\right\} \\ &= 2\theta\sum_{k=0}^{\infty}\frac{k(-\tau\|\theta\|^2)^k}{\|\theta\|^2k!}f^{(2k)}(\theta^\top\tilde{x}\tilde{y}) + \tilde{x}\tilde{y}\sum_{k=0}^{\infty}\frac{(-\tau\|\theta\|^2)^k}{k!}f^{(2k+1)}(\theta^\top\tilde{x}\tilde{y}) \\ &= -2\tau\theta\sum_{k=1}^{\infty}\frac{(-\tau\|\theta\|^2)^{k-1}}{(k-1)!}f^{(2k)}(\theta^\top\tilde{x}\tilde{y}) + \tilde{x}\tilde{y}\sum_{k=0}^{\infty}\frac{(-\tau\|\theta\|^2)^k}{k!}f^{(2k+1)}(\theta^\top\tilde{x}\tilde{y}) \\ &= -2\tau\theta\sum_{k=0}^{\infty}\frac{(-\tau\|\theta\|^2)^k}{k!}f^{(2k+2)}(\theta^\top\tilde{x}\tilde{y}) + \tilde{x}\tilde{y}\sum_{k=0}^{\infty}\frac{(-\tau\|\theta\|^2)^k}{k!}f^{(2k+1)}(\theta^\top\tilde{x}\tilde{y}).\end{aligned}$$

Recognizing the Weierstrass transforms and replacing τ, we obtain

$$\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top\tilde{x}\tilde{y}\right) = -\sigma^2\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f'']\left(\theta^\top\tilde{x}\tilde{y}\right) + \tilde{x}\tilde{y}\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f']\left(\theta^\top\tilde{x}\tilde{y}\right). \tag{16}$$

D.6. Uniform Bounds of $\|\mathbb{W}^{-1}_{2s}\nabla_\theta\nabla_x\ell(\theta,x,y)\|$

In this subsection, we derive uniform bounds on $\left\|\mathbb{W}^{-1}_{2s}\nabla_\theta\nabla_x\ell(\theta,x,y)\right\|$. From the expression (16), we have

$$\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right) = xy\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f']\left(\theta^\top xy\right) - \sigma^2\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f'']\left(\theta^\top xy\right).$$

It remains to differentiate again with respect to x:

$$\begin{aligned}\nabla_x\nabla_\theta\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right) &= \nabla_x\left[xy\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f']\left(\theta^\top xy\right)\right] - \sigma^2\theta\nabla_x\left[\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f'']\left(\theta^\top xy\right)\right] \\ &= y\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f']\left(\theta^\top xy\right)\mathrm{I}_p + x\theta^\top\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f'']\left(\theta^\top xy\right) - \sigma^2y\theta\theta^\top\mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f'''](\theta^\top xy). \end{aligned}\tag{17}$$

We can now distinguish multiple choices for the function f.

Quadratic: $f(z) = \frac{1}{2}(z-1)^2$. Here, $f'(z) = (z-1)$, $f''(z) = 1$ and $f'''(z) = 0$, thus (17) gives

$$\nabla_x \nabla_\theta \mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right) = x\theta^\top + y(\theta^\top xy - 1)\mathrm{I}_p.$$

So we can derive the following bound:

$$\begin{aligned}\left\|\nabla_x \nabla_\theta \mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right)\right\| &\leq \|x\|\,\|\theta\| + p(\|x\|\,\|\theta\| + 1)\\ &\leq \mathcal{O}\left(p\,\|\mathcal{X}\|\,\|\Theta\|\right),\end{aligned}$$

where the matrix norm is the Frobenius norm.

Exponential: $f(z) = \exp(-z)$. Here, $f'(z) = f'''(z) = -\exp(-z)$ and $f''(z) = \exp(-z)$, thus (17) gives

$$\nabla_x \nabla_\theta \mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right) = e^{-\sigma^2\|\theta\|^2/2} e^{-\theta^\top xy}\left(x\theta^\top - y\mathrm{I}_p + \sigma^2\theta\theta^\top y\right).$$

So we can derive the following bound:

$$\begin{aligned}\left\|\nabla_x \nabla_\theta \mathbb{W}^{-1}_{\sigma^2\|\theta\|^2}[f]\left(\theta^\top xy\right)\right\| &\leq e^{-\sigma^2\|\theta\|^2/2} e^{\|\theta\|\|x\|}\left(\|x\|\,\|\theta\| + p + \sigma^2\|\theta\|^2\right)\\ &\quad\downarrow \text{ Using } \|x\|\,\|\theta\| - \sigma^2\|\theta\|^2/2 \leq \|x\|^2/2\sigma^2.\\ &\leq \exp\left(\frac{\|x\|^2}{2\sigma^2}\right)\left(\|x\|\,\|\theta\| + p + \sigma^2\|\theta\|^2\right)\\ &\leq \exp\left(\frac{\|\mathcal{X}\|^2}{2\sigma^2}\right)\left(p + \|\mathcal{X}\|\,\|\Theta\| + \sigma^2\|\Theta\|^2\right)\\ &\quad\downarrow \text{ Replacing } \sigma^2 \text{ with its expression, the } \|\mathcal{X}\|^2 \text{ simplify.}\\ &\leq \mathcal{O}\left(\exp\left(\frac{\epsilon_x^2}{\log(1.25/\delta)}\right)\left(p + \|\mathcal{X}\|\,\|\Theta\| + \sigma^2\|\Theta\|^2\right)\right),\end{aligned}$$

where the matrix norm is the Frobenius norm.

D.7. In Absence of Closed-form Expression for $\mathbb{W}^{-1}$

When no closed-form expression of $\mathbb{W}^{-1}[f]$ is given, we fall into two cases

1. f is in $\Phi_{M,a}(\mathbb{R}^p)$,
2. f is not in $\Phi_{M,a}(\mathbb{R}^p)$.

In the two cases, we use the following approximation of $\mathbb{W}^{-1}_{2t}[f]$:

$$g_t^K(x) = \sum_{k=0}^{K} \frac{\Delta^k f(x)}{k!}(-t)^k.$$

In case 1 (f is in $\Phi_{M,a}(\mathbb{R}^p)$), we can bound the bias of g_t^K uniformly on $\mathcal{X}$ using the fact that $\lim_{K\to\infty} g_t^K(x) = \mathbb{W}^{-1}_{2t}[f](x)$,

$$\begin{aligned}\text{bias}_K &= \sup_{x\in\mathcal{X}} \left|\mathbb{W}_{2t}\left[g_t^K\right](x) - f(x)\right| = \sup_{x\in\mathcal{X}} \left|\mathbb{W}_{2t}\left[g_t^K - \mathbb{W}^{-1}_{2t}[f]\right](x)\right|\\ &\leq \sup_{x\in\mathcal{X}} \left|g_t^K(x) - \mathbb{W}^{-1}_{2t}[f](x)\right| = \sup_{x\in\mathcal{X}} \left|\sum_{k=K+1}^{\infty} \frac{\Delta^k f(x)}{k!}(-t)^k\right|\\ &\quad\downarrow \text{ By Equation (3), it exist } D > 0 \text{ such that}\\ &\leq D \sup_{x\in\mathcal{X}} A_x(4at)^{K+1}.\end{aligned}$$

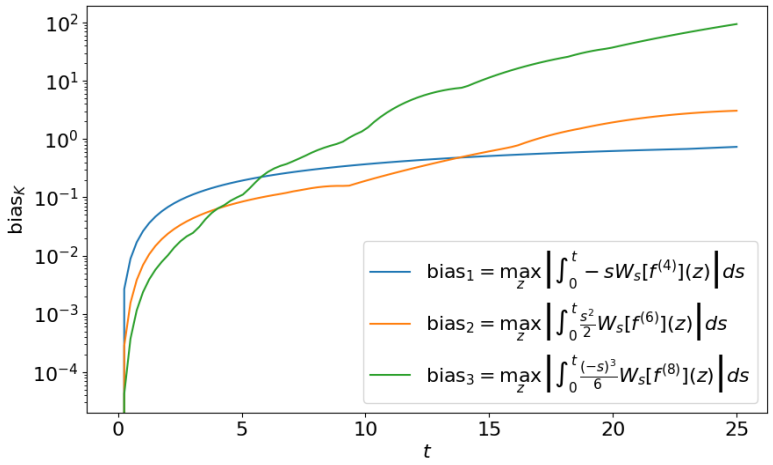


Figure 3. Approximation of the truncation error $bias_K$ for $K \in \{1, 2, 3\}$.

Then, the bias of using g_t^K instead of $\mathbb{W}_{2t}^{-1}[f]$ is exponentially decreasing with K since $a < 1/4t$.
In case 2 (f is not in $\Phi_{M,a}(\mathbb{R}^p)$), we do not have any guarantee that increasing K will result in a better approximation. Instead, we can estimate the $bias_K$ for small $K \in \{1, 2, 3, \dots\}$ and take the optimal truncation (Boyd, 1999):

$$K^* = \arg\min_K \left\{ bias_K = \sup_{x \in \mathcal{X}} \left| \mathbb{W}_{2t}\left[g_t^K\right](x) - f(x) \right| \right\}$$
$$= \arg\min_K \left\{ \sup_{x \in \mathcal{X}} \left| \int_0^t \frac{(-s)^K}{K!} \mathbb{W}_{2s}\left[\Delta^{K+1} f\right](x) ds \right| \right\} .$$

We present this case for the example of the log loss.
Example 4 (log loss). Consider $f\left(\theta^\top xy\right) = \log(1 + \exp(-\theta^\top xy))$ for any $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$. f is not in $\Phi_{M,a}(\mathbb{R}^p)$. Figure 3 shows the estimate of the truncation error ε_K via numerical integration and Monte-Carlo sampling approximation. We restrict the study to $t \leq 25$ which, for the unit ball $\|\mathcal{X}\| \leq 1$, is true for any $\epsilon \geq 1$. For low t, choosing $K \in \{2, 3\}$ can be better and for larger one $K = 1$ is showing a smaller bias.

D.8. Convergence of IWP-SGD – Proof of Theorem 5.5

We provide a proof of the following theorem.

Theorem 5.5 (Convergence guarantees of IWP-SGD)**.** *Let ℓ satisfy Assumption 4.1 and be such that $\mathcal{R}$ satisfies Assumption 5.4. Let the privacy budget be $\epsilon = \epsilon_x + \epsilon_y$, $\delta > 0$ such that $\sigma^2 < 1/4a$. Denote $\theta^* = \arg\min_\theta \mathcal{R}(\theta)$. Assume $\mathcal{X}$ and Θ are bounded convex sets and let C be as defined in Equation (8). For any $n \in \mathbb{N}$ the number of training samples, initial model $\theta_0 \in \Theta$ and step-size $\gamma \leq \frac{1}{2\mathcal{K}}$, Algorithm 1 is (ϵ, δ)-LDP and its output θ_n satisfies*

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leq (1 - \gamma\mu)^n \|\theta_0 - \theta^*\|^2 + \mathcal{O}\left(\frac{\gamma C^2}{\mu} \tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y) - 1\right) \frac{\log(1.25/\delta)}{\epsilon_x^2}\right).$$

In addition, for an appropriate step size $\gamma = \mathcal{O}(\log(n)/n)$,

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leq \tilde{\mathcal{O}}\left(\|\theta_0 - \theta^*\|^2 \exp\left(-\frac{\mu n}{2\mathcal{K}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{C^2}{\mu^2 n} \tilde{S}(\epsilon_y)\left(\tilde{S}(\epsilon_y) - 1\right) \frac{\log(1.25/\delta)}{\epsilon_x^2}\right),$$

where $\tilde{\mathcal{O}}$ hides logarithmic terms in n.

Proof. First, the algorithm is (ϵ, δ)-LDP by the post-processing theorem as the data is first privatized using the (ϵ, δ)-LDP release (1) before being used for the SGD iterations. The rest of the proof is about the convergence guarantees. For any

$t \geq 0$ the iterates of SGD are given by:

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \gamma g_t), \tag{18}$$

where $\gamma > 0$ denotes a step-size and g_t is the IWP gradient estimator computed on the (ϵ, δ)-LDP release defined in Equation (1)

$$g_t = \mathbb{T}^{-1}_{\epsilon,\delta}[\nabla_\theta \ell(\theta, \cdot, \cdot)](\tilde{x}, \tilde{y}).$$

For any $t \geq 0$, the expectation of g_t is

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[g_t \mid \theta_0, \ldots, \theta_t\right] = \nabla\mathcal{R}(\theta_t), \tag{19}$$

and the squared gradient satisfies

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\|g_t\|^2 \mid \theta_0, \ldots, \theta_t\right] = \|\nabla\mathcal{R}(\theta_t)\|^2 + \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}[\|g_t - \nabla\mathcal{R}(\theta_t)\|^2 \mid \theta_0, \ldots, \theta_t] \ .$$

Denote $\mathcal{R}^\star = \min_{\theta\in\Theta}\mathcal{R}(\theta)$. Following Theorem 5.3, and bounding $\|\nabla\mathcal{R}(\theta_t)\|^2 \leq 2\mathcal{K}(\mathcal{R}(\theta_t) - \mathcal{R}^\star)$, we obtain

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\|g_t\|^2 \mid \theta_0, \ldots, \theta_t\right] \\
&\leq \|\nabla\mathcal{R}(\theta_t)\|^2 + 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta, \cdot, y)]\|^2(x)ds \\
&\quad + \frac{2e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta, \cdot, 1) - \nabla_x\nabla_\theta\ell(\theta, \cdot, -1)]\|^2(x)ds \\
&\quad + \frac{e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\|\nabla_\theta\ell(\theta, x, 1) - \nabla_\theta\ell(\theta, x, -1)\|^2 \\
&\leq 2\mathcal{K}(\mathcal{R}(\theta_t) - \mathcal{R}^\star) + 2\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta, \cdot, y)]\|^2(x)ds \\
&\quad + \frac{2e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\int_0^{\sigma^2/2} \mathbb{W}_{2s}\|\mathbb{W}_{2s}^{-1}[\nabla_x\nabla_\theta\ell(\theta, \cdot, 1) - \nabla_x\nabla_\theta\ell(\theta, \cdot, -1)]\|^2(x)ds \\
&\quad + \frac{e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\|\nabla_\theta\ell(\theta, x, 1) - \nabla_\theta\ell(\theta, x, -1)\|^2 \ .
\end{aligned}$$

Using the assumption that $\sup_{\theta,x,y}\sup_{0<s<\sigma^2/2}\|\mathbb{W}_{2s}^{-1}\nabla_\theta\nabla_x\ell(\theta, x, y)\| \leq C$ and $\sup_{\theta,x,y}\|\nabla_\theta\ell(\theta, x, y)\| \leq C$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{(\tilde{x},\tilde{y})}\left[\|g_t\|^2 \mid \theta_0, \ldots, \theta_t\right] \leq 2\mathcal{K}(\mathcal{R}(\theta_t) - \mathcal{R}^\star) + \frac{4C^2e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2} + C^2\sigma^2\left(1 + \frac{4e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\right) \ . \tag{20}$$

We denote $A = \frac{4C^2e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2} + C^2\sigma^2\left(1 + \frac{4e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\right)$ in the following.

Deriving a Recursion. Let $t \geq 0$. Then

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{t+1} - \theta^\star\|^2 \mid \theta_0, \ldots, \theta_t\right] &\overset{(18)}{=} \mathbb{E}\left[\|\Pi_\Theta(\theta_t - \gamma g_t) - \theta^\star\|^2 \mid \theta_0, \ldots, \theta_t\right] \\
&\quad \downarrow \text{ As } \Theta \text{ is a convex bounded set and } \theta^* \in \Theta, \text{ we use contraction of the projection.} \\
&\leq \mathbb{E}\left[\|\theta_t - \gamma g_t - \theta^\star\|^2 \mid \theta_0, \ldots, \theta_t\right] \\
&= \mathbb{E}\left[\|\theta_t - \theta^\star\|^2 - 2\gamma\langle g_t, \theta_t - \theta^\star\rangle + \gamma^2\|g_t\|^2 \mid \theta_0, \ldots, \theta_t\right] \\
&\overset{(19)}{=} \|\theta_t - \theta^\star\|^2 - 2\gamma\langle\nabla\mathcal{R}(\theta_t), \theta_t - \theta^\star\rangle + \gamma^2\mathbb{E}\left[\|g_t\|^2 \mid \theta_0, \ldots, \theta_t\right] \\
&\overset{(20)}{\leq} \|\theta_t - \theta^\star\|^2 - 2\gamma\left(\frac{\mu}{2}\|\theta_t - \theta^\star\|^2 + \mathcal{R}(\theta_t) - \mathcal{R}^\star\right) + \gamma^2\big(2\mathcal{K}(\mathcal{R}(\theta_t) - \mathcal{R}^\star) + \gamma^2 A,
\end{aligned}$$

where we also used μ-strong convexity in the last inequality. By re-arranging and taking expectation on both sides, we get:

$$\mathbb{E}\|\theta_{t+1} - \theta^\star\|^2 \leq (1 - \mu\gamma)\mathbb{E}\|\theta_t - \theta^\star\|^2 - 2\gamma(1 - \mathcal{K}\gamma)(\mathbb{E}\mathcal{R}(\theta_t) - \mathcal{R}^\star) + \gamma^2 A,$$

and by observing $(1 - \mathcal{K}\gamma) \geq \frac{1}{2}$ for $\gamma \leq \frac{1}{2\mathcal{K}}$,

$$\mathbb{E}\|\theta_{t+1} - \theta^\star\|^2 \leq (1 - \mu\gamma)\mathbb{E}\|\theta_t - \theta^\star\|^2 - \gamma(\mathbb{E}\mathcal{R}(\theta_t) - \mathcal{R}^\star) + \gamma^2 A \ . \tag{21}$$

Unrolling the Recurrence. We can relax (21) to $\mathbb{E}\|\theta_{t+1} - \theta^\star\|^2 \leq (1-\mu\gamma)\mathbb{E}\|\theta_t - \theta^\star\|^2 + \gamma^2 A$ and obtain after unrolling the recurrence for any $n \geq 1$,

$$\begin{aligned}
\mathbb{E}\|\theta_n - \theta^\star\|^2 &\leq (1-\mu\gamma)^n \|\theta_0 - \theta^\star\|^2 + \gamma^2 A \sum_{i=0}^{n-1} (1-\mu\gamma)^i \\
&\leq (1-\mu\gamma)^n \|\theta_0 - \theta^\star\|^2 + \frac{\gamma}{\mu} A \\
&\leq (1-\mu\gamma)^n \|\theta_0 - \theta^\star\|^2 + \frac{\gamma C^2}{\mu}\left(\frac{4e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2} + \sigma^2\left(1 + \frac{4e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}\right)\right) \quad . \qquad (22)
\end{aligned}$$

This intermediate results shows that SGD with constant stepsizes reduces the initial error term $\|\theta_0 - \theta^\star\|^2$ linearly, but only converges towards a $\mathcal{O}\left(\frac{\gamma}{\mu}\frac{C^2 e^{\epsilon_y}}{(e^{\epsilon_y}-1)^2}(\sigma^2+1)\right)$-neighborhood of $\theta^\star$.

Choosing the Step-size. To obtain a convergence guarantee that holds for arbitrary accuracy, we need to choose the stepsize γ carefully:

- If $\frac{1}{2\mathcal{K}} \geq \frac{1}{\mu n}\log\max\left(2, \frac{\mu^2\|\theta_0-\theta^\star\|^2 n}{A}\right)$ then we choose $\gamma = \frac{1}{\mu n}\log\max\left(2, \frac{\mu^2\|\theta_0-\theta^\star\|^2 n}{A}\right)$.
- If otherwise $\frac{1}{2\mathcal{K}} < \frac{1}{\mu n}\log\max\left(2, \frac{\mu^2\|\theta_0-\theta^\star\|^2 n}{A}\right)$ then we pick $\gamma = \frac{1}{2\mathcal{K}}$.

With these choices of γ, we can show

$$\begin{aligned}
\mathbb{E}\|\theta_n - \theta^\star\|^2 &= \tilde{\mathcal{O}}\left(\|\theta_0 - \theta^\star\|^2 \exp\left[-\frac{\mu n}{2\mathcal{K}}\right] + \frac{A}{\mu^2 n}\right) \\
&= \tilde{\mathcal{O}}\left(\|\theta_0 - \theta^\star\|^2 \exp\left[-\frac{\mu n}{2\mathcal{K}}\right] + \frac{C^2 e^{\epsilon_y}}{\mu^2 n (e^{\epsilon_y}-1)^2}(\sigma^2+1)\right). \qquad (23)
\end{aligned}$$

Where the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors in n. Replacing σ^2 with its value $\frac{8C^2\log(1.25/\delta)}{\epsilon_x^2}$ in Equations (22) and (23) yields the desired results. □

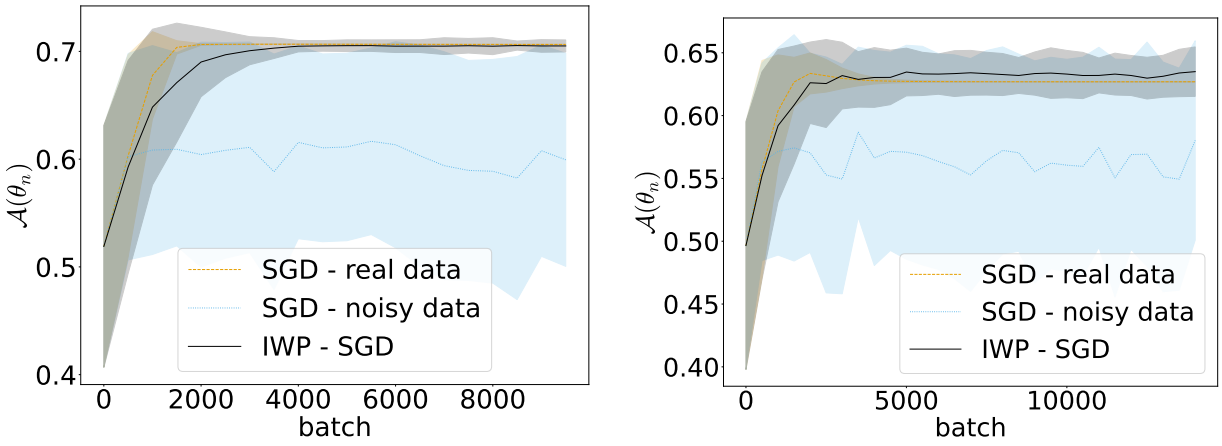
E. Experiments

In this section, we give more details about experiments of Section 6. Given a test dataset of m samples $D'_m = \{(x_i, y_i)\}_{i=1}^m$, the accuracy of the linear classification model θ on the test set is denoted $\mathcal{A}(\theta)$ and defined as

$$\mathcal{A}(\theta) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}(y_i \theta^\top x_i > 0).$$

Synthetic Data. Recall that we study two synthetic binary classification problems in dimension $p = 2$ and $p = 10$ generated with the `make_classification` routine of `scikit-learn` having features within $[-1, 1]^p$. We conduct the experiments on $n = 10^6$ samples for two privacy guarantees : $(2, 10^{-5})$-LDP for $p = 2$ and $(5, 10^{-5})$-LDP for $p = 10$. The ℓ_2 regularization constant is $\lambda = 5$ with the regularized loss $\ell(\theta, x, y) + \lambda\|\theta\|^2/2$. We average batches of size 128 and use a common learning rate of $\gamma = 10^{-4}$.

Real Data. Recall that, we study the ACSIncome and ACSPublicCoverage problems of the Folktables dataset (Ding et al., 2021). Both are based on ACS data (like UCI Adult), illustrating the fact that we can reuse, in a task-agnostic way, the same private releases when reusing the same data points. ACSIncome consists of predicting whether an individual's income is above $50 000 and ACSPublicCoverage consists of predicting individual coverage from health insurance. For both problems, we select the two variables *AGEP* (age in years) and *SCHL* (educational attainment). For ACSIncome we add *WKHP* (usual hours worked per week over the past year) and for ACSPublicCoverage we add *PINCP* (total annual income). All features are continuous or ordinal, allowing the use of the Gaussian mechanism. We merge the data of the five



(a) ACSPublicCoverage *(b)* ACSIncome

Figure 4. Comparison of Accuracy convergence of the model fitted on exp loss under $(2, 10^{-5})$-LDP on ACSPublicCoverage and ACSIncome.

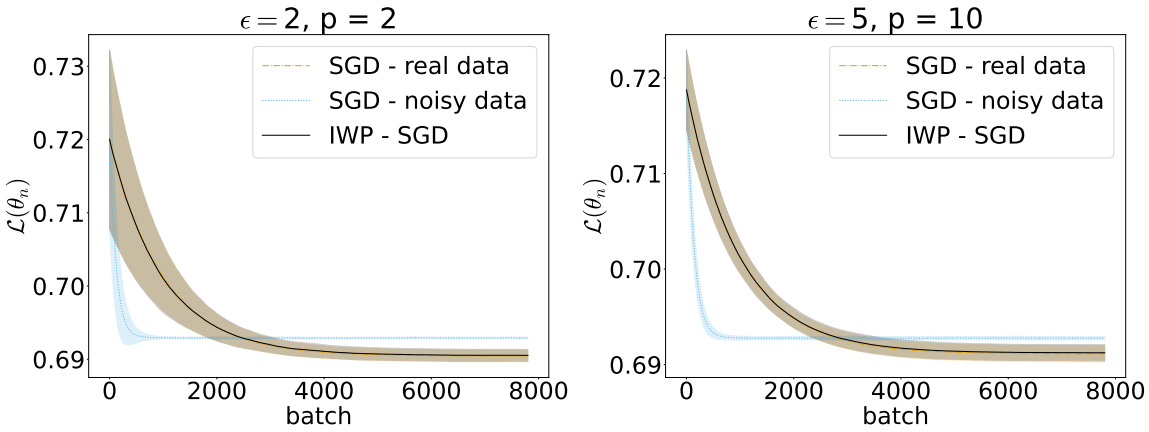


Figure 5. Comparison of SGD convergence of the log loss under $(2, 10^{-5})$-LDP for the 2-dimensional synthetic data and $(5, 10^{-5})$-LDP for the 10-dimensional synthetic data.

largest states yielding datasets of respectively 668 859 rows and 883 984 rows for ACSIncome and ACSPublicCoverage. The data is then randomly split into training (80%) and test (20%) sets. The ℓ_2 regularization constant is $\lambda = 10$ with the regularized loss $\ell(\theta, x, y) + \lambda \|\theta\|^2 / 2$. We average batches of size 50 and use a common learning rate of $\gamma = 2 \cdot 10^{-5}$ for ACSPublicCoverage and ACSIncome. Figure 4 is showing the accuracy convergence across batches for these experiments.

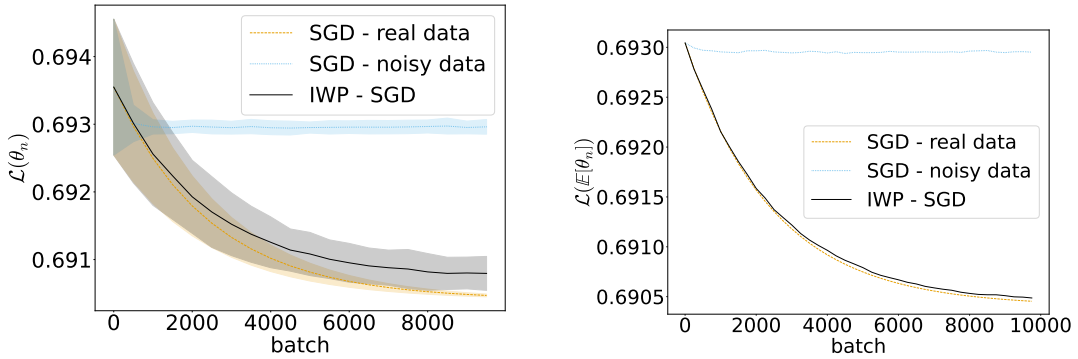
E.1. Experiments Using the Log Loss

Using the approximation of $\mathbb{W}^{-1}$ described in Appendix D.7, we applied our experiments on synthetic and real-world datasets to the log loss (with same batch sizes, regularization constants, and learning rates). Figures 5, 6 and 7 show similar results compared to the experiments on the exponential loss in Section 6.

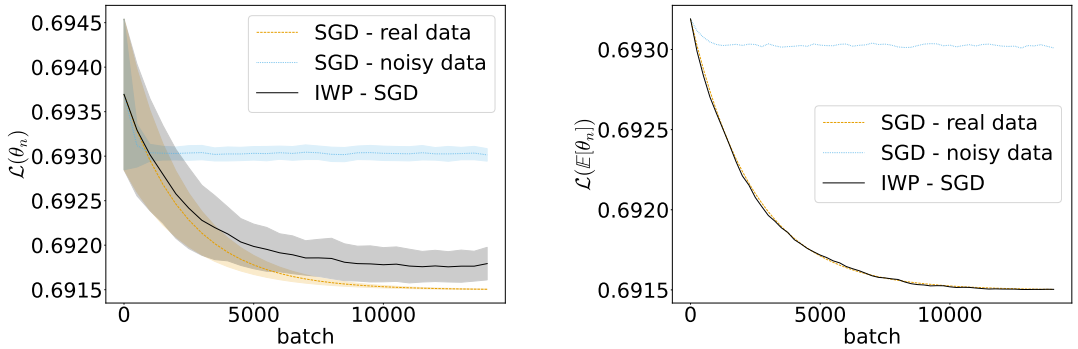
E.2. Experiments on Regression

The model presented on the paper can be generalized to regression. In this case, we only use the Weierstrass transform because the target is also continuous, and we have:

$$\mathbb{T}_{\epsilon,\delta}[h](x, y) = \mathbb{W}_{\sigma^2_{\epsilon_x,\delta_x}}\left[\mathbb{W}_{\sigma^2_{\epsilon_y,\delta_y}}[h](\cdot, y)\right](x)$$

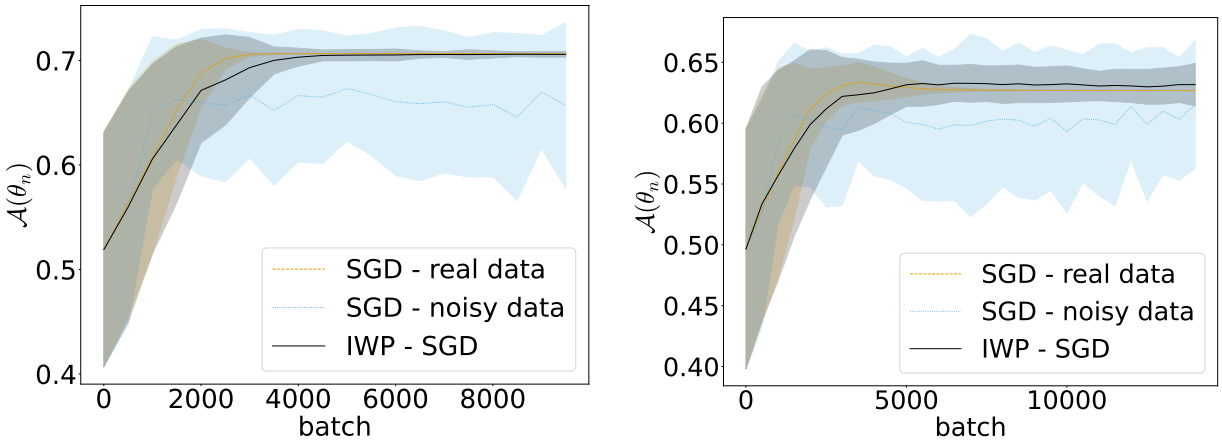


(a) ACSPublicCoverage



(b) ACSIncome

Figure 6. Comparison of SGD convergence of the log loss under $(2, 10^{-5})$-LDP on ACSPublicCoverage and ACSIncome.



(a) ACSPublicCoverage

(b) ACSIncome

Figure 7. Comparison of Accuracy convergence of the model fitted on log loss under $(2, 10^{-5})$-LDP on ACSPublicCoverage and ACSIncome.

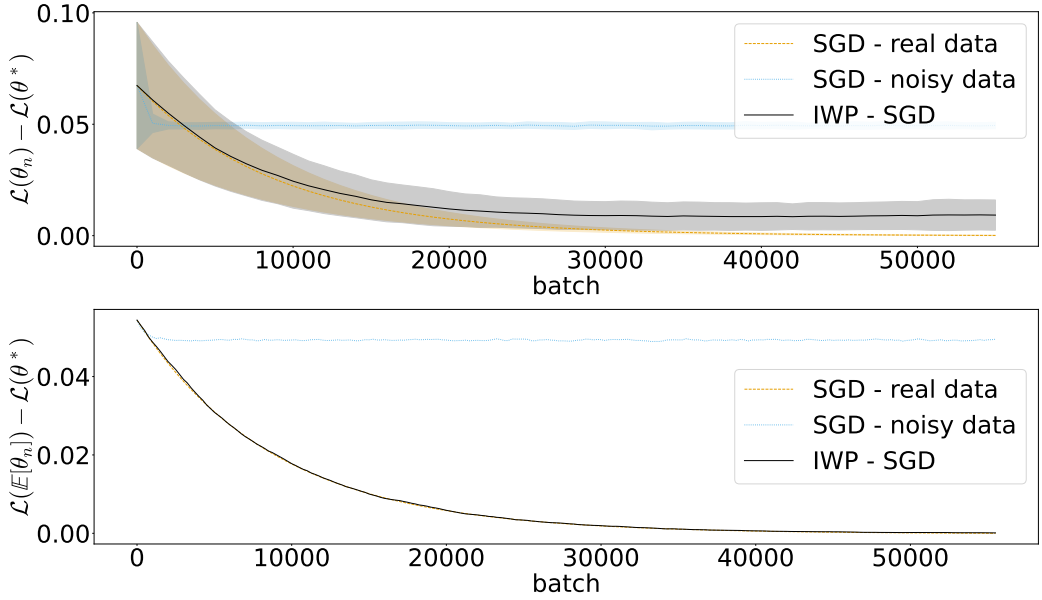


Figure 8. Comparison of SGD convergence under $(2, 10^{-5})$-LDP on ACSIncome linear regression variant.

with $\epsilon = \epsilon_x + \epsilon_y$ and $\delta = \delta_x + \delta_y$. In the linear regression model where $\mathcal{Y} \subset \mathbb{R}$, we have $\ell(\theta, x, y) = \frac{1}{2}(\theta^\top x - y)^2$ and there is no bias in the gradient to correct with respect to the labels. Indeed, it is linear with respect to y:

$$\nabla_\theta \ell(\theta, x, y) = x\theta^\top x - xy.$$

We then study a variant of the ACSIncome consisting in the prediction of the individual's income as a continuous value instead of the threshold at \$50 000. For this, we adapt our method to continuous output by replacing the Randomized Response transform by a second Weierstrass transform and we consider the Mean Square Error for the loss. We use ℓ_2 regularization with the constant $\lambda = 10$ forming a regularized loss $\ell(\theta, x, y) + \lambda \|\theta\|^2 / 2$. We average batches of size 128 and use a learning rate of $\gamma = 5 \cdot 10^{-6}$. Figure 8 shows the results of this experiment. The conclusions are the same as in binary classification, IWP-SGD converges to the same model as SGD - real data but with an increased variance whereas SGD - noisy data converges to a different solution, illustrating the presence of a bias.