Rényi Pufferfish Privacy: General Additive Noise Mechanisms and Privacy Amplification by Iteration via Shift Reduction Lemmas

Clément Pierquin [1 2] **Aurélien Bellet** [3] **Marc Tommasi** [2] **Matthieu Boussard** [1]

Abstract

Pufferfish privacy is a flexible generalization of differential privacy that allows to model arbitrary secrets and adversary's prior knowledge about the data. Unfortunately, designing general and tractable Pufferfish mechanisms that do not compromise utility is challenging. Furthermore, this framework does not provide the composition guarantees needed for a direct use in iterative machine learning algorithms. To mitigate these issues, we introduce a Rényi divergence-based variant of Pufferfish and show that it allows us to extend the applicability of the Pufferfish framework. We first generalize the Wasserstein mechanism to cover a wide range of noise distributions and introduce several ways to improve its utility. Finally, as an alternative to composition, we prove privacy amplification results for contractive noisy iterations and showcase the first use of Pufferfish in private convex optimization. A common ingredient underlying our results is the use and extension of shift reduction lemmas.

1. Introduction

Differential privacy (DP) (Dwork & Roth, 2014) is now considered as the gold standard for privacy-preserving data analysis. However, despite its many desirable properties, DP does not suit all types of data effectively. Specifically, the guarantees it offers are based on the underlying assumption that individuals in the dataset being analyzed are statistically independent. In reality, data often exhibit correlations, and when two correlated individuals are present in a dataset, performing the same analysis with and without one of these individuals could leak more knowledge about the individual than the conventional differential privacy framework assumes (Humphries et al., 2023).

To address these situations, specialized privacy definitions have been designed. Certain direct extensions of DP, like group privacy (Dwork & Roth, 2014) or entry privacy (Hardt & Roth, 2013), protect entire instances or groups, which results in strong privacy guarantees but often much poorer utility. More flexible frameworks allow to tailor the privacy definition to a set of distributions which could have plausibly generated the dataset, and thereby allow a tighter privacy analysis. In this work, we focus on the general framework of Pufferfish privacy (Kifer & Machanavajjhala, 2014), which is closely related to other similar definitions like Blowfish privacy (He et al., 2014) and distribution privacy (Kawamoto & Murakami, 2019; Chen & Ohrimenko, 2023).

Pufferfish privacy however comes with new challenges, first and foremost in the design of general and computationally tractable Pufferfish private mechanisms. Indeed, the sensitivity of the query, which is critical in DP to design additive noise mechanisms, has no direct use in Pufferfish privacy. Moreover, while various ways to measure and efficiently track the privacy loss have been proposed for DP, see for instance Rényi differential privacy (RDP) (Mironov, 2017), this flexibility is lacking in Pufferfish privacy. As a result, previous work on the design of Pufferfish mechanisms has focused on specific noise distributions and applications (Kifer & Machanavajjhala, 2014; Ou et al., 2018; Kessler et al., 2015; Niu et al., 2019; Song et al., 2017). For instance, Song et al. (2017) proposed the Wasserstein mechanism for the Laplace noise, which relies on the computation of ∞-Wasserstein distances. Another recent work proposes an exponential mechanism-based approach which provides a more computationally tractable approach but relies on (potentially loose) sufficient conditions for Pufferfish privacy (Ding, 2022). The Pufferfish framework thus lacks a unified theory that subsumes the original worst-case definition and allows for the design of general additive mechanisms compatible with a wide range of noise distributions.

Another key limitation of Pufferfish privacy is that it does not always compose when the same data is used across multiple computations. Existing sequential and adaptive compositions theorems hold only for some particular Pufferfish

[1]Craft AI, Paris, France [2]Université de Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France [3] Inria, Université de Montpellier. Correspondence to: Clément Pierquin <clement.pierquin@craft-ai.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

instantiations and mechanisms, sometimes without a closed-form that can be used in practice (Kifer & Machanavajjhala, 2014; Nuradha & Goldfeld, 2023). For instance, a sequential (but non-adaptive) composition result exists for the Markov Quilt Mechanism (Song et al., 2017), but it is limited to Bayesian networks. The lack of a universal adaptive composition theorem currently makes Pufferfish privacy unfit for the analysis of iterative algorithms such as those used in differentially private machine learning (Abadi et al., 2016).

In this paper, we mitigate the above limitations of Pufferfish privacy by making the following contributions:

- We define the Rényi Pufferfish privacy framework and show that it preserves the main desirable properties of Pufferfish while providing additional flexibility.
- We introduce the General Wasserstein Mechanism (GWM), a generalization of the Wasserstein mechanism of Song et al. (2017). Our mechanism allows to derive (Rényi) Pufferfish privacy guarantees for all additive noise distributions that are absolutely continuous with respect to the Lebesgue measure.
- We propose two ways to improve the utility of GWM by relaxing the ∞-Wasserstein distance used to calibrate the noise. Our first approach relies on a δ-approximation allowing the tail of the distribution of the mechanism to be disregarded, similar to what has been proposed by Chen & Ohrimenko (2023) for the distribution privacy framework. Incidentally, we demonstrate an equivalence between Pufferfish privacy and distribution privacy. Our second approach enables the use of p-Wasserstein distances, yielding the first general Pufferfish mechanism with better utility than the Wasserstein mechanism at the same privacy cost.
- Inspired by Feldman et al. (2018), we prove privacy amplification by iteration results for Pufferfish, allowing to bypass the use of composition in the analysis of contractive noisy iterations. This technique is particularly useful to analyze convex optimization with stochastic gradient descent, allowing the integration of Pufferfish privacy in machine learning pipelines.
- We provide examples of concrete instantiations of our framework where the proposed mechanisms are computationally efficient and provide better utility than (Group) DP.

One of our key technical contributions lies in the novel use and generalization of shift reduction lemmas (Feldman et al., 2018; Altschuler & Talwar, 2022) in the context of Pufferfish privacy. We argue that shift reduction is the right tool to analyze Pufferfish privacy, and believe this view may yield more results in the future.

All proofs and some additional content can be found in the supplementary material.

2. Rényi Pufferfish Privacy

Rényi differential privacy (RDP) ensures that an adversary cannot gain too much knowledge about whether an individual point is in the dataset or not by observing the output of the mechanism. In the original definition, it is implied that the elements of the dataset are statistically independent (see Appendix A.1 for definitions). A more general framework, Pufferfish privacy, has been designed to handle possibly correlated data and other types of secrets than the presence of an individual in a dataset (Kifer & Machanavajjhala, 2014). In a Pufferfish instantiation, we denote by $\mathcal{S}$ the set of possible secrets to be protected, and by $\mathcal{Q} \subseteq \mathcal{S}^2$ the specific pairs of secrets we aim to make indistinguishable. In contrast to differential privacy, the variable X representing the dataset is not deterministic in Pufferfish privacy. Instead, it is sampled from a certain distribution $\theta \in \Theta$. The set Θ represents the possible prior knowledge of an adversary.

Definition 2.1 (Pufferfish privacy, PP (Kifer & Machanavajjhala, 2014; Ding, 2022))**.** Let $\varepsilon \geq 0$ and $\delta \in (0, 1)$. A privacy mechanism $\mathcal{M}$ is said to be (ε, δ)-Pufferfish private in a framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $\theta \in \Theta$, for all secret pairs $(s_i, s_j) \in \mathcal{Q}$, and for all $w \in \mathrm{Range}(\mathcal{M})$, we have:

$$P\left(\mathcal{M}(X) = w \mid s_i, \theta\right) \leq e^{\varepsilon} P\left(\mathcal{M}(X) = w \mid s_j, \theta\right) + \delta,$$

where $X \sim \theta$ and (s_i, s_j) is such that $P\left(s_i \mid \theta\right) \neq 0, P\left(s_j \mid \theta\right) \neq 0$. If $\delta = 0$, $\mathcal{M}$ satisfies ε-Pufferfish privacy.

In this work, we introduce a Rényi divergence-based version of Pufferfish privacy. Using Rényi divergences in privacy definitions has several advantages. Especially relevant to our work will be the quantification of privacy guarantees by bounding certain moments of the exponential of the privacy loss (Mironov, 2017), and the ability to leverage a large body of results on Rényi divergences such as shift reduction lemmas (Feldman et al., 2018; Altschuler & Chewi, 2023).

Definition 2.2 (Rényi Pufferfish privacy, RPP)**.** Let $\alpha > 1$ and $\varepsilon \geq 0$. A privacy mechanism $\mathcal{M}$ is said to be (α, ε)-Rényi Pufferfish private in a framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $\theta \in \Theta$ and for all secret pairs $(s_i, s_j) \in \mathcal{Q}$, we have:

$$D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta\right), P\left(\mathcal{M}(X) \mid s_j, \theta\right)\right) \leq \varepsilon,$$

where $X \sim \theta$, (s_i, s_j) is such that $P\left(s_i \mid \theta\right) \neq 0$ and $P\left(s_j \mid \theta\right) \neq 0$, and $D_\alpha(\mu, \nu) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim \nu}\left[\left(\frac{\mu(x)}{\nu(x)}\right)^{\alpha}\right]$ is the Rényi divergence of order α.

Rényi Pufferfish privacy upholds the post-processing inequality, which is a key attribute for any effective privacy framework.

Proposition 2.1 (Post-processing). *Let $\mathcal{M}_1$ be a randomized algorithm and $\mathcal{M}$ be (α, ε)-RPP. Then,*

$$D_\alpha\left(P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_i, \theta\right), P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_j, \theta\right)\right) \leq D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta\right), P\left(\mathcal{M}(X) \mid s_j, \theta\right)\right) \leq \varepsilon.$$

It is easy to see that (∞, ε)-RPP corresponds to ε-PP. Furthermore, (α, ε)-RPP can be converted to (ε, δ)-PP.

Proposition 2.2 (RPP implies PP). *If $\mathcal{M}$ is (α, ε)-RPP, it also satisfies $\left(\varepsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-PP $\forall \delta \in (0, 1)$.*

Guarantees against close adversaries. In Pufferfish, the set Θ represents the possible beliefs of the adversary. It needs to be large enough to prevent harmful privacy leaks, but there is also a no free lunch theorem that states that if Θ is too large then the resulting mechanism will have poor utility (Kifer & Machanavajjhala, 2014). Hence, it is important to quantify the privacy protection offered by a mechanism $\mathcal{M}$ when the belief θ' of the adversary is not in Θ. This question has been addressed for ϵ-PP by Song et al. (2017). The theorem derived by Song et al. (2017), which we recall in Appendix A.4 for completeness, shows that if θ' is Δ-close to some $\theta \in \Theta$, then $\mathcal{M}$ retains its Pufferfish privacy guarantees for θ' up to an additive penalty 2Δ. However, Δ is measured in ∞-Rényi divergence, which corresponds to a worst-case scenario, and can thus be very large. We extend this result to our RPP framework, allowing the use of α-Rényi divergences (see Appendix A.4). Our result can provide better privacy guarantees in situations where the original one gives poor guarantees.

Running examples. We introduce here some examples of RPP instantiations which we will use throughout the paper to illustrate our private mechanisms. Let $n > 0$ be the total number of participants in a study. Let $\mathcal{X}$ be the potential values of an individual's private features. Let $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ describing the private properties of the n individuals. An adversary anticipates correlations among individuals within the study with a prior $\theta \in \Theta$. We define the set of secrets for this adversary as $\mathcal{S} = \{s_i^a \triangleq \{X_i = a\};\ a \in \mathcal{X}, i \in [\![1, n]\!]\}$ and define $\mathcal{Q} = \{(s_i^a, s_j^b);\ a, b \in \mathcal{X}, i, j \in [\![1, n]\!]\}$. Consider the following simple instantiations of this setting for datasets of size 2:

Example 1 *(Counting query with correlation). Each individual i holds a binary value $X_i \in \{0, 1\}$ and we consider a counting query $f(X) = X_1 + X_2$. For $p \in (0, 1), \rho \in [-1, 1]$, the adversary has the following prior: $P(X_1 = 1) = P(X_2 = 1) = p$, where X_1 and X_2 are drawn with correlation ρ.*

Example 2 *(Average salary query). Each individual i holds her salary $X_i \geq 0$ and we consider an average query $f(X) = \frac{1}{2}(X_1 + X_2)$. The adversary has the following prior for the marginals: for $i \in \{1, 2\}$,*

$$X_i = \begin{cases} 1 & \text{with prob. } 1/2 \\ 2 & \text{with prob. } 499/1000, \quad \text{for } i \in \{1, 2\} \\ 100 & \text{with prob. } 1/1000 \end{cases}$$

Here, X_1 and X_2 are thus considered independent.

Example 3 *(Sum query with user-dependent prior). We consider $\mathcal{X} = (0, r)$ and a sum query $f(X) = X_1 + X_2$. The adversary has an arbitrary prior about the distribution of (X_1, X_2) but assumes that each individual i holds a different value $X_i \in (0, r_i)$ with $0 < r_i \leq r$.*

3. A General Additive Mechanism for Rényi Pufferfish Privacy

In this section, we present a general approach to obtain Rényi Pufferfish privacy guarantees. Specifically, we introduce the General Wasserstein Mechanism (GWM), a generalization of the Laplacian-based Wasserstein mechanism of Song et al. (2017) to a wide range of noise distributions, and derive the corresponding RPP guarantees. We also highlight that the shift reduction lemma and its variants, introduced by Feldman et al. (2018) in the context of privacy amplification by iteration, provide the right framework for analyzing Rényi Pufferfish privacy.

We first introduce ∞-Wasserstein distances and couplings.

Definition 3.1 (Couplings). Let μ and ν be two distributions on a measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with $\mathcal{B}(\mathbb{R}^d))$ the Borel σ-algebra. A coupling π is a joint distribution on the product space $(\mathbb{R}^{d\times 2}, \mathcal{B}(\mathbb{R}^d)^2)$ with marginals μ and ν, where $\mathcal{B}(\mathbb{R}^d)^2$ is the product σ-algebra.

Definition 3.2 (∞-Wasserstein distance). Let μ and ν be two distributions on $\mathbb{R}^d$. We note Γ the set of the couplings between μ and ν. We define the ∞-Wasserstein distance between μ and ν as:

$$W_\infty(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \sup_{(x,y) \in supp(\pi)} \|x - y\|.$$

Throughout the paper, $\|\cdot\|$ represents a norm of $\mathbb{R}^d$. When necessary, in later results, the type of norm will be specified.

We now recall the shift reduction lemma, a result that allows to split the Rényi divergence between two noised distributions into two distinct components: one involving the two original distributions, and one involving the noise. Let μ, ν, ζ be three distributions on $\mathbb{R}^d$ and $z, a \geq 0$. We define the following quantities:

$$D_\alpha^{(z)}(\mu, \nu) = \inf_{W_\infty(\mu,\mu') \leq z} D_\alpha(\mu', \nu),$$

$$R_\alpha(\zeta, z) = \sup_{\|x\| < z} D_\alpha(\zeta_{-x}, \zeta),$$

where $\zeta_{-x} : y \mapsto \zeta(y-x)$, and denote by $*$ the convolution product.

Lemma 3.1 (Shift reduction (Feldman et al., 2018))**.** *Let μ, ν, ζ be three distributions on $\mathbb{R}^d$ and $z, a \geq 0$. Then,*

$$D_\alpha^{(a)}(\mu * \zeta, \nu * \zeta) \leq D_\alpha^{(z+a)}(\mu, \nu) + R_\alpha(\zeta, z).$$

We now show that the shift reduction lemma allows to obtain a unified approach for RPP analysis. In fact, it gives a closed formula for the privacy guarantees of releasing a query with additive noise. This yields our General Wasserstein Mechanism (GWM) and its associated privacy guarantees.

Theorem 3.3 (General Wasserstein mechanism, GWM)**.** *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a numerical query and denote:*

$$\Delta_G = \max_{\substack{(s_i,s_j)\in S \\ \theta\in\Theta}} W_\infty\left(P(f(X)|s_i,\theta), P(f(X)|s_j,\theta)\right).$$

Let $N = (N_1, \ldots, N_d) \sim \zeta$ drawn independently of the dataset X. Then, $\mathcal{M}(X) = f(X) + N$ satisfies $(\alpha, R_\alpha(\zeta, \Delta_G))$-RPP for all $\alpha \in (1, +\infty)$ and $R_\infty(\zeta, \Delta_G)$-PP.

While Theorem 3.3 is very general, we can easily derive explicit results for specific choices of noise distributions. Instantiating GWM with Laplacian noise, we recover the results of Song et al. (2017) for PP as a special case where $d = 1$. More interestingly, we also directly obtain a novel Gaussian mechanism and a novel Laplacian mechanism for RPP.

Corollary 3.1 (Privacy guarantees for usual noise distributions)**.** *We note I_d the identity matrix of size d. Plugging the expressions of $R_\infty(\zeta, z)$ and $R_\alpha(\zeta, z)$ for Laplacian and Gaussian distributions, we obtain:*

- *$\mathcal{M}(X) = f(X) + N$ with $N \sim \mathcal{N}\left(0, \frac{\alpha\Delta_G^2}{2\varepsilon} I_d\right)$ and Δ_G computed w.r.t. the l_2 norm is (α, ε)-RPP.*
- *$\mathcal{M}(X) = f(X) + L$ with $L \sim \mathrm{Lap}(0, \rho I_d)$ and Δ_G computed w.r.t. the l_1 norm is $\left(\alpha, \frac{1}{\alpha-1}\log\left(\frac{\alpha}{2\alpha-1}e^{\Delta_G(\alpha-1)/\rho} + \frac{\alpha-1}{2\alpha-1}e^{-\Delta_G\alpha/\rho}\right)\right)$-RPP.*
- *$\mathcal{M}(X) = f(X) + L$ with $L \sim \mathrm{Lap}\left(0, \frac{\Delta_G}{\varepsilon} I_d\right)$ with Δ_G computed w.r.t. the l_1 norm is ε-PP.*

The results of Corollary 3.1 are analogous to the results of Mironov (2017) for RDP, where the sensitivity of the query is replaced by Δ_G. It enables us to directly compare the utility of a RDP mechanism in the group privacy setting and the GWM in RPP. Considering Example 3, we have $\Delta_G \leq r_1 + r_2$, which is smaller than $\Delta_{\text{GROUP}} = 2r$. Therefore, GWM achieves better utility than group RDP in this case. This observation can be generalized to other settings as the utility guarantees of the Wasserstein mechanism of Song et al. (2017) extend to the GWM.

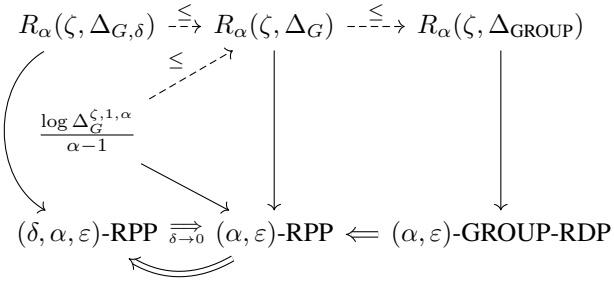

Figure 1. Relations between the mechanisms and privacy notions studied in the paper. The values on the top of the graph represent the value ε of the privacy budget guaranteed by the mechanisms. Δ_G corresponds to the sensitivity of the GWM (Section 3), and $\Delta_{G,\delta}$ corresponds to the sensitivity of the GAWM (Section 4.1), $\Delta_G^{\zeta,1,\alpha}$ corresponds to the sensitivity of the DAGWM (Section 4.2). Δ_{GROUP} corresponds to the sensitivity of mechanisms in the group privacy framework. The plain arrows indicate the privacy guarantees offered by the mechanisms. The dashed arrows compare the privacy budget offered by the mechanisms. The implication arrows illustrate the relations between the different frameworks.

Proposition 3.1 (Utility of the GWM, informal)**.** *Under mild conditions, an additive mechanism offers better utility in the GWM setting than in the group privacy setting (see Appendix B.3 for details).*

One drawback of GWM is that in some cases, Δ_G may be large, as it depends on ∞-Wasserstein distances. In Example 1, $\Delta_G = \Delta_{\text{GROUP}} = 2$, thus GWM gives no utility advantage compared to group RDP. In Example 2, $\Delta_G = 98$ is large although the event $X_i = 100$ is rare. We deal with this issue in the next section.

4. Improving Utility by Relaxing the W_∞ Constraint

In this section, we propose two ways to improve the utility of GWM by relaxing the ∞-Wasserstein constraint in the calibration of the noise. Figure 4 summarizes the relations between the different mechanisms and privacy definitions that we introduce.

4.1. δ-Approximation of (α, ε)-RPP

Our first approach is to define an approximation of Rényi Pufferfish Privacy that allows a low probability set of values to be disregarded.

Definition 4.1 (Approximate Rényi Pufferfish privacy)**.** A privacy mechanism $\mathcal{M}$ is said to be $(\alpha, \varepsilon, \delta)$-approximate Rényi Pufferfish private in a framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $\theta \in \Theta$ and for all secret pairs $(s_i, s_j) \in \mathcal{Q}$, there exists

E, E' such that $P(E) \geq 1-\delta, P(E') \geq 1-\delta$ and:

$$D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta, E\right), P\left(\mathcal{M}(X) \mid s_j, \theta, E'\right)\right) \leq \varepsilon,$$
$$D_\alpha\left(P\left(\mathcal{M}(X) \mid s_j, \theta, E'\right), P\left(\mathcal{M}(X) \mid s_i, \theta, E\right)\right) \leq \varepsilon,$$

where $X \sim \theta$ and (s_i, s_j) is such that $P(s_i \mid \theta) \neq 0, P(s_j \mid \theta) \neq 0$.

Note that similar privacy definitions have been proposed for versions of differential privacy in (Bun & Steinke, 2016, Definition 8.1) and (Papernot & Steinke, 2021, Definition 18). This definition implies (ε, δ)-PP when $\alpha \to +\infty$. It also implies $(\varepsilon', 2\delta)$-RPP for a specific value ε'.

Proposition 4.1. *If $\mathcal{M}$ is $(\alpha, \varepsilon, \delta)$-approximate RPP, then it is $(\varepsilon', 2\delta)$-PP, with $\varepsilon' = \varepsilon + \frac{\log(1/\delta)}{\alpha-1}$.*

We now design an approximate Wasserstein mechanism for Rényi Pufferfish privacy. To do so, we rely on the notion of (z, δ)-proximity (named *closeness* in Chen & Ohrimenko, 2023).

Definition 4.2 ((z, δ)-proximity)**.** Let μ, ν two distributions on $\mathbb{R}^d$ and $z \geq 0, \delta \in (0,1)$. We say that μ and ν are (z, δ)-near if there exists a coupling π between μ and ν and $\mathcal{R} \subset supp(\pi)$ such that $\int_{\mathcal{R}} d\pi(x,y) \geq 1-\delta$ and $\forall (x,y) \in \mathcal{R}, \|x-y\| \leq z$.

We also need to extend the shift reduction lemma of Feldman et al. (2018) to account for shifts that are (z, δ)-near to the original distribution μ, instead of shifts μ' such that $W_\infty(\mu, \mu') \leq z$.

Lemma 4.1 (Approximate shift reduction)**.** *Let μ, ν, ζ be three distributions on $\mathbb{R}^d$. We denote $D_\alpha^{(z,\delta)}(\mu, \nu) = \inf_{\mu, \mu' \ (z,\delta)\text{-near}} D_\alpha(\mu', \nu)$. Then, for all $\delta \in (0,1)$, there exists an event E such that $P(E) \geq 1-\delta$ and:*

$$D_\alpha\left((\mu * \zeta)_{|E}, (\nu * \zeta)\right) \leq D_\alpha^{(z,\delta)}(\mu, \nu) + R_\alpha(\zeta, z) + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{1-\delta}\right).$$

This approximate shift reduction lemma provides a general mechanism to achieve approximate RPP.

Theorem 4.3 (General approximate Wasserstein mechanism, GAWM)**.** *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a numerical query. For all $\delta \in (0,1)$, let us denote:*

$$\Delta_{G,\delta} > \inf\{z \in \mathbb{R}; \forall (s_i, s_j) \in S, \forall \theta \in \Theta, (P((f(X)|s_i, \theta), P(f(X)|s_j, \theta)) \text{ are } (z,\delta)\text{-near}\}.$$

Let $N = (N_1, \ldots, N_d) \sim \zeta$ drawn independently of the dataset X. Then, $\mathcal{M} = f(X) + N$ satisfies $(\alpha, R_\alpha(\zeta, \Delta_{G,\delta}) + \frac{\alpha}{\alpha-1}\log\frac{1}{1-\delta}, \delta)$-approximate RPP for all $\alpha \in (1, +\infty)$ and $(R_\infty(\zeta, \Delta_{G,\delta}) + \log\frac{1}{1-\delta}, \delta)$-PP.

From this general result, we can then design approximate RPP mechanisms for usual noise distributions. These results are similar to those of the general Wasserstein mechanism (see Corollary 3.1) but with an additive term that depends on δ. We refer to Appendix C.4 for details. Using this new mechanism, we can obtain better utility at a small privacy cost for queries that take large values with small probability. In Example 2, we have $\Delta_G = 98$ while for $\delta = 3 \cdot 10^{-3}$, $\Delta_{G,\delta} = 1$, which yields a major improvement in utility. This observation also holds in a more general case.

Proposition 4.2 (Utility of the GAWM, informal)**.** *At a privacy cost of $\delta \in (0,1)$, the GAWM offers more utility than the GWM (see Appendix C.6 for details).*

Remark 4.4 (Relation to distribution privacy). A related result has been shown by Chen & Ohrimenko (2023) for the distribution privacy framework (see Appendix C.5 for the definition of distribution privacy and the result). The formulation of the results are similar, despite employing a different proof technique to get the conclusions. We prove a connection between the two results by establishing a formal equivalence between Pufferfish privacy and distribution privacy, which appears to be novel and could be of independent interest. In the interest of space, we refer to Appendix C.5 for the formal result and its proof. While our approximate shift reduction result (Lemma 4.1) induces an additional term which prevents us from recovering exactly the results of Chen & Ohrimenko (2023) in the particular case of the Laplace mechanism for PP, our result can be used with a wide range of noise distributions and in the RPP framework, which is more general than PP (and thus more general than distribution privacy).

4.2. Leveraging p-Wasserstein Metrics

As another way to improve the utility of the GWM, we propose to use shifts constrained by p-Wasserstein metrics instead of ∞-Wasserstein metrics, thereby replacing the worst case transportation cost between $P(f(X)|s_i, \theta)$ and $P(f(X)|s_j, \theta)$ by moments of the transportation cost. This idea was explored in a different context by Altschuler & Chewi (2023), who considered Orlicz-Wasserstein shifts for Gaussian noise and identified a dependency between the noise distribution and the selected Wasserstein shift constraint. They argue that the Orlicz-Wasserstein metric is the "right" metric to use for the shifted Rényi analysis because the original shift reduction lemma fails for weaker shifts. Inspired by these considerations, we broaden the applicability of the Orlicz-Wasserstein shift reduction lemma of Altschuler & Chewi (2023) by adapting their result to a wider range of noise distributions.

Lemma 4.2 (Generalized shift reduction)**.** *Let ζ be a noise distribution of $\mathbb{R}^d$. Let $z, p, q > 0$ such that $1/p + 1/q = 1$. We note:*

$$D^{(z)}_{\alpha,\alpha',\zeta}(\mu,\nu) = \inf_{\xi;\, \mathop{\mathbb{E}}_{W\sim\xi}[\exp((\alpha'-1)D_{\alpha'}(\zeta,\zeta*W))]\leq z} D_\alpha(\mu*\xi,\nu).$$

Then, we have:

$$D_\alpha(\mu*\zeta,\nu*\zeta) \leq D^{(z)}_{p(\alpha-1)+1,q(\alpha-1)+1,\zeta}(\mu,\nu)+\frac{\log(z)}{q(\alpha-1)}.$$

In the case $q=1$:

$$D_\alpha(\mu*\zeta,\nu*\zeta) \leq D^{(z)}_{\infty,\alpha,\zeta}(\mu,\nu)+\frac{\log(z)}{\alpha-1}.$$

This lemma yields a general Wasserstein mechanism that incorporates the noise distribution within the shift.

Theorem 4.5 (Distribution Aware General Wasserstein Mechanism, DAGWM)**.** *Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query and ζ noise distribution of $\mathbb{R}^d$. Let $q\geq 1$. For $(s_i,s_j)\in\mathcal{Q},\theta\in\Theta$, we note $\mu_i^\theta = P(f(X)|s_i,\theta)$. We denote:*

$$\Delta_G^{\zeta,q,\alpha} = \max_{\substack{(s_i,s_j)\in S\\ \theta\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \mathbb{E}\left[e^{q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*(X-Y))}\right].$$

Let $N=(N_1,\dots,N_d)\sim\zeta$ drawn independently of the dataset X. Then, $\mathcal{M}(X)=f(X)+N$ satisfies $(\alpha,\frac{\log(\Delta_G^{\zeta,q,\alpha})}{q(\alpha-1)})$-RPP for all $\alpha\in(1,+\infty)$ and $\lim_{\alpha\to+\infty}\frac{\log(\Delta_G^{\zeta,q,\alpha})}{q(\alpha-1)}$-PP.

Leveraging this result allows for the design of mechanisms with sensitivity constrained by p-Wasserstein distances (W_p). In particular, we will consider noise drawn from generalized Cauchy distributions, originally introduced by Rider (1957).

Definition 4.6 (Generalized Cauchy Distributions)**.** Let $k\geq 2,\lambda>0$. We say that the real random variable $V\sim\mathrm{GCauchy}(\lambda,k)$ if it has the following density: $\zeta_{k,\lambda}(x)=\frac{\beta_{k,\lambda}}{((1+(\lambda x)^2)^{k/2}}, x\in\mathbb{R}$ and $\int\zeta_{k,\lambda}(x)dx=1$. The Cauchy distribution is the special case $k=2$.

Using generalized Cauchy noise enables to consider W_p shifts while ensuring the existence of moments for large values of k.

Corollary 4.1 (Cauchy Mechanism)**.** *Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query. We denote Q_α the Legendre polynomial of integer index $\alpha>1$ and $\overline{Q_\alpha}$ as the polynomial derived from Q_α by retaining only its non-negative coefficients. Let $k\geq 2$ and $q\geq 1$ such that $kq(\alpha-1)/2$ is an integer. We note: $\Delta_G^{dkq(\alpha-1)} = \max_{\substack{(s_i,s_j)\in S\\ \theta\in\Theta}} W_{dkq(\alpha-1)}\left(P(f(X)|s_i,\theta),P(f(X)|s_j,\theta)\right)$, with $W_{dkq(\alpha-1)}$ computed with the l_2 norm. Then, $\mathcal{M}(X)=f(X)+V$ with $V=(V_1,\dots,V_d)\overset{iid}{\sim}\mathrm{GCauchy}(\lambda,k)$ is $\left(\alpha,\frac{d\log\frac{\beta_{k,\lambda}\pi}{\lambda}\overline{Q}_{kq(\alpha-1)/2}\left(1+\left(\frac{\Delta_G^{dkq(\alpha-1)}}{d\lambda}\right)^2\right)}{q(\alpha-1)}\right)$-RPP.*

In Example 1, for $q=d=1$ and $\alpha=k=2$, we have $\Delta_G^{\zeta,2,2}=\sqrt{1+3\rho}$ and noising with $V\sim\mathrm{Cauchy}(\lambda)$ in DAGWM ensures $\left(\alpha,\frac{\log\left(1+\frac{1+3\rho}{\lambda^2}\right)}{\alpha-1}\right)$-RPP, while the GWM for the same noise distribution gives $\left(\alpha,\frac{\log\left(1+\frac{4}{\lambda^2}\right)}{\alpha-1}\right)$-RPP. Hence, in this case DAGWM is better than GWM, as it allows to capture the correlation between the attributes. In the general case, DAGWM consistently outperforms GWM.

Proposition 4.3 (Utility of the DAGWM, informal)**.** *The DAGWM always offers more utility than the GWM at no additional privacy cost (see Appendix D.4 for details).*

5. Privacy Amplification by Iteration

Analyzing the privacy guarantees of Pufferfish privacy under composition is known to be challenging (Kifer & Machanavajjhala, 2014). While Pufferfish satisfies a form of parallel composition (see Appendix E for the result in RPP), to our knowledge there does not exist any theorem providing mechanism-agnostic guarantees for sequential composition in Pufferfish privacy. As an alternative to composition, we show in this section that RPP is amenable to privacy amplification by iteration, providing a way to analyze iterative gradient descent algorithms for convex optimization.

In differential privacy, privacy amplification by iteration (PABI) allows to evaluate the privacy loss of applying multiple contractive noisy iterations to a dataset and releasing only the output of the last iteration (Feldman et al., 2018; Altschuler & Talwar, 2022). PABI has often been employed in private machine learning to analyze the privacy cost of projected noisy stochastic gradient descent (DP-SGD), bypassing the use of composition (Feldman et al., 2018). However, existing PABI results for differential privacy cannot be used in Pufferfish privacy. These results consider the distribution shift between two processes performed on two neighboring datasets (equal up to one element) and how this additional shift propagates through the rest of the iterations. In Pufferfish, privacy is obtained by conditioning over secrets and the dataset is sampled from an adversary's prior. This means that two datasets with different secrets might share no common elements. Hence, the original worst case PABI analysis must be adapted to account for shifts at each iteration, while measuring these shifts based on the dataset distribution conditioned by the secrets.

We start by defining contractive noisy iterations.

Definition 5.1 (Contractive noisy iteration (CNI))**.** Let $\mathcal{Z}\subset\mathbb{R}^d$. Given an initial random state $W_0\in\mathcal{Z}$, a sequence of

random variables $\{X_t\}$, a sequence of contractive maps in their first argument $\psi_t : \mathcal{Z} \times \mathcal{D} \to \mathcal{Z}$ and a sequence of noise distributions $\{\zeta_t\}$, we define the Contractive Noisy Iteration (CNI) by the following update rule:

$$W_{t+1} = \psi_{t+1}(W_t, X_{t+1}) + N_{t+1},$$

where $N_{t+1} \sim \zeta_{t+1}$. For brevity, we refer to the result W_T of the CNI at the time step T by $CNI_T(W_0, \{X_t\}, \{\psi_t\}, \{\zeta_t\})$.

As opposed to the work of Feldman et al. (2018), we make an explicit reference to the dataset distribution modeled by the random sequence $\{X_t\}$ in the CNI definition. The original PABI analysis leverages a contraction lemma that we need to adapt to the Pufferfish setting. We prove a new contraction lemma which incorporates the ∞-Wasserstein distance to take into account the dataset distribution.

Lemma 5.1 (Dataset Dependent Contraction lemma)**.** *Let ψ be a contractive map in its first argument on $(\mathcal{Z}, \|\cdot\|)$. Let X, X' be two r.v's. Suppose that $\sup_w W_\infty(\psi(w, X), \psi(w, X')) \leq s$. Then, for $z > 0$:*

$$D_\alpha^{(z+s)}(\psi(W, X), \psi(W', X')) \leq D_\alpha^{(z)}(W, W').$$

Coupled with the original shift reduction lemma (Lemma 3.1), this contraction lemma yields a relaxation of the original PABI bounds, allowing take into account the dataset distribution in the measurement of the shifts.

Theorem 5.2 (Dataset Dependent PABI)**.** *Let X_T and X'_T denote the output of $CNI_T(W_0, \{\psi_t\}, \{\zeta_t\}, X)$ and $CNI_T(W_0, \{\psi_t\}, \{\zeta_t\}, X')$. Let $s_t = \sup_w W_\infty(\psi(w, X_t), \psi(w, X'_t))$. Let $a_1, \ldots, a_T$ be a sequence of reals and let $z_t = \sum_{i \leq t} s_i - \sum_{i \leq t} a_i$. If $z_t \geq 0$ for all t, then, we have:*

$$D_\alpha^{(z_T)}(X_T, X'_T) \leq \sum_{t=1}^{T} R_\alpha(\zeta_t, a_t).$$

This new PABI bound allow for an RPP analysis of noisy gradient descent, as developed in the next section.

6. Applications

In this section, we focus on concrete applications of our RPP mechanisms and PABI for specific instantiations. Our generic mechanisms can be hard to compute in the general case, as they rely on the computation of Wasserstein distances between arbitrary distributions. Below, we present specific instances for which the sensitivity of the GWM has a simple closed form. We also apply our generic PABI result to convex optimization, bypassing the lack of adaptive composition theorems and avoiding the cost of group privacy.

6.1. Weakly Dependent Data

The sensitivity Δ_G of the GWM can be bounded in a straightforward way for *near-independent* data distributions, where the dependence level is quantified via a generalization of Wasserstein dependence metrics (Ozair et al., 2019). Below, we consider $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ with $\mathcal{X} \subset \mathbb{R}^d$ and, for any distribution $\theta \in P(\mathcal{X}^n)$, we denote by $\theta^\otimes$ the product distribution of the marginals of θ.

Proposition 6.1. *Let $\lambda > 0$, $(\mathcal{S}, \mathcal{Q}, \Theta)$ a Pufferfish framework. Let Δ be the sensitivity of a numerical query f, denote $\mu_i^\theta = P(f(X)|s_i, \theta)$, and let*

$$\Theta_\lambda = \{\theta \in P(\mathcal{X}^n); \sup_{s_i \in \mathcal{S}} W_\infty(\mu_i^\theta, \mu_i^{\theta^\otimes})) \leq \lambda\}.$$

Then, if $\Theta \subseteq \Theta_\lambda$, $\Delta_G \leq 2\lambda + \Delta$.

In other words, for instances with low dependencies, the GWM avoids the extra privacy cost of Group DP.

6.2. Attribute Privacy

Attribute privacy (Zhang et al., 2022) is a specific instantiation or Pufferfish. In this setting, each record X_i in a dataset $X = (X_i^j)_{i \in [\![1,n]\!], j \in [\![1,m]\!]}$ is viewed as independent, while an adversary possesses prior knowledge, denoted as θ, about the distribution generating each record. The columns, representing each attribute, are denoted by X^j. The objective is to reveal the answer of a query $f(X)$ while protecting some summary statistics $g_j(X^j)$ of each attribute X^j. A formal definition of attribute privacy is recalled in Appendix F.2.1. Below, we show that the GWM can be efficiently computed for Gaussian data, and we also conduct experiments to empirically show that RPP mechanisms have better utility than RDP mechanisms on real datasets.

Special case of Gaussian data. Attribute privacy guarantees for Gaussian priors can be obtained through specific attribute privacy mechanisms (Zhang et al., 2022). It is also possible to derive closed form bounds with the GWM for linear queries. A simple example is the case of releasing an attribute X^j while protecting another attribute X^i.

Proposition 6.2 (Attribute privacy with Gaussian data)**.** *Consider a dataset $X = (X_1^1, \ldots, X_n^m)$. Let $j \in [\![1, m]\!]$ be an attribute. We assume that each record X_i is independently sampled from $\theta = \mathcal{N}(\mu, \Sigma)$. Let the secrets $s_i^a = \{X_i^j = a\}$, with $a \in K$ and K a compact of $\mathbb{R}^m$, the pairs of secrets $\mathcal{Q} = \{(s_i^a, s_i^b); a, b \in K, i \in [\![1, n]\!]\}$ and the numerical query $f : x = (x^1, \ldots, x^m) \mapsto x^j$. We have*

$$\Delta_G \leq \max_{a,b \in K} \mathrm{Var}(X_1^i)^{-1} \,\mathrm{Cov}(X_1^i, X_1^j) \|a - b\|$$

for the Pufferfish framework $(\mathcal{S}, \mathcal{Q}, \{\theta\})$.

This proposition shows that if X^i and X^j are weakly correlated, then Δ_G is small. A more general version of this result can be found in Appendix F.3.

Experiments. Table 1 presents some experimental results showing that GWM provides strictly better utility than DP mechanisms in attribute privacy scenarios on three real-world datasets. We obtain lower sensitivities for the DAGWM and the GWM than for DP. Figures, discussions and detailed results can be found in Appendix F.2.2.

Table 1. Sensitivities for DP (Δ), GWM (Δ_G) and Cauchy mechanism ($\Delta_{G,2}$) in attribute privacy scenarios on 3 real datasets.

	Sensitivity		
Dataset	Δ	Δ_G	$\Delta_{G,2}$
Student Scores	20	8	2.76
Heart	$\geq$100	8	7.80
Adult	1	1	0.42

6.3. Privacy in Diffusion Processes

The GWM is also tractable for special cases of temporally correlated data. We consider the setting where one wants to release the output of a query $f(X_{t_1}, \dots, X_{t_n})$ performed on a time series $(X_t)_{t\geq 0}$ with $t_1 < \dots < t_n$ while protecting the privacy of the initial point X_0. With some assumptions, it is possible to derive contraction results that enable to compute Δ_G. We concentrate on the case where the adversary's prior can be modeled by a Langevin dynamic system.

Proposition 6.3. *Let $V : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla^2 V \succcurlyeq CI_d$. For $\theta_0 \in P(\mathbb{R}^d)$, we note θ_t the distribution of X_t, with $(X_t)_{t\geq 0}$ solution of the stochastic differential equation: $dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$, where $(B_t)_{t\geq 0}$ is a brownian motion. We note $\theta_{t_1,\dots,t_n}$ the distribution generating $X = (X_{t_1}, \dots, X_{t_n})$ from the distribution of $(X_t)_{t\geq 0}$. We consider the secrets $s^a = \{X_0 = a\}$, with $a \in K$ and K a compact of $\mathbb{R}^d$, and the pairs of secrets $\mathcal{Q} = \{(s^a, s^b); a, b \in K\}$. Then, the GWM of any L-Lipschitz query f performed on X has a sensitivity for the l_1 norm:*

$$\Delta_G \leq L\text{Diam}(K) \sum_{i=1}^{n} \exp(-2Ct_i)$$

for the Pufferfish framework $(\mathcal{S}, \mathcal{Q}, \{\theta_{t_1,\dots,t_n}\})$.

This result demonstrates that for some well-behaved diffusion processes, the GWM drastically mitigates the sensitivity compared to the Group DP scenario. This reduction is indicated by an exponential term which depends on the convexity of V and the released timestamps.

6.4. Application of PABI to Convex Optimization

We show an application of our general PABI result (Theorem 5.2) to the RPP analysis of the celebrated DP-SGD algorithm for private machine learning.

Let $m, d, T > 0$. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish framework. We note $\mathcal{X}$ the set of values taken by the elements of the dataset. Let the secrets $s_i^a = \{X_i = a\}, s_i^b = \{X_i = b\}$, $i \in [\![1, T]\!], a, b \in \mathcal{X}$. We note $X = (X_1, \dots, X_T) \sim P(X|s_i^a, \theta)$ and $X' = (X'_1, \dots, X'_T) \sim P(X|s_i^b, \theta)$. We assume that $\mathcal{X} \subset \mathbb{R}^m$. Let $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ be an objective function. We assume that f is convex, L-Lipschitz in its first argument, β-smooth in its second argument (see Appendix F.5.1 for definitions) and f satisfies the following condition: $\forall x_1, x_2 \in \mathcal{X}, w_1 \in \mathbb{R}^d, \exists C_{w_1} > 0$ such as :

$$\|\nabla_w f(w_1, x_1) - \nabla_w f(w_1, x_2)\| \leq C_{w_1}\|x_1 - x_2\|.$$

The last assumption, which is used in the adversarial training literature (see e.g., Liu et al., 2020), is satisfied in certain simple settings as linear regression. It enables to take into account the distribution of the gradients as a function of the distribution of the data in our PABI analysis. Let $\Pi : \mathbb{R}^d \to \mathbb{R}^d$ be a projection over a compact $\mathcal{K} \subset \mathbb{R}^d$ and $\eta > 0$ such that $\eta < 2/\beta$. By Proposition 18 of Feldman et al. (2018), the weight update function: $\psi : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}^d, (v, x) \mapsto \Pi(v - \eta\nabla_w f(v, x))$ is contractive. Let $W_0 = W'_0 \in \mathcal{K}$ be the initial weight and $\zeta = \mathcal{N}(0, \sigma^2 I_d)$ be a noise distribution. We note $(N_1, \dots, N_T) \sim \zeta^{\otimes T}$ and for all $t \in [\![1, T]\!]$, $W_t = \psi(W_{t-1}, X_t) + N_t$, $W'_t = \psi(W'_{t-1}, X'_t) + N_t$, as in DP-SGD. Then, we note $s_t = \eta \sup_{v\in\mathcal{K}} W_\infty(\nabla_w f(v, X_t), \nabla_w f(v, X'_t))$. As an example of application of Theorem 5.2, taking $(a_t) = (s_t)$, we have:

$$D_\alpha(W_T, W'_T) \leq \frac{\alpha\eta^2}{2\sigma^2} \sum_{t=1}^{T} \min(2L, \sup_{v\in\mathcal{K}} C_v W_\infty(X_t, X'_t))^2.$$

To interpret this formula, we can look at some extreme cases. If the adversary has a prior of high correlations, such as for example $X_1 = \dots = X_t$, $X'_1 = \dots = X'_t$, we get:

$$D_\alpha(W_T, W'_T) \leq \frac{T\alpha\eta^2}{2\sigma^2} \min(2L, \|a - b\| \sup_{v\in\mathcal{K}} C_v)^2,$$

which is no better than the group privacy analysis. On the other hand, when data points are independent as in differential privacy, we get:

$$D_\alpha(W_T, W'_T) \leq \frac{\alpha\eta^2}{2\sigma^2} \min(2L, \|a - b\| \sup_{v\in\mathcal{K}} C_v)^2.$$

In this case, the upper bound is independent of T and we thus obtain much better results than with group privacy. In fact, our result is general enough to recover the original results of Feldman et al. (2018) for DP-SGD as a special case.

Remark 6.1 (DP as a special case, informal). Theorem 5.2 allows to recover the same privacy bounds as Theorem 23 of Feldman et al. (2018) (see Appendix F.5.2 for details).

In the Gaussian case, our results allow to derive PABI bounds that explicitly depend on correlations in the dataset.

Proposition 6.4. *Assume that the adversary has a Gaussian prior θ. Then,*

$$D_\alpha(W_T, W_T') \leq \frac{\alpha\eta^2}{2\sigma^2}\big(\min(2L, \sup_{v\in\mathcal{K}} C_v\|a-b\|)^2 + \sum_{t\neq i}^{T} \min(2L, \sup_{v\in\mathcal{K}} C_v\|\operatorname{Cov}(X_t, X_i)\operatorname{Cov}(X_i)^{-1}(a-b)\|)^2\big).$$

This result is the sum of two terms: the first one is the same as for DP (i.e., the case where data points are independent), while the second one accounts for the dependence by summing, for each step t, the correlation between X_t and X_i.

This bound can be improved when $W_\infty(X_t, X_t')$ is non-increasing, leading to settings where the privacy loss converges to 0 as $T \to +\infty$. This consideration is discussed and illustrated numerically in Appendix F.5.4.

7. Conclusion

We presented a new framework, called Rényi Pufferfish privacy, which extends the original Pufferfish privacy definition. We designed general additive noise mechanisms for achieving (approximate) Rényi Pufferfish privacy and discussed their applicability for specific instantiations. As a way to use Pufferfish privacy to analyze sequential algorithms, we derived a privacy amplification by iteration result which allows to bypass the lack of adaptive composition theorems. We put forward a first application of this analysis for convex optimization with gradient descent, allowing the integration of Pufferfish in machine learning algorithms. Potential areas for future work include a tighter PABI analysis with other shift reduction lemmas, and a numerical analysis of Rényi Pufferfish privacy mechanisms to optimize utility in more complex practical use-cases.

Impact Statement

This paper presents work whose goal is to advance privacy in machine learning, offering tools to make it more secure. We provide methods to protect specific types of secrets and to manage correlations in datasets, which are frequently found in practice. Properly designed Pufferfish instantiations can provide greater utility than usual Group Differential Privacy mechanisms. However, the data curator must carefully design its Pufferfish instantiation in order to ensure adequate robust privacy protection.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, 2016. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Altschuler, J. M. and Chewi, S. Faster high-accuracy log-concave sampling via algorithmic warm starts, 2023. arXiv:2302.10249.

Altschuler, J. M. and Talwar, K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In *NeurIPS*, 2022.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, pp. 635–658. Springer Berlin Heidelberg, 2016.

Chen, M. and Ohrimenko, O. Protecting global properties of datasets with distribution privacy mechanisms. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7472–7491. PMLR, 2023. URL https://proceedings.mlr.press/v206/chen23f.html.

Cortez, P. Student Performance. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TG7T.

Ding, N. Kantorovich mechanism for pufferfish privacy. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5084–5103. PMLR, 2022. URL https://proceedings.mlr.press/v151/ding22b.html.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. Privacy amplification by iteration. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532, 2018.

Hardt, M. and Roth, A. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pp. 331–340,

2013. ISBN 9781450320290. doi: 10.1145/2488608.2488650. URL https://doi.org/10.1145/2488608.2488650.

He, X., Machanavajjhala, A., and Ding, B. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pp. 1447–1458. Association for Computing Machinery, 2014. ISBN 9781450323765. doi: 10.1145/2588555.2588581. URL https://doi.org/10.1145/2588555.2588581.

Humphries, T., Oya, S., Tulloch, L., Rafuse, M., Goldberg, I., Hengartner, U., and Kerschbaum, F. Investigating membership inference attacks under data dependencies. In *2023 2023 IEEE 36th Computer Security Foundations Symposium (CSF) (CSF)*, pp. 194–209. IEEE Computer Society, 2023. doi: 10.1109/CSF57540.2023.00013. URL https://doi.ieeecomputersociety.org/10.1109/CSF57540.2023.00013.

Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

Kawamoto, Y. and Murakami, T. Local obfuscation mechanisms for hiding probability distributions. In *Computer Security – ESORICS 2019*, pp. 128–148, Cham, 2019. Springer International Publishing.

Kessler, S., Buchmann, E., and Böhm, K. Deploying and evaluating pufferfish privacy for smart meter data. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 229–238, 2015. doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.55.

Kifer, D. and Machanavajjhala, A. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1), 2014. ISSN 0362-5915. doi: 10.1145/2514689. URL https://doi.org/10.1145/2514689.

Liu, C., Salzmann, M., Lin, T., Tomioka, R., and Süsstrunk, S. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. In *Advances in Neural Information Processing Systems*, pp. 21476–21487, 2020.

Markelle Kelly, Rachel Longjohn, K. N. The uci machine learning repository. https://archive.ics.uci.edu.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Niu, C., Zheng, Z., Tang, S., Gao, X., and Wu, F. Making big money from small sensors: Trading time-series data under pufferfish privacy. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 568–576, 2019. doi: 10.1109/INFOCOM.2019.8737579.

Nuradha, T. and Goldfeld, Z. Pufferfish privacy: An information-theoretic study. *IEEE Trans. Inf. Theory*, 69(11):7336–7356, 2023.

Ou, L., Qin, Z., Liao, S., Yin, H., and Jia, X. An optimal pufferfish privacy mechanism for temporally correlated trajectories. *IEEE Access*, 6:37150–37165, 2018. doi: 10.1109/ACCESS.2018.2847720.

Ozair, S., Lynch, C., Bengio, Y., van den Oord, A., Levine, S., and Sermanet, P. Wasserstein dependency measure for representation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Papernot, N. and Steinke, T. Hyperparameter tuning with renyi differential privacy. *CoRR*, abs/2110.03620, 2021. URL https://arxiv.org/abs/2110.03620.

Rider, P. R. Generalized cauchy distributions. *Annals of the Institute of Statistical Mathematics*, 9:215–223, 1957. URL https://api.semanticscholar.org/CorpusID:122913729.

Saumard, A. and Wellner, J. A. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45–114, 2014. URL https://api.semanticscholar.org/CorpusID:23773316.

Song, S., Wang, Y., and Chaudhuri, K. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pp. 1291–1306, 2017. ISBN 9781450341974. doi: 10.1145/3035918.3064025. URL https://doi.org/10.1145/3035918.3064025.

Verdú, S. The cauchy distribution in information theory. *Entropy*, 25(2), 2023. ISSN 1099-4300. doi: 10.3390/e25020346. URL https://www.mdpi.com/1099-4300/25/2/346.

Villani, C. Optimal transport: Old and new. 2008. URL https://api.semanticscholar.org/CorpusID:118347220.

Zhang, W., Ohrimenko, O., and Cummings, R. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 757–766, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533139. URL https://doi.org/10.1145/3531146.3533139.

This appendix provides some useful background, as well as more detailed versions of our results, along with their proofs.

A. Properties of Rényi Pufferfish Privacy (Section 2)

A.1. Definitions

Rényi differential privacy relies on Rényi divergences, which are defined as follows.

Definition A.1. Let μ and ν be two distributions on a measurable space $(E, \mathcal{A})$ and $\alpha > 1$. We define the Rényi divergence of order α between μ and ν as:

$$D_\alpha(\mu, \nu) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim\nu}\left[\left(\frac{\mu(x)}{\nu(x)}\right)^\alpha\right].$$

The definition extends to the case $\alpha = +\infty$ by continuity.

Definition A.2 (Rényi differential privacy, RDP (Mironov, 2017))**.** Let $\alpha > 1$ and $\varepsilon \geq 0$. A randomized algorithm $\mathcal{M} \colon \mathcal{D} \to \mathcal{R}$ satisfies (α, ε)-Rényi differential privacy if for any two adjacent datasets $X_1, X_2 \in \mathcal{D}$ differing by one element, it holds:

$$D_\alpha\left(P(\mathcal{M}(X_1)), P(\mathcal{M}(X_2)\right) \leq \varepsilon.$$

A.2. Proof of Proposition 2.1

Proposition 2.1 (Post-processing)**.** *Let $\mathcal{M}_1$ be a randomized algorithm and $\mathcal{M}$ be (α, ε)-RPP. Then,*

$$\begin{aligned}&D_\alpha\left(P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_i, \theta\right), P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_j, \theta\right)\right)\\ &\leq D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta\right), P\left(\mathcal{M}(X) \mid s_j, \theta\right)\right) \leq \varepsilon.\end{aligned}$$

Proof. Let $\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish framework. Let $(s_i, s_j) \in \mathcal{Q}, \theta \in \Theta$, $\alpha > 1$ and $\epsilon > 0$. Let $\mathcal{M}_1$ be a randomized algorithm and $\mathcal{M}$ satisfying (α, ε)-RPP. Then,

$$\begin{aligned}
&D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta\right), P\left(\mathcal{M}(X) \mid s_j, \theta\right)\right) = \mathbb{E}_{Z\sim P(\mathcal{M}(X)|s_j,\theta)}\left[\left(\frac{P\left(\mathcal{M}(X) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}(X) = Z \mid s_j, \theta\right)}\right)^\alpha\right]\\
&= \mathbb{E}_{(Z',Z)\sim(P(\mathcal{M}_1(\mathcal{M}(X))|s_j,\theta),P(\mathcal{M}(X)|s_j,\theta))}\left[\left(\frac{P\left(\mathcal{M}(X) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}(X) = Z \mid s_j, \theta\right)}\right)^\alpha\right]\\
&= \mathbb{E}_{(Z',Z)\sim(P(\mathcal{M}_1(\mathcal{M}(X))|s_j,\theta),P(\mathcal{M}(X)|s_j,\theta))}\left[\left(\frac{P\left(\mathcal{M}(X) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}(X) = Z \mid s_j, \theta\right)}\frac{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z' \mid \mathcal{M}(X) = Z\right)}{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z' \mid \mathcal{M}(X) = Z\right)}\right)^\alpha\right]\\
&= \mathbb{E}_{Z'\sim P(\mathcal{M}_1(\mathcal{M}(X))|s_j,\theta)}\left[\mathbb{E}_{Z\sim P(\mathcal{M}(X)|\mathcal{M}_1(\mathcal{M}(X)),s_j,\theta)}\left[\left(\frac{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z', \mathcal{M}(X) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z', \mathcal{M}(X) = Z \mid s_j, \theta\right)}\right)^\alpha\right]\right]\\
&\geq \mathbb{E}_{Z'\sim P(\mathcal{M}_1(\mathcal{M}(X))|s_j,\theta)}\left[\left(\mathbb{E}_{Z\sim P(\mathcal{M}(X)|\mathcal{M}_1(\mathcal{M}(X)),s_j,\theta)}\left[\frac{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z', \mathcal{M}(X) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z', \mathcal{M}(X) = Z \mid s_j, \theta\right)}\right]\right)^\alpha\right] \text{ Jensen inequality}\\
&= \mathbb{E}_{Z'\sim P(\mathcal{M}_1(\mathcal{M}(X))|s_j,\theta)}\left[\left(\frac{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z \mid s_i, \theta\right)}{P\left(\mathcal{M}_1(\mathcal{M}(X)) = Z \mid s_j, \theta\right)}\right)^\alpha\right]\\
&= D_\alpha\left(P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_i, \theta\right), P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_j, \theta\right)\right).
\end{aligned}$$

Thus,

$$D_\alpha\left(P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_i, \theta\right), P\left(\mathcal{M}_1(\mathcal{M}(X)) \mid s_j, \theta\right)\right) \leq D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta\right), P\left(\mathcal{M}(X) \mid s_j, \theta\right)\right) \leq \varepsilon. \qquad \square$$

A.3. Proof of Proposition 2.2

Proposition 2.2 (RPP implies PP)**.** *If $\mathcal{M}$ is (α, ε)-RPP, it also satisfies $\left(\varepsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta\right)$-PP for all $\delta \in (0, 1)$.*

Proof. The proof technique of (Mironov, 2017) remains applicable in the context of Rényi Pufferfish privacy. For clarity and completeness, we showcase it here. Let $\varepsilon \geq 0, \alpha > 1$. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish privacy framework and $\mathcal{M}$ an (α, ε)-RPP mechanism. Let $\delta \in (0, 1)$, $\theta \in \Theta$, $(s_i, s_j) \in \mathcal{Q}$ and $z \in Range(\mathcal{M})$. Then, we have:

$$P\left(\mathcal{M}(X) = z|s_i, \theta\right)^\alpha \leq e^{(\alpha-1)D_\alpha(P(\mathcal{M}(X)|s_i,\theta),P(\mathcal{M}(X)|s_j,\theta))}P\left(\mathcal{M}(X) = z|s_j, \theta\right)^{\alpha-1} \leq e^{\varepsilon(\alpha-1)}P\left(\mathcal{M}(X) = z|s_j, \theta\right)^{\alpha-1},$$

where the first inequality is obtained by Hölder inequality applied to the functions $\left(\frac{f^\alpha}{g^{\alpha-1}}\right)^{\frac{1}{\alpha}}$ and $g^{\frac{\alpha-1}{\alpha}}$. We then consider two cases:

- Case 1: $e^\varepsilon P\left(\mathcal{M}(X)=z|s_j,\theta\right) \leq \delta^{\frac{\alpha}{\alpha-1}}$. Then, $P\left(\mathcal{M}(X)=z|s_i,\theta\right) \leq \delta \leq e^{\varepsilon+\frac{\log(1/\delta)}{\alpha-1}} P\left(\mathcal{M}(X)=z|s_j,\theta\right)+\delta$.

- Case 2: $e^\varepsilon P\left(\mathcal{M}(X)=z|s_j,\theta\right) > \delta^{\frac{\alpha}{\alpha-1}}$. Then,

$$\begin{aligned} P\left(\mathcal{M}(X)=z|s_i,\theta\right) &\leq \left(e^\varepsilon P\left(\mathcal{M}(X)=z|s_j,\theta\right)\right)\left(e^\varepsilon P\left(\mathcal{M}(X)=z|s_j,\theta\right)\right)^{\frac{-1}{\alpha}} \\ &\leq e^\varepsilon P\left(\mathcal{M}(X)=z|s_j,\theta\right)\delta^{\frac{-1}{\alpha-1}} \\ &\leq e^{\varepsilon+\frac{\log(1/\delta)}{\alpha-1}} P\left(\mathcal{M}(X)=z|s_j,\theta\right)+\delta. \end{aligned}$$

□

A.4. Guarantees Against Close Adversaries

A.4.1. Original result from Song et al. (2017)

For completeness, we recall here the original theorem from Song et al. (2017) on the robustness of the Pufferfish privacy framework.

Theorem A.3 (Protection against close adversaries (Song et al., 2017))**.** *Let $\mathcal{M}$ be a mechanism that satisfies ε-PP in a framework $(\mathcal{S},\mathcal{Q},\Theta)$. Let $\theta' \notin \Theta$ and*

$$\begin{aligned} \Delta = \inf_{\theta\in\Theta}\sup_{s_i\in\mathcal{Q}} \max\{&D_\infty\left(P(X|s_i,\theta),P(X|s_i,\theta')\right), \\ &D_\infty\left(P(X|s_i,\theta),P(X|s_i,\theta')\right)\}. \end{aligned}$$

Then, $\mathcal{M}$ is $(\varepsilon+2\Delta)$-PP for the framework $(\mathcal{S},\mathcal{Q},\Theta')$ with $\Theta'=\Theta\cup\{\theta'\}$.

A.4.2. Protection against close adversaries in the RPP framework

Theorem A.4 (RPP protection against close adversaries)**.** *Let $p,q,r>0$ such that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1$, and let $\mathcal{M}$ be a mechanism that satisfies $(q(\alpha-1/p),\varepsilon)$-RPP in a framework $(\mathcal{S},\mathcal{Q},\Theta)$. Let $\theta'\notin\Theta$ and*

$$\begin{aligned} \Delta_p^1 &= \inf_{\theta\in\Theta}\sup_{s_i\in S} D_{\alpha p}\left(P(X|s_i,\theta'),P(X|s_i,\theta)\right), \\ \Delta_r^2 &= \inf_{\theta\in\Theta}\sup_{s_i\in S} D_{(\alpha-1)r+1}\left(P(X|s_i,\theta),P(X|s_i,\theta')\right). \end{aligned}$$

Then, for all $\alpha\in(1,\infty)$, $\mathcal{M}$ satisfies:

$$\left(\alpha,\left(1+\frac{1}{r(\alpha-1)}\right)\varepsilon+\left(1+\frac{\frac{1}{r}+\frac{1}{q}}{\alpha-1}\right)\Delta_p^1+\Delta_r^2\right)\text{-RPP}$$

for $(\mathcal{S},\mathcal{Q},\Theta')$ with $\Theta'=\Theta\cup\{\theta'\}$.

Proof. Let $(\mathcal{S},\mathcal{Q},\Theta)$ be a Pufferfish privacy instance and $\mathcal{M}$ a randomized mechanism. Let $\theta'\notin\Theta$ and $p,q,r>0$ such

that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1$. Let $s_i, s_j \in \mathcal{Q}$. We have:

$$\begin{aligned}
&\exp(\alpha-1)D_\alpha\left(P(\mathcal{M}(X)|s_i,\theta'), P(\mathcal{M}(X)|s_j,\theta')\right)\\
&= \int \frac{P(\mathcal{M}(X)=z|s_i,\theta')^\alpha}{P(\mathcal{M}(X)=z|s_j,\theta')^{\alpha-1}}dz\\
&= \int \frac{P(\mathcal{M}(X)=z|s_i,\theta')^\alpha}{P(\mathcal{M}(X)=z|s_i,\theta)^{\alpha-1/p}}\frac{P(\mathcal{M}(X)=z|s_i,\theta)^{\alpha-1/p}}{P(\mathcal{M}(X)=z|s_i,\theta)^{\alpha-1/p-1/q}}\frac{P(\mathcal{M}(X)=z|s_j,\theta)^{\alpha-1/p-1/q}}{P(\mathcal{M}(X)=z|s_j,\theta')^{\alpha-1}}dz\\
&\leq \left(\int \frac{P(\mathcal{M}(X)=z|s_i,\theta')^{\alpha p}}{P(\mathcal{M}(X)=z|s_i,\theta)^{\alpha p-1}}dz\right)^{\frac{1}{p}}\cdot\left(\int \frac{P(\mathcal{M}(X)=z|s_i,\theta)^{q(\alpha-1/p)}}{P(\mathcal{M}(X)=z|s_i,\theta)^{q(\alpha-1/p)-1}}dz\right)^{\frac{1}{q}}\\
&\cdot \left(\int \frac{P(\mathcal{M}(X)=z|s_j,\theta)^{\alpha-1/p-1/q}}{P(\mathcal{M}(X)=z|s_j,\theta')^{\alpha-1}}dz\right)^{\frac{1}{r}}\\
&\leq \exp\left(\alpha-1/p\right)D_{\alpha p}\left(P(\mathcal{M}(X)|s_i,\theta'), P(\mathcal{M}(X)|s_i,\theta)\right)\\
&\quad+\exp\left((\alpha-1+1/r)D_{q(\alpha-1/p)}\left(P(\mathcal{M}(X)|s_i,\theta), P(\mathcal{M}(X)|s_j,\theta)\right)\right)\\
&\quad+\exp\left((\alpha-1)D_{(\alpha-1)r+1}\left(P(\mathcal{M}(X)|s_j,\theta), P(\mathcal{M}(X)|s_j,\theta')\right)\right)
\end{aligned}$$

by using the generalized Hölder inequality: for $p,q,r,t>0$ such that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=\frac{1}{t}$ and $f\in L^p, g\in L^q, h\in L^r$,

$$\|fgh\|_t \leq \|f\|_p\|g\|_q\|h\|_r.$$

Then, the post-processing property of RPP (Proposition 2.1) gives the result. □

This theorem employs α-Rényi divergences and can be viewed as a generalization of the result of Song et al. (2017), which we recover as a special case for $\alpha=+\infty$. Note that neither Theorem A.4 nor the original result of Song et al. (2017) exploit the characteristics of the particular mechanism $\mathcal{M}$ of interest in the quantification of the additional privacy loss. As a matter of fact, it is likely that a mechanism with large variance would yield more robust guarantees. Interestingly, we can address this issue by refining our result to additive noise mechanisms using the shift reduction lemma.

A.4.3. Refinement of Theorem A.4 for Additive Mechanisms

Leveraging the shift reduction lemma (Lemma 3.1), we refine Theorem A.4 for additive mechanisms.

Theorem A.5 (RPP protection against close adversaries for additive noise mechanisms)**.** *Let $p,q,r>0$ such that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1$. Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query. Let $\mathcal{M}(X)=f(X)+N$ with $N\sim\zeta$ be an additive noise mechanism that satisfies $(q(\alpha-1/p),\varepsilon)$-RPP for $(\mathcal{S},\mathcal{Q},\Theta)$. Let $\theta'\notin\Theta$ and*

$$\Delta_{\theta'}=\inf_{\theta\in\Theta}\sup_{s_i\in S}W_\infty\left(P(f(X)|s_i,\theta'), P(f(X)|s_i,\theta)\right).$$

Then, for all $\alpha\in(1,\infty)$ and denoting

$$K=\left(1+\frac{\frac{1}{r}+\frac{1}{q}}{\alpha-1}\right)R_{\alpha p}(\zeta,\Delta_{\theta'})+R_{(\alpha-1)r+1}(\zeta,\Delta_{\theta'}),$$

$\mathcal{M}$ satisfies:

$$\left(\alpha,\left(1+\frac{1}{r(\alpha-1)}\right)\varepsilon+K\right)\text{-RPP}$$

for $(\mathcal{S},\mathcal{Q},\Theta')$ with $\Theta'=\Theta\cup\{\theta'\}$.

This theorem enables us to take into account the characteristics of the mechanism when examining the robustness of a RPP instance. We illustrate this below with the Gaussian mechanism.

Corollary A.1 (RPP protection against close adversaries for the Gaussian mechanism)**.** *We note I_d the identity matrix of size d. Let $p,q,r>0$ such that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1$. Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query. Let $\mathcal{M}(X)=f(X)+N$ with $N\sim\mathcal{N}\left(0,\frac{q(\alpha-1/p)\Delta_G^2}{2\varepsilon}I_d\right)$, where Δ_G is defined in Theorem 3.3. Let $\theta'\notin\Theta$.*

Then, for all $\alpha \in (1, \infty)$, $\mathcal{M}$ satisfies:

$$\left(\alpha, \left(\left(1 + \frac{1}{r(\alpha-1)}\right) + \left(\alpha\left(p + \frac{p-1}{\alpha-1}\right) + (\alpha-1)r + 1\right)\frac{\Delta_{\theta'}^2}{\Delta_G^2}\frac{1}{q(\alpha-1/p)}\right)\varepsilon\right) \text{-RPP}$$

for $(\mathcal{S}, \mathcal{Q}, \Theta')$ with $\Theta' = \Theta \cup \{\theta'\}$.

One can see that the additive penalty vanishes proportionally to $\frac{1}{\Delta_G^2}$. It establishes a trade-off between the utility of the mechanism and the robustness of the Pufferfish privacy framework when designing Θ. Remarkably, this consideration could not have been derived from our Theorem A.4 for RPP nor from the original result from Song et al. (2017) (Theorem A.3).

B. General Wasserstein Mechanism (Section 3)

B.1. Proof of Theorem 3.3

Theorem 3.3 (General Wasserstein mechanism, GWM)**.** *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a numerical query and denote:*

$$\Delta_G = \max_{\substack{(s_i,s_j)\in S \\ \theta\in\Theta}} W_\infty\left(P(f(X)|s_i,\theta), P(f(X)|s_j,\theta)\right).$$

Let $N = (N_1, \ldots, N_d) \sim \zeta$, where $N_1, \ldots, N_d$ are iid real random variables independent of the data X. Then, $\mathcal{M}(X) = f(X) + N$ satisfies $(\alpha, R_\alpha(\zeta, \Delta_G))$-RPP for all $\alpha \in (1, +\infty)$ and $R_\infty(\zeta, \Delta_G)$-PP.

Proof. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish privacy instance. Let $f : \mathcal{D} \to \mathbb{R}^d$ be a numerical query and denote:

$$\Delta_G = \max_{\substack{(s_i,s_j)\in S \\ \theta_i\in\Theta}} W_\infty\left(P(f(X)|s_i,\theta), P(f(X)|s_j,\theta)\right).$$

Let $N = (N_1, \ldots, N_d) \sim \zeta$, where $N_1, \ldots, N_d$ are iid real random variables independent of the data X. We use the abuse of notation $D_\alpha(X_{|E}, Y_{|E}) = D_\alpha(P(X|E), P(Y|E))$. Let $\alpha > 1$, $z > 0$, $(s_i, s_j) \in \mathcal{Q}$ and $\theta \in \Theta$. By the shift reduction lemma (Lemma 3.1), we have:

$$D_\alpha\left((f(X)+N)_{|s_i,\theta}, (f(X)+N)_{|s_j,\theta}\right) \leq D_\alpha^{(z)}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right) + R_\alpha(\zeta, z).$$

By definition,

$$D_\alpha^{(z)}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right) = \inf_{W\in\mathcal{P}(\mathbb{R}^d); W_\infty\left(W, f(X)_{|s_i,\theta}\right)\leq z} D_\alpha\left(W, f(X)_{|s_j,\theta}\right),$$

and

$$D_\alpha^{(W_\infty(P(f(X)|s_i,\theta),P(f(X)|s_j,\theta)))}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right) = 0.$$

Then,

$$D_\alpha\left((f(X)+N)_{|s_i,\theta}, (f(X)+N)_{|s_j,\theta}\right) \leq R_\alpha\left(\zeta, W_\infty\left(P(f(X)|s_i,\theta), P(f(X)|s_j,\theta)\right)\right) \leq R_\alpha\left(\zeta, \Delta_G\right). \qquad \square$$

B.2. Proof of Corollary 3.1

In order to compute R_α for usual distributions, we use the following result.

Lemma B.1 (R_α calculation criterion)**.** *Let $\alpha > 1$, $d \in \mathbb{N}^*$. Let ζ be a distribution on $\mathbb{R}$. If $\zeta = e^{-g}$ is even, non-decreasing on $\mathbb{R}^+$ and $z \mapsto D_\alpha(\zeta_{-g^{-1}(z)}, \zeta)$ is convex on $\mathbb{R}^+$, then $\sup_{g^{-1}(\sum_{i=1}^d g(x_i))\leq z} D_\alpha(\zeta_{-x}^{\otimes d}, \zeta^{\otimes d}) = D_\alpha(\zeta_{-z}, \zeta)$, where $\zeta^{\otimes d}$ is the joint distribution of d independent random variables drawn from ζ, and g^{-1} is the inverse of g on $\mathbb{R}^+$.*

Proof. We start by proving the following characterization of convex functions: if g is a convex function of $\mathbb{R}$, then for $z_1 \leq z_2$, the function $x \mapsto g(x - z_1) - g(x - z_2)$ is non-decreasing. A similar statement and its proof can be found in (Saumard & Wellner, 2014).

Let $x \leq x'$. By taking $\lambda = \frac{x-x'}{x-x'+z_1-z_2}$, we have:

$$x - z_1 = (1-\lambda)(x' - z_1) + \lambda(x - z_2) \text{ and } x' - z_2 = \lambda(x' - z_1) + (1-\lambda)(x - z_2).$$

By convexity:

$$g(x - z_1) \leq (1-\lambda)g(x' - z_1) + \lambda g(x - z_2) \text{ and } g(x' - z_2) = \lambda g(x' - z_1) + (1-\lambda)g(x - z_2).$$

Then, $g(x' - z_1) - g(x' - z_2) \geq g(x - z_1) - g(x - z_2)$.

We also prove that if g is convex, $g(0) = 0$ and $y_1, \dots, y_d \in \mathbb{R}^+$, $g(\sum_{i=1}^d y_i) \geq \sum_{i=1}^d g(y_i)$. We prove this assertion by induction. Obviously, $g(y_1) \geq g(y_1)$. Also, $\sum_{i=1}^d y_i \geq 0$. Then, $x \mapsto g(x + \sum_{i=1}^d y_i) - g(x)$ is non-decreasing. For $y_{d+1} \geq 0$, we have $g(\sum_{i=1}^{d+1} y_i) - g(y_{d+1}) \geq g(\sum_{i=1}^d y_i) - g(0)$, and by induction $g(\sum_{i=1}^{d+1} y_i) \geq \sum_{i=1}^{d+1} g(y_i)$.

Then, if $z \mapsto D_\alpha(\zeta_{-g^{-1}(z)}, \zeta)$ is convex, for $x \in \mathbb{R}^{d+}$, we have:

$$D_\alpha(\zeta_{-x}^{\otimes d}, \zeta^{\otimes d}) = \sum_{i=1}^d D_\alpha(\zeta_{-g^{-1}(g(x_i))}, \zeta) \geq D_\alpha\left(\zeta_{-g^{-1}(\sum_{i=1}^d g(x_i))}, \zeta\right),$$

which proves the statement. □

As we only considered shifts in $\mathbb{R}^{d+}$, we also need $z \mapsto D_\alpha(\zeta_{-z}, \zeta)$ to be even, which is achieved for symmetrical densities: if ζ is symmetric, $z \mapsto D_\alpha(\zeta_{-z}, \zeta)$ is also symmetric. In fact:

$$\int_{-\infty}^{+\infty} \frac{\zeta(x-z)^\alpha}{\zeta(x)^{\alpha-1}} dx = \int_{-\infty}^{+\infty} \frac{\zeta(x+z)^\alpha}{\zeta(x)^{\alpha-1}} dx$$

by symmetry and changes of variable.

Corollary 3.1 (Privacy guarantees for usual noise distributions)**.** *We note I_d the identity matrix of size d. Plugging the expressions of $R_\infty(\zeta, z)$ and $R_\alpha(\zeta, z)$ for Laplacian and Gaussian distributions, we obtain:*

- $\mathcal{M}(X) = f(X) + N$ *with* $N \sim \mathcal{N}\big(0, \frac{\alpha\Delta_G^2}{2\varepsilon} I_d\big)$ *and* Δ_G *computed on the* l_2 *norm is* (α, ε)*-RPP.*
- $\mathcal{M}(X) = f(X) + L$ *with* $L \sim \mathrm{Lap}(0, \rho I_d)$ *and* Δ_G *computed on the* l_1 *norm is* $\Big(\alpha, \frac{1}{\alpha-1}\log\big(\frac{\alpha}{2\alpha-1}e^{\Delta_G(\alpha-1)/\rho} + \frac{\alpha-1}{2\alpha-1}e^{-\Delta_G\alpha/\rho}\big)\Big)$*-RPP.*
- $\mathcal{M}(X) = f(X) + L$ *with* $L \sim \mathrm{Lap}\left(0, \frac{\Delta_G}{\varepsilon} I_d\right)$ *with* Δ_G *computed on the* l_1 *norm is* ε*-PP.*

Proof. The result is directly obtained by plugging Rényi divergences into the GWM and using Lemma B.1. Let $\alpha > 1, z \geq 0$.

- $\mathrm{Lap}(0, \rho I_d)$ is symmetric and $z \mapsto g^{-1}(\sum_{i=1}^d g(z_i))$ is the l_1 norm for $g : z \mapsto |z|$. For $L \sim \mathrm{Lap}(0, \rho I_d)$,

  $$D_\alpha(L + z, L) = \frac{1}{\alpha-1}\log\left(\frac{\alpha}{2\alpha-1}e^{|z|(\alpha-1)/\rho} + \frac{\alpha-1}{2\alpha-1}e^{-|z|\alpha/\rho}\right).$$

  Also, for $z \geq 0$:

  $$\frac{d}{dz^2}D_\alpha(L + z, L) = \frac{\alpha(2\alpha^2-1)}{\rho^2(2\alpha-1)}\frac{e^{-z/\rho}}{\alpha e^{z(\alpha-1)/\rho} + (\alpha-1)e^{-z\alpha/\rho}} \geq 0.$$

- $\mathcal{N}(0, \sigma^2 I_d)$ is symmetric and $z \mapsto g^{-1}(\sum_{i=1}^d g(z_i))$ is the l_2 norm for $g : z \mapsto z^2$. For $N \sim \mathcal{N}(0, \sigma^2 I_d)$, $D_\alpha(N + z, N) = \frac{\alpha z^2}{2\sigma^2}$.

□

B.3. Utility of the GWM (Proposition 3.1)

Below, we make the informal result of Proposition 3.1 precise and provide its proof.

Proposition B.1 (Utility of the GWM)**.** *Let $n, d, d_1, \dots, d_n \in \mathbb{N}^*$, $\mathcal{X} \subset \mathbb{R}^d$. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish framework such that, for each $\theta \in \Theta$, $\theta = \otimes_{k=1}^n \theta_k$, with $\theta_k \in \mathcal{P}(\mathcal{X}^{d_k})$. We note $X = (X_1^1, \dots, X_{d_1}^1, \dots, X_{d_n}^n) \sim \theta$. We assume that $s_{i,k}^a = \{X_i^k = a\} \in \mathcal{S}$ and $\mathcal{Q} = \{(s_{i,k}^a, s_{i,k}^b); k \in \{1, \dots, n\}, i \in \{1, \dots, d_k\}, a, b \in \mathcal{X}\}$. Following Song et al. (2017), we define the corresponding group differential privacy of the Pufferfish framework as: $G_k = (x_1^k, \dots, x_{d_k}^k) \in \mathcal{X}^{d_k}$ and $D_k = \{(x, x') \in \mathcal{X}^{d_k}$ such that x and x' only differ in $G_k\}$.*

$$\Delta_{GROUP}(f) = \max_{k \in \{1,\dots,n\}} \max_{(x,x') \in D_k} \|f(x) - f(x')\|.$$

Then, $\Delta_G \leq \Delta_{GROUP}(f)$.

Proof. Let $(s_{i,l}^a, s_{i,l}^b) \in \mathcal{Q}, \theta \in \Theta$, with $\theta = \otimes_{k=1}^n \theta_k$. Let $Y \sim P(f(X)|s_{i,l}^a, \theta)$. Let $Z \sim \theta_l | s_{i,l}^b$ drawn independently from Y. For $k \in [\![1, n]\!], i \in [\![1, d_k]\!]$. We define $Y_i^{'k} = \begin{cases} Y_i^k \text{ if } k \neq l \\ Z_i \text{ else} \end{cases}$ and $Y' = (Y_1^{'1}, \dots, Y_{d_1}^{'1}, \dots, Y_{d_n}^{'n})$.

Then, $(Y, Y') \in D_l, Y' \sim P(f(X)|s_{i,l}^b)$ and:

$$\|Y - Y'\| \leq \max_{(x,x') \in D_l} \|f(x) - f(x')\| \leq \Delta_{GROUP}(f).$$

Then, $W_\infty(P(f(X)|s_{i,l}^a), P(f(X)|s_{i,l}^b)) \leq \Delta_{GROUP}(f)$ and $\Delta_G \leq \Delta_{GROUP}(f)$. □

C. Approximate General Wasserstein Mechanism (Section 4.1)

Our result relies on the following characterization of (z, δ)-proximity.

Lemma C.1. *μ and ν are (z, δ)-near iff $\exists X \sim \mu, Y \sim \nu$ and $V \in \mathcal{P}(\mathbb{R}^d)$ such that $X + V = Y$ and $P(\|V\| > z) < \delta$.*

Proof. Let $z \geq 0$, $\delta \in (0, 1)$, μ, ν two distributions on $\mathbb{R}^d$ such that μ and ν are (z, δ)-near. Then, there exists π a coupling between μ and ν such that $\int_{\mathcal{R}} d\pi(x, y) \geq 1 - \delta$ and $\forall (x, y) \in \mathcal{R}, \|x - y\| \leq z$. We note $V = Y - W$ where (W, Y) is drawn from the coupling π. We observe that $\mathcal{R} \subset \{(x, y); \|x - y\| \leq z\}$.

Then, $P(\|V\| > z) \leq P((W, Y) \notin \mathcal{R}) = \int_{\mathbb{R}^d \setminus \mathcal{R}} d\pi(x, y) < \delta$.

For the opposite side, consider the coupling π of the pair (W, Y) such that $W \sim \mu, Y \sim \nu$ and $W + V = Y$ with $P(\|V\| > z) < \delta$.

Then, $P(\|V\| \leq z) = \int_{\|x-y\| < z} d\pi(x, y) \geq 1 - \delta$. □

C.1. Proof of Lemma 4.1

Lemma 4.1 (Approximate shift reduction)**.** *Let μ, ν, ζ be three distributions on $\mathbb{R}^d$. We denote $D_\alpha^{(z,\delta)}(\mu, \nu) = \inf_{\mu, \mu' \ (z,\delta)\text{-near}} D_\alpha(\mu', \nu)$. Then, for all $\delta \in (0, 1)$, there exists an event E such that $P(E) \geq 1 - \delta$ and:*

$$D_\alpha\left((\mu * \zeta)_{|E}, (\nu * \zeta)\right) \leq D_\alpha^{(z,\delta)}(\mu, \nu) + R_\alpha(\zeta, z) + \frac{\alpha}{\alpha - 1} \log\left(\frac{1}{1 - \delta}\right).$$

Proof. Let $\alpha > 1$, $z > 0$, $X \sim \mu, Y \sim \nu, N \sim \zeta$ and $W \sim \xi \in \mathcal{P}(\mathbb{R}^d)$ such that $P(\|W\| \geq z) = \delta$ and N is independent of X, Y and W. We use the abuse of notation $D_\alpha(\mu, \nu) = D_\alpha(X, Y)$, with $X \sim \mu$, $Y \sim \nu$. We consider the event $E = \{\|W\| \leq z\}$. Like in the original proof of the shift reduction lemma of Feldman et al. (2018), we have:

$$D_\alpha((X + N)_{|E}, Y + N) = D_\alpha((X + W + N - W)_{|E}, Y + N) \leq D_\alpha((X + W, N - W)_{|E}, (Y, N)).$$

by post-processing (Proposition 2.1) for $\mathcal{M}_1(x,y) = x + y$. Then, we have:

$$\begin{aligned}
&D_\alpha((X+W, N-W)_{|E}, (Y,N)) \\
&= \frac{1}{\alpha-1}\log\left(\int \frac{P_{(X+W,N-W)_{|E}}(x,y)^\alpha}{P_{Y,N}(x,y)^{\alpha-1}}dxdy\right) \\
&= \frac{1}{\alpha-1}\log\left(\int \frac{P_{X+W|E}(x)^\alpha P_{N-W|E,X+W=x}(y)^\alpha}{\nu(x)^{\alpha-1}\zeta(y)^{\alpha-1}}dxdy\right) \\
&= \frac{1}{\alpha-1}\log\left(\int \frac{P_{X+W|E}(x)^\alpha}{\nu(x)^{\alpha-1}}\left(\int \frac{P_{N-W|E,X+W=x}(y)^\alpha}{\zeta(y)^{\alpha-1}}dy\right)dx\right) \\
&= \frac{1}{\alpha-1}\log\int \frac{P_{X+W|E}(x)^\alpha}{\nu(x)^{\alpha-1}}\left(\int \frac{\left(\int_{\|u\|\le z} P_{N-W|X+W=x,W=u}(y)\xi(u)du\right)^\alpha}{\zeta(y)^{\alpha-1}}dy\right)dx \\
&\le \frac{1}{\alpha-1}\log\left(\int \frac{P_{X+W|E}(x)^\alpha}{\nu(x)^{\alpha-1}}\left(\int_{\|u\|\le z} \frac{\zeta(y+u)^\alpha}{\zeta(y)^{\alpha-1}}\xi(u)dudy\right)dx\right) \\
&\le \frac{1}{\alpha-1}\log\left(\int \frac{P_{X+W|E}(x)^\alpha}{\nu(x)^{\alpha-1}}dx\right) + R_\alpha(\zeta,z).
\end{aligned}$$

Yet,

$$P_{X+W|E}(x)^\alpha = \left(\frac{P_{X+W}(x) - P(\bar{E})P_{X+W|\bar{E}}(x)}{P(E)}\right)^\alpha \le \frac{P_{X+W}(x)^\alpha}{(1-\delta)^\alpha}.$$

Thus:

$$\begin{aligned}
&D_\alpha((X+W, N-W)_{|E}, (Y,N)) \\
&\le D_\alpha(X+W, Y) + R_\alpha(\zeta,z) - \frac{\alpha}{\alpha-1}\log(1-\delta).
\end{aligned}$$

□

C.2. Relationship between $(+\infty, \varepsilon, \delta)$-approximate RPP and (ε, δ)-PP

Proposition C.1. *If $\mathcal{M}$ is $(+\infty, \varepsilon, \delta)$-approximate RPP, then it is (ε, δ)-PP.*

Proof. The proof uses the same approach as the Lemma 8.8 of Bun & Steinke (2016). Let $(s_i, s_j) \in \mathcal{Q}$, $\theta \in \Theta$. Without loss of generality, we assume that there exists E, E' such that $P(E) = 1-\delta, P(E') = 1-\delta$ and we have: $D_\infty\left(P\left(\mathcal{M}(X) = w \mid s_i, \theta, E\right), P\left(\mathcal{M}(X) = w \mid s_j, \theta, E'\right)\right) \le \varepsilon$. Then,

$$\sup_{w \in Range(\mathcal{M})} \log \frac{P\left(\mathcal{M}(X) = w \mid s_i, \theta, E\right)}{P\left(\mathcal{M}(X) = w \mid s_j, \theta, E'\right)} \le \varepsilon.$$

$$\begin{aligned}
P\left(\mathcal{M}(X) = w \mid s_j, \theta\right) &= P(E')P\left(\mathcal{M}(X) = w \mid s_j, \theta, E'\right) + P(\bar{E}')P\left(\mathcal{M}(X) = w \mid s_j, \theta, \bar{E}'\right) \\
&\ge (1-\delta)P\left(\mathcal{M}(X) = w \mid s_j, \theta, E'\right), \\
P\left(\mathcal{M}(X) = w \mid s_i, \theta\right) &= P(E)P\left(\mathcal{M}(X) = w \mid s_i, \theta, E\right) + P(\bar{E})P\left(\mathcal{M}(X) = w \mid s_i, \theta, \bar{E}\right) \\
&\le (1-\delta)P\left(\mathcal{M}(X) = w \mid s_i, \theta, E\right) + \delta \\
&\le (1-\delta)P\left(\mathcal{M}(X) = w \mid s_j, \theta, E'\right)e^\varepsilon + \delta \\
&\le P\left(\mathcal{M}(X) = w \mid s_j, \theta\right)e^\varepsilon + \delta.
\end{aligned}$$

□

Proposition 4.1. *If $\mathcal{M}$ is $(\alpha, \varepsilon, \delta)$-approximate RPP, then it is $(\varepsilon', 2\delta)$-PP, with $\varepsilon' = \varepsilon + \frac{\log(1/\delta)}{\alpha-1}$.*

Proof. We use the proof techniques of Proposition C.1 and Proposition 2.2. Let $\varepsilon \ge 0, \alpha > 1$. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish privacy framework and $\mathcal{M}$ an $(\alpha, \varepsilon, \delta)$-RPP mechanism. Let $\delta \in (0,1)$, $\theta \in \Theta$, $(s_i, s_j) \in \mathcal{Q}$ and $z \in Range(\mathcal{M})$.

There exists E, E' such that $D_\alpha\left(P\left(\mathcal{M}(X) \mid s_i, \theta, E\right), P\left(\mathcal{M}(X) \mid s_j, \theta, E'\right)\right) \leq \varepsilon$ and $P(E), P(E') \geq 1-\delta$, The proof technique of Proposition 2.2 allows to show that:

$$P\left(\mathcal{M}(X)=z|E, s_i, \theta\right) \leq e^{\varepsilon+\frac{\log(1/\delta)}{\alpha-1}} P\left(\mathcal{M}(X)=z|E', s_j, \theta\right)+\delta.$$

Then, the proof technique of Proposition C.1 allows to show that:

$$P\left(\mathcal{M}(X)=z|s_i, \theta\right) \leq e^{\varepsilon+\frac{\log(1/\delta)}{\alpha-1}} P\left(\mathcal{M}(X)=z|, s_j, \theta\right)+2\delta.$$

□

C.3. Proof of Theorem 4.3

Theorem 4.3 (General approximate Wasserstein mechanism)**.** *Let $f: \mathcal{D} \rightarrow \mathbb{R}^d$ be a numerical query. For all $\delta \in (0,1)$, let us denote:*

$$\begin{aligned}\Delta_{G,\delta} > \inf\{&z \in \mathbb{R}; \forall (s_i, s_j) \in S, \forall \theta \in \Theta, \\ &(P((f(X)|s_i, \theta), P(f(X)|s_j, \theta)) \text{ are } (z,\delta)\text{-near}\}.\end{aligned}$$

Let $N=(N_1, \ldots, N_d) \sim \zeta$, where $N_1, \ldots, N_d$ are iid real random variables independent of the dataset X. Then, $\mathcal{M}=f(X)+N$ satisfies $(\alpha, R_\alpha(\zeta, \Delta_{G,\delta})+\frac{\alpha}{\alpha-1} \log \frac{1}{1-\delta}, \delta)$-approximate RPP for all $\alpha \in (1,+\infty)$ and $(R_\infty(\zeta, \Delta_{G,\delta})+\log \frac{1}{1-\delta}, \delta)$-PP.

Proof. This proof is similar to Theorem 3.3 but we use the approximate shift reduction lemma (Lemma 4.1). We use the abuse of notation $D_\alpha(X_{|E}, Y_{|E})=D_\alpha(P(X|E), P(Y|E))$. Let $f: \mathcal{D} \rightarrow \mathbb{R}^d$ be a numerical query and $N=(N_1, \ldots, N_d) \sim \zeta$, where $N_1, \ldots, N_d$ are iid real random variables independent of the data X. Let $\delta \in (0,1)$. Let us denote:

$$\Delta_{G,\delta} > \inf\{z \in \mathbb{R}; \forall (s_i, s_j) \in S, \forall \theta \in \Theta, (P(f(X)|s_i, \theta), P(f(X)|s_j, \theta)) \text{ are } (z,\delta)\text{-near}\}.$$

By the approximate shift reduction lemma (Lemma 4.1), there exists E such that $P(E) \geq 1-\delta$ and:

$$D_\alpha\left((f(X)+N)_{|E,s_i,\theta}, (f(X)+N)_{|s_j,\theta}\right) \leq D_\alpha^{(z,\delta)}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right)+R_\alpha(\zeta, z)-\frac{\alpha}{\alpha-1} \log(1-\delta).$$

By definition,

$$D_\alpha^{(z,\delta)}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right)=\inf_{\mu \in \mathcal{P}(\mathbb{R}^d); \mu, P(f(X)|s_i,\theta) \text{ are } (z,\delta)\text{-near}} D_\alpha\left(\mu, P(f(X) \mid s_j, \theta)\right),$$

and

$$D_\alpha^{(\Delta_{G,\delta},\delta)}\left(f(X)_{|s_i,\theta}, f(X)_{|s_j,\theta}\right)=0.$$

Then,

$$D_\alpha\left((f(X)+N)_{|E,s_i,\theta}, (f(X)+N)_{|s_j,\theta}\right) \leq R_\alpha(\zeta, \Delta_{G,\delta})-\frac{\alpha}{\alpha-1} \log(1-\delta).$$

□

C.4. Result for Usual Noise Distributions

We provide below a corollary of Theorem 4.3 that gives closed formula for usual noise distributions to get approximate RPP guarantees.

Proposition C.2 (Approximate Wasserstein mechanism)**.** *We note I_d the identity matrix of size d. The results are similar to those of the general Wasserstein mechanism (Corollary 3.1), but with an additive term which depends on δ:*

- *$\mathcal{M}(X)=X+N$ with $N \sim \mathcal{N}\left(0, \frac{\alpha \Delta_{G,\delta}^2}{2\left(\varepsilon+\frac{\alpha}{\alpha-1} \log(1-\delta)\right)} I_d\right)$ is $(\alpha, \varepsilon, \delta)$-approximate RPP.*
- *$\mathcal{M}(X)=X+L$ with $L \sim \operatorname{Lap}(0, \rho I_d)$ is $\left(\alpha, \frac{1}{\alpha-1}\left(\log(b)-\alpha \log(1-\delta)\right), \delta\right)$-approximate RPP for $b=\frac{\alpha}{2\alpha-1} e^{\Delta_{G,\delta}(\alpha-1)/\rho}+\frac{\alpha-1}{2\alpha-1} e^{-\Delta_{G,\delta}\alpha/\rho}$.*
- *$\mathcal{M}(X)=X+L$ with $L \sim \operatorname{Lap}\left(0, \frac{\Delta_{G,\delta}}{\varepsilon+\log(1-\delta)} I_d\right)$ is (ε, δ)-PP.*

C.5. Relationship with Distribution Privacy Results of Chen & Ohrimenko (2023)

We start by recalling the definition of distribution privacy.

Definition C.1 (Distribution privacy (Chen & Ohrimenko, 2023))**.** A mechanism $\mathcal{M}$ satisfies (ε, δ)-distribution privacy with respect to a set of distribution pairs $\Psi \subset \Theta \times \Theta$ if for all pairs $(\psi_i, \psi_j) \in \Psi$ and all subsets $S \subset \text{Range}(\mathcal{M})$,

$$P(\mathcal{M}(X) \in S|\psi_i) \leq e^{\varepsilon} P(\mathcal{M}(X) \in S|\psi_j) + \delta,$$

where the expression $P(\mathcal{M}(X) \in S|\psi)$ denotes the probability that $\mathcal{M}(X)$ given $X \sim \psi$.

For completeness, we recall the original approximate Wasserstein mechanism Theorem for distribution privacy from (Chen & Ohrimenko, 2023).

Theorem C.2 (Approximate Wasserstein mechanism for distribution privacy (Chen & Ohrimenko, 2023))**.** *Let (Ψ, Θ) be a distribution privacy framework. Let $W > 0$, $\delta \in (0, 1)$. Suppose that for all $(\psi_i, \psi_j) \in \Psi$, $P(X \mid \psi_i)$ and $P(X \mid \psi_j)$ are (W, δ)-near. Then $\mathcal{M}(X) = X + L$ where $L \sim Lap\big(0, \frac{W}{\varepsilon} I\big)$ is (ε, δ)-distribution private.*

We now formally state and prove the equivalence between Pufferfish privacy and distribution privacy.

Proposition C.3. *Let $(E, \mathcal{B}(E))$ be a measurable space, where $|E| \leq \aleph_1$ is a topological space with its Borel σ-algebra $\mathcal{B}(E)$ and $\aleph_1$ is the cardinality of $\mathbb{R}$. Let $\Theta \subset \mathcal{P}(\mathcal{B}(E))$. Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish privacy instance and $\mathcal{M}$ a randomized mechanism. Then, there exists a distribution privacy instance (Ψ, Θ') such that $\mathcal{M}$ is (ε, δ)-PP iff $\mathcal{M}$ is (ε, δ)-distribution private. Conversely, let (Ψ, Θ) be a distribution privacy instance. Then, there exists a Pufferfish privacy instance $(\mathcal{S}, \mathcal{Q}, \Theta')$ such that $\mathcal{M}$ is (ε, δ)-PP iff $\mathcal{M}$ is (ε, δ)-distribution private.*

Remark C.3. The condition $|E| \leq \aleph_1$ is quite general. In particular, it allows the data space to be (a subset of) $\mathbb{R}^d$, thus covering typical data domains found in fields like data analysis, machine learning, text processing, computer vision, and database management.

Proof. We show the equivalence between the Pufferfish privacy framework and the distribution privacy framework. Let $\Theta \subset \mathcal{P}(\mathcal{B}(E))$, where $|E| \leq \aleph_1$.

- Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish privacy instance. We consider:

$$\Psi = \{(P(X|s_i, \theta), P(X|s_j, \theta)) \text{ such that } (s_i, s_j) \in \mathcal{Q}, \theta \in \Theta \text{ and } P(s_i \mid \theta) \neq 0, P(s_j \mid \theta) \neq 0\}.$$

  Then,

$$\begin{aligned}
&\forall w \in \text{Range}(\mathcal{M}), \forall (\psi_i, \psi_j) \in \Psi, \\
&P(\mathcal{M}(X) = w \mid \psi_i) \leq e^{\varepsilon} P(\mathcal{M}(X) = w \mid \psi_j) + \delta \\
&\iff \\
&\forall w \in \text{Range}(\mathcal{M}), \forall (s_i, s_j) \in \mathcal{Q}, \theta \in \Theta \text{ such that } P(s_i \mid \theta) \neq 0, P(s_j \mid \theta) \neq 0, \\
&P(\mathcal{M}(X) = w \mid s_i, \theta) \leq e^{\varepsilon} P(\mathcal{M}(X) = w \mid s_j, \theta) + \delta.
\end{aligned}$$

- Let (Ψ, Θ) be a distribution privacy instance. First, we consider the case where each $\psi \in \Theta$ is parametrized by a vector $\rho \in \mathbb{R}^d$, which means that there exists a bijection between a subset of $\mathbb{R}^d$ and Θ. For $\rho \in \mathbb{R}^d$, if it exists, we denote $\psi_\rho \in \Theta$ the corresponding distribution. Then, we denote $\Phi = \{\rho \in \mathbb{R}^d \text{ such that } \exists \psi \in \Psi; (\psi_\rho, \psi) \in \Psi \vee (\psi, \psi_\rho) \in \Psi\}$ and $\Omega = \{(\rho_1, \rho_2) \in \Phi \times \Phi \text{ such that } (\psi_{\rho_1}, \psi_{\rho_2)} \in \Psi\} \subset \mathbb{R}^{n \times 2}$ and $\Pi = \{\pi \in P(\mathcal{B}(\mathbb{R}^d)) \text{ such that } supp(\pi) = \Phi\}$. We consider:

$$\begin{aligned}
\mathcal{S} &= \{(s_\rho = \text{“}X \text{ has been generated from the distribution } \psi_\rho\text{”}), \forall \rho \in \Phi\}, \\
\mathcal{Q} &= \{(s_{\rho_1}, s_{\rho_2}) \text{ such that } (\rho_1, \rho_2) \in \Omega\}, \\
\Theta' &= \left\{\theta_\pi \in \mathcal{P}(\mathcal{B}(E)) \text{ such that } \pi \in \Pi \wedge P(X|\theta_\pi) = \int_\Phi \pi(\rho) P(X|\psi_\rho) d\rho\right\}.
\end{aligned}$$

Then, $\forall w \in \text{Range}(\mathcal{M}), \forall (s_i, s_j) \in \mathcal{Q}, \theta_\pi \in \Theta'$, $P(X|\theta_\pi, s_i) = P(X|\psi_i)$. Thus, we have:

$$\begin{aligned}
&\forall w \in \text{Range}(\mathcal{M}), \forall (\psi_i, \psi_j) \in \Psi, \\
&P(\mathcal{M}(X) = w \mid \psi_i) \leq e^\varepsilon P(\mathcal{M}(X) = w \mid \psi_j) + \delta \\
&\iff \\
&\forall w \in \text{Range}(\mathcal{M}), \forall (s_i, s_j) \in \mathcal{Q}, \theta \in \Theta \text{ such that } P(s_i \mid \theta) \neq 0, P(s_j \mid \theta) \neq 0, \\
&P(\mathcal{M}(X) = w \mid s_i, \theta) \leq e^\varepsilon P(\mathcal{M}(X) = w \mid s_j, \theta) + \delta.
\end{aligned}$$

In this proof, the case $|\Theta| = n \in \mathbb{N}^*$ is a case where $\psi \in \Theta$ can be parameterized. One such parameterization is to define $\Theta = \{\psi_1, \ldots, \psi_n\}$ and the mapping $i \in \mathbb{N} \mapsto \psi_i \in \Theta$.

The second part of the proof relies on the fact that the distributions of $\mathcal{P}(\mathcal{B}(E))$ are parameterizable. The hypothesis $|E| \leq \aleph_1$ allows us to reduce to the case $E = \mathbb{R}$, up to a bijection. Yet, every distribution of $\mathcal{B}(\mathbb{R})$ is entirely defined by its values taken on open intervals of $\mathbb{R}$ and each open interval of $\mathbb{R}$ is a countable union of open intervals with rational endpoints. Therefore, $|\mathcal{P}(\mathcal{B}(\mathbb{R}))| \leq 2^{\aleph_0} = \aleph_1$, where the notation $\aleph_0$ denotes the cardinal of $\mathbb{N}$ and we can map every distribution of $\mathbb{R}$ with elements of $\mathbb{R}$.

□

Remark C.4. The proof shows how to transition from the Pufferfish privacy framework to the distribution privacy framework. Thus, it is possible to use Pufferfish private mechanisms to achieve distribution privacy guarantees (and vice versa).

This equivalence result allows us to precisely compare our result (Theorem 4.3) to the result of Chen & Ohrimenko (2023). Our approximate shift reduction result (Lemma 4.1) induces an additional term which prevents us from recovering exactly the results of Chen & Ohrimenko (2023) in the particular case of the Laplace mechanism for PP. However, we believe that our analysis can be improved and lead to better results. More generally, our result can be used with a wide range of noise distributions and in the RPP framework, which is more general than PP (and thus more general than distribution privacy).

C.6. Utility of the GAWM (Proposition 4.2)

Below, we make the informal result of Proposition 4.2 precise and provide its proof.

Proposition C.4 (Utility of the GAWM). *Let $(\mathcal{S}, \mathcal{Q}, \Theta)$ be a Pufferfish framework, $\delta \in (0,1), \alpha > 1$ and let $\mathcal{M}(X) = f(X) + N$, where $X \sim \theta \in \Theta$, $N \sim \zeta$ and f is a numerical query. Then, Δ_G as defined in Theorem 3.3 is greater or equal than $\Delta_{G,\delta}$ defined in Theorem 4.3. Moreover, if $R_\alpha(\Delta_{G,\delta}, \zeta) \leq R_\alpha(\Delta_G, \zeta) + \frac{\alpha}{\alpha-1}\log(1-\delta)$ then the GAWM achieves better utility than the GWM with $(\alpha, \varepsilon, \delta)$-RPP, without additional privacy cost on the ε. It happens when Δ_G is sufficiently larger than $\Delta_{G,\delta}$, which happens when there exists $(s_i, s_j) \in \mathcal{Q}, \theta \in \Theta$ and $(Y, Y') \sim \pi \in \Gamma(P(f(X)|s_i, \theta), P(f(X)|s_j, \theta))$ such that $\|Y - Y'\|$ is large with small probability.*

Proof. Let $(s_i, s_j) \in \mathcal{Q}, \theta \in \Theta$. Then, there exists $(Y, Y') \sim \pi \in \Gamma(P(f(X)|s_i, \theta), P(f(X)|s_j, \theta))$ such that $P(\|Y - Y'\| > \Delta_G) = 0$. Then, for any $\delta \in (0,1)$, $P(\|Y - Y'\| > \Delta_G) < \delta$ and by Lemma C.1, Y and Y' are (Δ_G, δ)-near. Finally, $\Delta_{G,\delta} \leq \Delta_G$. □

D. Leveraging W_p metrics (Section 4.2)

D.1. Proof of Lemma 4.2

Lemma 4.2 (Generalized shift reduction). *Let ζ be a noise distribution of $\mathbb{R}^d$. Let $z, p, q > 0$ such that $1/p + 1/q = 1$. We note :*

$$D^{(z)}_{\alpha,\alpha',\zeta}(\mu, \nu) = \inf_{\xi;\ \mathop{\mathbb{E}}_{W\sim\xi}[\exp((\alpha'-1)D_{\alpha'}(\zeta, \zeta * W))] \leq z} D_\alpha(\mu * \xi, \nu).$$

Then, we have :

$$D_\alpha(\mu * \zeta, \nu * \zeta) \leq D^{(z)}_{p(\alpha-1)+1, q(\alpha-1)+1, \zeta}(\mu, \nu) + \frac{\log(z)}{q(\alpha-1)}.$$

In the case $q = 1$:

$$D_\alpha(\mu * \zeta, \nu * \zeta) \leq D^{(z)}_{\infty,\alpha,\zeta}(\mu, \nu) + \frac{\log(z)}{\alpha - 1}.$$

Proof. The proof construction is similar to the one developed in (Chen & Ohrimenko, 2023). We do not apply Jensen inequality at the last step of the proof to obtain Orlicz-Wasserstein metrics, and keep the result general and working for a broader range of distributions. We use the abuse of notation $D_\alpha(\mu,\nu) = D_\alpha(X,Y)$, with $X \sim \mu$, $Y \sim \nu$. Let $z > 0$, $X \sim \mu, Y \sim \nu, N \sim \zeta \in \mathcal{P}(\mathbb{R}^d)$ be a noise distribution and $W \sim \xi \in \mathcal{P}(\mathbb{R}^d)$ such that:

$$\mathbb{E}_W[\exp(q(\alpha-1)D_{q(\alpha-1)+1}(\zeta, \zeta * W))] \leq z.$$

Let $p, q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$. We want to compute : $D_\alpha(X+N, Y+N)$. By the post processing theorem applied on the map $f : (x,y) \to x+y$, and the fact that $X+N = X+W-W+N$, we have :

$$D_\alpha(X+N, Y+N) \leq D_\alpha((X+W, N-W), (Y,N)).$$

We have:

$$\begin{aligned}
D_\alpha((X+W,N-W),(Y,N)) &= \frac{1}{\alpha-1}\log\left(\int \frac{P_{(X+W,N-W)}(x,y)^\alpha}{P_{Y,N}(x,y)^{\alpha-1}}dxdy\right) \\
&= \frac{1}{\alpha-1}\log\left(\int \frac{P_{X+W}(x)^\alpha P_{N-W|X+W=x}(y)^{\alpha-1}}{\nu(x)^{\alpha-1}\zeta(y)^\alpha}dxdy\right) \\
&= \frac{1}{\alpha-1}\log \mathbb{E}_{\substack{U\sim X+W \\ V\sim N-W|X+W=U}}\left[\left(\frac{P_{X+W}(U)}{\nu(U)}\right)^{\alpha-1}\left(\frac{P_{N-W|X+W=x}(V)}{\zeta(V)}\right)^{\alpha-1}\right] \\
&\leq \frac{1}{p(\alpha-1)}\log \mathbb{E}_{U\sim X+W}\left[\left(\frac{P_{X+W}(U)}{\nu(U)}\right)^{p(\alpha-1)}\right] (1) \\
&+ \frac{1}{q(\alpha-1)}\log \mathbb{E}_{\substack{U\sim X+W \\ V\sim N-W|X+W=U}}\left[\left(\frac{P_{N-W|X+W=x}(V)}{\zeta(V)}\right)^{q(\alpha-1)}\right] (2) \text{ by Hölder inequality}
\end{aligned}$$

Immediately $(1) = D_{p(\alpha-1)+1}(X+W, Y)$ and, given that

$$\begin{aligned}
P_{N-W|X+W=x}(y)^{q(\alpha-1)+1} &= \left(\int P_{N-W|W=z}(y)\xi(z)dz\right)^{q(\alpha-1)+1} \\
&= \mathbb{E}_W\left[\zeta(y+W)\right]^{q(\alpha-1)+1} \\
&\leq \mathbb{E}_W\left[\zeta(y+W)^{q(\alpha-1)+1}\right],
\end{aligned}$$

we have:

$$\begin{aligned}
(2) &= \frac{1}{q(\alpha-1)}\log\int\left(\frac{P_{N-W|X+W=x}(y)}{\zeta(y)}\right)^{q(\alpha-1)} P_{X+W}(x)P_{N-W|X+W=x}(y)dxdy \\
&\leq \frac{1}{q(\alpha-1)}\log\int\frac{\zeta(y+u)^{q(\alpha-1)+1}}{\zeta(y)^{q(\alpha-1)}}\xi(u)P_{X+W}(x)dxdydu \\
&\leq \frac{1}{q(\alpha-1)}\log \mathbb{E}_{W\sim\xi}\left[\exp(q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*W))\right] \\
&\leq \frac{\log(z)}{q(\alpha-1)}.
\end{aligned}$$

In the case $p = +\infty$, let $W \sim \xi \in \mathcal{P}(\mathbb{R}^d)$ such that:

$$\mathbb{E}_W[\exp((\alpha-1)D_\alpha(\zeta,\zeta*W))] \leq z.$$

$$\begin{aligned}
D_\alpha((X+W,N-W),(Y,N)) &\leq \sup_{U\sim X+W}\frac{1}{\alpha-1}\log\left(\frac{P_{X+W}(U)}{\nu(U)}\right)^{\alpha-1} (3) \\
&+ \frac{1}{(\alpha-1)}\log \mathbb{E}_{\substack{U\sim X+W \\ V\sim N-W|X+W=U}}\left[\left(\frac{P_{N-W|X+W=x}(V)}{\zeta(V)}\right)^{\alpha-1}\right] (4)
\end{aligned}$$

Yet, $(3) = D_\infty(P_{X+W}, \nu)$ and:

$$\begin{aligned}(4) &= \frac{1}{\alpha-1}\log\int\left(\frac{P_{N-W|X+W=x}(y)}{\zeta(y)}\right)^{\alpha-1}P_{X+W}(x)P_{N-W|X+W=x}(y)dxdy\\ &\leq \frac{1}{\alpha-1}\log\mathbb{E}_{W\sim\xi}\left[\exp((\alpha-1)D_\alpha(\zeta,\zeta*W))\right]\\ &\leq \frac{\log(z)}{\alpha-1}.\end{aligned}$$

□

D.2. Proof of Theorem 4.5

Theorem 4.5 (Distribution Aware General Wasserstein Mechanism)**.** *Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query and ζ a noise distribution of $\mathbb{R}^d$. Let $q\geq 1$. For $(s_i,s_j)\in\mathcal{Q},\theta\in\Theta$, we note $\mu_i^\theta = P(f(X)|s_i,\theta)$. We denote:*

$$\Delta_G^{\zeta,q,\alpha} = \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}}\inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)}\mathbb{E}\left[e^{q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*(X-Y))}\right].$$

Let $N=(N_1,\dots,N_d)\sim\zeta$, where $N_1,\dots,N_d$ are iid real random variables independent of the data X. Then, $\mathcal{M}(X) = f(X)+N$ satisfies $(\alpha,\frac{\log(\Delta_G^{\zeta,q,\alpha})}{q(\alpha-1)})$-RPP for all $\alpha\in(1,+\infty)$ and $\lim_{\alpha\to+\infty}\frac{\log(\Delta_G^{\zeta,q,\alpha})}{q(\alpha-1)}$-PP.

Proof. The proof is similar to Theorem 3.3 but we use the generalized shift reduction lemma (Lemma 4.2). Let $(\mathcal{S},\mathcal{Q},\Theta)$ be a Pufferfish privacy instance. Let $f:\mathcal{D}\to\mathbb{R}^d$ be a numerical query and denote:

$$\Delta_G^{\zeta,q,\alpha} = \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}}\inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)}\mathbb{E}\left[e^{q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*(X-Y))}\right].$$

Let $N=(N_1,\dots,N_d)\sim\zeta$, where $N_1,\dots,N_d$ are iid real random variables independent of the data X. Let $\alpha>1$, $z>0$, $(s_i,s_j)\in\mathcal{Q}$ and $\theta\in\Theta$. We use the abuse of notation $D_\alpha(X_{|E},Y_{|E}) = D_\alpha(P(X|E),P(Y|E))$. By the shift reduction lemma (Lemma 4.2), we have:

$$D_\alpha\left((f(X)+N)_{|s_i,\theta},(f(X)+N)_{|s_j,\theta}\right)\leq D^{(z)}_{p(\alpha-1)+1,q(\alpha-1)+1,\zeta}\left(f(X)_{|s_i,\theta},f(X)_{|s_j,\theta}\right)+\frac{\log(z)}{q(\alpha-1)}.$$

By definition,

$$D^{(z)}_{p(\alpha-1)+1,q(\alpha-1)+1,\zeta}\left(f(X)_{|s_i,\theta},f(X)_{|s_j,\theta}\right) = \inf_{W\in\mathcal{P}(\mathbb{R}^d);\ \mathbb{E}_{W\sim\xi}[\exp(q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*(W-f(X)_{|s_i,\theta})))]\leq z} D_\alpha\left(W,f(X)_{|s_j,\theta}\right),$$

and

$$D^{(\exp(q(\alpha-1)D_{q(\alpha-1)+1}(\zeta,\zeta*(f(X)_{|s_i,\theta}-f(X)_{|s_j,\theta}))))}_{p(\alpha-1)+1,q(\alpha-1)+1,\zeta}\left(f(X)_{|s_i,\theta},f(X)_{|s_j,\theta}\right)=0.$$

Then,

$$D_\alpha\left((f(X)+N)_{|s_i,\theta},(f(X)+N)_{|s_j,\theta}\right)\leq\frac{\log(\Delta_G^{\zeta,q,\alpha})}{q(\alpha-1)}.$$

□

D.3. Proof of Corollary 4.1

Divergences of shifts in Cauchy distributions have been discussed in (Verdú, 2023). We generalize their results for certain types of generalized Cauchy distributions in the following lemma.

Lemma D.1 (Shifts of generalized Cauchy distributions)**.** *Let $k\in\mathbb{N}^*,\alpha>1,\lambda>0$ and $\beta_{k,\lambda}>0$ such that $\zeta_{k,\lambda}:x\mapsto\beta_{k,\lambda}(\frac{1}{1+(\lambda x)^2})^{\frac{k}{2}}$ verifies $\int\zeta_{k,\lambda}(x)dx=1$. Let $X\sim\zeta_{k,\lambda}$ and $z\geq 0$. Then,*

$$D_\alpha(X+z,X)\leq\frac{1}{\alpha-1}\log\frac{\beta_{k,\lambda}\pi}{\lambda}Q_{k(\alpha-1)/2}\left(1+\frac{z^2}{\lambda^2}\right),$$

where $Q_{k(\alpha-1)/2}$ is the Legendre function of the first kind of index $k(\alpha-1)/2$.

Proof. Let $\lambda, z > 0, k \in \mathbb{N}^*$. We have:

$$\begin{aligned}
\int \frac{\zeta_{k,\lambda}(x+z)^\alpha}{\zeta_{k,\lambda}(x)^{\alpha-1}}dx &= \beta_{k,\lambda}\int \frac{\left(1+(\lambda(x-z))^2\right)^{(\alpha-1)k/2}}{(1+(\lambda x)^2)^{\alpha k/2}}dx \\
&= \frac{\beta_{k,\lambda}}{\lambda}\int \frac{\left(1+(u-\lambda z)^2\right)^{(\alpha-1)k/2}}{(1+u^2)^{\alpha k/2}}du \\
&= \frac{\beta_{k,\lambda}}{\lambda}\int_{-\pi/2}^{\pi/2} \frac{\left(1+(\tan(t)-\lambda z)^2\right)^{(\alpha-1)k/2}}{(1+\tan(t)^2)^{\alpha k/2}}(1+\tan^2(t))dt \\
&= \frac{\beta_{k,\lambda}}{\lambda}\int_{-\pi/2}^{\pi/2} \left(1+\tan^2(t)-2\tan(t)\lambda z+\lambda^2 z^2\right)^{(\alpha-1)k/2}(\cos^2(t))^{\alpha k/2-1}dt \\
&= \frac{\beta_{k,\lambda}}{\lambda}\int_{-\pi/2}^{\pi/2} \left(\cos^2(t)(1+\tan^2(t)-2\tan(t)\lambda z+\lambda^2 z^2)\right)^{(\alpha-1)k/2}(\cos^2(t))^{k/2-1}dt \\
&\leq \frac{\beta_{k,\lambda}}{\lambda}\int_{-\pi/2}^{\pi/2} \left(1-2\sin(t)\cos(t)\lambda z+\cos^2(t)\lambda^2 z^2)\right)^{(\alpha-1)k/2}dt \\
&\leq \frac{\beta_{k,\lambda}}{2\lambda}\int_{-\pi}^{\pi} \left(1-\sin(t)\lambda z+(\cos(t)+1)\lambda^2 z^2/2)\right)^{(\alpha-1)k/2}dt \\
&\leq \frac{\beta_{k,\lambda}}{2\lambda}\int_{-\pi}^{\pi} \left(1+\lambda^2 z^2/2-\sin(t)\lambda z+\cos(t)\lambda^2 z^2/2)\right)^{(\alpha-1)k/2}dt \\
&\leq \frac{\beta_{k,\lambda}}{2\lambda}\int_{-\pi}^{\pi} \left(1+\lambda^2 z^2/2+\sqrt{\lambda z+\lambda^2 z^2/2}\cos(t)\right)^{(\alpha-1)k/2}dt,
\end{aligned}$$

And $Q_\alpha(z)$ is defined by:

$$Q_\alpha(z) = \frac{1}{\pi}\int_0^\pi \left(z+\sqrt{z^2-1}\cos(t)\right)^\alpha dt.$$

□

We are now ready to prove Corollary 4.1.

Corollary 4.1 (Cauchy Mechanism)**.** *Let $d \in \mathbb{N}^*$. We denote Q_α the Legendre polynomial of integer index $\alpha > 1$ and $\overline{Q_\alpha}$ as the polynomial derived from Q_α by retaining only its non-negative coefficients. Let $k \geq 2$ and $q \geq 1$ such that $kq(\alpha-1)/2$ is an integer. We note:*

$$\Delta_G^{dkq(\alpha-1)} = \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} W_{dkq(\alpha-1)}\left(P(f(X)|s_i,\theta), P(f(X)|s_j,\theta)\right),$$

with $W_{dkq(\alpha-1)}$ computed with the l_2 norm. Then, $\mathcal{M}(X) = f(X) + V$ with $V = (V_1,\dots,V_d) \overset{iid}{\sim} \mathrm{GCauchy}\,(0,\lambda,k)$ is $\left(\alpha, \frac{d\log\frac{\beta_{k,\lambda}\pi}{\lambda}\overline{Q}_{kq(\alpha-1)/2}\left(1+\left(\frac{\Delta_G^{dkq(\alpha-1)}}{d\lambda}\right)^2\right)}{q(\alpha-1)}\right)$-RPP.

Proof. We apply Theorem 4.5 and compute $\Delta_G^{\zeta,q,\alpha}$ to prove our claim. We start by noticing that $h : z \mapsto \log Q_{k(\alpha-1)/2}(z)$ is convave for $z \geq 0$. It is shown by factorizing the polynomial $Q_{k(\alpha-1)/2}$. It is known that all roots $r_1,\dots,r_{k(\alpha-1)/2}$ of Legendre polynomials are real, distinct from each other and lie in $(-1,1)$. Then, for $z \in \mathbb{R}$, $Q_{k(\alpha-1)/2}(z) = \Pi_{i=1}^{k(\alpha-1)/2}(z-r_i)$.

Then, $h(z) = \log Q_{k(\alpha-1)/2}(z) = \sum_{i=1}^{k(\alpha-1)/2} \log(z - r_i)$, which is concave for $z \geq 0$. Then, for $z_1, \ldots, z_d \geq 0$, we have $\sum_{i=1}^{d} h(z_i) \leq dh(\sum_{i=1}^{d} z_i/d)$. We have:

$$
\begin{aligned}
\Delta_G^{\zeta,q,\alpha} &= \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \mathbb{E}\left[e^{q(\alpha-1)D_{q(\alpha-1)+1}(\zeta^{\otimes d},\zeta^{\otimes d}*(X-Y))}\right] \\
&= \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \left[e^{q(\alpha-1)\sum_{i=1}^{d} D_{q(\alpha-1)+1}(\zeta*(Y_i-X_i),\zeta)}\right] \text{ independence of the noise} \\
&\leq \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \left[e^{\sum_{i=1}^{d}\log\frac{\beta_{k,\lambda}\pi}{\lambda} Q_{kq(\alpha-1)/2}\left(1+\frac{(X_i-Y_i)^2}{\lambda^2}\right)}\right] \text{ Lemma D.1} \\
&\leq \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \mathbb{E}\left[\left(\frac{\beta_{k,\lambda}\pi}{\lambda} Q_{kq(\alpha-1)/2}\left(1+\frac{\|X-Y\|^2}{d^2\lambda^2}\right)\right)^d\right] \text{ Concavity inequality} \\
&\leq \left(\frac{\beta_{k,\lambda}\pi}{\lambda}\right)^d \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \sum_{i=0}^{dkq(\alpha-1)/2} a_i \mathbb{E}\left[\left(1+\frac{\|X-Y\|^2}{d^2\lambda^2}\right)^i\right] \ \overline{Q}^d_{kq(\alpha-1)/2} \text{ is a polynomial (degree } dkq(\alpha-1)/2) \\
&\leq \left(\frac{\beta_{k,\lambda}\pi}{\lambda}\right)^d \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \sum_{i=0}^{dkq(\alpha-1)/2} \sum_{l=0}^{i} \binom{i}{l} a_i \mathbb{E}\left[\frac{\|X-Y\|^{2i}}{\lambda^{2i}}\right] \\
&\leq \left(\frac{\beta_{k,\lambda}\pi}{\lambda}\right)^d \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \sum_{i=0}^{dkq(\alpha-1)/2} \sum_{l=0}^{i} \binom{i}{l} a_i \frac{\mathbb{E}\left[\|X-Y\|^{dkq(\alpha-1)}\right]^{\frac{2i}{dkq(\alpha-1)}}}{d^{2i}\lambda^{2i}} \ \text{Jensen inequality } (2i \leq dkq(\alpha-1)) \\
&\leq \left(\frac{\beta_{k,\lambda}\pi}{\lambda}\right)^d \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \sum_{i=0}^{dkq(\alpha-1)/2} \sum_{l=0}^{i} \binom{i}{l} a_i \frac{W_{dkq(\alpha-1)}(\mu_i^\theta,\mu_j^\theta)^{2i}}{d^{2i}\lambda^{2i}} \text{ by definition of } W_{dkq(\alpha-1)} \\
&\leq \left(\frac{\beta_{k,\lambda}\pi}{\lambda} \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \overline{Q}_{kq(\alpha-1)/2}\left(1+\frac{W_{dkq(\alpha-1)}(\mu_i^\theta,\mu_j^\theta)^2}{d^2\lambda^2}\right)\right)^d. \qquad \square
\end{aligned}
$$

D.4. Utility of the DAGWM (Proposition 4.3)

In order to prove analyze the utility of the DAWGM, we resort to the following lemma.

Lemma D.2. *Let $\alpha > 1$. Let ζ be a distribution of $\mathbb{R}^d$ and W a random variable of $\mathbb{R}^d$ such that $\|W\| \leq z$ a.s. For N drawn from ζ and independent of W, we have:*

$$D_\alpha(N+W, N) \leq R_\alpha(\zeta, z).$$

Proof. We have:

$$\begin{aligned}
e^{(\alpha-1)D_\alpha(N+W,N)} &= \int \frac{P_{W+N}(x)^\alpha}{P_N(x)^{\alpha-1}}dx \\
&= \int \frac{\mathbb{E}[P_N(x-W)]^\alpha}{P_N(x)^{\alpha-1}}dx \\
&\leq \int\int_{\|u\|\leq z} \frac{P_N(x-u)^\alpha}{P_N(x)^{\alpha-1}}P_W(u)dxdu \\
&= \int_{\|u\|\leq z} e^{(\alpha-1)D_\alpha(N+u,N)}P_W(u)du \\
&\leq \int_{\|u\|\leq z} \sup_{\|x\|\leq z} e^{(\alpha-1)D_\alpha(N+z,N)}P_W(u)du \\
&\leq \sup_{\|x\|\leq z} e^{(\alpha-1)D_\alpha(N+W,N)}.
\end{aligned}$$

□

Below, we make the informal result of Proposition 4.3 precise and provide its proof.

Proposition D.1 (Utility of the DAGWM)**.** *Let $(\mathcal{S},\mathcal{Q},\Theta)$ be a Pufferfish framework, and let $\mathcal{M}(X)=f(X)+N$, where $X\sim\theta\in\Theta$, $N\sim\zeta\in\mathcal{P}(\mathbb{R}^d)$ and f is a numerical query. Let $\alpha>1$. Then, $R_\alpha(\zeta,\Delta_G)$ as defined in Theorem 3.3 is greater or equal to $\frac{\log(\Delta_G^{\zeta,1,\alpha})}{\alpha-1}$ defined in Theorem 4.5.*

Proof. By definition: for $(s_i,s_j)\in\mathcal{Q},\theta\in\Theta$, we note $(\mu_i^\theta,\mu_j^\theta)=(P(f(X)|s_i,\theta),P(f(X|s_j,\theta)))$, and if $(f(X)_{|s_i,\theta},f(X)_{|s_j,\theta})\sim\pi^*\in\Gamma(\mu_i^\theta,\mu_j^\theta)$ realises the optimal transport plan for $W_\infty(\mu_i^\theta,\mu_j^\theta)$:

$$\|f(X)_{|s_i,\theta}-f(X)_{|s_j,\theta}\|\leq W_\infty(\mu_i^\theta,\mu_j^\theta) \text{ a.s.}$$

Using Lemma D.2, We have:

$$\begin{aligned}
\mathbb{E}\left[e^{(\alpha-1)D_\alpha(\zeta,\zeta*(f(X)_{|s_i,\theta}-f(X)_{|s_j,\theta}))}\right] &= \mathbb{E}\left[e^{(\alpha-1)D_\alpha(\zeta*(f(X)_{|s_j,\theta}-f(X)_{|s_i,\theta}),\zeta)}\right] \\
&\leq e^{(\alpha-1)R_\alpha(\zeta,W_\infty(\mu_i^\theta,\mu_j^\theta))}.
\end{aligned}$$

It follows:

$$\begin{aligned}
\Delta_G^{\zeta,1,\alpha} &= \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \inf_{P(X,Y)\in\Gamma(\mu_i^\theta,\mu_j^\theta)} \mathbb{E}\left[e^{(\alpha-1)D_\alpha(\zeta,\zeta*(X-Y))}\right] \\
&\leq \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} \mathbb{E}\left[e^{(\alpha-1)D_\alpha(\zeta,\zeta*(f(X)_{|s_i,\theta}-f(X)_{|s_j,\theta}))}\right] \\
&\leq e^{(\alpha-1)R_\alpha(\zeta,W_\infty(\mu_i^\theta,\mu_j^\theta))}.
\end{aligned}$$

Finally :

$$\frac{\log(\Delta_G^{\zeta,1,\alpha})}{\alpha-1}\leq \max_{\substack{(s_i,s_j)\in S\\ \theta_i\in\Theta}} R_\alpha(\zeta,W_\infty(\mu_i^\theta,\mu_j^\theta))=R_\alpha(\zeta,\Delta_G).$$

□

E. Privacy Amplification by Iteration (Section 5)

E.1. Parallel Composition

Assessing the privacy guarantees of composition in RPP may be challenging. As a matter of fact, there does not exist, to our knowledge, any theorem stating the mechanism-agnostic privacy guarantees of sequential composition in Pufferfish privacy. However, we can recover a staightforward result of parallel composition for the RPP framework.

Proposition E.1 (RPP parallel composition for queries performed over independent datasets)**.** *Let $m > 0$ and $(\mathcal{S}, \mathcal{Q}, \Theta_k)$ be Pufferfish frameworks corresponding to each dataset $X_k \sim P(\cdot|s_i^k, \theta_k)$. We assume that each secret s_i^k is independent of the distributions θ_l, for $l \neq k$ and that $\mathcal{Q}$ only contains pairs of the form (s_i^k, s_j^k). For all $k \in \{1, \dots, n\}$, let $\mathcal{M}_k(X_k)$ be mechanisms that satisfy (α, ε_k)-RPP. Let $\Theta = \{\bigotimes_{k=1}^m \theta_k; \forall k \in \{1, \dots, m\}, \theta_k \in \Theta_k\}$. Then, the mechanism $(\mathcal{M}_1, \dots, \mathcal{M}_m)$ satisfies $(\alpha, \max_k \varepsilon_k)$-RPP for $(\mathcal{S}, \mathcal{Q}, \Theta)$.*

Proof. Let $s_i^l, s_j^l \in \mathcal{Q}, \theta = \bigotimes_{k=1}^m \theta_k \in \Theta$.

$$\begin{aligned} D_\alpha(P(\mathcal{M}(X)|s_i^l, \theta), P(\mathcal{M}(X)|s_j^l, \theta)) &= D_\alpha\big(P\left((\mathcal{M}_1(X_1), \dots, \mathcal{M}_n(X_n))|s_i^l, \otimes_{k=1}^m \theta_k\right), \\ &\quad P\left((\mathcal{M}_1(X_1), \dots, \mathcal{M}_n(X_n))|s_j^l, \otimes_{k=1}^m \theta_k\right)\big) \\ &= \sum_{k=1}^n D_\alpha\left(P\left(\mathcal{M}_k(X_k)|s_i^l, \theta_l\right), P\left(\mathcal{M}_k(X_k)|s_j^l, \theta_k\right)\right) \\ &= D_\alpha\left(P\left(\mathcal{M}_l(X_l)|s_i^l, \theta_l\right), P\left(\mathcal{M}_l(X_l)|s_j^l, \theta_l\right)\right) \leq \varepsilon_l. \end{aligned}$$

□

This theorem states that if an adversary assumes that the dataset can be split into independent parts and if the secrets have some form of separability, such as in our Example 2, it is possible to apply a different RPP mechanisms to each independent part while paying only for the maximum privacy loss, similar to the parallel composition result for differential privacy.

E.2. Proof of Lemma 5.1

Lemma 5.1 (Dataset Dependent Contraction lemma)**.** *Let ψ be a contractive map in its first argument on $(\mathcal{Z}, \|\cdot\|)$. Let X, X' be two r.v's. Suppose that $\sup_w W_\infty(\psi(w, X), \psi(w, X')) \leq s$. Then, for $z > 0$:*

$$D_\alpha^{(z+s)}(\psi(W, X), \psi(W', X')) \leq D_\alpha^{(z)}(W, W').$$

Proof. This proof is similar to the contraction lemma of (Feldman et al., 2018). Let $s > 0$ such that $\sup_w W_\infty(\psi(W, X), \psi(W, X')) \leq s$, we have, for Y a v.a. such that $D_\alpha^{(z)}(W, W') = D_\alpha(Y, W')$ and $W_\infty(W, Y) \leq z$:

$$\begin{aligned} W_\infty(\psi(W, X), \psi(Y, X')) &\leq W_\infty(\psi(W, X), \psi(W, X')) + W_\infty(\psi(W, X'), \psi(Y, X')) \\ &\leq s + W_\infty(W, Y) \\ &\leq s + z. \end{aligned}$$

It follows that:

$$D_\alpha^{(z+s)}(\psi(W, X), \psi(W', X')) \leq D_\alpha(\psi(Y, X'), \psi(W', X')) \leq D_\alpha(Y, W') = D_\alpha^{(z)}(W, W').$$

□

E.3. Proof of Theorem 5.2

Theorem 5.2 (Dataset Dependent PABI)**.** *Let X_T and X'_T denote the output of $CNI_T(W_0, \{\psi_t\}, \{\zeta_t\}, X)$ and $CNI_T(W_0, \{\psi_t\}, \{\zeta_t\}, X')$. Let $s_t = \sup_w W_\infty(\psi(w, X_t), \psi(w, X'_t))$. Let $a_1, \dots, a_T$ be a sequence of reals and let $z_t = \sum_{i \leq t} s_i - \sum_{i \leq t} a_i$. If $z_t \geq 0$ for all t, then, we have:*

$$D_\alpha^{(z_T)}(X_T, X'_T) \leq \sum_{t=1}^T R_\alpha(\zeta_t, a_t).$$

Proof. The proof is similar to the original PABI proof of (Feldman et al., 2018). It is obtained by induction by replacing in the original PABI proof $s_t = \sup_{w \in \mathbb{R}^d, x, x' \in \mathcal{X}} \|\psi(w, x) - \psi(w, x')\|$ by $s_t = \sup_w W_\infty(\psi(w, X_t), \psi(w, X'_t))$ and using the dataset dependent contraction lemma (Lemma 5.1). □

F. Applications (Section 6)

F.1. Proof of Proposition 6.1

Proposition 6.1. *Let $\lambda > 0$, $(\mathcal{S}, \mathcal{Q}, \Theta)$ a Pufferfish framework. We note Δ the sensitivity of a numerical mechanism f and:*

$$\Theta_\lambda = \{\theta \in P(\mathcal{X}^n); \\ \sup_{s_i \in \mathcal{S}} W_\infty(P(f(X)|s_i, \theta), P(f(X)|s_i, \theta^\otimes)) \leq \lambda\}.$$

Then, if $\Theta \subseteq \Theta_\lambda$, $\Delta_G \leq 2\lambda + \Delta$.

Proof. It is a direct consequence of triangle inequality for the W_∞ distance: for $(s_i, s_j) \in \mathcal{Q}$,

$$\begin{aligned} W_\infty(P(f(X)|s_i, \theta), P(f(X)|s_j, \theta)) &\leq W_\infty(P(f(X)|s_i, \theta), P(f(X)|s_j, \theta)) \\ &+ W_\infty(P(f(X)|s_i, \theta^\otimes), P(f(X)|s_j, \theta^\otimes)) \\ &+ W_\infty(P(f(X)|s_j, \theta^\otimes), P(f(X)|s_j, \theta)) \\ &\leq \Delta + \sup_{s_i \in \mathcal{S}} W_\infty(P(f(X)|s_i, \theta), P(f(X)|s_i, \theta^\otimes)) \end{aligned}$$

□

F.2. Attribute Inference Setting

F.2.1. Definitions

We recall the definition of Dataset Attribute Privacy from (Zhang et al., 2022):

Definition F.1 (Dataset Attribute Privacy (Zhang et al., 2022))**.** Let $X = (X_i^1, \ldots, X_i^m)$ a record with m attributes that is sampled from an unknown distribution $\mathcal{D}$, and let $X = (X^1, \ldots, X^m)$ be a dataset of n records sampled i.i.d from $\mathcal{D}$, where X^i denotes the (column) vector containing values of ith attribute of every record. Let $C \subseteq [m]$ be the set of indices of sensitive attributes, and for each $i \in C$, let $g_i(X^i)$ be a function with codomain $\mathcal{U}_i$. A mechanism $\mathcal{M}$ satisfies (ε, δ)-dataset attribute privacy if it is (ε, δ)-Pufferfish private for the following framework $(\mathcal{S}, \mathcal{Q}, \{\theta\})$:

- Set of secrets $\mathcal{S} = \{s_i^a \overset{def}{=} \mathbf{1}[g_i(X^i) \in \mathcal{U}_i^a \subseteq \mathcal{U}_i; i \in C]\}$.
- Set of secret pairs $\mathcal{Q} = \{(s_i^a, s_i^b) \in \mathcal{S} \times \mathcal{S}; i \in C\}$.
- Θ is a set of possible distributions θ over the dataset X. For each possible distribution $\mathcal{D}$ over records, there exists a $\theta_\mathcal{D} \in \Theta$ that corresponds to the distribution over n i.i.d. samples from $\mathcal{D}$.

F.2.2. EXPERIMENTS

We make some experiments in order to highlight strictly better utility than DP mechanisms for real world datasets. We compare the sensitivity of DP, noted Δ, the sensitivity of the GWM, noted Δ_G and the sensitivity of the DAGWM for the Cauchy distribution, noted $\Delta_{G,2}$.

F.2.3. ATTRIBUTE INFERENCE SETTING

Datasets The datasets are taken from the UCI Machine Learning Repository (Markelle Kelly).

- Student grade prediction: the Student Performance dataset (Cortez, 2014) has been collected to predict student grades. We want to release the column X of final math grades (0-20) while protecting the privacy of the values of the column S representing attendance in extra paid classes. We find Here, $W_\infty(P(X|S = \text{"no"}), P(X|S = \text{"yes"})) = 8$, $W_2(P(X|S = \text{"no"}), P(X|S = \text{"yes"})) \approx 2.76$. The distribution of X conditioned on S is shown in the accompanying figure.

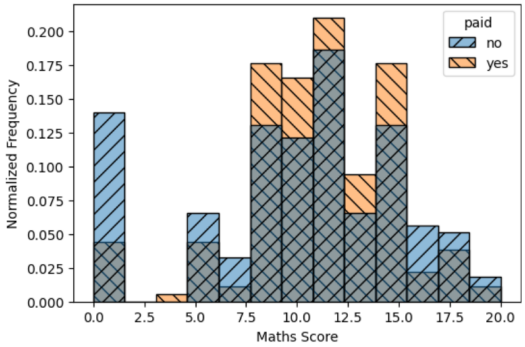

Figure 2. Distribution of the student scores on attendance in extra paid classes from the Student Performance dataset.

- Heart disease prediction: the Heart Disease dataset (Janosi et al., 1988) has been collected to predict heart disease diagnosis. We want to release the column X of ages (integer) while protecting the privacy of the values of the column S representing heart disease diagnosis (represented by integer values in (0-4). $\max_{\text{disease}_i,\text{disease}_j} W_\infty(P(X|S = \text{disease}_i), P(X|S = \text{disease}_j)) = 8$, $\max_{\text{disease}_i,\text{disease}_j} W_2(P(X|S = \text{disease}_i), P(X|S = \text{disease}_j)) \approx 7.80$. The distribution of X conditioned on S is shown in the accompanying figure.

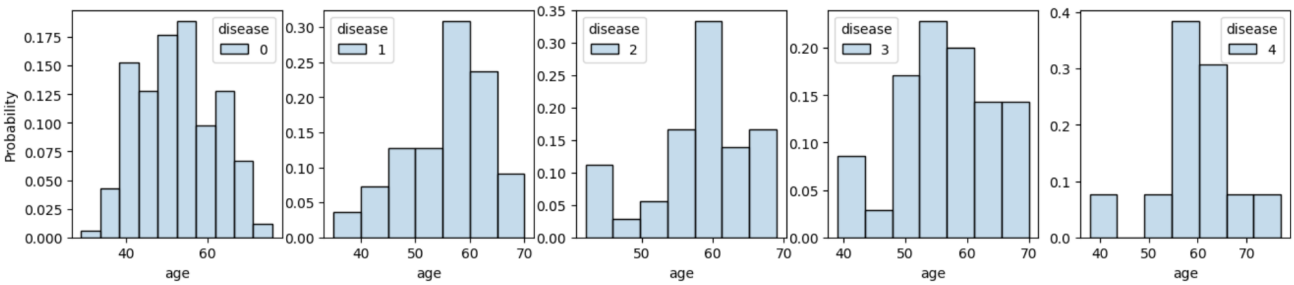

Figure 3. Distribution of ages conditioned on heart diagnosis from the Heart Disease dataset

- Salary prediction: the Adult dataset (Becker & Kohavi, 1996) is a popular dataset allowing to predict the salary of an individual. We want to release the column X of salaries in $\{\leq 50K, > 50K\}$ ($\{0, 1\}$ for privacy analysis) while protecting the privacy of the values of the column S representing the individual race. We find $\max_{\text{race}_i,\text{race}_j} W_\infty(P(X|S = \text{race}_i), P(X|S = \text{race}_j)) = 1$, $\max_{\text{race}_i,\text{race}_j} W_2(P(X|S = \text{race}_i), P(X|S = \text{race}_j)) \approx 0.42$. The distribution of X conditioned on S is shown in the accompanying figure.

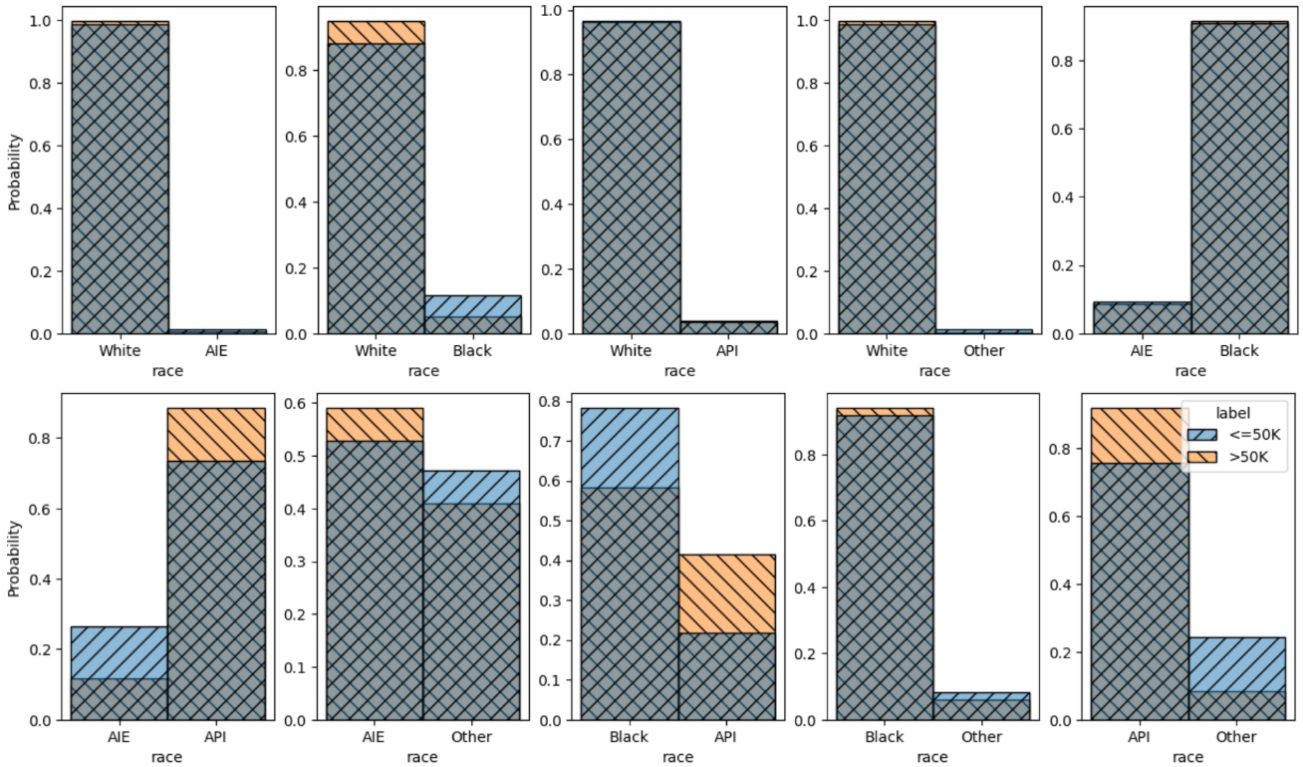

Figure 4. Distribution of the salary label conditioned on every pair of races from the Adult dataset.

F.3. Attribute Inference for Gaussian Data

Proposition F.1 (Multiple attribute inference with Gaussian data)**.** *Let $M \in \mathbb{R}^{l_1 \times m}, N \in \mathbb{R}^{l_2 \times m}$, with $l_1, l_2 \leq m$. We assume that the adversary has a prior $\theta = \mathcal{N}(\mu, \Sigma) \in P(\mathbb{R}^m)$, with $X = (X_1^1, \ldots, X_n^m)$. Then, considering the secrets $s_i^a = \{NX_i = a\}$, with $a \in K$ a compact of $\mathbb{R}^m$, the pairs of secrets $\mathcal{Q} = \{(s_i^a, s_i^b); i \in [\![1, d]\!], a, b \in K\}$ and the linear numerical query $f : x = (x_1, \ldots, x_n) \mapsto (Mx_1, \ldots, Mx_n)$,*

$$\Delta_G \leq \max_{a,b \in K} \| \operatorname{Cov}(MX_1, NX_1) \operatorname{Cov}(NX_1)^{-1}(a - b)\|$$

for the Pufferfish framework $(\mathcal{S}, \mathcal{Q}, \{\theta\})$.

Proof. For $i \in [\![1, n]\!]$, we have: $\begin{pmatrix} M \\ N \end{pmatrix} X_i \sim \mathcal{N}\left(\begin{pmatrix} M\mu \\ N\nu \end{pmatrix}, \begin{pmatrix} M\Sigma M^T & M\Sigma N^T \\ N\Sigma M^T & N\Sigma N^T \end{pmatrix}\right)$. Then,

$$MX_i | NX_i = a \sim \mathcal{N}\left(M\mu + M\Sigma N^T (N\Sigma N^T)^{-1}(a - N\mu), M\Sigma M^T - M\Sigma N^T (N\Sigma N^T)^{-1} N\Sigma M^T\right)$$

We note $\operatorname{Cov}(MX_i, NX_i) = M\Sigma N^T$. Drawing $Y \sim P(MX_i | NX_i = a)$ and noting:

$$Z = Y + \operatorname{Cov}(MX_i, NX_i)(N\Sigma N^T)^{-1}(b - a),$$

we have $Z \sim P(MX_i | NX_i = b)$ and:

$$\|Y - Z\| = \| \operatorname{Cov}(MX_i, NX_i)(N\Sigma N^T)^{-1}(b - a)\|.$$

□

F.4. Proof of Proposition 6.3

Proposition 6.3. *Let $V : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla^2 V \succcurlyeq CI_d$. For $\theta_0 \in P(\mathbb{R}^d)$, we note θ_t the distribution of X_t, with $(X_t)_{t\geq 0}$ solution of the stochastic differential equation: $dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$, where $(B_t)_{t\geq 0}$ is a brownian motion. We note $\theta_{t_1,\dots,t_n}$ the distribution generating $X = (X_{t_1}, \dots, X_{t_n})$ from the distribution of $(X_t)_{t\geq 0}$. We consider the secrets $s^a = \{X_0 = a\}$, with $a \in K$ a compact of $\mathbb{R}^d$, the pairs of secrets $\mathcal{Q} = \{(s^a, s^b); a, b \in K\}$ Then, the GWM of any L-Lipschitz query f performed on X has a sensitivity for the l_1 norm:*

$$\Delta_G \leq LDiam(K) \sum_{i=1}^{n} \exp(-2Ct_i)$$

for the Pufferfish framework $(\mathcal{S}, \mathcal{Q}, \{\theta_{t_1,\dots,t_n}\})$.

Proof. The proof can be obtained via a synchronous coupling argument, which can for example be found in (Villani, 2008). Let $a, b \in K$. Let $(B_t)_{t\geq 0}$ be a brownian motion and we define:

$$X_t = a - \int \nabla V(X_s)ds + \sqrt{2}B_t,$$
$$Y_t = b - \int \nabla V(Y_s)ds + \sqrt{2}B_t,$$

with the same realization of $(B_t)_{t\geq 0}$ for the two processes. Noting $\alpha_t = W_t - V_t$, we have: $\frac{d\alpha_t}{dt} = -(\nabla R_\lambda(W_t) - \nabla R_\lambda(V_t))$ and, by convexity of V:

$$\frac{d\|\alpha_t\|_1^2}{dt} = -2\langle \nabla R_\lambda(W_t) - \nabla R_\lambda(V_t), W_t - V_t\rangle \leq -2C\|\alpha_t\|_1.$$

Gronwall's lemma implies that $\forall t \geq 0, \|\alpha_t\|_1 \leq e^{-Ct}\|a - b\|_1$. Then,

$$\|f(X_{t_1}, \dots, X_{t_n}) - f(Y_{t_1}, \dots, Y_{t_n})\|_1 \leq L \sum_{i=1}^{n} \|\alpha_{t_i}\|_1 \leq L\|a - b\|_1 \sum_{i=1}^{n} \exp(-2Ct_i)$$

□

F.5. Application of PABI to Convex Optimization (Section 6.4)

F.5.1. Setup of convex optimization for PABI

Here is the setup for projected noisy stochastic gradient descent in the convex setting:

- f is L-Lipschitz in its first argument: there exists $L > 0$ such that $\forall x \in \mathcal{X}, w_1, w_2 \in \mathbb{R}^d$,
$$\|f(w_1, x) - f(w_2, x)\| \leq L\|w_1 - w_2\|.$$
- f is β-smooth in its first argument: there exists $\beta > 0$ such that $\forall x \in \mathcal{X}, w_1, w_2 \in \mathbb{R}^d$,
$$\|\nabla_w f(w_1, x) - \nabla_w f(w_2, x)\| \leq \beta\|w_1 - w_2\|.$$
- f satisfies the following condition: $\forall x_1, x_2 \in \mathcal{X}, w_1 \in \mathbb{R}^d, \exists C_{w_1} > 0$ such as :
$$\|\nabla_w f(w_1, x_1) - \nabla_w f(w_1, x_2)\| \leq C_{w_1}\|x_1 - x_2\|.$$

F.5.2. Application to DP

Lemma F.1 (Example: DP as a special case)**.** *In the case of DP, each distribution $\theta \in \Theta$ corresponds to a prior of independence between the elements of the dataset. Let $\beta, \eta, \sigma, L, T > 0, \alpha > 1$ such that $\eta > 2/\beta$. We set the secrets $\mathcal{S} = \left\{s_i^a \overset{def}{=} \{X_i = a\}; a \in \mathcal{X}\right\}$ and the pairs of secrets : $\mathcal{Q} = \{(s_i^a, s_i^b); a, b \in \mathcal{X}\}$. Let $(X, X') \sim \pi \in \Gamma(P(X|s_i^a), P(X|s_i^b))$. Let f be an objective function which is convex, β-smooth and L-Lipschitz. Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact set. Let $W_0 = W_0' \in \mathcal{K}$ be the original weight of the stochastic gradient descent and ψ the update function of the projected noisy stochastic gradient descent of learning rate η. Let $\zeta = \mathcal{N}(0, \sigma^2\eta^2 I_d)$ be the noising distribution. For $t \in [\![0, T]\!]$, we define $W_t = CNI_t(W_0, \psi, \zeta, X), W_t = CNI_t(W_0', \psi, \zeta, X')$. Then, Theorem 5.2 allows to obtain:*
$$D_\alpha^{(z_T)}(X_T, X_T') \leq \frac{2\alpha L^2}{\sigma^2(T - i + 1)}.$$
This recovers the results of Feldman et al. (2018) for the case of DP-SGD.

Proof. Let $\sigma > 0$. Let $(s_i^a, s_i^b) \in \mathcal{Q}, \theta \in \Theta$, with θ representing a prior of independence. Then, for $t \in [\![1, T]\!]$, $(X, X') \sim \pi \in \Gamma(P(X|s_i^a), P(X|s_i^b))$, $s_t = \sup_w W_\infty(\psi(w, X_t), \psi(w, X_t')) = \begin{cases} \sup_w \|\psi(w, a) - \psi(w, b)\| \text{ if } t = i \\ 0 \text{ else} \end{cases}$, $\zeta_t = \mathcal{N}(0, (\eta\sigma)^2 I_d)$. Then, setting $a_t = \begin{cases} \frac{s_i}{T-i+1} \text{ if } t \geq i \\ 0 \text{ else} \end{cases}$, we get:

$$\begin{aligned} D_\alpha^{(z_T)}(X_T, X_T') &\leq \sum_{t=i}^{T} R_\alpha\left(\zeta_t, \frac{\sup_w \|\psi(w, a) - \psi(w, b)\|}{T - i + 1}\right) \\ &\leq \sum_{t=i}^{T} \frac{\alpha \sup_w \|\psi(w, a) - \psi(w, b)\|}{2\eta^2\sigma^2(T - i + 1)^2} \\ &\leq \frac{2\alpha L^2}{\sigma^2(T - i + 1)}, \end{aligned}$$

which is the bound of Theorem 23 of Feldman et al. (2018). □

F.5.3. PABI bounds for Gaussian datasets

Proposition 6.4. *Assume that the adversary has a Gaussian prior θ. Then,*
$$\begin{aligned} D_\alpha(W_T, W_T') &\leq \frac{\alpha\eta^2}{2\sigma^2} \min(2L, \sup_{v \in \mathcal{K}} C_v\|a - b\|)^2 \\ &+ \frac{\alpha\eta^2}{2\sigma^2} \sum_{t \neq i}^{T} \min(2L, \sup_{v \in \mathcal{K}} C_v\| \operatorname{Cov}(X_t, X_i) \operatorname{Cov}(X_i)^{-1}(a - b)\|)^2. \end{aligned}$$

Proof. Let $t \in [\![1, T]\!]$ such that $t \neq i$. We want to find a upper bound to $W_\infty P(X_t|X_i = a), P(X_t|X_i = b))$. We note $X = (X_1^1, \dots, X_1^d, \dots, X_T^d) \sim \mathcal{N}(\mu, \Sigma)$, $X_t = (X_t^1, \dots, X_t^d) \sim \mathcal{N}(\mu_t, \Sigma_t)$ and $M_t^i = \begin{pmatrix} 0_{(i-1)d} & I_d & 0_{(T-i)d} \\ 0_{(t-1)d} & I_d & 0_{(T-t)d} \end{pmatrix}$. Then, $\begin{pmatrix} X_i \\ X_t \end{pmatrix} = M_t^i X \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_t \end{pmatrix}, \begin{pmatrix} \Sigma_i & \Sigma_{it} \\ \Sigma_{ti} & \Sigma_t \end{pmatrix}\right)$, and:

$$X_t|X_i = a \sim \mathcal{N}(\mu_t + \Sigma_{ti}\Sigma_i^{-1}(a - \mu_i), \Sigma_t - \Sigma_{ti}\Sigma_i^{-1}\Sigma_{it}).$$

Then, for $Y \sim X_t|X_i = a$ and $Z = Y + \mathrm{Cov}(X_t, X_i)\,\mathrm{Cov}(X_i)^{-1}(b - a)$, $Z \sim X_t|X_i = b$. For $t = i$, we have $W_\infty P(X_i|X_i = a), P(X_i|X_i = b)) = \|b - a\|$. □

F.5.4. PABI BOUNDS FOR DECREASING DEPENDENCIES

The bounds of Section 6.4 can be improved in the case where $(W_\infty(X_t, X_t'))_t$ is non-increasing.

Proposition F.2. *Taking the notations from Theorem 5.2, let (X_t) and (X_t') be CNIs. Assume that $(s_t)_t = (W_\infty(X_t, X_t'))_t$ is non-increasing. Then,*

$$D_\alpha^{(z_T)}(X_T, X_T') \leq \sum_{t=1}^{T} R_\alpha\left(\zeta_t, \frac{\sum_{k=1}^{T} W_\infty(X_k, X_k')}{T}\right).$$

When $\zeta_t = \zeta$ for all $t \in \{1, \dots, T\}$, the bound becomes:

$$D_\alpha^{(z_T)}(X_T, X_T') \leq T R_\alpha\left(\zeta, \frac{\sum_{t=1}^{T} W_\infty(X_t, X_t')}{T}\right).$$

Proof. We assume that the sequence $(s_t)_t = (W_\infty(X_t, X_i))_t$ is decreasing. We apply Theorem 5.2 to get new bounds. Compared to the analysis of Section 6.4, it gives For $t \in \{1, \dots, T\}$, we take $a_t = \frac{\sum_{k=1}^{T} W_\infty(X_k, X_k')}{T}$, we have:

$$\begin{aligned} z_t = \sum_{k \leq t} s_i - \sum_{k \leq t} a_i &= \sum_{i \leq t} W_\infty(X_i, X_i') - \frac{1}{T}\sum_{i \leq t}\sum_{k=1}^{T} W_\infty(X_k, X_k') \\ &= \frac{T-t}{T}\sum_{i \leq t} W_\infty(X_i, X_i') - \frac{t}{T}\sum_{t < i \leq T} W_\infty(X_i, X_i') \\ &\geq \frac{T-t}{T}\left(\sum_{i \leq t} W_\infty(X_i, X_i') - t W_\infty(X_t, X_t')\right) \geq 0. \ (s_t) \text{ is non-increasing} \end{aligned}$$ □

We further analyze this new bound for the case of Gaussian noise : $\zeta_t = \zeta = \mathcal{N}(0, \sigma^2 I_d)$ for all t. Then, we have:

$$D_\alpha^{(z_T)}(X_T, X_T') \leq \frac{\alpha(\sum_{t=1}^{T} W_\infty(X_t, X_t'))^2}{2T\sigma^2}. \tag{1}$$

We can compare this bound with the PABI bound of Section 6.4:

$$D_\alpha^{(z_T)}(X_T, X_T') \leq \frac{\alpha \sum_{t=1}^{T} W_\infty(X_t, X_t')^2}{2\sigma^2}. \tag{2}$$

While the latter result (2) allows to derive privacy guarantees for composition in the Pufferfish framework, it does not ensure that privacy loss tends to 0 as $T \to +\infty$ even when $\sum_{t=1}^{T} W_\infty(X_t, X_t')^2$ converges. However, when dependencies are decreasing over time, the privacy loss analysis is improved with (1).

We now compare our PABI bounds with the DP and the Group DP (which represent two extreme cases of our analysis). In order to do this, we illustrate the privacy loss as a function of the number of iterations. We let the secrets $s_a = \{X_1 = a\}, s_b = \{X_1 = b\}$ and for simplicity and visualization, we stay in the Gaussian setting of Proposition 6.4. We assume that each X_t has a covariance matrix $\mathrm{Cov}(X_t) = I_d$ and the covariance between X_t and X_1 is $\mathrm{Cov}(X_1, X_t) = \rho_t I_d$. This

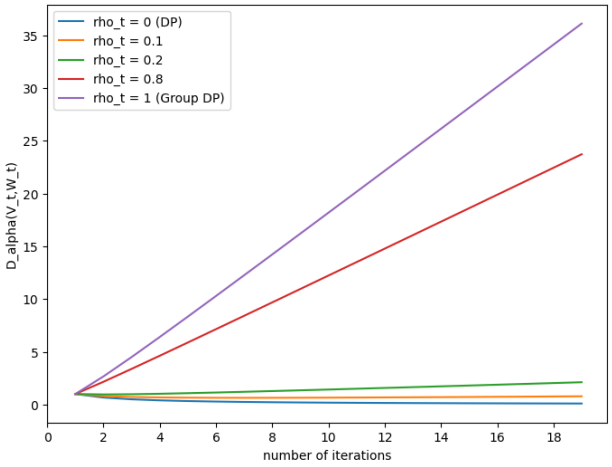

Figure 5. Privacy loss as a function of the number of iterations for the following values of ρ_t: 0 (DP), 0.1, 0.2, 0.8 and 1 (Group DP).

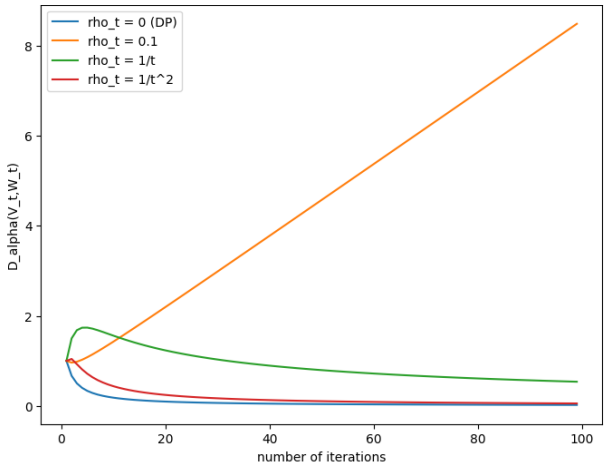

Figure 6. Privacy loss as a function of the number of iterations for the following values of ρ_t : 0 (DP), $0.1, 1/t$ and $1/t^2$.

corresponds to the case where each the columns of the dataset are independent of each other but dependencies within each column are controlled by the parameter ρ_t. $\rho_t \overset{t\to+\infty}{\to} 0$ means that X_t becomes increasingly independent of X_1 as t increases, indicating that X_t is less correlated with X_1 when they are far apart in the dataset. Using Proposition F.2 and Proposition 6.4, we have, for $t \in \{1, \dots, T\}$ and $\rho_1 = 1$: $\| \operatorname{Cov}(X_t, X_i) \operatorname{Cov}(X_i)^{-1}(a-b)\| = |\rho_t| \|a-b\|$. Then:

$$D_\alpha^{(z_T)}(X_T, X_T') \leq \frac{\alpha\eta^2}{2T\sigma^2}\left(\sum_{t=1}^{T} \min(2L, \sup_{v\in\mathcal{K}} C_v|\rho_t|\|a-b\|)\right)^2 \leq \frac{\alpha\eta^2\|a-b\|^2(\sup_{v\in\mathcal{K}} C_v)^2}{2T\sigma^2}\left(\sum_{t=1}^{T}|\rho_t|\right)^2.$$

We set the following parameters for visualization: $L = \sigma = \eta = \sup C_v = \|a-b\| = 1, \alpha = 2$. Recall that standard DP corresponds to the absence of correlation ($\rho_t = 0$), while Group DP corresponds to maximal correlation ($\rho_t = 1$). In Figure 5, we show how our PABI bounds compare with the DP setting in the case where all elements of the dataset are equally correlated ($\rho_t = \rho$), highlighting the privacy gains over Group DP. In Figure 6, we show the convergence of the privacy loss to 0 when the correlations vanish ($\rho_t \overset{t\to+\infty}{\to} 0$).