Kantorovich Mechanism for Pufferfish Privacy

Ni Ding
University of Melbourne

Abstract

Pufferfish privacy achieves ϵ-indistinguishability over a set of secret pairs in the disclosed data. This paper studies how to attain ϵ-pufferfish privacy by exponential mechanism, an additive noise scheme that generalizes the Laplace noise. It is shown that the disclosed data is ϵ-pufferfish private if the noise is calibrated to the sensitivity of the Kantorovich optimal transport plan. Such a plan can be obtained directly from the data statistics conditioned on the secret, the prior knowledge of the system. The sufficient condition is further relaxed to reduce the noise power. It is also proved that the Gaussian mechanism based on the Kantorovich approach attains the δ-approximation of ϵ-pufferfish privacy.

1 INTRODUCTION

Data privacy is about how to protect the confidential information when people are sharing data with each other. For a data user that is able to analyzes the disclosed dataset and computes statistics about it, the purpose is to prevent any prediction on the sensitive attributes or secrets, e.g., ID, age, gender, race, etc., that could be maliciously used for discrimination or unfair decision making. Thus, data protection nowadays is far beyond anonymization, but more of *inference control* (Dwork, 2008).

Differential privacy (Dwork et al., 2014) ensures the statistical indifference about the secret when the adversary is collecting the aggregated statistics about the underlying population. For all neighboring databases s_i and s_j that differ only in one entry, ϵ-differential privacy upper bounds the statistical distance between the probabilities $\Pr(\tilde{f}(s_i) = y)$ and $\Pr(\tilde{f}(s_j) = y)$ by a positive threshold ϵ. Here, $\tilde{f}$ denotes the noised, or randomized, query function f. Differential privacy is a rigorous mathematical definition of privacy that can be easily applied to machine learning tasks, e.g., the contingency table release (Barak et al., 2007), the privacy-preserving data mining (Dwork and Nissim, 2004).

On the other hand, information theorists study data privacy in a Bayesian inference setting (du Pin Calmon and Fawaz, 2012; Sankar et al., 2013; Ding and Farokhi, 2020). The secret is treated as a random variable S that is correlated with the public data X to be released. For $\Pr(S = s)$ being the adversary's prior belief of the secret, the purpose is to randomize X and generate the sanitized data Y to reduce the information gain on S, i.e., Y should reduce the difference between the posterior belief $\Pr(S = s|Y = y)$ and prior belief $\Pr(S = s)$. This can be translated, by Bayes' rule, to bounding the statistical distance of the conditional probabilities $\Pr(Y = y|S = s_i)$ and $\Pr(Y = y|S = s_j)$ by ϵ for each pair of secret instances s_i and s_j, which is called ϵ-local differential privacy (Mironov, 2017; Issa et al., 2020; Ding et al., 2021).

Pufferfish privacy: Kifer and Machanavajjhala (2012, 2014) introduced a more general privacy framework called 'pufferfish'. For ρ being the prior knowledge of the system, ϵ-pufferfish privacy enforces statistical indistinguishability between $\Pr(Y = y|S = s_i, \rho)$ and $\Pr(Y = y|S = s_j, \rho)$ over all pairs of secrets (s_i, s_j) in a discriminative pair set $\mathbb{S}$. This is a more flexible and practical setting in that: $\mathbb{S}$ can be specified by all secret pairs that actually raise the privacy concerns in the real application; ρ could denote the side information obtained by the adversary, which incorporates the concept of Bayesian inference in information-theoretic data privacy. It is also shown by Kifer and Machanavajjhala (2014) that, for specific $\mathbb{S}$ and ρ, ϵ-pufferfish privacy is equivalent to ϵ-differential privacy (Dwork, 2006) and ϵ-indistinguishability (Dwork et al., 2006).

Privatization: The question then is how to privatize

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

the data to attain pufferfish privacy. The information-theoretic solution is to determine a privacy-preserving encoding function $\Pr(Y = y|X = x)$ for each piece of message x and codeword y (Makhdoumi et al., 2014). But, this is only practical for discrete and finite alphabet and not as convenient as the additive noise (e.g., Laplace) mechanism, where we only need to calibrate the parameter of the noise distribution. There are other attempts in the literature, e.g., segmented noise mechanism for publishing counting query and histogram (Kifer and Machanavajjhala, 2014), Laplace mechanism for monitoring web browsing behavior (Liang et al., 2020). However, these methods only apply to specific applications, e.g., a particular query function, Markovian assumption about the prior knowledge ρ.

The *challenge* here is that the statistics of Y is caused by the randomness in the data regulation scheme, as well as the intrinsic correlation between S and X, i.e., the conditional probability $\Pr(X = x|S = s, \rho)$.[1] Song et al. (2017) proposed an additive noise scheme based on the Wasserstein metric in the probability space: calibrating Laplace noise to the maximum ∞-Wasserstein distance over all secret pairs in $\mathbb{S}$ attains pufferfish privacy. While this method generally applies to any system, computing the ∞-Wasserstein distance is hard.[2] Song et al. (2017) then resorted to a Markov quilt mechanism, which does not require Wasserstein metric, but only works in Bayesian network models.

1.1 Our Contributions

In this paper, we propose a Kantorovich mechanism for attaining pufferfish privacy that generally applies to any data S and X and the prior knowledge ρ.

Our main contributions are the following.

1. We consider the exponential mechanism, an additive noise scheme that generalizes Laplace noise. A sufficient condition is derived showing that pufferfish privacy is attained by calibrating noise to the sensitivity of the Kantorovich optimal transport plan. This transport plan can be directly determined by the conditional probabilities $\Pr(X = x|S = s_i, \rho)$ and $\Pr(X = x|S = s_i, \rho)$ for all pairs of secrets s_i and s_j in $\mathbb{S}$. It is also proved that the Gaussian mechanism based on this Kantorovich approach attains (ϵ, δ)-pufferfish privacy.
2. We relax the sufficient condition to reduce the noise power. Experimental results show that the relaxed sufficient condition improves data utility of the pufferfish private data regulation schemes.
3. In a multi-user system, where each user is assigned an independent random variable, we study the ϵ-indistinguishability as to whether a user is present in the system and the value of the random variable he/she obtains. It is shown that, for any deterministic query function f, the sensitivity of the Kantorovich optimal transport plan is equivalent to that of f, regardless of the randomness in the prior knowledge ρ. Therefore, pufferfish privacy is attained by calibrating noise to the sensitivity of f.

In this paper, we only present a proof sketch for each statement (incl. theorem, lemma and corollary). The full proof and detailed derivation can be found in the supplementary materials. We use capital letters, e.g, X, to denote a random variable and lower case letters, e.g., x, to denote the instance of this random variable. Notation $P_X(x)$ denotes the probability $\Pr(X = x)$.

2 PUFFERFISH PRIVACY

Let $\mathcal{S}$ be the alphabet of the secret S and Y be the sanitized version of X. We say Y is private if it attains a certain degree of statistical indistinguishability of the sensitive information S. Here, the indistinguishability refers to the bounded probability of Y evoked by two secrets $s_i, s_j \in \mathcal{S}$:

$$\left|\log \frac{P_{Y|S}(y|s_i)}{P_{Y|S}(y|s_j)}\right| \leq \epsilon \tag{1}$$

for some $\epsilon > 0$. We call ϵ the *privacy budget*. If (1) holds for all secret pairs $(s_i, s_j) \in \mathcal{S}^2$, Y is ϵ-local differentially private (Duchi et al., 2013; Sarwate and Sankar, 2014); if it holds for all neighboring (s_i, s_j),[3] Y attains ϵ-differential privacy Dwork et al. (2006).

Let $\mathbb{S} \subseteq \mathcal{S}^2$ be the *discriminative pair set* containing secret pairs (s_i, s_j). Pufferfish privacy attains ϵ-indistinguishability in $\mathbb{S}$.

Definition 1 (Pufferfish Privacy (Kifer and Machanavajjhala, 2014))**.** *The sanitized data Y attains $(\epsilon, \mathbb{S})$-pufferfish privacy if*

$$\left|\log \frac{P_{Y|S}(y|s_i, \rho)}{P_{Y|S}(y|s_j, \rho)}\right| \leq \epsilon, \tag{2}$$

[1] This is the reason that pufferfish privacy is considered a generalization of differential privacy for correlated data. An equivalent problem is how to attain differential privacy when the query answer f itself is a randomized function, for which the noise calibration by ℓ_1-sensitivity (Dwork et al., 2006) does not directly apply. See Section 3.4.

[2] This is due to the difficulty (non-convexity) in obtaining the optimal transport plan for the ∞-Wasserstein metric. See Champion et al. (2008); De Pascale and Louet (2019).

[3] The neighborhood is defined by Hamming distance: $d_H(s_i, s_j) \leq 1$

for all $\rho \in \mathbb{D}$ *and* $(s_i, s_j) \in \mathbb{S}$.

Here, ρ denotes the prior knowledge, e.g., the hyperparameter, that is sufficient to describe the probability distribution of X given the secret S, which we denote by $P_{X|S}(x|s,\rho)$. ρ could be the true knowledge of $P_{X|S}(x|s,\rho)$. Or, if there are more than one adversary in the system, each ρ can be used to denote the prior belief of an adversary and $\mathbb{D}$ is the set containing all adversaries in the system. In this case, the $(\epsilon, \mathbb{S})$-pufferfish privacy guarantees the statistical indistinguishability (2) against all adversaries $\rho \in \mathbb{D}$.

We show two examples of S, X and ρ below. They will be used to present the main results in this paper.

V independent user system Denote $\mathcal{V}$ the finite set that indexes $V = |\mathcal{V}|$ users or participants. Let each user throw a dice S_i independently and denote the outcome by a multiple random variable $S = (S_i \colon i \in \mathcal{V})$. Assume $X = f(S)$, where f is a deterministic query function. If $S_i \in \{0,1\}$ for all $i \in \mathcal{V}$, $X = f(S) = \sum_{i\in\mathcal{V}} S_i$ is a vote counting function. In this case, let $\rho = (p_i \colon i \in \mathcal{V})$ with each p_i denoting the probability of the event $S_i = 1$ for a Bernoulli distribution.[4] The conditional probability $P_{X|S}(\cdot|s_i,\rho)$ is fully determined by ρ.

Attributes in tabular data A tabular dataset needs to be denonymized or privatized before being disclosed to protect the sensitive attributes/columns, e.g., 'name', 'age', 'race'. Denote S and X the sensitive and public attributes,[5] respectively. This is the information-theoretic data privacy problem formulated in Sankar et al. (2013, Fig. 1). Let ρ be the empirical joint distribution of S and X, which determines the conditional probability $P_{X|S}(\cdot|s_i,\rho)$. In this case, we can still write $X = f(S)$, while f is not a deterministic but a randomized function.

2.1 Additive Noise Privatization Scheme

Consider the additive noise mechanism

$$Y = X + N$$

where the noise N is independent of X. Denote the probability of N by $P_N(\cdot)$. The conditional probability $P_{Y|S}(\cdot|s,\rho)$ is determined by the convolution

$$P_{Y|S}(y|s,\rho) = \int P_N(y-x) P_{X|S}(x|s,\rho)\,\mathrm{d}x. \qquad (3)$$

[4] In this case, each p_i could denote the local randomization scheme chosen by individual i, as in local differential privacy Duchi et al. (2013).

[5] Or, X could denote some deterministic function of the public attribute, as in Kifer and Machanavajjhala (2014); Song et al. (2017).

For zero-mean noise, we have $\mathbb{E}[Y] = \mathbb{E}[X]$ and the noise variance determines mean square error (MSE) $\mathbb{E}[(Y-X)^2] = \mathrm{VAR}[N]$. Here, the MSE denotes the average distortion of the released data Y, which can be used to quantify the data utility loss, e.g., He et al. (2014). It is clear that for two additive noise mechanisms both attaining pufferfish privacy, the one with less noise power is superior to the other. In this paper, we use the notation N_θ for noise, where θ denotes the parameter that determines the probability density function $P_{N_\theta}(\cdot)$.

3 KANTOROVICH MECHANISM

In this section, we propose a Kantorovich approach, based on the 1-Wasserstein metric, for attaining the pufferfish privacy. We first convert the ∞-Wasserstein metric in the existing randomization mechanism (Song et al., 2017) to a 1-Wasserstein distance and propose the Kantorovich solution for Laplace noise. Then, we extend this result to the exponential mechanism that generalizes the Laplace noise.

3.1 Preliminaries

We introduce the notation and definition for the Wasserstein metric and review the Kantororivich optimal transport plan as follows. For each ρ, a joint distribution $\pi \colon \mathbb{R}^2 \mapsto [0,1]$ is called a *coupling* of $P_{X|S}(\cdot|s_i,\rho)$ and $P_{X|S}(\cdot|s_j,\rho)$ if they are the marginals of π.[6] Denote $\Gamma(s_i,s_j)$ the set of all couplings for the secret pair $(s_i,s_j) \in \mathbb{S}$. The αth Wasserstein distance is $W_\alpha(s_i,s_j) = \big(\inf_{\pi\in\Gamma(s_i,s_j)} \int d^\alpha(x-x')\,\mathrm{d}\pi(x,x')\big)^{1/\alpha}$. For $\alpha = 1$,

$$W_1(s_i,s_j) = \inf_{\pi\in\Gamma(s_i,s_j)} \int d(x-x')\,\mathrm{d}\pi(x,x') \qquad (4)$$

is called the Kantorovich transportation problem for the mass transport cost d.

Kantorovich optimal transport plan π^* (Villani, 2009; Santambrogio, 2015) For convex d, the minimizer π^* of (4) can be directly determined by $P_{X|S}(\cdot|s_i,\rho)$ and $P_{X|S}(\cdot|s_j,\rho)$:

$$\pi^*(x,x') = \frac{\mathrm{d}^2}{\mathrm{d}x\,\mathrm{d}x'} \min\big\{F_{X|S}(x|s_i,\rho), F_{X|S}(x'|s_j,\rho)\big\},$$

where $F_{X|S}(\cdot|s_i,\rho)$ and $F_{X|S}(\cdot|s_j,\rho)$ are the cumulative density functions for $P_{X|S}(\cdot|s_i,\rho)$ and $P_{X|S}(\cdot|s_j,\rho)$, respectively.

[6] That is, $\int \pi(x,x')\,\mathrm{d}x' = P_{X|S}(s|s_i), \forall x$ and $\int \pi(x,x')\,\mathrm{d}x = P_{X|S}(x'|s_j), \forall x'$.

3.2 Kantorovich-Laplace mechanism

For Laplace noise N_θ with the noise distribution $P_{N_\theta}(z) = \frac{1}{2\theta}e^{-\frac{|z|}{\theta}}$, it is shown in Song et al. (2017) that calibrating noise power to

$$\theta = \frac{1}{\epsilon} \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} W_\infty(s_i, s_j) \tag{5}$$

for the ℓ_1-norm $d(z) = |z|$ attains $(\epsilon, \mathbb{S})$-pufferfish privacy in Y. Because the minimizer of $W_\infty(s_i, s_j) = \inf_{\pi\in\Gamma(s_i,s_j)} \sup_{(x,x')\in\mathrm{supp}(\pi)} |x - x'|$ is hard to obtain (Champion et al., 2008; De Pascale and Louet, 2019), we convert it to a W_1 metric and propose a Kantorovich approach below.

Lemma 1 (From W_∞ to Kantorovich)**.** *Adding Laplace noise N_θ with*

$$\theta = \frac{1}{\epsilon} \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \sup_{(x,x')\in supp(\pi^*)} |x - x'| \tag{6}$$

attains $(\epsilon, \mathbb{S})$-pufferfish privacy in Y.

Proof. We have θ in (5) equal to

$$\max_{\substack{\rho\in\mathbb{D},\\(s_i,s_j)\in\mathbb{S}}} \Big\{\theta\colon \inf_{\pi\in\Gamma(s_i,s_j)} \int \Big[\frac{|x-x'|}{\theta} - \epsilon\Big]_+ \mathrm{d}\pi(x, x') = 0\Big\} \tag{7}$$

where $[z]_+ = \max\{z, 0\}$ is convex. For each θ, the minimizer of $\inf_{\pi\in\Gamma(s_i,s_j)} \int \Big[\frac{|x-x'|}{\theta} - \epsilon\Big]_+ \mathrm{d}\pi(x, x')$ is the Kantorovich optimal transport plan π^*. Therefore, the maximum value of θ in (7) equals to (6). □

It is clear in the proof of Lemma 1 (see the full proof in Section B in the supplementary material) that the Wasserstein mechanism in the order of $\alpha = \infty$ proposed in Song et al. (2017) is equivalent to the Kantorovich-Laplace mechanism in Lemma 1. See Section C in the supplementary material how to efficiently compute the solutions to the two examples in Song et al. (2017) by Lemma 1.

3.3 Exponential Mechanism

Let d be a metric, i.e., d is nonnegative, symmetric $d(z) = d(-z), \forall z$, and satisfies the triangular inequality $d(z) \le d(\delta) + d(z-\delta), \forall z, \delta$. Consider the exponential mechanism (Dwork et al., 2006, Section 3.3), where the noise distribution is characterized by an exponential function $P_{N_\theta}(z) \propto e^{-\eta(\theta)d(z)}$. Assume $\eta \propto \frac{1}{\theta}$. For Laplace mechanism, $\eta(\theta) = 1/\theta$ and $d(z) = |z|$. By the triangular inequality,

$$P_{N_\theta}(y - x) \le e^{\eta(\theta)d(x-x')} P_{N_\theta}(y - x'), \quad \forall x, x', y \tag{8}$$

where $e^{\eta(\theta)d(x-x')}$ denotes an upper bound on the probability mass transport cost from x to x'. This cost upper bound is used in Dwork et al. (2006, 2014) to prove that the differential privacy is attained by calibrating the standard deviation of the noise to the sensitivity of the query function f. We obtain a similar result for attaining the pufferfish privacy in the V independent user system as follows.

3.3.1 Calibrating Noise to Sensitivity in V independent User System

For each user i, denote $S_i = a$ the event that user i is present in the system and reports the value a of random variable (i.e., the dice face) of S_i. We write $S_i = \perp_i$ for the event when user i is absent in the system. Consider the following two discriminative pair sets

$$\mathbb{S}_i = \{(S_i = a, S_i = b)\colon a, b \in \mathcal{S}_i\},$$
$$\mathbb{S}_{\perp_i} = \{(S_i = a, S_i = \perp_i)\colon a \in \mathcal{S}_i\}.$$

where $\mathcal{S}_i$ denotes the alphabet of S_i for user i. Using $\mathbb{S}_i$ and $\mathbb{S}_{\perp_i}$, Kifer and Machanavajjhala (2014) proved that $(\epsilon, \mathbb{S})$-pufferfish privacy is equivalent to ϵ-differential privacy (Dwork, 2006) and ϵ-indistinguishability (Dwork et al., 2006), respectively.

Denote $S_{-i} = (S_{i'}\colon i' \in \mathcal{V} \setminus \{i\})$ the multiple random variable excluding dimension/user i. For the deterministic query function f, let the sensitivity for metric d over the discriminative pair set $\mathbb{S}_i$ be

$$\triangle_f(\mathbb{S}_i) = \max_{a,b\in\mathcal{S}_i} \max_{s_{-i}} d\big(f(s_i = a, s_{-i}) - f(s_i = b, s_{-i})\big).$$

Note that for deterministic function f, the sensitivity $\triangle_f(\mathbb{S}_i)$ is independent of ρ.

Lemma 2. *In the V independent user system, for any deterministic query function f and any prior knowledge ρ, the following holds for all $i \in \mathcal{V}$:*

(a) adding noise N_θ with $\theta = \eta^{-1}(\epsilon/\triangle_f(\mathbb{S}_i))$ attains $(\epsilon, \mathbb{S}_i)$-pufferfish privacy in Y,

(b) if Y is $(\epsilon, \mathbb{S}_i)$-pufferfish private, then it is also $(\epsilon, \mathbb{S}_{\perp_i})$-pufferfish private.

Proof. For any $a, b \in \mathcal{S}_i$, we have the statistical indistinguishability in Y bounded by the mass transport cost upper bound function: $P_{Y|S_i}(y|a, \rho) \le e^{\eta(\theta)d(f(s_i=a,s_{-i})-f(s_i=b,s_{-i}))} P_{Y|S_i}(y|b, \rho)$. Then, for all $a, b \in \mathcal{S}_i$, y and ρ, $P_{Y|S_i}(y|a, \rho) \le e^{\eta(\theta)\triangle_f(\mathbb{S}_i)} P_{Y|S_i}(y|b, \rho)$. Therefore, (a) is a sufficient condition for attaining $(\epsilon, \mathbb{S}_i)$-pufferfish privacy.

Using the fact that $\min_{s_i} P_{Y|S_i}(y|s_i, \rho) \le P_{Y|S_i}(y|\ \perp_i, \rho) \le \max_{s_i} P_{Y|S_i}(y|s_i, \rho)$, for $(\epsilon, \mathbb{S}_i)$-pufferfish private Y, we have $|\log \frac{P_{Y|S_i}(y|\perp_i,\rho)}{P_{Y|S_i}(y|a,\rho)}| \le$

$\max_{s_i} |\log|\frac{P_{Y|S_i}(y|s_i,\rho)}{P_{Y|S_i}(y|a,\rho)}| \leq \epsilon$, i.e., $(\epsilon, \mathbb{S}_{\perp_i})$-pufferfish privacy attains simultaneously. □

See Sections A and D in the supplementary material for the detailed derivation and proof. In Section 3.4.1, we will verify Lemma 2 by the Kantorovich optimal transport plan π^*, where it is revealed that $\triangle_f(\mathbb{S}_i)$ coincides with the sensitivity of π^*.

3.4 Sufficient condition

Following Lemma 2, for f being a randomized function, adding noise N_θ with $\theta = \eta^{-1}(\epsilon/\triangle_f(\mathbb{S}))$ also attains pufferfish privacy. However, we should take into account the domain of f as well as the randomness in ρ. That is, for $X = f(S)$ where f is a randomized function, the sensitivity of f for metric d over the discriminative pair set $\mathbb{S}$ is

$$\triangle_f(\mathbb{S}) = \max_{(s_i,s_j)\in\mathbb{S}} \max_{\substack{x\in\text{supp}(P_{X|S}(\cdot|s_i,\rho)),\\ x'\in\text{supp}(P_{X|S}(\cdot|s_j,\rho))}} d\big(x-x'\big). \quad (9)$$

The probability mass $P_{X|S}(\cdot|s,\rho)$ could spread over a wide range of X that significantly increases $\triangle_f(\mathbb{S})$. Section 4 shows an example when $\text{supp}(P_{X|S}(\cdot|s_i)) = \text{supp}(P_{X|S}(\cdot|s_j)) = \mathcal{X}$, where $\mathcal{X}$ denotes the alphabet containing all possible values of X. In this case, the sensitivity f is as large as the pairwise distance in $\mathcal{X}$: $\triangle_f(\mathbb{S}) = \triangle_{\mathcal{X}} = \max_{x,x'\in\mathcal{X}} d(x-x')$ and the resulting $\theta = \eta^{-1}(\epsilon/\triangle_{\mathcal{X}})$ could make the noise power too large to convey any useful information of X in the disclosed dataset. The following theorem proposes another approach based on a distance metric over the probability space. The full proof is in Section E in the supplementary material

Theorem 1 (Kantorovich-exponential mechanism)**.** *For the exponential mechanism, adding noise N_θ with*[7]

$$\theta = \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \eta^{-1}\Big(\epsilon/\sup_{(x,x')\in supp(\pi^*)} d(x-x')\Big) \quad (10)$$

attains $(\epsilon,\mathbb{S})$-pufferfish privacy in Y.

Proof. For any pair $(s_i,s_j)\in\mathbb{S}$, we have

$$\begin{aligned} &P_{Y|S}(y|s_i,\rho) - e^\epsilon P_{Y|S}(y|s_j,\rho) \\ &= \int \big(P_{N_\theta}(y-x) - e^\epsilon P_{N_\theta}(y-x')\big)\,\mathrm{d}\pi^*(x,x') \\ &\leq \int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')} - e^\epsilon\big)\,\mathrm{d}\pi^*(x,x'), \forall y. \end{aligned} \quad (11)$$

[7] In (10), there is a Kantorovich optimal transport plan π^* for each $\rho\in\mathbb{D}$ and $(s_i,s_j)\in\mathbb{S}$.

For each y, (10) is a sufficient condition for $e^{\eta(\theta)d(x-x')} \leq e^\epsilon$ for all x,x', by which we have $\frac{P_{Y|S}(y|s_i,\rho)}{P_{Y|S}(y|s_j,\rho)} \leq e^\epsilon$. Due to the symmetric property $d(x-x') = d(x'-x), \forall x,x'$, (10) is also a sufficient condition for $\frac{P_{Y|S}(y|s_j,\rho)}{P_{Y|S}(y|s_i,\rho)} \leq e^\epsilon$. □

Theorem 1 essentially states that it is sufficient to only calibrating noise to the maximum pairwise distance over the support of the Kantorovich optimal transport plan π^*, which can be regarded as the *sensitivity of* π^*. It is clear that the maximum sensitivity of π^* over all $(s_i,s_j)\in\mathbb{S}$ is no greater than $\triangle_f(\mathbb{S})$. This has also been verified by Song et al. (2017, Theorem 3.3). In fact, in most cases, we have $\sup_{(x,x')\in\text{supp}(\pi^*)} d(x-x') \ll \triangle_f(\mathbb{S})$. See the experimental results in Section 4.

3.4.1 Interpretation of Lemma 2

Theorem 1 in return explains Lemma 2. For any deterministic query f on the V independent user system, we have the Kantorovich optimal transport plan $\pi^*(x,x') = 0$, for all x,x' such that $d(x-x') > \triangle_f(\mathbb{S}_i)$, i.e.,

$$\text{supp}(\pi^*) \subseteq \big\{(x,x')\colon d(x-x') \leq \triangle_f(\mathbb{S}_i)\big\}. \quad (12)$$

By Theorem 1, tuning θ to $\eta^{-1}\big(\epsilon/\triangle_f(\mathbb{S}_i)\big)$ attains $(\epsilon,\mathbb{S}_i)$-pufferfish privacy.

It is clear from (12) that $\triangle_f(\mathbb{S}_i)$ is in fact the sensitivity of the Kantorovich optimal transport plan π^* for metric d. This is the key point that Lemma 2(a) is valid.

Separable query function For all separable query functions f such that $X = f(S) = \sum_{i\in\mathcal{V}} f_i(S_i)$, we have

$$f(s_i=a,s_{-i}) - f(s_i=b,s_{-i}) = f_i(a) - f_i(b), \forall s_{-i}.$$

In this case, $\pi^*(x,x') = 0$, for all x,x' such that $x - x' \neq f_i(a) - f_i(b)$ and therefore

$$\text{supp}(\pi^*) = \big\{(x,x')\colon x - x' = f_i(a) - f_i(b)\big\}. \quad (13)$$

The sensitivity is $\triangle_f(\mathbb{S}_i) = \max_{a,b\in\mathcal{S}_i} d(f_i(a) - f_i(b))$. $(\epsilon,\mathbb{S}_i)$-pufferfish privacy attains by setting $\theta = \eta^{-1}\big(\epsilon/\max_{a,b\in\mathcal{S}_i} d(f_i(a)-f_i(b))\big)$.

A typical example of separable query function is the counting query, where $X = f(S) = \sum_{i\in\mathcal{V}} S_i$. In Figure 1, we show the Kantorovich optimal transport plan π^* for the counting query in a 25 independent user system, where each $S_i\in\{0,1\}$ follows Bernoulli(0.7) distribution. For discriminative secret pair $(S_i=0,S_i=1)$, $\text{supp}(\pi^*) = \big\{(x,x')\colon x-x'=1\big\}$; for discriminative pair $(S_i=0,S_i=\perp_i)$, $\text{supp}(\pi^*) =$

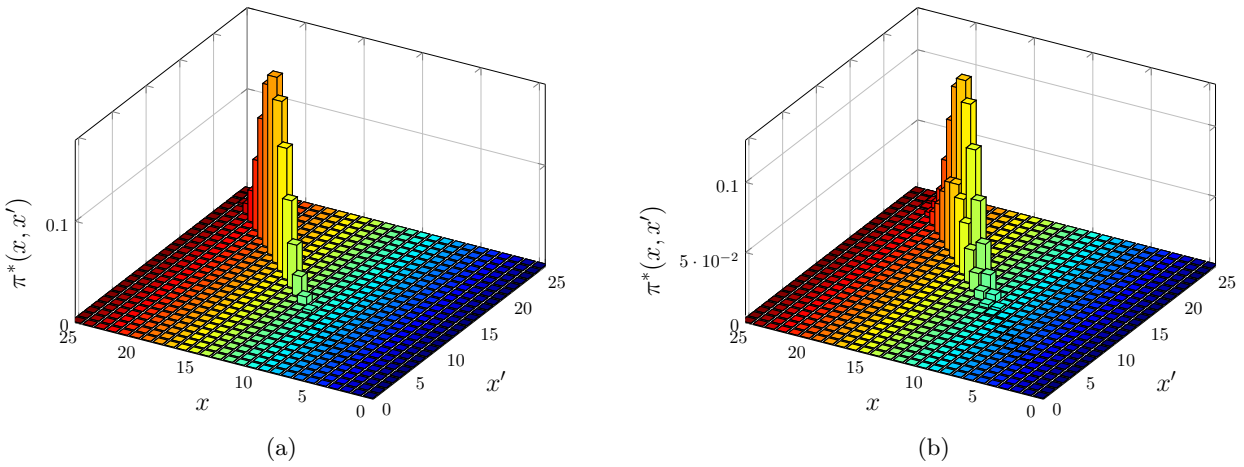


Figure 1: The Kantorovich optimal transportation plan π^* in the V independent user system for the discriminative pair set $\mathbb{S} = \{(S_i = 0, S_i = 1) \colon i \in \mathcal{V}\}$ in (a) and $\mathbb{S} = \{(S_i = 0, S_i = \perp_i) \colon i \in \mathcal{V}\}$ in (b). The function f is a counting query $X = f(S) = \sum_{i \in \mathcal{V}} S_i$ and $S = (S_i \colon i \in \mathcal{V}) \sim \text{Binomial}(V, p)$. We set $V = 25$ and $p = 0.7$.

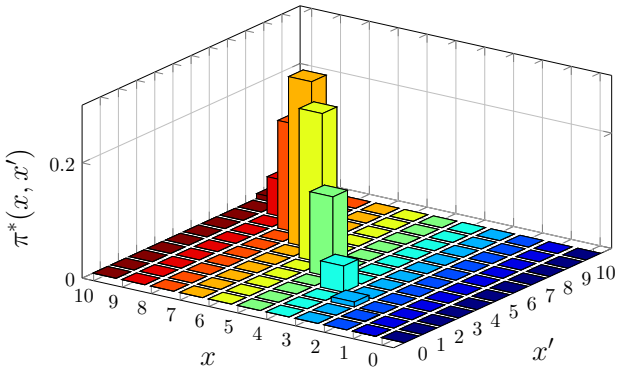


Figure 2: The Kantorovich optimal transportation plan π^* in the V independent user system for the discriminative pair $S_1 = 0$ and $S_1 = 1$ is in (a). There are $V = 10$ users. For each user i, $S_i \sim \text{Bernoulli}(p_i)$ and so the counting query $X = \sum_{i \in \mathcal{V}} S_i$ follows Poisson Binomial distribution.

$\big\{(x, x') \colon x - x' \le 1\big\}$. In this case, adding noise N_θ with $\theta = \eta^{-1}\big(\epsilon / d(1)\big)$ attains $(\epsilon, \mathbb{S}_i \cup \mathbb{S}_{\perp_i})$-pufferfish privacy.

This example is equivalent to the single-prior privacy for answering counting query f in Kifer and Machanavajjhala (2014, Section 7.1.1), which can be extended to the pufferfish private histogram publishing (see Kifer and Machanavajjhala, 2014, Section 7.1.2). The name 'single prior' refers to the same probability distributions of S_i for all i. Kifer and Machanavajjhala, 2014, Algorithm 1 proposes a segmented randomization scheme, in which the noise distribution in each segment needs to be determined by the prior knowledge ρ. Whereas, Lemma 2 and Theorem 1 work for different ρ and other query functions f. For example, assuming each user determines his/her own coin flipping probability, i.e., $S_i \sim \text{Bernoulli}(\rho_i)$, we still have $\text{supp}(\pi^*) = \big\{(x, x') \colon x - x' = 1\big\}$. See the Kantorovich optimal transport plan π^* in Figure 2.

See Section F in the supplementary materials for the full proof of Lemma 2(a) and (b) by Theorem 1 and the detailed derivation of (13).

3.5 Relaxed Sufficient Condition

The sufficient condition in Theorem 1 is strict in that it enforces the inequality $d(x - x') \le \epsilon$ to hold for each $(x, x') \in \text{supp}(\pi^*)$ in (11). However, the integral in (11) is a noised expected distance. Knowing that the randomization does not increase statistical differences, Theorem 1 may result in a larger θ that overkill the data utility.[8] We relax this sufficient condition in the following theorem. The full proof is in Section G in the supplementary material.

Theorem 2 (relaxed sufficient condition)**.** *Let $\theta(s_i, s_j)$ be the maximum value of θ that holds the equalities*

$$\int e^{\eta(\theta) d(x - x')} \pi^*(x, x') \, \mathrm{d}x = e^{\epsilon} p(x' | s_j), \tag{14a}$$

$$\int e^{\eta(\theta) d(x - x')} \pi^*(x, x') \, \mathrm{d}x' = e^{\epsilon} p(x | s_i), \tag{14b}$$

over all x and x'. For the exponential mechanism,

[8] In (11), the distance $e^{\eta(\theta) d(x - x')}$ is averaged w.r.t. the joint probability $\pi^*(x, x')$ and then randomized by $P_{Y|X}(y|x) = P_{N_\theta}(y - x)$. Therefore, it is randomized statistical distance.

adding noise N_θ with $\theta = \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \theta(s_i,s_j)$ attains $(\epsilon,\mathbb{S})$-pufferfish privacy in Y.

Proof. Rewrite (11) as

$$\int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\,\mathrm{d}\pi^*(x,x')$$
$$=\int P_{N_\theta}(y-x')\int\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\pi^*(x,x')\,\mathrm{d}x\,\mathrm{d}x'.$$

Relaxing the condition $e^{\eta(\theta)d(x-x')}\le e^\epsilon,\forall x,x'$ to

$$\int\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\pi^*(x,x')\,\mathrm{d}x\le 0$$
$$\Longrightarrow\int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x\le e^\epsilon p(x'|s_j),\forall x' \quad (15)$$

still holds the inequality $\frac{P_{Y|S}(y|s_i,\rho)}{P_{Y|S}(y|s_j,\rho)}\le e^\epsilon$. For η nonincreasing in θ, the minimum θ for the condition (15) is the one that holds (15) as an equality. We have (14a). It can be shown in the same way that $\int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x'\le e^\epsilon p(x|s_j),\forall x$ is a relaxed sufficient condition for $\frac{P_{Y|S}(y|s_j,\rho)}{P_{Y|S}(y|s_i,\rho)}\le e^\epsilon$ and therefore (14b). Maximize θ that holds (14a) and (14b) over x' and x, respectively, and over all $(s_i,s_j)\in\mathbb{S}$, we have Theorem 2. □

The relaxation in Theorem 2 produces an $(\epsilon,\mathbb{S})$-pufferfish privacy achieving θ that is smaller than the one in Theorem 1.[9] Though for continuous X solving the integral equations in (14) could be complex, it is convenient to apply Theorem 2 to the integer-valued metric d, where (14) reduces to

$$\sum_{x,x'} e^{\eta(\theta)d(x-x')}\pi^*(x,x')=e^\epsilon p(x'|s_j),\quad\forall x' \quad (16a)$$
$$\sum_{x,x'} e^{\eta(\theta)d(x-x')}\pi^*(x,x')=e^\epsilon p(x|s_i),\quad\forall x \quad (16b)$$

and $\theta(s_i,s_j)$ for each $(s_i,s_j)\in\mathbb{S}$ can be determined by solving the polynomial equations above.

4 EXPERIMENT

In the UCI machine learning repository (Asuncion and Newman, 2007), the adult dataset was extracted from the census bureau database to predict whether the participant's income exceeds 50K/yr. It contains 32652 records and 15 attributes. In this experiment, we use 2 attributes/columns. Let S and X denote the columns 'race' and 'education', respectively. That is, we want to publish the column 'education' while protecting the privacy of 'race' for all participants.

We assign a numeric index to X: value 1 denotes 'Bachelors', value 2 denotes 'Some-college' and so on. See horizontal axis in Figure 3(a) for the alphabet $\mathcal{X}$ containing all possible values of 'education'. Consider the events $S=$ 'White' and $S=$ 'Asian-Pac-Islander'. The probability of 'education' X conditioned on each event is plotted in Figure 3(a). The support of both $P_{X|S}(\cdot|\text{'White'},\rho)$ and $P_{X|S}(\cdot|\text{'Asian-Pac-Islander'},\rho)$ is $\mathcal{X}=\{1,\dots,14\}$. In Figure 3(a), we show the corresponding Kantorovich optimal transport plan π^*. Take the Laplace noise for example. We have ℓ_1-norm $d(z)=|z|$ and

$$\triangle_{\mathcal{X}}=\max_{x,x'\in\mathcal{X}}|x-x'|=14,$$
$$\max_{(x,x')\in\mathrm{supp}(\pi^*)}|x-x'|=2.$$

The noise power is much reduced if we calibrating θ to the sensitivity of $\mathrm{supp}(\pi^*)$ rather than $\triangle_{\mathcal{X}}$.

Increasing the privacy budget ϵ from 0.8, we apply the sufficient condition in Theorem 1 and the relaxed sufficient condition in Theorem 2 to obtain θ for Laplace noise. Figure 4 shows the noise variance as a function of ϵ. Here, the value of θ in Theorem 2 is obtained by solving the polynomial equations in (16). It can be seen that the resulting noise power by the relaxed sufficient condition in Theorem 2 is always less than Theorem 1.

5 δ-Approximation by Gaussian

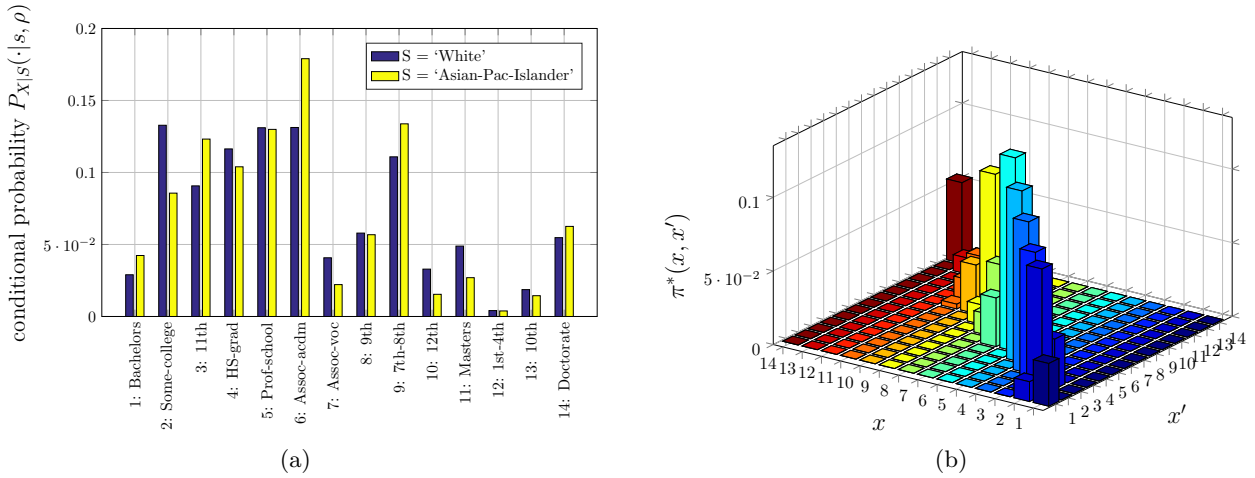
Mironov (2017) pointed out two reasons that Gaussian noise is preferred over Laplace noise: the noise variance proportional to the ℓ_2-norm is no larger than ℓ_1-norm; the tail probability of Gaussian distribution decays faster than Laplace distribution. This section considers zero-mean Gaussian noise N_θ: $P_{N_\theta}(z)=\frac{1}{\sqrt{2\pi\theta}}e^{-\frac{z^2}{2\theta^2}}$.

It is shown in Dwork et al. (2014, Theorem 3.22) that for $\delta\in(0,1)$, Gaussian mechanism attains δ-approximation of differential privacy: $\left|\log\frac{P_{Y|S}(y|s_i)}{P_{Y|S}(y|s_j)}\right|\le\epsilon$ for all neighboring s_i and s_j with probability at least $1-\delta$. Following the definition of (ϵ,δ)-differential privacy, we say that Y is a δ-approximation of $(\epsilon,\mathbb{S})$-pufferfish privacy if

$$\log\frac{P_{Y|S}(y|s_i,\rho)-\delta}{P_{Y|S}(y|s_j,\rho)}\le\epsilon,\ \log\frac{P_{Y|S}(y|s_j,\rho)-\delta}{P_{Y|S}(y|s_i,\rho)}\le\epsilon$$

for all $(s_i,s_j)\in\mathbb{S}$ and $\rho\in\mathbb{D}$. The theorem below states that to achieve this approximation, it suffices to calibrate the standard deviation θ of Gaussian noise

[9] For two functions $f_1(\theta)$ and $f_2(\theta)$ both nonincreasing in θ, $f_1^{-1}(a)\le f_2^{-1}(a),\forall a$.



(a) (b)

Figure 3: For S and X being the attributes 'race' and 'education', respectively, in the adult database (Asuncion and Newman, 2007), (a) shows the distribution of 'education' conditioned on two events: the 'race' is 'White' and the 'race' is 'Asian-Pac-Islander'. (b) shows the corresponding Kantorovich optimal transport plan π^*.

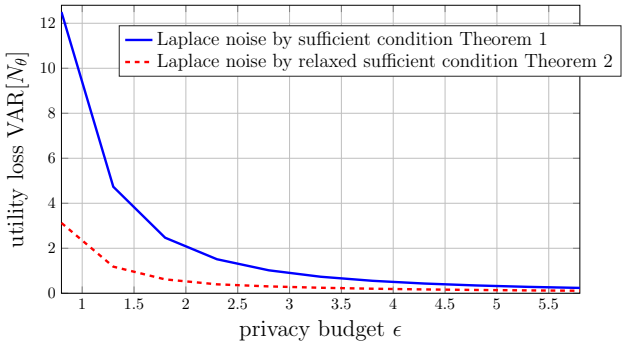


Figure 4: The variance of Laplace noise then Theorem 1 and Theorem 2 are applied to the optimal transportation plan π^* in Figure 3(b).

to the sensitivity of the Kantorovich optimal transport plan.

Theorem 3. *For $\delta \in (0,1)$ and N_θ being zero-mean Gaussian noise, Y attains δ-approximation of ϵ-pufferfish privacy*

(a) if $\theta \geq \frac{\sqrt{2\log(1.25/\delta)}}{\epsilon}\triangle$ for all $\epsilon \leq 1$;

(b) if $\theta = \frac{\triangle}{\epsilon}c$ for $c > 0.41\delta^{-\frac{1}{3}} + \sqrt{(0.41\delta^{-\frac{1}{3}})^2 + \frac{\epsilon}{2}}$,

where $\triangle = \sup_{(x,x')\in supp(\pi^*)} |x - x'|$ is the ℓ_1-sensitivity of the Kantorovich optimal transport plan π^*.

Proof. Let $\theta = \frac{\triangle}{\epsilon}c$. To have

$$P_{Y|S}(y|s_i,\rho) - e^\epsilon P_{Y|S}(y|s_j,\rho) = \int \frac{1}{\sqrt{2\pi}\theta}\left(e^{-\frac{(y-x)^2}{2\theta^2}} - e^{\epsilon-\frac{(y-x')^2}{2\theta^2}}\right)\mathrm{d}\pi^*(x,x') \leq 0,$$

it suffice to make $\frac{(y-x')^2-(y-x)^2}{2\theta^2} \leq \frac{2\triangle|y-x|+\triangle^2}{2\theta^2} \leq \epsilon \Longrightarrow \frac{|y-x|}{\theta} \leq c - \frac{\epsilon}{2c}$, where $\frac{|Y-X|}{\theta} = \frac{N_\theta}{\theta}$ is standard normally distributed. To have $\Pr(\frac{|y-x|}{\theta} \leq t) \geq 1-\delta$, we use the Gaussian tail bound $\Pr(\frac{y-x}{\theta} > t) < \frac{1}{\sqrt{2\pi}t}e^{-\frac{t^2}{2}} < \frac{\delta}{2}$, which can be written as

$$\log t + \frac{t^2}{2} > \log\sqrt{\frac{2}{\pi}}\frac{1}{\delta}. \tag{17}$$

Substituting $t = c - \frac{\epsilon}{2c}$, we have $c^2 > 2\log(1.25/\delta)$ and (a). To prove (b), we relieve the constraint $\epsilon \leq 1$ and apply the inequality $t - 1 \geq \log t, \forall t > 0$ to (17), i.e., request $\log t + \log\frac{t^2}{2} + 1 \geq \log t + \log\frac{t^2}{2} > \log\sqrt{\frac{2}{\pi}}\frac{1}{\delta}$. We have $t = c - \frac{\epsilon}{2c} > \left(\frac{2}{e}\right)^{\frac{1}{3}}\left(\frac{2}{\pi}\right)^{\frac{1}{6}}\delta^{-\frac{1}{3}}$, where $\left(\frac{2}{e}\right)^{\frac{1}{3}}\left(\frac{2}{\pi}\right)^{\frac{1}{6}} = 0.8373$. Solving the quadratic inequality $c - \frac{\epsilon}{2c} > 0.84\delta^{-\frac{1}{3}}$, we get $c > 0.41\delta^{-\frac{1}{3}} + \sqrt{(0.41\delta^{-\frac{1}{3}})^2 + \frac{\epsilon}{2}}$. See Section H in the supplementary material for the full proof of Theorem 3. □

The proof of Theorem 3(a) is similar to Dwork et al. (2014, Appendix A) for (ϵ,δ)-differential privacy, except that the noise should be calibrated to the sensitivity of the Kantorovich optimal transport plan π^*, instead of the query function f.

Remark 1. *Lemma 1, Theorem 1 and Theorem 3(a) parallel the well-known results on Laplace, exponential*

and Gaussian mechanisms for differential privacy: replacing the sensitivity of the query function f in Theorems 3.6, 3.10 and 3.22 in Dwork et al. (2014) by the maximum pairwise distance in the Kantorovich optimal transport plan π^* over all $(s_i, s_j) \in \mathbb{S}$ and $\rho \in \mathbb{D}$, the additive noise mechanism attains pufferfish privacy.

6 CONCLUSION

We studied the problem of how to attain pufferfish privacy, the ϵ-indistinguishability when the secret S is correlated with the public data X, by adding independent noise N to X. We proved that calibrating noise to the maximum pairwise distance over the support of the Kantorovich optimal transport plan π^* attains pufferfish privacy. Unlike the difficulty in determining the optimal transport plan in the existing ∞-Wasserstein mechanism, π^* is directly obtained by the conditional probabilities $P_{X|S}(\cdot|s_i, \rho)$ and $P_{X|S}(\cdot|s_j, \rho)$ for every secret pair (s_i, s_j). We also derived a relaxed sufficient condition and showed that it enhances data utility for integer-valued X.

This paper in fact proposes a method for attaining the pufferfish privacy based on the mass transport cost upper bound $C(x, x'; \theta)$: for any noise distribution $P_{N_\theta}(\cdot)$ such that $P_{N_\theta}(y - x) \leq C(x, x'; \theta) P_{N_\theta}(y - x'), \forall x, x', y$, pufferfish privacy attains if $\sup_{(x,x') \in \text{supp}(\pi^*)} C(x, x'; \theta) \leq e^\epsilon$, for all $(s_i, s_j) \in \pi^*$ and $\rho \in \mathbb{D}$. See (11). Here, $C(x, x'; \theta)$ does not need to be an exponential function. Therefore, it is worth exploring noise distributions other than the exponential mechanism. For $P_{X|S}(\cdot|s, \rho)$ being Gaussian distribution or Gaussian mixture model for all s, the Kantorovich optimal transport plan π^* is fully characterized by the mean and covariance matrix (Takatsu, 2010; Delon and Desolneux, 2020). Since Gaussian models are widely used in machine learning, it is of interest whether the Kantorovich mechanism can be apply to the privacy-preserving pattern recognition problems.

Acknowledgements

The author would like to thank A/Prof Olya Ohrimenko for helping her initiate the study on pufferfish privacy and Prof Ben Rubinstein for his useful advice on the dissemination of the research results in this paper.

References

Asuncion, A. and Newman, D. (2007). UCI machine learning repository https://archive.ics.uci.edu/ml/index.php.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '07, page 273–282, New York, NY, USA. Association for Computing Machinery.

Champion, T., De Pascale, L., and Juutinen, P. (2008). The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20.

De Pascale, L. and Louet, J. (2019). A study of the dual problem of the one-dimensional l_∞-optimal transport problem with applications. *Journal of Functional Analysis*, 276(11):3304–3324.

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970.

Ding, N. and Farokhi, F. (2020). Developing non-stochastic privacy-preserving policies using agglomerative clustering. *IEEE Transactions on Information Forensics and Security*, 15:3911–3923.

Ding, N., Liu, Y., and Farokhi, F. (2021). A linear reduction method for local differential privacy and log-lift. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 551–556, Melbourne.

du Pin Calmon, F. and Fawaz, N. (2012). Privacy against statistical inference. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1401–1408.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. (2008). Differential privacy: A survey of results. In Agrawal, M., Du, D., Duan, Z., and Li, A., editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. and Nissim, K. (2004). Privacy-preserving datamining on vertically partitioned databases. In Franklin, M., editor, *Advances in Cryptology – CRYPTO 2004*, pages 528–544, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

He, X., Machanavajjhala, A., and Ding, B. (2014). Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, page 1447–1458, Snowbird, Utah, USA. Association for Computing Machinery.

Issa, I., Wagner, A. B., and Kamath, S. (2020). An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657.

Kifer, D. and Machanavajjhala, A. (2012). A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '12, page 77–88, New York, NY, USA. Association for Computing Machinery.

Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1).

Liang, W., Chen, H., Liu, R., Wu, Y., and Li, C. (2020). A pufferfish privacy mechanism for monitoring web browsing behavior under temporal correlations. *Computers & Security*, 92:101754.

Makhdoumi, A., Salamatian, S., Fawaz, N., and Médard, M. (2014). From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pages 501–505.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Sankar, L., Rajagopalan, S. R., and Poor, H. V. (2013). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Sarwate, A. D. and Sankar, L. (2014). A rate-disortion perspective on local differential privacy. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 903–908.

Song, S., Wang, Y., and Chaudhuri, K. (2017). Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, page 1291–1306, New York, NY, USA.

Takatsu, A. (2010). On Wasserstein geometry of Gaussian measures. In *Probabilistic approach to geometry*, pages 463–472. Mathematical Society of Japan.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Supplementary Material: Kantorovich Mechanism for Pufferfish Privacy

A V INDEPENDENT USER SYSTEM

In the V independent user system, Let ρ be the prior knowledge about the probability distribution $P_{S_i}(\cdot)$ for all $i \in \mathcal{V}$. Since S_i's are independent random variables, $P_S(s) = \Pi_{i\in\mathcal{V}} P_{S_i}(s_i)$ and $P_{S|S_i}(s|s_i) = P_{S_{-i}}(s_{-i}), \forall i \in \mathcal{V}$, where $S_{-i} = (S_j \colon j \in \mathcal{V} \setminus \{i\})$ denote the multiple random variable excluding dimension i. We also have the conditional probabilities

$$\begin{aligned}
P_{X|S_i}(x|a,\rho) &= \Pr(f(S) = x | S_i = a) \\
&= \int_{\mathcal{S}_{-i}(x,a)} P_{S|S_i}(s|a,\rho)\,\mathrm{d}s_{-i} \\
&= \int_{\mathcal{S}_{-i}(x,a)} P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i} \\
&= \Pr(f(S_i = a, S_{-i}) = x), \\
P_{X|S_i}(x|\perp_i,\rho) &= \Pr(f(S) = x) \\
&= \int_{\mathcal{S}(x)} P_{S|S_i}(s|\perp_i)\,\mathrm{d}s \\
&= \int_{\mathcal{S}(x)} P_S(s)\,\mathrm{d}s \\
&= \int \Big(\int_{\mathcal{S}_{-i}(x,s_i)} P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i} \Big) P_{S_i}(s_i)\,\mathrm{d}s_i \\
&= \int P_{X|S_i}(x|s_i,\rho) P_{S_i}(s_i)\,\mathrm{d}s_i,
\end{aligned}$$

where $\mathcal{S}(x) = \{s \colon x = f(s)\}$ and $\mathcal{S}_{-i}(x,a) = \{s_{-i} \colon x = f(s_i = a, s_{-i})\}$. The last equality above means $\Pr(f(S) = x) = \mathbb{E}_{S_i}[\Pr(f(S) = x|S_i)]$.

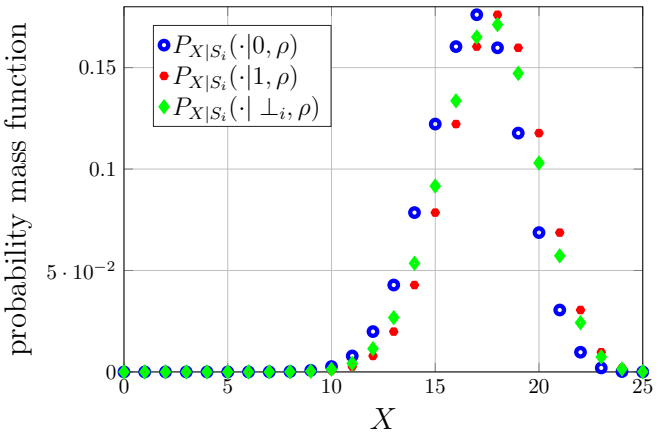


Figure 5: The conditional probability $P_{X|S_i}(x|s_i,\rho)$ for the events $S_i = 0$, $S_i = 1$ and $S_i = \perp_i$ in the V independent user system. The number of users is $V = 25$. Each $S_i \sim$ Bernoulli(0.7). The function f is a counting query $X = f(S) = \sum_{i\in\mathcal{V}} S_i$ and $X \sim$ Binomial$(25, 0.7)$.

In Figure 5, we show the conditional probability mass function $P_{X|S_i}(x|\cdot,\rho)$ in an $V = 25$ independent user

system, where $\mathcal{S}_i \in \{0, 1, \perp_i\}$. Each dimension S_i in the multiple random variable $S = (S_i \colon i \in \mathcal{V})$ follows Bernoulli(p) distribution, where $p = 0.7$. The corresponding Kantorovich optimal transport plan π^* is shown in Figure 1 in the main submission.

For $Y = X + N$, where N is independent of X,

$$\begin{aligned}
P_{Y|S_i}(y|a,\rho) &= \int P_N(y-x) P_{X|S_i}(x|a,\rho)\,\mathrm{d}x \\
&= \int P_N(y - f(s_i = a, s_{-i})) P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i}, \\
P_{Y|S_i}(y|\perp_i,\rho) &= \int P_N(y-x) P_{X|S_i}(x|\perp_i,\rho)\,\mathrm{d}x \\
&= \int P_N(y - f(s)) P_S(s)\,\mathrm{d}s \\
&= \iint P_N(y - f(s_i, s_{-i})) P_{S_{-i}}(s_{-i}) P_{S_i}(s_i)\,\mathrm{d}s_{-i}\,\mathrm{d}s_i \\
&= \int P_{Y|S_i}(y|s_i,\rho) P_{S_i}(s_i)\,\mathrm{d}s_i.
\end{aligned}$$

The last equality above means $P_{Y|S_i}(y|\perp_i,\rho) = \mathbb{E}_{S_i}[P_{Y|S_i}(y|\cdot,\rho)]$.

B PROOF of LEMMA 1

Denote $[z]_+ = \max\{z, 0\}$ for all $z \in \mathbb{R}$. First, for all $\pi \in \Gamma(s_i, s_j)$, $\inf_{\pi\in\Gamma(s_i,s_j)} \sup_{(x,x')\in\mathrm{supp}(\pi)} \left(\frac{|x-x'|}{\theta} - \epsilon\right) = 0$ is equivalent to

$$\inf_{\pi\in\Gamma(s_i,s_j)} \int \Big[\frac{|x-x'|}{\theta} - \epsilon\Big]_+ \mathrm{d}\pi(x,x') = 0.$$

Then, for the ∞-Wasserstein mechanism with $\theta = \frac{1}{\epsilon}\max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} W_\infty(s_i, s_j)$ proposed by Song et al. (2017), we have

$$\begin{aligned}
\theta &= \frac{1}{\epsilon} \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} W_\infty(s_i, s_j) && (18)\\
&= \frac{1}{\epsilon} \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \inf_{\pi\in\Gamma(s_i,s_j)} \sup_{(x,x')\in\mathrm{supp}(\pi)} |x-x'| \\
&= \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \Big\{\theta \colon \inf_{\pi\in\Gamma(s_i,s_j)} \sup_{(x,x')\in\mathrm{supp}(\pi)} \left(\frac{|x-x'|}{\theta} - \epsilon\right) = 0\Big\} \\
&= \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \Big\{\theta \colon \inf_{\pi\in\Gamma(s_i,s_j)} \int \Big[\frac{|x-x'|}{\theta} - \epsilon\Big]_+ \mathrm{d}\pi(x,x') = 0\Big\} && (19)\\
&= \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \Big\{\theta \colon \int \Big[\frac{|x-x'|}{\theta} - \epsilon\Big]_+ \mathrm{d}\pi^*(x,x') = 0\Big\} && (20)\\
&= \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \Big\{\theta \colon \sup_{(x,x')\in\mathrm{supp}(\pi^*)} \big(\frac{|x-x'|}{\theta} - \epsilon\big) = 0\Big\} \\
&= \frac{1}{\epsilon} \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \sup_{(x,x')\in\mathrm{supp}(\pi^*)} |x-x'|.
\end{aligned}$$

The infinum in (19) is a Kantorovich optimal transport problem that determines W_1 distance. Here, $\left[\frac{|\cdot|}{\theta} - \epsilon\right]_+$ is a convex function for all θ and ϵ. Therefore, in (20), we substitute by the Kantorovich optimal transport plan π^*, i.e., the minimizer of $\inf_{\pi\in\Gamma(s_i,s_j)} \int \big[\frac{|x-x'|}{\theta} - \epsilon\big]_+ \mathrm{d}\pi(x,x')$. □

C EFFICIENT SOLUTION TO TWO PROBLEMS IN Song et al. (2017) BY LEMMA 1

For the two examples in Song et al. (2017, Section 3.1), we show how to determine θ by Lemma 1. Table 1 shows the probabilities of X conditioned on two instances s_i and s_j of S.

Table 1: 1st example of $P_{X|S}$

	$X=1$	$X=2$	$X=3$	$X=4$
$P_{X\|S}(\cdot\|s_i,\rho)$	1/3	1/6	1/3	1/6
$P_{X\|S}(\cdot\|s_j,\rho)$	1/4	1/4	1/6	1/3

We first obtain the joint cumulative mass function (CMF) of the Kantorovich optimal transport plan $\pi^*((-\infty,x],(-\infty,x'])=\min\{F(x|s_i),F(x'|s_j)\}$ in Table 2 and then the join probability mass function (PMF) $\pi^*(x,x')$ in Table 3.[10]

Table 2: $\pi^*((-\infty,x],(-\infty,x'])$

	$X'=1$	$X'=2$	$X'=3$	$X'=4$
$X=1$	1/4	1/3	1/3	1/3
$X=2$	1/4	1/2	1/2	1/2
$X=3$	1/4	1/2	2/3	5/6
$X=4$	1/4	1/2	2/3	1

Table 3: $\pi^*(x,x')$

	$X'=1$	$X'=2$	$X'=3$	$X'=4$
$X=1$	1/4	1/12	0	0
$X=2$	0	1/6	0	0
$X=3$	0	0	1/6	1/6
$X=4$	0	0	0	1/6

We have $\sup_{(x,x')\in\text{supp}(\pi^*)}|x-x'|=1$. Applying Lemma 1, adding Laplace noise with $\theta=\frac{1}{\epsilon}$ attains pufferfish privacy.

Table 4 shows the second example of the conditional probabilities $P_{X|S}(\cdot|s_i,\rho)$ and $P_{X|S}(x|s_j,\rho)$, for which we have the joint CMF and PMF in Tables 5 and 6, respectively. By Lemma 2, adding Laplace noise with $\theta=\frac{2}{\epsilon}$ attains pufferfish privacy.

Table 4: 2nd example of $P_{X|S}$

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$
$P_{X\|S}(\cdot\|s_i,\rho)$	0.2	0.225	0.5	0.075	0
$P_{X\|S}(\cdot\|s_j,\rho)$	0	0.075	0.5	0.225	0.2

[10] Recall that for $x_1,x_2,x_1',x_2'\in\mathcal{X}$ such that $x_1<x_2$ and $x_1'<x_2'$, $\pi^*([x_1,x_2],[x_1',x_2'])=\pi^*((-\infty,x_2],(-\infty,x_2'])-\pi^*((-\infty,x_1],(-\infty,x_2'])-\pi^*((-\infty,x_2],(-\infty,x_1'])+\pi^*((-\infty,x_1],(-\infty,x_1'])$.

Table 5: $\pi^*((-\infty, x], (-\infty, x'])$

	$X'=1$	$X'=2$	$X'=3$	$X'=4$	$X'=5$
$X=1$	0	0.075	0.2	0.2	0.2
$X=2$	0	0.075	0.425	0.425	0.425
$X=3$	0	0.075	0.575	0.8	0.925
$X=4$	0	0.075	0.575	0.8	1
$X=5$	0	0.075	0.575	0.8	1

Table 6: $\pi^*(x, x')$

	$X'=1$	$X'=2$	$X'=3$	$X'=4$	$X'=5$
$X=1$	0	0.075	0.125	0	0
$X=2$	0	0	0.225	0	0
$X=3$	0	0	0.15	0.225	0.125
$X=4$	0	0	0	0	0.075
$X=5$	0	0	0	0	0

D PROOF of LEMMA 2

In the V independent user system, for the discriminative pair set $\mathbb{S}_i = \{(S_i = a, S_i = b) \colon a, b \in \mathcal{S}_i\}$, the sensitivity for the query function f and metric d is

$$\triangle_f(\mathbb{S}_i) = \max_{a,b \in \mathcal{S}_i} \max_{s_{-i}} d\big(f(s_i = a, s_{-i}) - f(s_i = b, s_{-i})\big)$$

for all $\rho \in \mathbb{D}$. Using the probabilities derived in Section A,

$$\begin{aligned}
\frac{P_{Y|S_i}(y|a,\rho)}{P_{Y|S_i}(y|b,\rho)} &= \frac{\int P_{N_\theta}(y - f(s_i = a, s_{-i}))P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i}}{\int P_{N_\theta}(y - f(s_i = b, s_{-i}))P_{S_{-i}}(z_{-i})\,\mathrm{d}s_{-i}} \\
&\le \frac{\int P_{N_\theta}(y - f(s_i = b, s_{-i}))e^{\eta(\theta)d(f(s_i=a,s_{-i})-f(s_i=b,s_{-i}))}P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i}}{\int P_{N_\theta}(y - f(s_i = b, s_{-i}))P_{S_{-i}}(s_{-i})\,\mathrm{d}s_{-i}} \\
&\le e^{\eta(\theta)\triangle_f(\mathbb{S}_i)}, \quad \forall a, b \in \mathcal{S}_i.
\end{aligned}$$

Therefore, $\theta = \eta^{-1}(\frac{\epsilon}{\triangle_f(\mathbb{S}_i)})$ is a sufficient condition for $\frac{P_{Y|S_i}(y|a,\rho)}{P_{Y|S_i}(y|b,\rho)} \le e^\epsilon$ for all $a, b \in \mathcal{S}_i$, y and ρ. This proves (a).

For $(\epsilon, \mathbb{S}_i)$-pufferfish private Y, we have $P_{Y|S_i}(y|a,\rho) \le e^\epsilon P_{Y|S_i}(y|b,\rho)$ and $P_{Y|S_i}(y|b,\rho) \le e^\epsilon P_{Y|S_i}(y|a)$ for all $a, b \in \mathcal{S}_i$, y and ρ. Using the fact that $P_{Y|S_i}(y|\perp_i, \rho) = \int P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i$,

$$\begin{aligned}
\frac{P_{Y|S_i}(y|\perp_i,\rho)}{P_{Y|S_i}(y|a,\rho)} &= \frac{\int P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i}{P_{Y|S_i}(y|a,\rho)} \\
&\le \frac{\int e^\epsilon P_{Y|S_i}(y|a,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i}{P_{Y|S_i}(y|a,\rho)} = e^\epsilon, \\
\frac{P_{Y|S_i}(y|a,\rho)}{P_{Y|S_i}(y|\perp_i,\rho)} &= \frac{P_{Y|S_i}(y|a,\rho)}{\int P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i} \\
&= \frac{\int P_{Y|S_i}(y|a,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i}{\int P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i} \\
&\le \frac{\int e^\epsilon P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i}{\int P_{Y|S_i}(y|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i} = e^\epsilon,
\end{aligned}$$

for all $a \in \mathcal{S}_i$, y and ρ, i.e., Y is $(\epsilon, \mathbb{S}_{\perp_i})$-pufferfish private. This proves (b). □

E PROOF of THEOREM 1

We derive the following results for each $\rho \in \mathbb{D}$. For any pair $(s_i, s_j) \in \mathbb{S}$,

$$\begin{aligned} P_{Y|S}(y|s_i,\rho) - e^{\epsilon} P_{Y|S}(y|s_j,\rho) &= \int P_{N_\theta}(y-x) P_{X|S}(x|s_i,\rho)\,\mathrm{d}x - e^{\epsilon}\int P_{N_\theta}(y-x') P_{X|S}(x'|s_j,\rho)\,\mathrm{d}x' \\ &= \int \big(P_{N_\theta}(y-x) - e^{\epsilon} P_{N_\theta}(y-x')\big)\,\mathrm{d}\pi(x,x') \\ &\le \int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')} - e^{\epsilon}\big)\,\mathrm{d}\pi(x,x'), \qquad \forall y. \end{aligned} \tag{21}$$

Requesting $e^{\eta(\theta)d(x-x')} - e^{\epsilon} \le 0$ for each pair of x and x', we derive a sufficient condition

$$\inf_{\pi\in\Gamma(s_i,s_j)} \sup_{(x,x')\in\mathrm{supp}(\pi)} \big(\eta(\theta)d(x-x') - \epsilon\big) \le 0 \tag{22}$$

for $\frac{P_{Y|S}(y|s_i,\rho)}{P_{Y|S}(y|s_j,\rho)} \le e^{\epsilon}$. Note that the infimum in (22) is for the purpose of searching the minimum value of θ (over all couplings) that holds the sufficient condition, knowing $\mathrm{VAR}[N_\theta] \propto \theta$.

The sufficient condition (22) is equivalent to

$$\inf_{\pi\in\Gamma(s_i,s_j)} \int \big[\eta(\theta)d(x-x') - \epsilon\big]_+ \,\mathrm{d}\pi(x,x') = \int \big[\eta(\theta)d(x-x') - \epsilon\big]_+ \,\mathrm{d}\pi^*(x,x') \le 0 \tag{23}$$

where π^* denotes the Kantorovich optimal transport plan. We further convert (23) to

$$\begin{aligned} \int \big[\eta(\theta)d(x-x') - \epsilon\big]_+ \,\mathrm{d}\pi^*(x,x') \le 0 &\Longrightarrow \sup_{(x,x')\in\mathrm{supp}(\pi^*)} \eta(\theta)d(x-x') - \epsilon \le 0 \\ &\Longrightarrow \eta(\theta) \le \epsilon \Big/ \sup_{(x,x')\in\mathrm{supp}(\pi^*)} d(x-x'). \end{aligned} \tag{24}$$

For invertible η that is nonincreasing in θ, the minimum value of θ that holds the inequality (24) is

$$\eta^{-1}\Big(\epsilon \Big/ \sup_{(x,x')\in\mathrm{supp}(\pi^*)} d(x-x')\Big).$$

Taking the maximum of this value over all $(s_i, s_j) \in \mathbb{S}$ and $\rho \in \mathbb{D}$, we have the sufficient condition

$$\theta = \max_{\rho\in\mathbb{D},(s_i,s_j)\in\mathbb{S}} \eta^{-1}\Big(\epsilon \Big/ \sup_{(x,x')\in\mathrm{supp}(\pi^*)} d(x-x')\Big). \tag{25}$$

To make $\frac{P_{Y|S}(y|s_j,\rho)}{P_{Y|S}(y|s_i,\rho)} \le e^{\epsilon}$, we have the sufficient condition $\inf_{\pi\in\Gamma(s_j,s_i)} \sup_{(x',x)\in\mathrm{supp}(\pi)} \big(\eta(\theta)d(x'-x) - \epsilon\big) \le 0$ the minimizer for which is $\pi^{*\intercal}$ such that $\pi^{*\intercal}(x',x) = \pi^*(x,x')$. Based on the symmetry property of d, i.e., $d(x-x') = d(x'-x)$, $\sup_{(x,x')\in\pi^*} d(x-x') = \sup_{(x',x)\in\pi^{*\intercal}} d(x'-x)$, i.e., (24) is also a sufficient condition for $\frac{P_{Y|S}(y|s_j,\rho)}{P_{Y|S}(y|s_i,\rho)} \le e^{\epsilon}$. This completes the proof. □

F PROOF of LEMMA 2 BY THEOREM 1

For the V independent user system, the result that $(\epsilon, \mathbb{S})$-pufferfish privacy can be attained by calibrating noise to the sensitivity of the query function f in Lemma 2 is explained by the support of the Kantorovich optimal transport plan π^* in Theorem 1.

To obtain the cumulative density distribution (CDF) $\pi^*((-\infty,x],(-\infty,x']) = \min\{F_{X|S_i}(x|a,\rho), F_{X|S_i}(x'|\perp_i, \rho)\}$, first consider any x and x' such that $x > x'$. Let $\delta = d(x-x')$. Then, $x = x' + d^{-1}(\delta)$. For any $a \in \mathcal{S}_i$, we derive the CDFs $F_{X|S_i}(x|a,\rho)$ and $F_{X|S_i}(x|\perp_i,\rho)$ for the conditional probabilities $P_{X|S_i}(x|a,\rho)$ and $P_{X|S_i}(x|\perp_i,\rho)$, respectively, as follows.

$$
\begin{aligned}
F_{X|S_i}(x|a,\rho) &= \int_{-\infty}^{x} P_{X|S_i}(l|a,\rho)\,\mathrm{d}l \\
&= \int_{-\infty}^{x} \Pr(f(S_i=a,S_{-i})=l)\,\mathrm{d}l \\
&= \int_{-\infty}^{x'+d^{-1}(\delta)} \Pr(f(S_i=a,S_{-i})=l)\,\mathrm{d}l \\
&= \int_{-\infty}^{x'} \Pr(f(S_i=a,S_{-i})=l+d^{-1}(\delta))\,\mathrm{d}l, \\
F_{X|S_i}(x|\perp_i,\rho) &= \int_{-\infty}^{x} P_{X|S_i}(l|\perp_i,\rho)\,\mathrm{d}l \\
&= \int_{-\infty}^{x} \int P_{X|S_i}(l|s_i,\rho)P_{S_i}(s_i)\,\mathrm{d}s_i\,\mathrm{d}l \\
&= \int \Big(\int_{-\infty}^{x} \Pr(f(S_i=s_i,S_{-i})=l)\,\mathrm{d}l\Big)P_{S_i}(s_i)\,\mathrm{d}s_i.
\end{aligned}
$$

Comparing $F_{X|S_i}(x|a,\rho)$ and $F_{X|S_i}(x'|b,\rho)$, we have

$$
\begin{aligned}
F_{X|S_i}(x|a,\rho) - F_{X|S_i}(x'|b,\rho) &= \int_{-\infty}^{x'} \Pr(f(S_i=a,S_{-i})=l+d^{-1}(\delta))\,\mathrm{d}l - \int_{-\infty}^{x'} \Pr(f(S_i=b,S_{-i})=l)\,\mathrm{d}l \\
&= \int_{-\infty}^{x'} \Big(\Pr(f(S_i=a,S_{-i})=l+d^{-1}(\delta)) - \Pr(f(S_i=b,S_{-i})=l)\Big)\,\mathrm{d}l \\
&< 0, \qquad \forall \delta > \triangle_f(\mathbb{S}_i).
\end{aligned}
$$

Similarly, for all x and x' such that $x < x'$, due to the symmetry property of d, $d(x-x') = d(x'-x) = \delta \Longrightarrow x' = x + d^{-1}(\delta)$ and

$$
\begin{aligned}
&F_{X|S_i}(x|a,\rho) - F_{X|S_i}(x'|b,\rho) \\
&\qquad = \int_{-\infty}^{x'} \Big(\Pr(f(S_i=a,S_{-i})=l) - \Pr(f(S_i=b,S_{-i})=l+d^{-1}(\delta))\Big)\,\mathrm{d}l > 0, \qquad \forall \delta > \triangle_f(\mathbb{S}_i).
\end{aligned}
$$

That is, for all x, x' such that $d(x-x') > \triangle_f(\mathbb{S}_i)$, $\pi^*((-\infty,x],(-\infty,x']) = \min\{F_{X|S_i}(x|a,\rho), F_{X|S_i}(x'|\perp_i,\rho)\}$ is independent on either x or x' and so

$$
\pi^*(x,x') = \frac{\mathrm{d}^2}{\mathrm{d}x\,\mathrm{d}x'}\pi^*((-\infty,x],(-\infty,x']) = 0, \qquad \forall x,x' \colon d(x-x') > \triangle_f(\mathbb{S}_i).
$$

Equivalently,

$$
\mathrm{supp}(\pi^*) \subseteq \big\{(x,x') \colon d(x-x') \le \triangle_f(\mathbb{S}_i)\big\}
$$

and so

$$
\sup_{(x,x')\in \mathrm{supp}(\pi^*)} d(x-x') \le \triangle_f(\mathbb{S}_i).
$$

Therefore, by Theorem 1, adding noise θ with

$$
\theta = \eta^{-1}\Big(\frac{\epsilon}{\triangle_f(\mathbb{S}_i)}\Big).
$$

attains $(\epsilon, \mathbb{S}_i)$-pufferfish privacy, which proves Lemma 2(a).

Consider the discriminative pair set $\mathbb{S}_{\perp_i} = \{(S_i = a, S_i = \perp_i) \colon a \in \mathcal{S}_i\}$. For all $a \in \mathcal{S}_i$, we have

$$
\begin{aligned}
F_{X|S_i}(x|a,\rho) - F_{X|S_i}(x'|\perp_i,\rho) &= \int \Big(\int_{-\infty}^{x} \Pr(f(S_i=a,S_{-i}=l))\,\mathrm{d}l - \int_{-\infty}^{x'} \Pr(f(S_i=s_i,S_{-i}=l))\,\mathrm{d}l\Big)P_{S_i}(s_i)\,\mathrm{d}s_i \\
&= \int \Big(F_{X|S_i}(x|s_i,\rho) - F_{X|S_i}(x|a,\rho)\Big)P_{S_i}(s_i)\,\mathrm{d}s_i.
\end{aligned}
$$

Again, we have $\mathrm{supp}(\pi^*) \subseteq \big\{(x,x')\colon d(x-x') \le \triangle_f(\mathbb{S}_i)\big\}$ and $\theta = \max_{(s_i,s_j)\in\mathbb{S}} \eta^{-1}\big(\epsilon/\triangle_f(\mathbb{S}_i)\big)$. This proves Lemma 2(b).

F.1 Separable Query Function

In addition, we obtain an extra result for the separable query function. Assume f is separable, i.e., $f(s) = \sum_{i\in\mathcal{V}} f_i(s_i), \forall s = (s_i\colon i\in\mathcal{V})$. One example is the counting query $f(s) = \sum_{i\in\mathcal{V}} s_i$.

For any $a, b \in \mathcal{S}_i$,

$$f(s_i = a, s_{-i}) - f(s_i = b, s_{-i}) = f_i(a) - f_i(b), \qquad \forall s_{-i}.$$

Let $\delta = f_i(a) - f_i(b)$, we have

$$\begin{aligned} F_{X|S_i}(x|a,\rho) &= \int_{-\infty}^{x} \Pr(f(S_i = a, S_{-i}) = l)\,\mathrm{d}l \\ &= \int_{-\infty}^{x} \Pr(f(S_i = b, S_{-i}) = l-\delta)\,\mathrm{d}l \\ &= \int_{-\infty}^{x-\delta} \Pr(f(S_i = b, S_{-i}) = l)\,\mathrm{d}l \end{aligned}$$

and so

$$F_{X|S_i}(x|a,\rho) - F_{X|S_i}(x'|b,\rho) = \int_{-\infty}^{x-\delta} \Pr(f(S_i = b, S_{-i}) = l)\,\mathrm{d}l - \int_{-\infty}^{x'} \Pr(f(S_i = b, S_{-i}) = l)\,\mathrm{d}l$$
$$\begin{cases} > 0 & x - x' > \delta \\ < 0 & x - x' < \delta \\ = 0 & x - x' = \delta \end{cases}.$$

Therefore,

$$\begin{aligned} \pi^*(x,x') &= \frac{\mathrm{d}^2}{\mathrm{d}x\,\mathrm{d}x'}\pi^*(x,x') \\ &= \frac{\mathrm{d}^2}{\mathrm{d}x\,\mathrm{d}x'} \min\{F_{X|S_i}(x|a,\rho), F_{X|S_i}(x'|b,\rho)\} = 0, \qquad \forall x, x'\colon x - x' \neq \delta. \end{aligned}$$

That is, the support of π^* is

$$\mathrm{supp}(\pi^*) = \big\{(x,x')\colon x - x' = f_i(a) - f_i(b)\big\}.$$

Also note that for separable query functions f,

$$\begin{aligned} \triangle_f(\mathbb{S}_i) &= \max_{a,b\in\mathcal{S}_i} \max_{s_{-i}} d\big(f(s_i = a, s_{-i}) - f(s_i = b, s_{-i})\big) \\ &= \max_{a,b\in\mathcal{S}_i} d(f_i(a) - f_i(b)). \end{aligned}$$

is independent of S_{-i} and

$$\theta = \eta^{-1}\Big(\frac{\epsilon}{\max_{a,b\in\mathcal{S}_i} d(f_i(a) - f_i(b))}\Big).$$

G PROOF of THEOREM 2

The proof starts with the proposition below.

Proposition 1. *If there exists a nonnegative function $D_\epsilon(\cdot;\theta)$ such that*

$$e^{\eta(\theta)d(z)} - e^{\epsilon} \le D_\epsilon(z;\theta), \quad \forall\theta, z \tag{26}$$

and $D_\epsilon(\cdot;\theta)$ is convex in z and nonincreasing in θ, $(\epsilon,\mathbb{S})$-pufferfish privacy is attained by adding noise N_θ with any θ that holds the inequalities

$$\int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x \le e^\epsilon p(x'|s_j), \quad \forall x' \tag{27a}$$

$$\int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x \le e^\epsilon p(x|s_i), \quad \forall x \tag{27b}$$

for all $(s_i,s_j)\in\mathbb{S}$.

Proof. Recall the inequality $P_{Y|S}(y|s_i,\rho)-e^\epsilon P_{Y|S}(y|s_j,\rho)\le\int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\,\mathrm{d}\pi(x,x'),\forall y$ obtained in (21). Upon the condition (26), we have

$$\begin{aligned} P_{Y|S}(y|s_i,\rho)-e^\epsilon P_{Y|S}(y|s_j,\rho) &\le \int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\,\mathrm{d}\pi(x,x') &(28)\\ &\le \int P_{N_\theta}(y-x')D_\epsilon(x-x';\theta)\,\mathrm{d}\pi(x,x') &(29)\\ &\le \int D_\epsilon(x-x';\theta)\,\mathrm{d}\pi(x,x'). &(30)\end{aligned}$$

The inequalities above holds for all $\pi\in\Gamma(s_i,s_j)$. That is, for any coupling $\pi\in\Gamma(s_i,s_j)$, if $\int D_\epsilon(x-x';\theta)\,\mathrm{d}\pi(x,x')\le 0$, then $\frac{P_{Y|S}(y|s_i)}{P_{Y|S}(y|s_j)}\le e^\epsilon$.

For D_ϵ nonincreasing in θ, we take the infinum of the integral in (30) over all couplings and request

$$\inf_{\pi\in\Gamma(s_i,s_j)}\int D_\epsilon(x-x';\theta)\,\mathrm{d}\pi(x,x')\le 0. \tag{31}$$

The purpose is to find the smallest value of θ that holds the sufficient condition. It is clear that for $D_\epsilon(z;\theta)$ being convex in z for all θ, the minimizer of (31) is the Kantorovich optimal transport plan π^*, i.e., (31) is equivalent to $\int D_\epsilon(x-x';\theta)\,\mathrm{d}\pi^*(x,x')\le 0$. Instead, we request the integral in (28) under π^* to be nonpositive,[11] i.e.,

$$\begin{aligned}\int P_{N_\theta}(y-x')\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\,\mathrm{d}\pi^*(x,x') &= \iint P_{N_\theta}(y-x)\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\pi^*(x,x')\,\mathrm{d}x\,\mathrm{d}x'\\ &= \int P_{N_\theta}(y-x')\int\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\pi^*(x,x')\,\mathrm{d}x\,\mathrm{d}x'\le 0.\end{aligned}$$

A sufficient condition to hold this inequality is to make $\int\big(e^{\eta(\theta)d(x-x')}-e^\epsilon\big)\pi^*(x,x')\,\mathrm{d}x\le 0$ for all x', which is equivalent to (27a).

Similarly, a sufficient condition for $\frac{P_{Y|S}(y|s_j)}{P_{Y|S}(y|s_i)}\le e^\epsilon$ is

$$\int e^{\eta(\theta)d(x'-x)}\pi^{*\intercal}(x',x)\,\mathrm{d}x \le e^\epsilon p(x|s_i), \quad \forall x, \tag{32}$$

where $\pi^{*\intercal}(x',x)=\pi^*(x,x'),\forall x,x'$. Due to the symmetry property of d, the condition (32) is equivalent to (27b). □

Note that the value of θ in Proposition 1 is not determined by the upper bound function $D_\epsilon(z;\theta)$ in (26). That is, we only require the existence of such a function $D_\epsilon(z;\theta)$ regardless of the tightness of this upper bound. We show below that one example of this upper bound function is the piecewise linear function $[\cdot]_+=\max\{\cdot,0\}$.

For the exponential mechanism, we have

$$\begin{aligned} e^{\eta(\theta)d(z)}-e^\epsilon &\le \frac{e^{\eta(\theta)d(z)}+e^\epsilon}{2}\big[\eta(\theta)d(z)-\epsilon\big]_+\\ &\le \frac{e^{M\eta(\theta)}+e^\epsilon}{2}\big[\eta(\theta)d(z)-\epsilon\big]_+,\end{aligned}$$

where $M=\max_{x,x'}d(x-x')$ assuming $d(\cdot)$ is a bounded measure.[12] Here, $[f(\cdot)]_+$ for convex f is convex and

[11] This is also for the purpose of searching the minimum value of θ that holds the sufficient condition.

[12] Here, we apply the inequality for exponential function: $e^{\frac{x+y}{2}}\le\frac{e^y-e^x}{y-x}\le\frac{e^x+e^y}{2}$ for all $x,y\in\mathbb{R}$.

$\big[\eta(\theta)d(z)-\epsilon\big]_+$ is nonincreasing in θ since $\eta \propto \frac{1}{\theta}$.

Consider (27a). For $\eta(\theta)$ nonincreasing in θ, $\inf\{\theta\colon \int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x \le e^{\epsilon}p(x'|s_j)\}$ equals to the value of θ that holds the equality

$$\int e^{\eta(\theta)d(x-x')}\pi^*(x,x')\,\mathrm{d}x = e^{\epsilon}p(x'|s_j)$$

for each x'. Taking also the equality of (27b), we have Theorem 2. □

H PROOF of THEOREM 3

For zero-mean Gaussian noise, $P_{N_\theta}(z)=\frac{1}{\sqrt{2\pi}\theta}e^{-\frac{z^2}{2\theta^2}}$ and

$$\begin{aligned}
P_{Y|S}(y|s_i,\rho)-e^{\epsilon}P_{Y|S}(y|s_j,\rho) &= \int P_{N_\theta}(y-x)P_{X|S}(x|s_i,\rho)\,\mathrm{d}x - e^{\epsilon}\int P_{N_\theta}(y-x')P_{X|S}(x'|s_j,\rho)\,\mathrm{d}x' \\
&= \int \big(P_{N_\theta}(y-x)-e^{\epsilon}P_{N_\theta}(y-x')\big)\,\mathrm{d}\pi(x,x') \\
&= \int \frac{1}{\sqrt{2\pi}\theta}\big(e^{-\frac{(y-x)^2}{2\theta^2}}-e^{\epsilon-\frac{(y-x')^2}{2\theta^2}}\big)\,\mathrm{d}\pi(x,x'), \qquad \forall y. \qquad (33)
\end{aligned}$$

This equality holds for the Kantorovich optimal transport plan π^*.[13] In this case, to have (33)≤ 0, we only need to request

$$\begin{aligned}
\frac{(y-x')^2-(y-x)^2}{2\theta^2} &= \frac{(y-x+x-x')^2-(y-x)^2}{2\theta^2} \\
&= \frac{2(x-x')(y-x)+(x-x')^2}{2\theta^2} \\
&\le \frac{2\triangle|y-x|+\triangle^2}{2\theta^2} \le \epsilon, \qquad (34)
\end{aligned}$$

where $\triangle=\sup_{(x,x')\in\mathrm{supp}(\pi^*)}|x-x'|$. We follow the same approach in Dwork et al. (2014, Appendix A) to prove (a). Let $\theta=\frac{\triangle}{\epsilon}c$, where $c\ge 0$. Then, $\frac{\theta}{\triangle}=\frac{c}{\epsilon}$. Rewriting inequality (34) to

$$\begin{aligned}
\frac{\triangle|y-x|}{\theta^2} \le \epsilon-\frac{\triangle^2}{2\theta^2} \quad &\Longrightarrow \quad \frac{|y-x|}{\theta} \le \epsilon\frac{\theta}{\triangle}-\frac{\triangle}{2\theta} \\
&\Longrightarrow \quad \frac{|y-x|}{\theta} \le c-\frac{\epsilon}{2c}.
\end{aligned}$$

Here, $\frac{Y-X}{\theta}$ follows standard normal distribution.

Recall that for standard normal distributed random variable Z, we have a lower bound on tail probability: $\Pr(Z>t)>\frac{1}{\sqrt{2\pi}t}e^{-\frac{t^2}{2}}$. For $t\ge 0$, we are seeking the condition on t that holds inequality $\Pr(|Z|>t)<\delta$, which (due to the symmetry of Gaussian distribution) can be enforced on the positive range:

$$\begin{aligned}
\Pr(Z>t)<\frac{1}{\sqrt{2\pi}t}e^{-\frac{t^2}{2}}<\frac{\delta}{2} \quad &\Longrightarrow \quad te^{\frac{t^2}{2}}>\sqrt{\frac{2}{\pi}}\frac{1}{\delta} \\
&\Longrightarrow \quad \log t+\frac{t^2}{2}>\log\sqrt{\frac{2}{\pi}}\frac{1}{\delta} \qquad (35)
\end{aligned}$$

So, for $Z=\frac{Y-X}{\theta}$, $t=c-\frac{\epsilon}{2c}$ with $c\ge\sqrt{\frac{\epsilon}{2}}$, we need to determine c such that

$$\underbrace{\log\big(c-\frac{\epsilon}{2c}\big)}_{A}+\frac{1}{2}\underbrace{\big(c^2-\epsilon+\frac{\epsilon^2}{4c}\big)}_{B}>\log\sqrt{\frac{2}{\pi}}\frac{1}{\delta} \qquad (36)$$

[13] Choosing π^* will necessarily reduce the sensitivity $\triangle$.

For $\epsilon \leq 1$, we set $c \geq \frac{3}{2}$ to have $A > 0$, for which, $B \geq c^2 - \frac{8}{9}$. We instead request $c^2 - \frac{8}{9} \geq 2\log\sqrt{\frac{2}{\pi}}\frac{1}{\delta}$ and have $c^2 \geq 2\log(1.25/\delta)$.

We use the inequality $\log t \leq t - 1, \forall t > 0$ to prove (b). Alternative to (35), request

$$\log t + \frac{t^2}{2} \geq \log t + \log\frac{t^2}{2} + 1 > \log\sqrt{\frac{2}{\pi}}\frac{1}{\delta} \quad \implies \quad \log\frac{t^3}{2}\sqrt{\frac{\pi}{2}}\delta > -1 \tag{37}$$

$$\implies \quad t > \left(\frac{2}{e}\right)^{\frac{1}{3}}\left(\frac{2}{\pi}\right)^{\frac{1}{6}}\delta^{-\frac{1}{3}}. \tag{38}$$

As $(2/e)^{\frac{1}{3}}(2/\pi)^{1/6} = 0.8373$, we need to have $t > 0.84\delta^{-\frac{1}{3}}$. For $Z = \frac{Y-X}{\theta}$ and $t = c - \frac{\epsilon}{2c}$ with $c \geq \sqrt{\frac{\epsilon}{2}}$,

$$c - \frac{\epsilon}{2c} > 0.84\delta^{-\frac{1}{3}} \quad \implies \quad c^2 - 0.84\delta^{-\frac{1}{3}}c - \frac{\epsilon}{2} > 0 \tag{39}$$

$$\implies \quad c > 0.41\delta^{-\frac{1}{3}} + \sqrt{(0.41\delta^{-\frac{1}{3}})^2 + \frac{\epsilon}{2}} \tag{40}$$

This proves (b). □