Optimal Distribution Mapping for Inference Privacy

Ruochi Zhang, Parv Venkitasubramaniam,
Electrical and Computer Engineering
Lehigh University, USA
{ruz614, parv.v}@lehigh.edu

***Abstract*—Information sanitization to protect an underlying label from being inferred through a data stream is investigated in this work. The problem is posed as an optimal mapping from an underlying distribution that reveals a class/label for the data to a target distribution with minimum distortion. The optimal sanitization operation are transformed to convex optimization problems corresponding to the domain of the source and target distributions. In particular, when one of the distributions is discrete, a parallel is drawn to a biased quantization method and an efficient sub-gradient method is proposed to derive the optimal transformation. The method is extended to a real time scenario when multiple source distributions are to be mapped to a fixed target distribution without prior knowledge of the label of the streaming data, in order to defeat any hypothesis test between the labels. It is shown that even when the source label is unknown to the sanitizer, optimal distortion is possible with perfect privacy.**

I. Introduction

Data collection systems are ubiquitous, and machine learning algorithms that derive useful classifiers for these large and diverse swathes of information have grown fast and sharp. These systems gather data often for legitimate purposes, and the learning algorithms developed usually arise out of commercial needs of the individual. For instance social networks provide individuals a virtual stage to reveal themselves to their own community and commercial information is targeted at these individuals through learned preferences. With the powerful computing infrastructures and enormous data collected [1], supervised learning increasingly affects human decisions in various domains, including spam classifiers of e-mail, face recognizers over images, and medical diagnosis systems for patients [2].

However, data, even if gathered legitimately, can lead to undesirable inferences compromising private sensitive information of an individual or an entity. For instance, web access patterns can reveal anything from political preferences to social biases [3], packet transmission timing can reveal websites accessed, driving patterns extracted from GPS data can reveal demographic information. Furthermore, with decreasing difficulty of deploying powerful machine learning tools, an adversary with intent can with ease derive sensitive information from individuals and organizations [4]. Supervised learning can also affect decisions in domains protected by anti-discrimination law [5], and its effect of existing biases is not well understood. The primary focus of this work is to develop a data *sanitization* mechanism that prevents the inference of sensitive information whilst minimally distorting the data and thus could be used for other legitimate processing or learning purposes. In other words, we develop methods for minimum distortion inference suppression.

Supervised learning [6] systems make predictions by collecting a large number of (X, θ) pairs, where X is a feature information of user and θ is a label with practical sense. After sufficient observations, the system understands the connection between X and θ, and is capable of making prediction $\theta^\star$ base on new arrival samples with only feature X. In this work, we investigate the problem from the reverse perspective —given enough statistical knowledge, what is the best method to remove the label information from X? Or, given a sample from one class, how can we minimally change X such that the sample looks like a sample from another class?

We call such techniques *information sanitization* where a sanitizer receives user-generated data and outputs label-resistant sanitized data. Not surprisingly, one can always produce independently generated samples from some target distribution and retains complete privacy, or refuse any output to the system and do not release any information. Such a naive data sanitization mechanism however limits any legitimate utility that can be derived from the data. When the utility is measurable as a function of the sanitization induced distortion from the source data, users can carefully craft sanitization schemes to minimize cost whilst rendering the data label-resistant. Since the pre-sanitize data is not revealed to the system, the cost can be reduced by utilizing the dependency — as long as the output data has desired distribution, complete privacy is still retained.

Mathematically, this work investigates the mapping scheme to transform a set of distributions to a single target distribution and thus lose the label that defines the source distribution. The goal is to find the optimal joint distribution —or, equivalently, conditional distribution — such that the distortion between pre and post sanitized data is minimized. Throughout this paper, we assume statistical knowledge about the underlying data sources in the form of probability distribution functions. Under such an approach, perfect privacy is guaranteed against any inference mechanism, and thus provides a performance benchmark for any sanitization scheme.

Related Works

There are several techniques to protect privacy against potential sensitive information leakage in a supervised learning system. One classical example would be random noise-addition methods as a patch to existing algorithms [7], [8], [9]. Regularization techniques, which aim to avoid overfitting to the examples used for training, may also hide details of those examples [10]. Another approach is designing a decentralized learning system which limits the power of any individual adversary [11], [12].

A similar problem to this work is the distribution matching or signal shaping problem, where we want to transform independently generated input bits into a sequence of output with a desired distribution [13][14]. Distribution matchers are used for rate adaption, or to achieve the capacity of the additive white Gaussian noise channel [15]. Although seems similar, the work we present here is a different problem. The distribution matching problem aims at generating a approximate target distribution with a limited code length, while the goal of data sanitization problem is to generate an identical target distribution with a limited distortion cost.

Privacy in streaming data, and in particular, within the context of machine learning algorithms has been studied using different measures and methodologies. Information theoretic approaches such as in [16], [17] use conditional entropy to measure privacy of an underlying source whilst transforming the measurements to guarantee privacy with minimum distortion. Differential privacy is a common approach in the context of supervised learning [8], with dominant applications in static databases, and limited applications to time series data which is the focus of this work. We note that, when considering perfect privacy, as in this work, the solutions would guarantee perfect privacy under an entropic framework (conditional entropy equal to unconditional entropy), and a differential privacy framework (epsilon equal to zero). In the context of inference based privacy, stealthy attacks on dynamical systems [18], [19], [20] are close to this work. In those works, an adversary maintains stealth by preventing inference about his presence whilst achieving a target objective (akin to distortion).

In the present work, we primarily look into the optimal sanitization problem under several sets of assumptions. In Section II, we discuss the general problem formulation. In Section III, we investigate the optimal single source sanitization problem, including continuous-to-continuous in $\mathbb{R}$, and continuous-to-discrete in $\mathbb{R}^n$. Although there are still other cases to be studied, we do believe that these assumptions are the most common random variable types one may encounter in practical. In Section IV, we demonstrate an sanitization method in real time systems which obtains perfect privacy and optimal performance asymptomatically.

II. Problem Formulation

Consider a sanitizer receiving data from m data sources $\{X_1, \cdots, X_m\}$. Each data source has a known probability distribution P_k. The goal of the sanitizer, is to reshape these distributions to the same distribution, that an adversary cannot infer which sources the data comes from. In the meanwhile, the sanitizer suffers a distortion measured by $\mathbb{E}f(X_k, Y_k)$.

The problem we investigate is, given a set of distributions $\{P_k\}_{k=1}^m$, how to derive conditional distributions $P(Y_k|X_k)$ such that the resulting distributions P_{Y_k} are identically distributed and the distortions $\mathbb{E}f(X_k, Y_k)$ are minimized.

When the source label k of an arriving stream of data is known to the sanitizer, the problem reduces to k independent optimization problems given that the target distribution P_Y is known. In other words, the key challenge that remains is deriving a conditional distribution that maps a source to target with minimum distortion. The derivation of the optimal mapping is one of the foci of this work, and in particular, in Section III, where the mapping mechanism depends on whether the source or target distributions are discrete or continuous valued. The optimal mapping, thus derived, can be applied to sanitize the data stream in real time assuming prior knowledge of the label (distribution) of the data.

When the source label is unknown, but the underlying set of distributions from where the data stream is derived is known to the sanitizer, the problem cannot be reduced to independent optimization problems apriori, and a real time dynamic strategy is required to sanitize the data with minimum distortion. This problem will be defined formally in Section IV where we show that even without prior knowledge of the distribution, perfect inference privacy is achievable in real time with minimum distortion identical to a sanitizer with perfect knowledge of the stream label.

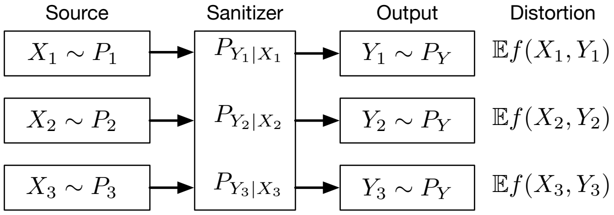


Fig. 1. The Sanitization Model

III. Optimal Single Source-to-Target Mapping

In this section, we consider the case when there is only one source distribution and the target distribution is known and fixed. Consider a user generating message $X \in \mathcal{X}$ from a source distribution P_1, which reveals the label of user. The user wants to output a message to a system, while hiding its label information, so that the adversary cannot learn the label from the message. Motivated by this, the user outputs a sanitized message $Y \in \mathcal{Y}$ with a target distribution P_2, which is considered to be label-resistant. In the meanwhile, the user suffers a distortion cost. The goal of this work is to find the minimal distortion sanitization scheme for the user, which is a random mapping $X \to Y$. Such a sanitizer can be modeled as a memoryless channel.

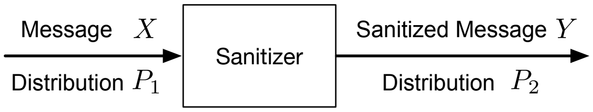


Fig. 2. Single Source-to-Target Sanitization

We formally state the problem as follows.

Problem 1. Let X be an $\mathcal{X}$-valued random variable and Y be a $\mathcal{Y}$-valued random variable, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be a cost function, we want to find the best joint distribution P_{XY} such that

$$\begin{aligned} \underset{P_{XY}}{\text{minimize}} \quad & \mathbb{E} f(X,Y) \\ \text{s.t.} \quad & P_X = P_1 \\ & P_Y = P_2, \end{aligned}$$

where P_X and P_Y be the marginal distributions of P_{XY} corresponding to X and Y respectively. We do note that this is equivalent to find a random mapping since one can compute conditional distribution with joint distribution. For example, for discrete random variables, $P(y|X = x) = P_{XY}(x,y)/P_1(x)$.

If X and Y were discrete random variables, the optimization problem as stated above would reduce to a straightforward linear programming problem which are easily solved using solution techniques such as criss-cross [21] or affine scaling [22]. When one or both of the variables are continuous valued, it is computationally infeasible to run a traditional optimization problem. In the subsequent subsections, we utilize the dual version of the above optimization and under certain conditions on the distortion metric $f(X,Y)$ we propose specialized algorithms catering to the specific subclasses of the problem.

The following theorem provides the dual of the proposed optimization problem, and reduces the feasible set of solutions using complementary slackness.

Theorem 1. *Let $u : \mathcal{X} \to \mathbb{R}$ and $v : \mathcal{Y} \to \mathbb{R}$ be two measurable functions. We define the dual problem be*

$$\begin{aligned} d = & \max_{u,v} \left(\int_{\mathcal{X}} u dP_1 + \int_{\mathcal{Y}} v dP_2 \right) \\ & \text{s.t.} \quad u(x) + v(y) \leq f(x,y) \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \end{aligned} \tag{1}$$

The strong duality holds, i.e. $p = d$ where p is the optimal value of problem 1.

*(**Complementary slackness**) Furthermore, let $P^\star_{XY}$ be the solution of primal problem and $u^\star, v^\star$ be the solution of dual problem. Then $P^\star_{XY}(A) = 0$, where*

$$A := \{(x,y) \in \mathcal{X} \times \mathcal{Y} : u^\star(x) + v^\star(y) < f(x,y)\}$$

Proof. See appendix A. □

In the following subsections, we provide solutions and algorithms for sanitizing continuous random variables to a target distribution when the distortion constraint is expressionless as a semi-norm.

A. Optimal Sanitizers for Continuous Random Variables

a) Continuous Target Distribution: When both the source and target distributions are continuous, the optimization problem in either primal or dual problem is generally hard to solve. Under certain conditions on the distortion function, however, we show that the optimal transformation is a simple CDF mapping typically used for random variable generation. In particular, let X and Y be real-valued random variables and let the distortion function $f(x,y) = h(x-y)$, where the function $h : \mathbb{R} \to \mathbb{R}^+$ satisfies the following conditions:

- $h(x) = 0$ if and only if $x = 0$.
- $h(x)$ is symmetric, and strictly increasing on $[0, \infty)$
- Quadrilateral inequality: For any $x_1 < x_2$ and $y_1 < y_2$,

  $$h(x_1 - y_1) + h(x_2 - y_2) < h(x_1 - y_2) + h(x_2 - y_1)$$

  A simple example that satisfies this is $h(x) = |x|^\alpha$ with $\alpha > 1$. In particular, $\alpha = 2$ leads to a minimal mean square error (MMSE) sanitization.

The optimal solution of the problem is always a deterministic mapping that maps any x_0 to a y_0 with the same "percentile". We formally states the result in the following Theorem, and illustrate the meaning of percentile mapping in Figure 3.

Theorem 2. *The optimal solution of this problem is the CDF mapping. That is, the optimal $P_{Y|X}$ is the deterministic mapping $X \to g(X)$ where $g(x) = F_Y^{-1}(F_X(x))$ and F_Y^{-1} is the generalized inverse distribution function:*
$F_Y^{-1}(y) = \inf \{z \in \mathbb{R} : F_Y(z) \geq y\}$

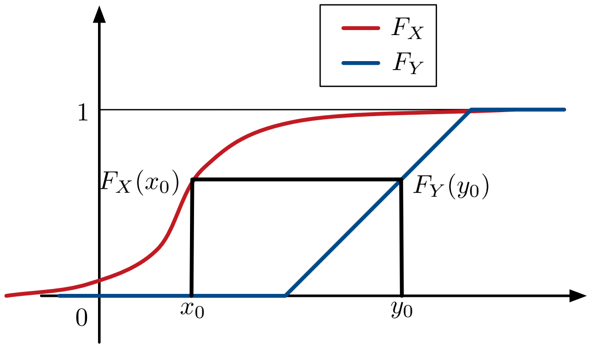


Fig. 3. The CDF mapping

Proof. Our first observation is that, for any $x_0 \in \mathbb{R}$ such that $p_1(x_0) > 0$, there always exists a y_0 such that (x_0, y_0) is on the boundary of the constraint, i.e., $u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)$. Such condition will "block" any mapping possibility between x_1 and y_1 if $(x_1 - x_0)(y_1 - y_0) < 0$, i.e., we cannot have crossing mappings, as is shown in Figure 4. Base on this result, we can prove that the "percentile" of x_0 in distribution P_1 is the same as the "percentile" of y_0 in distribution P_2. A detailed proof is available in Appendix B. □

In the following, we provide the optimal MMSE sanitizers and their performance for a few example distributions. Note that the transformation as described in Theorem 2 is independent of the actual distortion metric, as long as the metric is expressible as a convex function of norm. Furthermore, if the CDF of the source and target distribution are expressible in functional form, it is easy to obtain closed form sanitizers for the data.

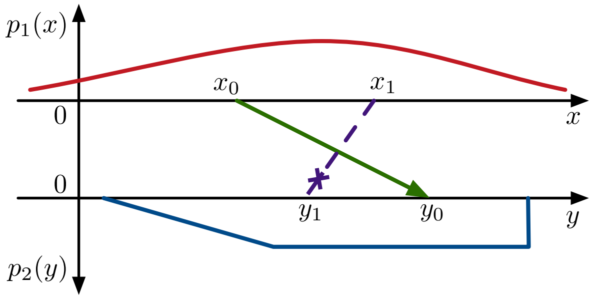


Fig. 4. $u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)$ blocks any crossing mappings.

Example (Gaussian)**.** Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. The MMSE sanitizer is given by deterministic mapping $Y = (X - \mu_1)\frac{\sigma_2}{\sigma_1} + \mu_2$, with the minimal square error $E(X-Y)^2 = (\mu_2 - \mu_1)^2 + (\sigma_2 - \sigma_1)^2$. In comparison, a naive independently generated Y will induce error $\mathbb{E}(X - Y_{naive})^2 = (\mu_2 - \mu_1)^2 + \sigma_1^2 + \sigma_2^2$

Example (Step to Exponential)**.** Let X be distributed according to a piecewise constant function with $p(x) = \frac{6}{\pi^2 n^2}, (n-1 < x < n)$. Let Y be exponential distribution with PDF $p_2(y) = \lambda e^{-\lambda y}, (y > 0)$. The MMSE sanitizer is given by deterministic mapping $Y = -\lambda^{-1}\log(1-U)$, where

$$U = \frac{6}{\pi^2}\left(\sum_{n=1}^{[x]} \frac{1}{n^2} + \frac{x - [x]}{[x+1]^2}\right)$$

b) Discrete Sanitizers for Continuous Sources: The continuous to discrete sanitization is a quantization process. If there were freedom to choose the best target distribution, then the optimal sanitization is a classical quantization or rate distortion problem which is well studied in literature (Lloyd's Algorithm)[23], which will generate a Voronoi tessellation as a partition of space $\mathbb{R}^n$. When a target distribution is specified, we will show that the result is similar to a "biased" quantization problem, with a sub-gradient method to find the optimal bias vector. The problem states as follows:

Problem 2. Let X be an $\mathbb{R}^n$-valued continuous random variable with distribution measure P_1, and Y be a $\mathcal{Y}$-valued discrete random variable with distribution measure P_2. We assume that

- P_1 is absolutely continuous w.r.t. Lebesgue measure, with $q(x)$ be the PDF.
- P_2 is distributed in finite points $\mathcal{Y} = \{y_k\}_{k=1}^d$, with $P_2(y_k) = p_k$.

Let $f : \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}^+$ be a cost function, we want to find the best joint distribution $P(x, y)$ such that minimize $\mathbb{E}f(X, Y)$.

This problem can be solved by vectorizing the dual problem given by Theorem 1. Since P_2 is distributed on a finite set, the function $v : \mathcal{Y} \to \mathbb{R}$ can be treated as a d-dimensional vector $\mathbf{v} = \{v_1, \cdots, v_d\}$ where $v_k = v(y_k)$. Then, we apply a diminishing step size subgradient method to solve the problem numerically. A detailed algorithm exploiting this idea is provided in Algorithm 1, and the following theorem proves that the algorithm converges to the optimal solution of the dual problem.

Theorem 3. *Algorithm 1 converges to the optimal solution to Problem 2. i.e.* $g(\mathbf{v}_{best}) \to \min_{P_{XY}} \mathbb{E}f(X,Y)$ *as* $N \to \infty$*. Here* $\mathbf{v}_{best}$ *is the best vector* $\mathbf{v}$ *within the first* N *iterations, and* $\mathbf{v}_{best} \to \mathbf{v}^\star = (v_1^\star, \cdots, v_d^\star)$ *as* $N \to \infty$*. Furthermore, the optimal mapping is always a deterministic mapping that maps* x *to* y_k*, where*

$$k = \arg\min_{1 \le j \le d}\,[f(x, y_k) - v_k^\star]$$

If there is more than one such indices, one can choose any one from them.

Algorithm 1 Sub-gradient method for C→D sanitization

```
Define:
    $g(\mathbf{v}) = \mathbf{v}^T\mathbf{p} + \int_{\mathbb{R}^n} \min_{1\le k\le d}\,[f(x,y_k) - v_k]\,q(x)dx$
    $\mathbf{v} = (v_1, \cdots, v_d)^T$
    $\mathbf{p} = (p_1, \cdots, p_d)^T$
Initialize: $\mathbf{v} = \mathbf{0}$, $g_{best} = g(\mathbf{0})$, $\mathbf{v}_{best} = \mathbf{v}(\mathbf{0})$, pick maximum step $N$, pick step size $\{\alpha_k\}_{k=1}^\infty$ s.t., (One example is $\alpha_k = 1/k$)
    $\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^\infty \alpha_k = +\infty$
for $k = 1$ to $N$ do
    Partition $\mathbb{R}^n$ by
        $\tilde{S}_k = \left\{ s \in S : f(x,y_k) - v_k = \min_{j=1,\cdots,d} (f(x,y_j) - v_j) \right\}$
        $S_1 = \tilde{S}_1$
        $S_k = \tilde{S}_k \setminus \cup_{j=1}^{k-1} \tilde{S}_j \quad \forall k = 2, \cdots, d$
    Compute the supergradient $\mathbf{w} = (w_1, \cdots, w_d)$, where
        $w_k = p_k - \int_{S_k} dP_1 = p_k - \int_{S_k} q(x)dx$
    $\mathbf{v} \leftarrow \mathbf{v} - \alpha_k \mathbf{w}$
    if $g(\mathbf{v}) > g_{best}$ then
        $\mathbf{v}_{best} \leftarrow \mathbf{v}, \quad g_{best} \leftarrow g(\mathbf{v})$
    end if
end for
Output: $\mathbf{v}_{best}, \quad g_{best}$
```

Proof. We show that when the target distribution is discrete, then the dual problem can be stated as:

$$d = \max_{\mathbf{v}\in\mathbb{R}^d} g(\mathbf{v})$$

It is easy to see that $g(\mathbf{v})$ is concave with respect to $\mathbf{v}$. Indeed, $\min_{1\le k\le d}[f(x,y_k) - v_k]$ is concave w.r.t. $\mathbf{v}$ since it is the minimal over finite linear functions. The algorithm we proposed here is a diminishing step size subgradient algorithm, the proof of convergence can be found in Shor's book [24]. The optimal mapping rule is a direct corollary of Theorem 1. The detailed proof is in Appendix C. □

We are especially interested in the partition method introduced in Algorithm 1, since one may use the partition scheme to find the optimal deterministic mapping rule. This partition can be seen as an generalization of Voronoi tessellation. Here we use the following example as an illustration.

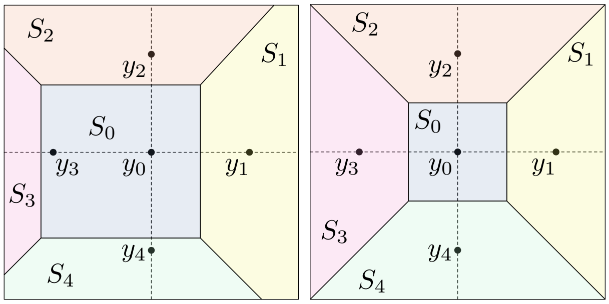
Example. Let X and Y be random variable in $\mathbb{R}^2$, with

- $X \sim \mathcal{N}(0, I)$ be normalized Gaussian distribution
- Y is a discrete random variable $P(Y = y_k) = p_k$:

$$y_0 = (0,0), y_1 = (1,0), y_2 = (0,1), y_3 = (-1,0), y_4 = (0,-1)$$
$$p_0 = 0.32, p_1 = 0.25, p_2 = 0.18, p_3 = 0.1, p_4 = 0.15$$

- The cost function is $f(x, y) = ||x - y||_2^2$ (MMSE).

We run the subgradient algorithm and find the optimal $\mathbf{v}^\star = (1.17, 1.16, 0.76, 0, 0.35)$. Then we partition $\mathbb{R}^2$ and generate the following tessellation in Figure 4(a). Figure 4 illustrate "biased Voronoi tessellation" generated by the partition method. A regular Voronoi tessellation, as is shown in Figure 4(b), is a partitioning of a plane into regions based on distance to points in a specific subset. However, a sanitizer using such a Voronoi tessellation as mapping scheme will not generate a sanitized signal Y with desired probability distribution. To address the statistical requirement, we need to use a generalized Voronoi tessellation partition with the bias vector $\mathbf{v}^\star$. As a result, the space is no longer partitioned by a sequence of line segment bisectors. Interestingly, unlike classical quantizers, note that the region S_k may not contain y_k, see S_3 in Figure 4(a).



(a) Biased Voronoi Tessellation (b) Regular Voronoi Tessellation

Fig. 5. Comparison of Biased and Regular Voronoi Tessellations

Remark. Since the sanitizer problem is symmetric, the optimal discrete-to-continuous sanitizer can be obtained by looking at the optimal continuous-to-discrete sanitizer reversely. Specifically, if X is continuous, Y is discrete, and we want to map Y to X, the optimal sanitizer is given by

$$P_{X|Y}(x|Y = y_k) = \frac{q(x)}{\int_{S_k} q(x)dx}$$

where $q(x)$ is the PDF of X.

IV. Real Time Mapping

In this section, we consider a real time data processing system where the sanitizer receives data from one of two source classes. This formulation is useful when the sanitizer needs to process streaming data without perfect knowledge of the source label apriori. Here, the user generates i.i.d. data $\{X_t\}_{t=1}^N$ from a fixed class $j \in \{0, 1\}$. The probability distribution of each class $P_\theta, \theta \in \{0, 1\}$ is known, however the correct class label j is unknown to the sanitizer. The sanitizer needs to generate a mapping rule $P_{Y|X}^t$ at every time slot t. We do note that the mapping rule can be generated by the previous data observations.

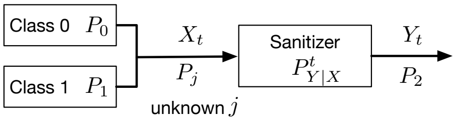


Fig. 6. Real Time Sanitization Model

Let N be the time horizon. The goal of the sanitizer is to map $X_t \to Y_t$ such that Y_t has distribution P_2, and minimize the average cost

$$c_N = \frac{1}{N} \sum_{t=1}^{N} \mathbb{E}_{P_{Y|X}^t} f(X_t, Y_t)$$

while keeps a perfect-privacy, which is defined as follows.

Definition 1 (perfect-privacy)**.** If a time-series mapping policy $\left\{P_{Y_t|X_t}\right\}_{t=1}^N$ leaks label information j with a probability of 0, we call the policy $\left\{P_{Y_t|X_t}\right\}_{t=1}^N$ be perfect-private, wherein leaking is defined to be the event $A := \left\{\text{there exists } 1 \le t \le N \text{ s.t. } P_{Y_t|\theta=j} \neq P_2\right\}$.

If there were a sanitizer with the knowledge of j apriori, it could use the static single-source-to-target method in Section III and find a optimal mapping $P_{Y|X}^{\star,j}$, which would yield an average error

$$c^\star = \mathbb{E}_{P_{Y|X}^{\star,j}} f(X, Y), \text{ where } X \sim P_j, Y \sim P_2$$

This can be considered as a optimal sanitizer, yet it require the knowledge of j. Here we prove that perfect-private and optimal performance $c^\star$ can be reached asymptotically without such an known apriori, as is shown in the following theorem.

Theorem 4. *If the distribution of the data stream is unknown save for a set of 2 distributions $\{P_0, P_1\}$, a perfectly private mapping is still possible as $N \to \infty$ such that the average distortion converges to the optimal distortion where the source label known apriori. i.e.* $\lim_{N\to\infty} \frac{1}{N}\mathbb{E}f(X, Y) = c^\star$

Proof. We propose a mapping rule generating method that asymptotically reaches perfect-privacy with an average cost asymptotically converges to $c^\star$, which is defined as follows,

- For the first fixed horizon of size $s = [\log N]$, the sanitizer output independently generated Y_t. That is $P_{Y_t|X} = P_2$ for all $t \leq s$. Note that in this stage the sanitizer is not possible to leak information.
- In the meanwhile, run a likelihood ratio test to determine the estimated label $\hat{j}$, specifically

$$\sum_{t=1}^{s} \log \frac{P_1(X_t)}{P_0(X_t)} \underset{H_0}{\overset{H_1}{\gtrless}} 0$$

- For $s < t \leq N$, the sanitizer output the optimal static single-source-to-target sanitization scheme by assuming the data comes from source $\hat{j}$.

Let $P_e = P(\hat{j} \neq j)$ be the probability of error in the hypothesis testing stage. The proposed likelihood ratio test has a exponentially decaying error rate [25], that is

$$\lim_{n\to\infty} -\frac{1}{n} \log P_e = C(P_0||P_1)$$

where $C(\cdot||\cdot)$ is the Chernoff distance between the two distributions $C(P_0||P_1) = -\min_{0\leq u\leq 1} \log \int_{\mathcal{X}} P_0^u(x) P_1^{1-u}(x)dx$. Therefore, for sufficiently large N, $P_e \simeq A/Ne^{C(P_0||P_1)}$. Moreover, the average cost

$$\begin{aligned} c_N = \frac{1}{N}\Bigg[& \sum_{t=1}^{s} \mathbb{E}_{P_Y} f(X_t, Y_t) + P_e \sum_{t=s+1}^{N} \mathbb{E}_{P^{\star,1-j}_{Y|X}} f(X_t, Y_t) \\ & + (1-P_e) \sum_{t=s+1}^{N} \mathbb{E}_{P^{\star,j}_{Y|X}} f(X_t, Y_t) \Bigg] \\ = & \frac{sc_{ind}}{N} + \frac{N-s}{N}\left[P_e c_e + (1-P_e)c^\star\right] \end{aligned}$$

where $c_e = \mathbb{E}_{P^{\star,1-j}_{Y|X}} f(X_t, Y_t)$ is the single step average cost if the estimated label $\hat{j}$ were incorrect, and c_{ind} is the single step average cost if we generate Y independently. Now if we let $N \to \infty$, we have $s/N \to 0$ and $P_e \to 0$, thus we have $\lim_{N\to\infty} c_N = c^\star$, i.e. the proposed method is asymptotically optimal. In the meanwhile, the probability of information leakage equals P_e, therefore the proposed method is asymptotically perfect-private. □

That being said, in asymptotically sense, one can always obtains a sanitizer near-optimal performance and near-perfect privacy by attaching a single-source sanitizer to a likelihood ratio test.

V. Conclusion

In this work, we consider the problem where a user from a specific class wants to hide his/her class label completely from any potential adversarial supervised learning system. We investigate the optimal sanitization that takes data from a source class so that the sanitized data has identical probability distribution with data from a target class. The optimality is evaluated by the cost induced by sanitization distortion. While the discrete-to-discrete transformation is easily solved by linear programming, for continuous space problems, a primal-dual methodology and complementary slackness theorem is crucial for verifying claims about optimal solutions. In particular, the continuous-to-discrete problem can be solved with sub-gradient method and biased Voronoi tessellation akin to quantizer design albeit with a target value distribution. The work primarily focuses on i.i.d. distributions and known target distributions. The natural expansions to the scope would be relaxing those assumptions.

Appendix A
Proof of Theorem 1

(**Strong Duality**) Let $\mathcal{M}$ be the collection of all σ-finite measures on the set $\mathcal{X} \times \mathcal{Y}$. We rewrite the primal problem as

$$\begin{aligned} p = & \min_{Q\in\mathcal{M}} \int_{\mathcal{X}\times\mathcal{Y}} f dQ \\ \text{s.t.} \quad & Q_1 = P_1 \\ & Q_2 = P_2 \end{aligned}$$

where Q_1 and Q_2 be the "marginal measures" generated by Q. Specifically, we define Q_1 and Q_2 be the measures satisfy $Q_1(A) = Q(A \times \mathcal{Y})$ and $Q_2(B) = Q(\mathcal{X} \times B)$ for any measurable set $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. It is noteworthy that any feasible Q is a probability measure. Indeed, $Q(\mathcal{X} \times \mathcal{Y}) = Q_1(\mathcal{X}) = P_1(\mathcal{X}) = 1$.

Let $u : \mathcal{X} \to \mathbb{R}$ and $v : \mathcal{Y} \to \mathbb{R}$ be two measurable functions. We define the Lagrangian L be a functional $L : \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{Y}} \times \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ defined as

$$L(u, v, Q) = \int_{\mathcal{X}\times\mathcal{Y}} f dQ + \int_{\mathcal{X}} u(dP_1 - dQ_1) + \int_{\mathcal{Y}} v(dP_2 - dQ_2) \tag{2}$$

It is easy to see that

$$\sup_{u,v} L(u, v, Q) = \begin{cases} \int_{\mathcal{X}\times\mathcal{Y}} f dQ & \text{if } Q \text{ is feasible} \\ +\infty & \text{otherwise} \end{cases}$$

That is, $p = \inf_{Q\in\mathcal{M}} \sup_{u,v} L(u, v, Q)$. Note that $L(u, v, Q)$ is linear with respect to u, v and Q. By the Sion's minimax theorem [26], $p = d := \sup_{u,v} \min_{Q\in\mathcal{M}} L(u, v, Q)$. Moreover,

$$\begin{aligned} \inf_{Q\in\mathcal{M}} L(u, v, Q) = & \int_{\mathcal{X}} u dP_1 + \int_{\mathcal{Y}} v dP_2 \\ & + \inf_{Q\in\mathcal{M}} \left(\int_{\mathcal{X}\times\mathcal{Y}} f dQ - \int_{\mathcal{X}} u dQ_1 - \int_{\mathcal{Y}} v dQ_2 \right) \\ = & \int_{\mathcal{X}} u dP_1 + \int_{\mathcal{Y}} v dP_2 \\ & + \inf_{Q\in\mathcal{M}} \int_{\mathcal{X}\times\mathcal{Y}} \left[f(x, y) - u(x) - v(y)\right] dQ \end{aligned}$$

If $u(x) + v(y) \leq f(x, y) \;\; \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \;\; (*)$, the Q that minimize $L(u, v, Q)$ will be the zero measure $Q(A) = 0, \forall A \subset S \times T$. Otherwise, it is easy to see that $\inf_{Q\in\mathcal{M}} L(u, v, Q)$ is not lower-bounded. Therefore,

$$\inf_{Q\in\mathcal{M}} L(u, v, Q) = \begin{cases} \int_{\mathcal{X}} u dP_1 + \int_{\mathcal{Y}} v dP_2 & \text{if } u, v \text{ satisfy } (*) \\ -\infty & \text{otherwise} \end{cases}$$

By the definition of dual problem we finished the proof.

(**Complementary Slackness**) We first show that

$$\inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} [f(x,y) - u^\star(x) - v^\star(y)] = 0$$

If not, let $a = \inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} [f(x,y) - u^\star(x) - v^\star(y)] > 0$ and define function $u' : \mathcal{X} \to \mathbb{R}$ as

$$u' = u^\star + a$$

It is easy to see that $\int_{\mathcal{X}} u' dP_1 = \int_{\mathcal{X}} u^\star dP_1 + a$ and the function pair $(u', v^\star)$ is feasible, which contradicts with the optimality of $(u^\star, v^\star)$.

By Eq.(2) and minimax Theorem, we have

$$L(u^\star, v^\star, P^\star_{XY}) = \inf_{Q\in\mathcal{M}} L(u^\star, v^\star, Q)$$

Therefore,

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{Y}} [f(x,y) - u^\star(x) - v^\star(y)]\, dP^\star_{XY}\\ =& \inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} [f(x,y) - u^\star(x) - v^\star(y)] = 0\end{aligned}$$

Since $(u^\star, v^\star)$ is feasible, $f(x,y) - u^\star(x) - v^\star(y) \geq 0$. Thus

$$\begin{aligned}&\int_{A} [f(x,y) - u^\star(x) - v^\star(y)]\, dP^\star_{XY}\\ \leq& \int_{\mathcal{X}\times\mathcal{Y}} [f(x,y) - u^\star(x) - v^\star(y)]\, dP^\star_{XY} = 0\end{aligned}$$

Combine with the fact that $f(x,y) - u^\star(x) - v^\star(y) > 0$ on the set A, the proof is complete. □

Appendix B
Proof of Theorem 2

Let $p_1(x)$ and $p_2(y)$ be the PDF of X and Y respectively. By Theorem 4, the dual problem is given by

$$\begin{aligned}d = \max_{u,v} &\left(\int_{-\infty}^{+\infty} u(x)p_1(x)dx + \int_{-\infty}^{+\infty} v(y)p_2(y)dy\right)\\ \text{s.t.}\quad & u(x) + v(y) \leq h(x-y) \quad \forall (x,y) \in \mathbb{R}^2\end{aligned}$$

Although both the primal problem and the dual problem seems not easy to solve, the strong duality and complementary slackness still hold. One thing we can use complementary slackness for is to verify claims about optimal solutions. Our first observation is that, for any $x_0 \in \mathbb{R}$ such that $p_1(x_0) > 0$, there always exists a y_0 such that (x_0, y_0) is on the boundary of the constraint, i.e., $u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)$.

Indeed, since $p_1(x_0) > 0$, the conditional distribution $P_{Y|X}(y|X = x_0)$ is well-defined. By Theorem 1, the support of $P_{Y|X}(y|X = x_0)$ is a subset of $\{y_0 \in \mathbb{R} : u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)\}$, which yields the result.

Furthermore, for real-valued continuous random variables, the fact (x_0, y_0) is on the boundary of the constraint will "block" any mapping possibility between x_1 and y_1 if $(x_1 - x_0)(y_1 - y_0) < 0$, i.e., we cannot have crossing mappings, as is shown in Figure 4. Base on this result, we can claim that the "percentile" of x_0 in distribution P_1 is the same as the "percentile" of y_0 in distribution P_2.

Lemma 1. *Let $(u^\star, v^\star)$ be the optimal solution of the dual problem and $u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)$. Let $F_X(\cdot)$ and $F_Y(\cdot)$ be the cumulative distribution of X and Y. Then $F_X(x_0) = F_Y(y_0)$.*

Proof: For any $x > x_0$ and $y < y_0$, we claim that $u^\star(x) + v^\star(y) < h(x - y)$. Indeed

$$u(x_0) + v(y) \leq h(x_0 - y) \quad (3)$$

$$u(x) + v(y_0) \leq h(x - y_0) \quad (4)$$

$$u(x_0) + v(y_0) = h(x_0 - y_0) \quad (5)$$

By (3)+(4)-(5) we have

$$\begin{aligned}u(x) + v(y) &\leq h(x_0 - y) + h(x - y_0) - h(x_0 - y_0)\\ &< h(x - y)\end{aligned}$$

Similarly, for any $x < x_0$ and $y > y_0$, we have the same result. By the Theorem 1, we have

$$P^\star_{XY}(X < x_0, Y > y_0) = 0,\; P^\star_{XY}(X > x_0, Y < y_0) = 0$$

Which yields

$$\begin{aligned}F_X(x_0) &= P_1(X < x_0)\\ &= P^\star_{XY}(X < x_0, Y < y_0) + P^\star_{XY}(X < x_0, Y > y_0)\\ &= P^\star_{XY}(X < x_0, Y < y_0)\end{aligned}$$

Similarly we have

$$F_Y(y_0) = P^\star_{XY}(X < x_0, Y < y_0) = F_X(x_0)$$

■

Lemma 1 reveals the connection between the boundary condition on (x_0, y_0) and their percentile in distributions P_1 and P_2. Now, for any $x_0 \in \mathbb{R}$ with $P_1(x_0) > 0$, consider the set

$$\mathcal{T}(x_0) = \{y_0 \in \mathbb{R} : u^\star(x_0) + v^\star(y_0) = h(x_0 - y_0)\}$$

By Lemma 1, $\mathcal{T}(x_0) \subset \{y_0 \in \mathbb{R} : F_Y(y_0) = F_X(x_0)\}$. Since X and Y are real-valued continuous random variables, the CDF functions $F_X(\cdot)$ and $F_Y(\cdot)$ are continuous and non-decreasing. Thus, the set $\{y_0 \in \mathbb{R} : F_Y(y_0) = F_X(x_0)\}$ is either a single point $\{g(x_0)\}$ or an interval starting at $g(x_0)$. By the Complementary slackness Theorem, the optimal sanitizer should map x_0 to a subset of $\mathcal{T}(x_0)$. Moreover, if $\{y_0 \in \mathbb{R} : F_Y(y_0) = F_X(x_0)\}$ is an interval, the probability $P_2\{y_0 \in \mathbb{R} : F_Y(y_0) = F_X(x_0)\} = 0$. Therefore, we can always map x_0 to $g(x_0)$ without consider the rest of the interval. □

Appendix C
Proof of Theorem 3

Firstly, we show that the given problem is equivalent to the dual problem given by Theorem 1. Let $\mathcal{C}_D$ be the collection of all function pair $u : \mathbb{R}^n \to \mathbb{R}$ and $v : \mathcal{Y} \to \mathbb{R}$ such that satisfy the constrain of dual problem. That is,

$\mathcal{C}_D = \{(u,v) : u(x) + v(y) \le f(x,y) \quad \forall (x,y) \in \mathbb{R}^n \times \mathcal{Y}\}$. Then, for any $(u,v) \in \mathcal{C}_D$, consider the following function pair $(u^\star, v^\star)$,

$$u^\star(x) = \min_{1 \le k \le d} [f(x,y_k) - v(y_k)]$$

$$v^\star(y) = v(y)$$

We claim that $u^\star(x) \ge u(x)$. Indeed,

$$\begin{aligned} u^\star(x) - u(x) \ge 0 &= \min_{1 \le k \le d} \left[f(x,y_k)^2 - v(y_k) \right] - u(x) \\ &= \min_{1 \le k \le d} \left[f(x,y_k)^2 - u(x) - v(y_k) \right] \ge 0 \end{aligned}$$

Moreover, it is easy to see $(u^\star, v^\star) \in \mathcal{C}_D$ by the definition. Now, for any $(u,v) \in \mathcal{C}_D$, we pick

$$\mathbf{v} = (v_1, \cdots, v_d)^T = (v(y_1), \cdots, v(y_d))^T$$

We have

$$\begin{aligned} & g(\mathbf{v}) - \left[\int_{\mathbb{R}^n} u dP_1 + \int_{\mathcal{Y}} v dP_2 \right] \\ &= \left[\sum_{k=1}^d v_k p_k + \int_{\mathbb{R}^n} u^\star(x) dP_1 \right] - \left[\sum_{k=1}^d v_k p_k + \int_{\mathbb{R}^n} u(x) dP_1 \right] \\ &= \int_{\mathbb{R}^n} [u^\star(x) - u(x)] \, dP_1 \ge 0 \end{aligned}$$

Therefore, $\max_{\mathbf{v} \in \mathbb{R}^d} g(\mathbf{u}) \ge \max_{(u,v) \in \mathcal{C}_D} \left[\int_{\mathbb{R}^n} u dP_1 + \int_{\mathcal{Y}} v dP_2 \right]$. Conversely, we can prove that $\max_{\mathbf{v} \in \mathbb{R}^d} g(\mathbf{u}) \le \max_{(u,v) \in \mathcal{C}_D} \left[\int_{\mathbb{R}^n} u dP_1 + \int_{\mathcal{Y}} v dP_2 \right]$ by picking $u(x) = \min_{1 \le k \le d} [f(x,y_k) - v_k]$ and $v(y_k) = v_k$ for any $\mathbf{v} \in \mathbb{R}^d$. Combine these inequalities we conclude that the given problems are equivalent.

Then we show that the vector $\mathbf{w} = (w_1, \cdots, w_d) \in \mathbb{R}^d$ is a super-gradient of $g(\cdot)$ at $\mathbf{v}$, That is, $g(\mathbf{v}') - g(\mathbf{v}) \le \mathbf{w}^T(\mathbf{v}' - \mathbf{v})$ for all $\mathbf{v} \in \mathbb{R}^d$. Let $\{S_k\}_{k=1}^d$ be the partition with respect to $\mathbf{v}$ and $\{S'_k\}_{k=1}^d$ be the partition with respect to $\mathbf{u}'$. Then

$$g(\mathbf{v}) = \sum_{k=1}^d v_k \left[p_k - \int_{S_k} dP_1 \right] + \sum_{k=1}^d \int_{S_k} f(x,y_k) dP_1$$

$$g(\mathbf{v}') = \sum_{k=1}^d v'_k \left[p_k - \int_{S'_k} dP_1 \right] + \sum_{k=1}^d \int_{S'_k} f(x,y_k) dP_1$$

$$\begin{aligned} & g(\mathbf{v}') - g(\mathbf{v}) = \sum_{k=1}^d (v'_k - v_k) \left[p_k - \int_{S_k} dP_1 \right] \\ & + \sum_{k=1}^d \int_{S'_k} [f(x,y_k) - v'_k] \, dP_1 - \sum_{k=1}^d \int_{S_k} [f(x,y_k) - v'_k] \, dP_1 \end{aligned}$$

Note that $\sum_{k=1}^d (v'_k - v_k) \left[p_k - \int_{S_k} dP_1 \right] = \mathbf{w}^T(\mathbf{v}' - \mathbf{v})$. Therefore,

$$\begin{aligned} & g(\mathbf{v}') - g(\mathbf{v}) - \mathbf{w}^T(\mathbf{v}' - \mathbf{v}) \\ &= \int_S \min_{j=1,\cdots,d} \left(f(x,y_j) - v'_j \right) dP_1 - \sum_{k=1}^d \int_{S_k} (f(x,y_k) - v'_k) dP_1 \\ &= \sum_{k=1}^d \int_{S_k} \left[\min_{j=1,\cdots,d} \left(f(x,y_j) - v'_j \right) - \left(f(x,y_k) - v'_k \right) \right] dP_1 \le 0 \end{aligned}$$

Now, we move to the convergence of the sub-gradient method. It is easy to see that $g(\mathbf{v})$ is concave with respect to $\mathbf{v}$. Indeed, $\min_{1 \le k \le d} [f(x,y_k) - v_k]$ is concave w.r.t. $\mathbf{v}$ since it is the minimal over finite linear functions. The algorithm we proposed here is a diminishing step size subgradient algorithm, which is well studied, and guaranteed to converge to the optimal value. The proof of convergence of diminishing step size subgradient algorithm can be found in Shor's book [24].

Finally we prove the optimality of proposed mapping. By $u^\star(x) = \min_{1 \le k \le d} [f(x,y_k) - v_k^\star]$, the value of $u^\star(x) + v_k^\star$ will reach the dual constraint boundary $f(x,y_k)$ if and only if x is in the region w.r.t. v_k, otherwise we have $u^\star(x) + v_k^\star < f(x,y_k)$ and therefore $P^\star_{XY}(x,y_k) = 0$ by Theorem 1. □

Acknowledgment

The authors would like to gratefully acknowledge the support of the National Science Foundation through the grants CCF-1149495 and CCF-1617889.

References

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[3] E. B. Weiser, "Gender differences in internet use patterns and internet application preferences: A two-sample comparison," *Cyberpsychology and behavior*, vol. 3, no. 2, pp. 167–178, 2000.

[4] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, 2016.

[5] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[6] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[7] B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *arXiv preprint arXiv:0911.5708*, 2009.

[8] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[9] M. Z. Islam and L. Brankovic, "Noise addition for protecting privacy in data mining," in *Proceedings of The 6th Engineering Mathematics and Applications Conference (EMAC2003), Sydney*, 2003, pp. 85–90.

[10] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Advances in Neural Information Processing Systems*, 2009, pp. 289–296.

[11] F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2483–2493, 2013.

[12] O. Hasan, L. Brunie, E. Bertino, and N. Shang, “A decentralized privacy preserving reputation protocol for the malicious adversarial model,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 949–962, 2013.

[13] P. Schulte and G. Böcherer, “Constant composition distribution matching,” *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 430–434, 2016.

[14] R. A. Amjad and G. Böcherer, “Fixed-to-variable length distribution matching,” in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1511–1515.

[15] G. Böcherer, F. Steiner, and P. Schulte, “Bandwidth efficient and rate-matched low-density parity-check coded modulation,” *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4651–4665, 2015.

[16] A. Mishra and P. Venkitasubramaniam, “Admissible length study in anonymous networking: A detection theoretic perspective,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1957–1969, 2013.

[17] L. Sankar, S. R. Rajagopalan, and H. V. Poor, “Utility-privacy tradeoffs in databases: An information-theoretic approach,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[18] R. Zhang and P. Venkitasubramaniam, “Stealthy control signal attacks in linear quadratic gaussian control systems: detectability reward tradeoff,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1555–1570, 2017.

[19] P. Pradhan and P. Venkitasubramaniam, “Stealthy attacks in dynamical systems: Tradeoffs between utility and detectability with application in anonymous systems,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 779–792, 2017.

[20] C.-Z. Bai and V. Gupta, “On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds,” in *American Control Conference (ACC), 2014*. IEEE, 2014, pp. 3029–3034.

[21] S. Zionts, “The criss-cross method for solving linear programming problems,” *Management Science*, vol. 15, no. 7, pp. 426–445, 1969.

[22] T. Tsuchiya, “Affine scaling algorithm,” in *Interior point methods of mathematical programming*. Springer, 1996, pp. 35–82.

[23] S. Lloyd, “Least squares quantization in pcm,” *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[24] N. Z. Shor, *Minimization methods for non-differentiable functions*. Springer Science & Business Media, 2012, vol. 3.

[25] B. C. Levy, *Principles of signal detection and parameter estimation*. Springer Science & Business Media, 2008.

[26] M. Sion, “On general minimax theorems,” *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.