entropy

MDPI

Article

Bottleneck Problems: An Information and Estimation-Theoretic View †

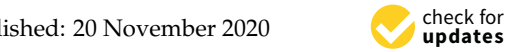
Shahab Asoodeh * and **Flavio P. Calmon**

School of Engineering and Applied Science, Harvard University, Cambridge, MA 02138, USA; flavio@seas.harvard.edu

* Correspondence: shahab@seas.harvard.edu

† Part of the Results in This Paper Was Presented at the International Symposium on Information Theory 2018, Vail, CO, USA, 17–22 June 2018.

Received: 15 October 2020; Accepted: 17 November 2020; Published: 20 November 2020



Abstract: Information bottleneck (IB) and privacy funnel (PF) are two closely related optimization problems which have found applications in machine learning, design of privacy algorithms, capacity problems (e.g., Mrs. Gerber's Lemma), and strong data processing inequalities, among others. In this work, we first investigate the functional properties of IB and PF through a unified theoretical framework. We then connect them to three information-theoretic coding problems, namely hypothesis testing against independence, noisy source coding, and dependence dilution. Leveraging these connections, we prove a new cardinality bound on the auxiliary variable in IB, making its computation more tractable for discrete random variables. In the second part, we introduce a general family of optimization problems, termed "bottleneck problems", by replacing mutual information in IB and PF with other notions of mutual information, namely f-information and Arimoto's mutual information. We then argue that, unlike IB and PF, these problems lead to easily interpretable guarantees in a variety of inference tasks with statistical constraints on accuracy and privacy. While the underlying optimization problems are non-convex, we develop a technique to evaluate bottleneck problems in closed form by equivalently expressing them in terms of lower convex or upper concave envelope of certain functions. By applying this technique to a binary case, we derive closed form expressions for several bottleneck problems.

Keywords: information bottleneck; privacy funnel; mutual information; data processing inequality

1. Introduction

Optimization formulations that involve information-theoretic quantities (e.g., mutual information) have been instrumental in a variety of learning problems found in machine learning. A notable example is the information bottleneck (IB) method [1]. Suppose Y is a target variable and X is an observable correlated variable with joint distribution P_{XY}. The goal of IB is to learn a "compact" summary (aka bottleneck) T of X that is maximally "informative" for inferring Y. The bottleneck variable T is assumed to be generated from X by applying a random function F to X, i.e., $T = F(X)$, in such a way that it is conditionally independent of Y given X, that we denote by

$$Y \multimap X \multimap T. \tag{1}$$

The IB quantifies this goal by measuring the "compactness" of T using the mutual information $I(X;T)$ and, similarly, "informativeness" by $I(Y;T)$. For a given level of compactness $R \geq 0$, IB extracts the bottleneck variable T that solves the constrained optimization problem

$$\mathsf{IB}(R) \coloneqq \sup\ I(Y;T) \qquad \text{subject to} \qquad I(X;T) \leq R, \tag{2}$$

where the supremum is taken over all randomized functions $T = F(X)$ satisfying $Y \multimap X \multimap T$.

The optimization problem that underlies the information bottleneck has been studied in the information theory literature as early as the 1970's—see [2–5]—as a technique to prove impossibility results in information theory and also to study the common information between X and Y. Wyner and Ziv [2] explicitly determined the value of $\mathsf{IB}(R)$ for the special case of binary X and Y—a result widely known as Mrs. Gerber's Lemma [2,6]. More than twenty years later, the information bottleneck function was studied by Tishby et al. [1] and re-formulated in a data analytic context. Here, the random variable X represents a high-dimensional observation with a corresponding low-dimensional feature Y. IB aims at specifying a compressed description of image which is maximally informative about feature Y. This framework led to several applications in clustering [7–9] and quantization [10,11].

A closely-related framework to IB is the privacy funnel (PF) problem [12–14]. In the PF framework, a bottleneck variable T is sought to maximally preserve "information" contained in X while revealing as little about Y as possible. This framework aims to capture the inherent trade-off between revealing X perfectly and leaking a sensitive attribute Y. For instance, suppose a user wishes to share an image X for some classification tasks. The image might carry information about attributes, say Y, that the user might consider as sensitive, even when such information is of limited use for the tasks, e.g., location, or emotion. The PF framework seeks to extract a representation of X from which the original image can be recovered with maximal accuracy while minimizing the privacy leakage with respect to Y. Using mutual information for both privacy leakage and informativeness, the privacy funnel can be formulated as

$$\mathsf{PF}(r) \coloneqq \inf \; I(Y;T) \qquad \text{subject to} \qquad I(X;T) \geq r, \tag{3}$$

where the infumum is taken over all randomized function $T = F(X)$ and r is the parameter specifying the level of informativeness. It is evident from the formulations (2) and (3) that IB and PF are closely related. In fact, we shall see later that they correspond to the upper and lower boundaries of a two-dimensional compact convex set. This duality has led to design of greedy algorithms [12,15] for estimating PF based on the agglomerative information bottleneck [9] algorithm. A similar formulation has recently been proposed in [16] as a tool to train a neural network for learning a private representation of data X; see [17,18] for other closely-related formulations. Solving IB and PF optimization problems analytically is challenging. However, recent machine learning applications, and deep learning algorithms in particular, have reignited the study of both IB and PF (see Related Work).

In this paper, we first give a cohesive overview of the existing results surrounding the IB and the PF formulations. We then provide a comprehensive analysis of IB and PF from an information-theoretic perspective, as well as a survey of several formulations connected to the IB and PF that have been introduced in the information theory and machine learning literature. Moreover, we overview connections with coding problems such as remote source-coding [19], testing against independence [20], and dependence dilution [21]. Leveraging these connections, we prove a new cardinality bound for the bottleneck variable in IB, leading to more tractable optimization problem for IB. We then consider a broad family of optimization problems by going beyond mutual information in formulations (2) and (3). We propose two candidates for this task: Arimoto's mutual information [22] and f-information [23]. By replacing $I(Y;T)$ and/or $I(X;T)$ with either of these measures, we generate a family of optimization problems that we referred to as the bottleneck problems. These problems are shown to better capture the underlying trade-offs intended by IB and PF (see also the short version [24]). More specifically, our main contributions are listed next.

- Computing IB and PF are notoriously challenging when X takes values in a set with infinite cardinality (e.g., X is drawn from a continuous probability distribution). We consider three different scenarios to circumvent this difficulty. First, we assume that X is a Gaussian perturbation of Y, i.e., $X = Y + N^{\mathsf{G}}$ where N^{G} is a noise variable sampled from a Gaussian distribution

independent of Y. Building upon the recent advances in entropy power inequality in [25], we derive a sharp upper bound for $\mathsf{IB}(R)$. As a special case, we consider jointly Gaussian (X, Y) for which the upper bound becomes tight. This then provides a significantly simpler proof for the fact that in this special case the optimal bottleneck variable T is also Gaussian than the original proof given in [26]. In the second scenario, we assume that Y is a Gaussian perturbation of X, i.e., $Y = X + N^{\mathsf{G}}$. This corresponds to a practical setup where the feature Y might be perfectly obtained from a noisy observation of X. Relying on the recent results in strong data processing inequality [27], we obtain an upper bound on $\mathsf{IB}(R)$ which is tight for small values of R. In the last scenario, we compute second-order approximation of $\mathsf{PF}(r)$ under the assumption that T is obtained by Gaussian perturbation of X, i.e., $T = X + N^{\mathsf{G}}$. Interestingly, the rate of increase of $\mathsf{PF}(r)$ for small values of r is shown to be dictated by an asymmetric measure of dependence introduced by Rényi [28].

- We extend the Witsenhausen and Wyner's approach [3] for analytically computing IB and PF. This technique converts solving the optimization problems in IB and PF to determining the convex and concave envelopes of a certain function, respectively. We apply this technique to binary X and Y and derive a closed form expression for $\mathsf{PF}(r)$– we call this result Mr. Gerber's Lemma.
- Relying on the connection between IB and noisy source coding [19] (see [29,30]), we show that the optimal bottleneck variable T in optimization problem (2) takes values in a set $\mathcal{T}$ with cardinality $|\mathcal{T}| \leq |\mathcal{X}|$. Compared to the best cardinality bound previously known (i.e., $|\mathcal{T}| \leq |\mathcal{X}| + 1$), this result leads to a reduction in the search space's dimension of the optimization problem (2) from $\mathbb{R}^{|\mathcal{X}|^2}$ to $\mathbb{R}^{|\mathcal{X}|(|\mathcal{X}|-1)}$. Moreover, we show that this does not hold for PF, indicating a fundamental difference in optimizations problems (2) and (3).
- Following [14,31], we study the deterministic IB and PF (denoted by dIB and dPF) in which T is assumed to be a deterministic function of X, i.e., $T = f(X)$ for some function f. By connecting dIB and dPF with entropy-constrained scalar quantization problems in information theory [32], we obtain bounds on them explicitly in terms of $|\mathcal{X}|$. Applying these bounds to IB, we obtain that $\frac{\mathsf{IB}(R)}{I(X;Y)}$ is bounded by one from above and by $\min\{\frac{R}{H(X)}, \frac{e^R-1}{|\mathcal{X}|}\}$ from below.
- By replacing $I(Y;T)$ and/or $I(X;T)$ in (2) and (3) with Arimoto's mutual information or f-information, we generate a family of bottleneck problems. We then argue that these new functionals better describe the trade-offs that were intended to be captured by IB and PF. The main reason is three-fold: First, as illustrated in Section 2.3, mutual information in IB and PF are mainly justified when $n \gg 1$ independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ of P_{XY} are considered. However, Arimoto's mutual information allows for operational interpretation even in the single-shot regime (i.e., for $n = 1$). Second, $I(Y;T)$ in IB and PF is meant to be a proxy for the efficiency of reconstructing Y given observation T. However, this can be accurately formalized by probability of correctly guessing Y given T (i.e., Bayes risk) or minimum mean-square error (MMSE) in estimating Y given T. While $I(Y;T)$ bounds these two measures, we show that they are precisely characterized by Arimoto's mutual information and f-information, respectively. Finally, when P_{XY} is unknown, mutual information is known to be notoriously difficult to estimate. Nevertheless, Arimoto's mutual information and f-information are easier to estimate: While mutual information can be estimated with estimation error that scales as $O(\log n/\sqrt{n})$ [33], Diaz et al. [34] showed that this estimation error for Arimoto's mutual information and f-information is $O(1/\sqrt{n})$.

  We also generalize our computation technique that enables us to analytically compute these bottleneck problems. Similar as before, this technique converts computing bottleneck problems to determining convex and concave envelopes of certain functions. Focusing on binary X and Y, we derive closed form expressions for some of the bottleneck problems.

1.1. Related Work

The IB formulation has been extensively applied in representation learning and clustering [7,8,35–38]. Clustering based on IB results in algorithms that cluster data points in terms of the similarity of $P_{Y|X}$. When data points lie in a metric space, usually geometric clustering is preferred where clustering is based upon the geometric (e.g., Euclidean) distance. Strouse and Schwab [31,39] proposed the deterministic IB (denoted by dIB) by enforcing that $P_{T|X}$ is a deterministic mapping: $\mathsf{dIB}(R)$ denotes the supremum of $I(Y;f(X))$ over all functions $f:\mathcal{X}\to\mathcal{T}$ satisfying $H(f(X))\le R$. This optimization problem is closely related to the problem of scalar quantization in information theory: designing a function $f:\mathcal{X}\to[M]:=\{1,\dots,M\}$ with a pre-determined output alphabet with f optimizing some objective functions. This objective might be maximizing or minimizing $H(f(X))$ [40] or maximizing $I(Y;f(X))$ for a random variable Y correlated with X [32,41–43]. Since $H(f(X))\le\log M$ for $f:\mathcal{X}\to[M]$, the latter problem provides lower bounds for dIB (and thus for IB). In particular, one can exploit [44] (Theorem 1) to obtain $I(X;Y)-\mathsf{dIB}(R)\le O(e^{-2R/|\mathcal{Y}|-1})$ provided that $\min\{|\mathcal{X}|,2^R\}>2|\mathcal{Y}|$. This result establishes a linear gap between dIB and $I(X;Y)$ irrespective of $|\mathcal{X}|$.

The connection between quantization and dIB further allows us to obtain multiplicative bounds. For instance, if $Y\sim\mathsf{Bernoulli}(\frac{1}{2})$ and $X=Y+N^{\mathsf{G}}$, where $N^{\mathsf{G}}\sim\mathcal{N}(0,1)$ is independent of Y, then it is well-known in information theory literature that $I(Y;f(X))\ge\frac{2}{\pi}I(X;Y)$ for all non-constant $f:\mathcal{X}\to\{0,1\}$ (see, e.g., [45] (Section 2.11)), thus $\mathsf{dIB}(R)\ge\frac{2}{\pi}I(X;Y)$ for $R\le1$. We further explore this connection to provide multiplicative bounds on $\mathsf{dIB}(R)$ in Section 2.5.

The study of IB has recently gained increasing traction in the context of deep learning. By taking T to be the activity of the hidden layer(s), Tishby and Zaslavsky [46] (see also [47]) argued that neural network classifiers trained with cross-entropy loss and stochastic gradient descent (SGD) inherently aims at solving the IB optimization problems. In fact, it is claimed that the graph of the function $R\mapsto\mathsf{IB}(R)$ (the so-called the information plane) characterizes the learning dynamic of different layers in the network: shallow layers correspond to maximizing $I(Y;T)$ while deep layers' objective is minimizing $I(X;T)$. While the generality of this claim was refuted empirically in [48] and theoretically in [49,50], it inspired significant follow-up studies. These include (i) modifying neural network training in order to solve the IB optimization problem [51–55]; (ii) creating connections between IB and generalization error [56], robustness [51], and detection of out-of-distribution data [57]; and (iii) using IB to understand specific characteristic of neural networks [55,58–60].

In both IB and PF, mutual information poses some limitations. For instance, it may become infinity in deterministic neural networks [48–50] and also may not lead to proper privacy guarantee [61]. As suggested in [55,62], one way to address this issue is to replace mutual information with other statistical measures. In the privacy literature, several measures with strong privacy guarantee have been proposed including Rényi maximal correlation [21,63,64], probability of correctly recovering [65,66], minimum mean-squared estimation error (MMSE) [67,68], χ^2-information [69] (a special case of f-information to be described in Section 3), Arimoto's and Sibson's mutual information [61,70]—to be discussed in Section 3, maximal leakage [71], and local differential privacy [72]. All these measures ensure interpretable privacy guarantees. For instance, it is shown in [67,68] that if χ^2-information between Y and T is sufficiently small, then no functions of Y can be efficiently reconstructed given T; thus providing an interpretable privacy guarantee.

Another limitation of mutual information is related to its estimation difficulty. It is known that mutual information can be estimated from n samples with the estimation error that scales as $O(\log n/\sqrt{n})$ [33]. However, as shown by Diaz et al. [34], the estimation error for most of the above measures scales as $O(1/\sqrt{n})$. Furthermore, the recently popular variational estimators for mutual information, typically implemented via deep learning methods [73–75], presents some fundamental limitations [76]: the variance of the estimator might grow exponentially with the ground truth mutual information and also the estimator might not satisfy basic properties of mutual information such

as data processing inequality or additivity. McAllester and Stratos [77] showed that some of these limitations are inherent to a large family of mutual information estimators.

1.2. Notation

We use capital letters, e.g., X, for random variables and calligraphic letters for their alphabets, e.g., $\mathcal{X}$. If X is distributed according to probability mass function (pmf) P_X, we write $X \sim P_X$. Given two random variables X and Y, we write P_{XY} and $P_{Y|X}$ as the joint distribution and the conditional distribution of Y given X. We also interchangeably refer to $P_{Y|X}$ as a channel from X to Y. We use $H(X)$ to denote both entropy and differential entropy of X, i.e., we have

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$$

if X is a discrete random variable taking values in $\mathcal{X}$ with probability mass function (pmf) P_X and

$$H(X) = -\int \log f_X(x) \log f_X(x) \mathrm{d}x,$$

where X is an absolutely continuous random variable with probability density function (pdf) f_X. If X is a binary random variable with $P_X(1) = p$, we write $X \sim \mathsf{Bernoulli}(p)$. In this case, its entropy is called binary entropy function and denoted by $h_{\mathsf{b}}(p) \coloneqq -p \log p - (1-p) \log(1-p)$. We use superscript G to describe a standard Gaussian random variable, i.e., $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$. Given two random variables X and Y, their (Shannon's) mutual information is denoted by $I(X;Y) \coloneqq H(Y) - H(Y|X)$. We let $\mathcal{P}(\mathcal{X})$ denote the set of all probability distributions on the set $\mathcal{X}$. Given an arbitrary $Q_X \in \mathcal{P}(\mathcal{X})$ and a channel $P_{Y|X}$, we let $Q_X P_{Y|X}$ denote the resulting output distribution on $\mathcal{Y}$. For any $a \in [0,1]$, we use $\bar{a}$ to denote $1-a$ and for any integer $k \in \mathbb{N}$, $[k] \coloneqq \{1, 2, \ldots, k\}$.

Throughout the paper, we assume a pair of (discrete or continuous) random variables $(X,Y) \sim P_{XY}$ are given with a fixed joint distribution P_{XY}, marginals P_X and P_Y, and conditional distribution $P_{Y|X}$. We then use $Q_X \in \mathcal{P}(\mathcal{X})$ to denote an arbitrary distribution with $Q_Y = Q_X P_{Y|X} \in \mathcal{P}(\mathcal{Y})$.

2. Information Bottleneck and Privacy Funnel: Definitions and Functional Properties

In this section, we review the information bottleneck and its closely related functional, the privacy funnel. We then prove some analytical properties of these two functionals and develop a convex analytic approach which enables us to compute closed-form expressions for both these two functionals in some simple cases.

To precisely quantify the trade-off between these two conflicting goals, the IB optimization problem (2) was proposed [1]. Since any randomized function $T = F(X)$ can be equivalently characterized by a conditional distribution, the optimization problem (2) can be instead expressed as

$$\mathsf{IB}(P_{XY}, R) \coloneqq \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) \le R}} I(Y;T), \qquad \text{or} \qquad \widetilde{\mathsf{IB}}(P_{XY}, \tilde{R}) \coloneqq \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(Y;T) \ge \tilde{R}}} I(X;T). \tag{4}$$

where R and $\tilde{R}$ denote the level of desired compression and informativeness, respectively. We use $\mathsf{IB}(R)$ and $\widetilde{\mathsf{IB}}(\tilde{R})$ to denote $\mathsf{IB}(P_{XY}, R)$ and $\widetilde{\mathsf{IB}}(P_{XY}, \tilde{R})$, respectively, when the joint distribution is clear from the context. Notice that if $\mathsf{IB}(P_{XY}, R) = \tilde{R}$, then $\widetilde{\mathsf{IB}}(P_{XY}, \tilde{R}) = R$.

Now consider the setup where data X is required to be disclosed while maintaining the privacy of a sensitive attribute, represented by Y. This goal was formulated by PF in (3). As before, replacing randomized function $T = F(X)$ with conditional distribution $P_{T|X}$, we can equivalently express (3) as

$$\mathsf{PF}(P_{XY}, r) := \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) \geq r}} I(Y;T), \qquad \text{or} \qquad \widetilde{\mathsf{PF}}(P_{XY}, \tilde{r}) := \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(Y;T) \leq \tilde{r}}} I(X;T), \tag{5}$$

where $\tilde{r}$ and r denote the level of desired privacy and informativeness, respectively. The case $\tilde{r} = 0$ is particularly interesting in practice and specifies perfect privacy, see e.g., [13,78]. As before, we write $\widetilde{\mathsf{PF}}(\tilde{r})$ and $\mathsf{PF}(r)$ for $\widetilde{\mathsf{PF}}(P_{XY}, \tilde{r})$ and $\mathsf{PF}(P_{XY}, r)$ when P_{XY} is clear from the context.

The following properties of IB and PF follow directly from their definitions. The proof of this result (and any other results in this section) is given in Appendix A.

Theorem 1. *For a given P_{XY}, the mappings $\mathsf{IB}(R)$ and $\mathsf{PF}(r)$ have the following properties:*

- $\mathsf{IB}(0) = \mathsf{PF}(0) = 0$.
- $\mathsf{IB}(R) = I(X;Y)$ *for any* $R \geq H(X)$ *and* $\mathsf{PF}(r) = I(X;Y)$ *for* $r \geq H(X)$.
- $0 \leq \mathsf{IB}(R) \leq \min\{R, I(X;Y)\}$ *for any* $R \geq 0$ *and* $\mathsf{PF}(r) \geq \max\{r - H(X|Y), 0\}$ *for any* $r \geq 0$.
- $R \mapsto \mathsf{IB}(R)$ *is continuous, strictly increasing, and concave on the range* $(0, I(X;Y))$.
- $r \mapsto \mathsf{PF}(r)$ *is continuous, strictly increasing, and convex on the range* $(0, I(X;Y))$.
- *If* $P_{Y|X}(y|x) > 0$ *for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$, *then both* $R \mapsto \mathsf{IB}(R)$ *and* $r \mapsto \mathsf{PF}(r)$ *are continuously differentiable over* $(0, H(X))$.
- $R \mapsto \frac{\mathsf{IB}(R)}{R}$ *is non-increasing and* $r \mapsto \frac{\mathsf{PF}(r)}{r}$ *is non-decreasing.*
- *We have*

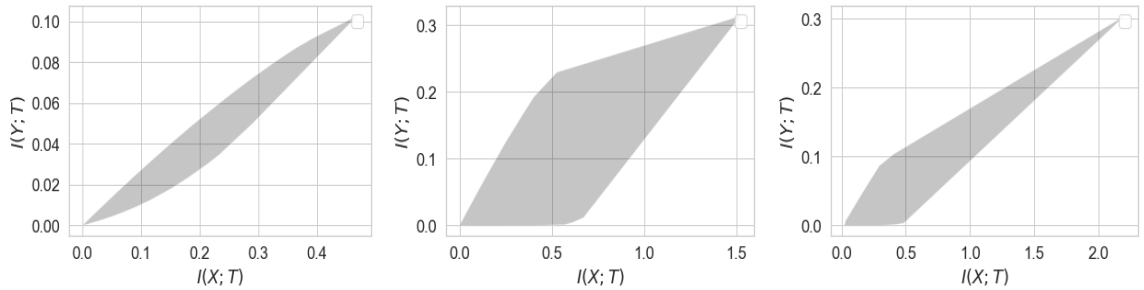
$$\mathsf{IB}(R) := \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) = R}} I(Y;T), \qquad \text{and} \qquad \mathsf{PF}(r) := \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) = r}} I(Y;T).$$

According to this theorem, we can always restrict both R and r in (4) and (5), respectively, to $[0, H(X)]$ as $\mathsf{IB}(R) = \mathsf{PF}(r) = I(X;Y)$ for all $r, R \geq H(X)$.

Define $\mathcal{M} = \mathcal{M}(P_{XY}) \subset \mathbb{R}^2$ as

$$\mathcal{M} := \left\{ (I(X;T), I(Y;T)) : Y \multimap X \multimap T, (X,Y) \sim P_{XY} \right\}. \tag{6}$$

It can be directly verified that $\mathcal{M}$ is convex. According to this theorem, $R \mapsto \mathsf{IB}(R)$ and $r \mapsto \mathsf{PF}(r)$ correspond to the upper and lower boundary of $\mathcal{M}$, respectively. The convexity of $\mathcal{M}$ then implies the concavity and convexity of IB and PF. Figure 1 illustrates the set $\mathcal{M}$ for the simple case of binary X and Y.



(a) $|\mathcal{X}| = |\mathcal{Y}| = 2$ **(b)** $|\mathcal{X}| = 3, |\mathcal{Y}| = 2$ **(c)** $|\mathcal{X}| = 5, |\mathcal{Y}| = 2$

Figure 1. Examples of the set $\mathcal{M}$, defined in (6). The upper and lower boundaries of this set correspond to information bottleneck (IB) and privacy funnel (PF), respectively. It is worth noting that, while $\mathsf{IB}(R) = 0$ only at $R = 0$, $\mathsf{PF}(r) = 0$ holds in general for r belonging to a non-trivial interval (only for $|\mathcal{X}| > 2$). Moreover, note that in general neither upper nor lower boundaries are smooth. A sufficient condition for smoothness is $P_{Y|X}(y|x) > 0$ (see Theorem 1), thus both IB and PF are smooth in the binary case.

While both $\mathsf{IB}(0) = 0$ and $\mathsf{PF}(0) = 0$, their behavior in the neighborhood around zero might be completely different. As illustrated in Figure 1, $\mathsf{IB}(R) > 0$ for all $R > 0$, whereas $\mathsf{PF}(r) = 0$ for $r \in [0, r_0]$ for some $r_0 > 0$. When such $r_0 > 0$ exists, we say perfect privacy occurs: there exists a variable T satisfying $Y \multimap X \multimap T$ such that $I(Y;T) = 0$ while $I(X;T) > 0$; making T a representation of X having perfect privacy (i.e., no information leakage about Y). A necessary and sufficient condition for the existence of such T is given in [21] (Lemma 10) and [13] (Theorem 3), described next.

Theorem 2 (Perfect privacy)**.** *Let $(X,Y) \sim P_{XY}$ be given and $\mathcal{A} \subset [0,1]^{|\mathcal{Y}|}$ be the set of vectors $\{P_{Y|X}(\cdot|x), x \in \mathcal{X}\}$. Then there exists $r_0 > 0$ such that $\mathsf{PF}(r) = 0$ for $r \in [0, r_0]$ if and only if vectors in $\mathcal{A}$ are linearly independent.*

In light of this theorem, we obtain that perfect privacy occurs if $|\mathcal{X}| > |\mathcal{Y}|$. It also follows from the theorem that for binary X, perfect privacy cannot occur (see Figure 1a).

Theorem 1 enables us to derive a simple bounds for IB and PF. Specifically, the facts that $\frac{\mathsf{PF}(r)}{r}$ is non-decreasing and $\frac{\mathsf{IB}(R)}{R}$ is non-increasing immediately result in the the following linear bounds.

Theorem 3 (Linear lower bound)**.** *For $r, R \in (0, H(X))$, we have*

$$\inf_{\substack{Q_X \in \mathcal{P}(\mathcal{X}) \\ Q_X \neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y \| P_Y)}{D_{\mathsf{KL}}(Q_X \| P_X)} \leq \frac{\mathsf{PF}(r)}{r} \leq \frac{I(X;Y)}{H(X)} \leq \frac{\mathsf{IB}(R)}{R} \leq \sup_{\substack{Q_X \in \mathcal{P}(\mathcal{X}) \\ Q_X \neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y \| P_Y)}{D_{\mathsf{KL}}(Q_X \| P_X)} \leq 1. \tag{7}$$

In light of this theorem, if $\mathsf{PF}(r) = r$, then $I(X;Y) = H(X)$, implying $X = g(Y)$ for a deterministic function g. Conversely, if $X = g(Y)$ then $\mathsf{PF}(r) = r$ because for all T forming the Markov relation $Y \multimap g(Y) \multimap T$, we have $I(Y;T) = I(g(Y);T)$. On the other hand, we have $\mathsf{IB}(R) = R$ if and only if there exists a variable T^* satisfying $I(X;T^*) = I(Y;T^*)$ and thus the following double Markov relations

$$Y \multimap X \multimap T^*, \qquad \text{and} \qquad X \multimap Y \multimap T^*.$$

It can be verified (see [79] (Problem 16.25)) that this double Markov condition is equivalent to the existence of a pair of functions f and g such that $f(X) = g(Y)$ and $(X,Y) \multimap f(X) \multimap T^*$. One special case of this setting, namely where g is an identity function, has been recently studied in details in [53] and will be reviewed in Section 2.5. Theorem 3 also enables us to characterize the "worst"

joint distribution P_{XY} with respect to IB and PF. As demonstrated in the following lemma, if $P_{Y|X}$ is an erasure channel then $\frac{\mathsf{PF}(r)}{r} = \frac{\mathsf{IB}(R)}{R} = \frac{I(X;Y)}{H(X)}$.

Lemma 1.

- *Let P_{XY} be such that $\mathcal{Y} = \mathcal{X} \cup \{\perp\}$, $P_{Y|X}(x|x) = 1-\delta$, and $P_{Y|X}(\perp|x) = \delta$ for some $\delta > 0$. Then*

$$\frac{\mathsf{PF}(r)}{r} = \frac{\mathsf{IB}(R)}{R} = 1-\delta.$$

- *Let P_{XY} be such that $\mathcal{X} = \mathcal{Y} \cup \{\perp\}$, $P_{X|Y}(y|y) = 1-\delta$, and $P_{X|Y}(\perp|y) = \delta$ for some $\delta > 0$. Then*

$$\mathsf{PF}(r) = \max\{r - H(X|Y), 0\}.$$

The bounds in Theorem 3 hold for all r and R in the interval $[0, H(X)]$. We can, however, improve them when r and R are sufficiently small. Let $\mathsf{PF}'(0)$ and $\mathsf{IB}'(0)$ denote the slope of $\mathsf{PF}(\cdot)$ and $\mathsf{IB}(\cdot)$ at zero, i.e., $\mathsf{PF}'(0) := \lim_{r\to 0^+} \frac{\mathsf{PF}(r)}{r}$ and $\mathsf{IB}'(0) := \lim_{R\to 0^+} \frac{\mathsf{IB}(R)}{R}$.

Theorem 4. *Given $(X,Y) \sim P_{XY}$, we have*

$$\inf_{\substack{Q_X \in \mathcal{P}(\mathcal{X}) \\ Q_X \neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)} = \mathsf{PF}'(0) \le \min_{\substack{x\in\mathcal{X}: \\ P_X(x)>0}} \frac{D_{\mathsf{KL}}(P_{Y|X}(\cdot|x)\|P_Y(\cdot))}{-\log P_X(x)}$$
$$\le \max_{\substack{x\in\mathcal{X}: \\ P_X(x)>0}} \frac{D_{\mathsf{KL}}(P_{Y|X}(\cdot|x)\|P_Y(\cdot))}{-\log P_X(x)} \le \mathsf{IB}'(0) = \sup_{\substack{Q_X \in \mathcal{P}(\mathcal{X}) \\ Q_X \neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)}.$$

This theorem provides the exact values of $\mathsf{PF}'(0)$ and $\mathsf{IB}'(0)$ and also simple bounds for them. While the exact expressions for $\mathsf{PF}'(0)$ and $\mathsf{IB}'(0)$ are usually difficult to compute, a simple plug-in estimator is proposed in [80] for $\mathsf{IB}'(0)$. This estimator can be readily adapted to estimate $\mathsf{PF}'(0)$. Theorem 4 reveals a profound connection between IB and the strong data processing inequality (SDPI) [81]. More precisely, thanks to the pioneering work of Anantharam et al. [82], it is known that the supremum of $\frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)}$ over all $Q_X \neq P_X$ is equal the supremum of $\frac{I(Y;T)}{I(X;T)}$ over all $P_{T|X}$ satisfying $Y \multimap X \multimap T$ and hence $\mathsf{IB}'(0)$ specifies the strengthening of the data processing inequality of mutual information. This connection may open a new avenue for new theoretical results for IB, especially when X or Y are continuous random variables. In particular, the recent non-multiplicative SDPI results [27,83] seem insightful for this purpose.

In many practical cases, we might have n i.i.d. samples $(X_1,Y_1),\ldots,(X_n,Y_n)$ of $(X,Y) \sim P_{XY}$. We now study how IB behaves in n. Let $X^n := (X_1,\ldots,X_n)$ and $Y^n := (Y_1,\ldots,Y_n)$. Due to the i.i.d. assumption, we have $P_{X^nY^n}(x^n,y^n) = \prod_{i=1}^n P_{XY}(x_i,y_i)$. This can also be described by independently feeding X_i, $i \in [n]$, to channel $P_{Y|X}$ producing Y_i. The following theorem, demonstrated first in [3] (Theorem 2.4), gives a formula for IB in terms of n.

Theorem 5 (Additivity). *We have*

$$\frac{1}{n}\mathsf{IB}(P_{X^nY^n}, nR) = \mathsf{IB}(P_{XY}, R).$$

This theorem demonstrates that an optimal channel $P_{T^n|X^n}$ for i.i.d. samples $(X^n,Y^n) \sim P_{XY}$ is obtained by the Kronecker product of an optimal channel $P_{T|X}$ for $(X,Y) \sim P_{XY}$. This, however, may not hold in general for PF, that is, we might have $\mathsf{PF}(P_{X^nY^n}, nr) < n\mathsf{PF}(P_{XY}, r)$, see [13] (Proposition 1) for an example.

2.1. Gaussian IB and PF

In this section, we turn our attention to a special, yet important, case where $X = Y + \sigma N^{\mathsf{G}}$, where $\sigma > 0$ and $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of Y. This setting subsumes the popular case of jointly Gaussian (X, Y) whose information bottleneck functional was computed in [84] for the vector case (i.e., (X, Y) are jointly Gaussian random vectors).

Lemma 2. *Let $\{Y_i\}_{i=1}^n$ be n i.i.d. copies of $Y \sim P_Y$ and $X_i = Y_i + \sigma N_i^{\mathsf{G}}$ where $\{N_i^{\mathsf{G}}\}$ are i.i.d samples of $\mathcal{N}(0,1)$ independent of Y. Then, we have*

$$\frac{1}{n}\mathsf{IB}(P_{X^nY^n}, nR) \le H(X) - \frac{1}{2}\log\left[2\pi e\sigma^2 + e^{2(H(Y)-R)}\right].$$

It is worth noting that this result was concurrently proved in [85]. The main technical tool in the proof of this lemma is a strong version of the entropy power inequality [25] (Theorem 2) which holds even if X_i, Y_i, and N_i are random vectors (as opposed to scalar). Thus, one can readily generalize Lemma 2 to the vector case. Note that the upper bound established in this lemma holds without any assumptions on $P_{T|X}$. This upper bound provides a significantly simpler proof for the well-known fact that for the jointly Gaussian (X, Y), the optimal channel $P_{T|X}$ is Gaussian. This result was first proved in [26] and used in [84] to compute an expression of IB for the Gaussian case.

Corollary 1. *If (X, Y) are jointly Gaussian with correlation coefficient ρ, then we have*

$$\mathsf{IB}(R) = \frac{1}{2}\log\frac{1}{1-\rho^2+\rho^2 e^{-2R}}. \tag{8}$$

Moreover, the optimal channel $P_{T|X}$ is given by $P_{T|X}(\cdot|x) = \mathcal{N}(0, \tilde{\sigma}^2)$ for $\tilde{\sigma}^2 = \sigma_Y^2 \frac{e^{-2R}}{\rho^2(1-e^{-2R})}$ where σ_Y^2 is the variance of Y.

In Lemma 2, we assumed that X is a Gaussian perturbation of Y. However, in some practical scenarios, we might have Y as a Gaussian perturbation of X. For instance, let X represent an image and Y be a feature of the image that can be perfectly obtained from a noisy observation of X. Then, the goal is to compress the image with a given compression rate while retaining maximal information about the feature. The following lemma, which is an immediate consequence of [27] (Theorem 1), gives an upper bound for IB in this case.

Lemma 3. *Let X^n be n i.i.d. copies of a random variable X satisfying $\mathbb{E}[X^2] \le 1$ and Y_i be the result of passing X_i, $i \in [n]$, through a Gaussian channel $Y = X + \sigma N^{\mathsf{G}}$, where $\sigma > 0$ and $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of X. Then, we have*

$$\frac{1}{n}\mathsf{IB}(P_{X^nY^n}, nR) \le R - \Psi(R, \sigma), \tag{9}$$

where

$$\Psi(R,\sigma) := \max_{x\in[0,\frac{1}{2}]} 2\mathsf{Q}\left(\sqrt{\frac{1}{x\sigma^2}}\right)\left(R - h_{\mathsf{b}}(x) - \frac{x}{2}\log\left(1+\frac{1}{x\sigma^2}\right)\right), \tag{10}$$

$\mathsf{Q}(t) := \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t$ is the Gaussian complimentary CDF and $h_{\mathsf{b}}(a) := -a\log(a) - (1-a)\log(1-a)$ for $a \in (0,1)$ is the binary entropy function. Moreover, we have

$$\frac{1}{n}\mathsf{IB}(P_{X^nY^n}, nR) \le R - e^{-\frac{1}{R\sigma^2}\log\frac{1}{R} + \Theta\left(\log\frac{1}{R}\right)}. \tag{11}$$

Note that that Lemma 3 holds for any arbitrary X (provided that $\mathbb{E}[X^2] \le 1$) and hence (9) bounds information bottleneck functionals for a wide family of P_{XY}. However, the bound is loose in general

for large values of R. For instance, if (X, Y) are jointly Gaussian (implying $Y = X + \sigma N^{\mathsf{G}}$ for some $\sigma > 0$), then the right-hand side of (9) does not reduce to (8). To show this, we numerically compute the upper bound (9) and compare it with the Gaussian information bottleneck (8) in Figure 2.

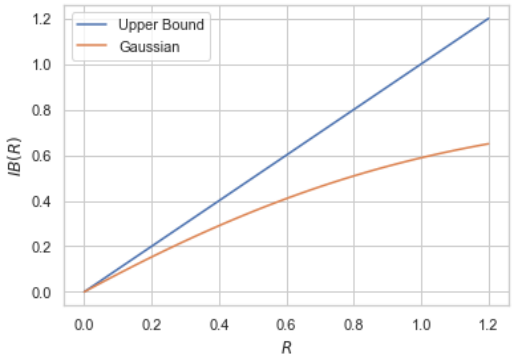


Figure 2. Comparison of (8), the exact value of IB for jointly Gaussian X and Y (i.e., $Y = X + \sigma N^{\mathsf{G}}$ with X and N^{G} being both standard Gaussian $\mathcal{N}(0,1)$), with the general upper bound (9) for $\sigma^2 = 0.5$. It is worth noting that while the Gaussian IB converges to $I(X;Y) \approx 0.8$, the upper bound diverges.

The privacy funnel functional is much less studied even for the simple case of jointly Gaussian. Solving the optimization in PF over $P_{T|X}$ without any assumptions is a difficult challenge. A natural assumption to make is that $P_{T|X}(\cdot|x)$ is Gaussian for each $x \in \mathcal{X}$. This leads to the following variant of PF

$$\mathsf{PF}^{\mathsf{G}}(r) \coloneqq \inf_{\substack{\sigma \geq 0, \\ I(X;T_\sigma) \geq r}} I(Y;T_\sigma),$$

where

$$T_\sigma \coloneqq X + \sigma N^{\mathsf{G}},$$

and $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of X. This formulation is tractable and can be computed in closed form for jointly Gaussian (X, Y) as described in the following example.

Example 1. *Let X and Y be jointly Gaussian with correlation coefficient ρ. First note that since mutual information is invariant to scaling, we may assume without loss of generality that both X and Y are zero mean and unit variance and hence we can write $X = \rho Y + \sqrt{1-\rho^2} M^{\mathsf{G}}$ where $M^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of Y. Consequently, we have*

$$I(X;T_\sigma) = \frac{1}{2} \log \left(1 + \frac{1}{\sigma^2}\right), \tag{12}$$

and

$$I(Y;T_\sigma) = \frac{1}{2} \log \left(1 + \frac{\rho^2}{1-\rho^2+\sigma^2}\right). \tag{13}$$

In order to ensure $I(X;T_\sigma) \geq r$, we must have $\sigma \leq \left(e^{2r} - 1\right)^{-\frac{1}{2}}$. Plugging this choice of σ into (13), we obtain

$$\mathsf{PF}^{\mathsf{G}}(r) = \frac{1}{2} \log \left(\frac{1}{1-\rho^2\left(1-e^{-2r}\right)}\right). \tag{14}$$

This example indicates that for jointly Gaussian (X, Y), we have $\mathsf{PF}^{\mathsf{G}}(r) = 0$ if and only if $r = 0$ (thus perfect privacy does not occur) and the constraint $I(X;T_\sigma) = r$ is satisfied by a unique σ. These two properties in fact hold for all continuous variables X and Y with finite second moments as demonstrated in Lemma A1 in Appendix A. We use these properties to derive a second-order approximation of $\mathsf{PF}^{\mathsf{G}}(r)$ when r is sufficiently small. For the following theorem, we use $\mathsf{var}(U)$

to denote the variance of the random variable U and $\mathsf{var}(U|V) \coloneqq \mathbb{E}[(U - \mathbb{E}[U|V])^2|V]$. We use $\sigma_X^2 = \mathsf{var}(X)$ for short.

Theorem 6. *For any pair of continuous random variables (X, Y) with finite second moments, we have as $r \to 0$*

$$\mathsf{PF}^{\mathsf{G}}(r) = \eta(X,Y)r + \Delta(X,Y)r^2 + o(r^2),$$

where $\eta(X,Y) \coloneqq \frac{\mathsf{var}(\mathbb{E}[X|Y])}{\sigma_X^2}$ and

$$\Delta(X,Y) \coloneqq \frac{2}{\sigma_X^4}\left[\mathbb{E}[\mathsf{var}^2(X|Y)] - \sigma_X^2\mathbb{E}[\mathsf{var}(X|Y)]\right].$$

It is worth mentioning that the quantity $\eta(X,Y)$ was first defined by Rényi [28] as an asymmetric measure of correlation between X and Y. In fact, it can be shown that $\eta(X,Y) = \sup_f \rho^2(X, f(Y))$, where supremum is taken over all measurable functions f and $\rho(\cdot,\cdot)$ denotes the correlation coefficient. As a simple illustration of Theorem 6, consider jointly Gaussian X and Y with correlation coefficient ρ for which PF^{G} was computed in Example 1. In this case, it can be easily verified that $\eta(X,Y) = \rho^2$ and $\Delta(X,Y) = -2\sigma_X^2\rho^2(1-\rho^2)$. Hence, for jointly Gaussian (X,Y) with correlation coefficient ρ and unit variance, we have $\mathsf{PF}^{\mathsf{G}}(r) = \rho^2 r - 2\rho^2(1-\rho^2)r^2 + o(r^2)$. In Figure 3, we compare the approximation given in Theorem 6 for this particular case.

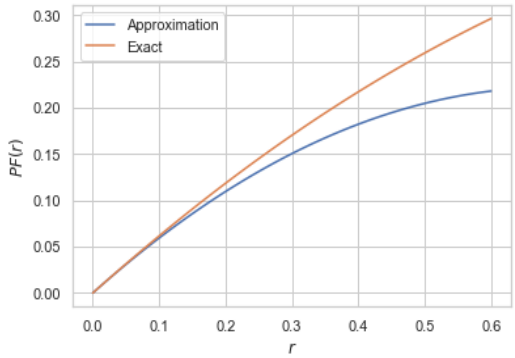


Figure 3. Second-order approximation of PF^{G} according to Theorem 6 for jointly Gaussian X and Y with correlation coefficient $\rho = 0.8$. For this particular case, the exact expression of PF^{G} is computed in (14).

2.2. Evaluation of IB and PF

The constrained optimization problems in the definitions of IB and PF are usually challenging to solve numerically due to the non-linearity in the constraints. In practice, however, both IB and PF are often approximated by their corresponding Lagrangian optimizations

$$\mathcal{L}_{\mathsf{IB}}(\beta) \coloneqq \sup_{P_{T|X}} I(Y;T) - \beta I(X;T) = H(Y) - \beta H(X) - \inf_{P_{T|X}} \left[H(Y|T) - \beta H(X|T)\right], \tag{15}$$

and

$$\mathcal{L}_{\mathsf{PF}}(\beta) \coloneqq \inf_{P_{T|X}} I(Y;T) - \beta I(X;T) = H(Y) - \beta H(X) - \sup_{P_{T|X}} \left[H(Y|T) - \beta H(X|T)\right], \tag{16}$$

where $\beta \in \mathbb{R}_+$ is the Lagrangian multiplier that controls the tradeoff between compression and informativeness in for IB and the privacy and informativeness in PF. Notice that for the computation of $\mathcal{L}_{\mathsf{IB}}$, we can assume, without loss of generality, that $\beta \in [0,1]$ since otherwise the maximizer of (15) is trivial. It is worth noting that $\mathcal{L}_{\mathsf{IB}}(\beta)$ and $\mathcal{L}_{\mathsf{PF}}(\beta)$ in fact correspond to lines of slope β supporting $\mathcal{M}$ from above and below, thereby providing a new representation of $\mathcal{M}$.

Let (X', Y') be a pair of random variables with $X' \sim Q_X$ for some $Q_X \in \mathcal{P}(\mathcal{X})$ and Y' is the output of $P_{Y|X}$ when the input is X' (i.e., $Y' \sim Q_X P_{Y|X}$). Define

$$F_\beta(Q_X) \coloneqq H(Y') - \beta H(X').$$

This function, in general, is neither convex nor concave in Q_X. For instance, $F(0)$ is concave and $F(1)$ is convex in P_X. The lower convex envelope of $F_\beta(Q_X)$ is defined as the largest convex function smaller than $F_\beta(Q_X)$. Similarly, the upper concave envelope of $F_\beta(Q_X)$ is defined as the smallest concave function larger than $F_\beta(Q_X)$. Let $\mathcal{K}_\cup[F_\beta(Q_X)]$ and $\mathcal{K}_\cap[F_\beta(Q_X)]$ denote the lower convex and upper concave envelopes of $F_\beta(Q_X)$, respectively. If $F_\beta(Q_X)$ is convex at P_X, that is $\mathcal{K}_\cup[F_\beta(Q_X)]\big|_{P_X} = F_\beta(P_X)$, then $F_\beta(Q_X)$ remains convex at P_X for all $\beta' \geq \beta$ because

$$\begin{aligned}\mathcal{K}_\cup[F_{\beta'}(Q_X)] &= \mathcal{K}_\cup[F_\beta(Q_X) - (\beta' - \beta)H(X')] \\ &\geq \mathcal{K}_\cup[F_\beta(Q_X)] + \mathcal{K}_\cup[-(\beta' - \beta)H(X')] \\ &= \mathcal{K}_\cup[F_\beta(Q_X)] - (\beta' - \beta)H(X'),\end{aligned}$$

where the last equality follows from the fact that $-(\beta' - \beta)H(X)$ is convex. Hence, at P_X we have

$$\mathcal{K}_\cup[F_{\beta'}(Q_X)]\big|_{P_X} \geq \mathcal{K}_\cup[F_\beta(Q_X)]\big|_{P_X} - (\beta' - \beta)H(X) = F_\beta(P_X) - (\beta' - \beta)H(X) = F_{\beta'}(P_X).$$

Analogously, if $F_\beta(Q_X)$ is concave at P_X, that is $\mathcal{K}_\cap[F_\beta(Q_X)]\big|_{P_X} = F_\beta(P_X)$, then $F_\beta(Q_X)$ remains concave at P_X for all $\beta' \leq \beta$.

Notice that, according to (15) and (16), we can write

$$\mathcal{L}_{\mathsf{IB}}(\beta) = H(Y) - \beta H(X) - \mathcal{K}_\cup[F_\beta(Q_X)]\big|_{P_X}, \tag{17}$$

and

$$\mathcal{L}_{\mathsf{PF}}(\beta) = H(Y) - \beta H(X) - \mathcal{K}_\cap[F_\beta(Q_X)]\big|_{P_X}. \tag{18}$$

In light of the above arguments, we can write

$$\mathcal{L}_{\mathsf{IB}}(\beta) = 0,$$

for all $\beta > \beta_{\mathsf{IB}}$ where β_{IB} is the smallest β such that $F_\beta(P_X)$ touches $\mathcal{K}_\cup[F_\beta(Q_X)]$. Similarly,

$$\mathcal{L}_{\mathsf{PF}}(\beta) = 0,$$

for all $\beta < \beta_{\mathsf{PF}}$ where β_{PF} is the largest β such that $F_\beta(P_X)$ touches $\mathcal{K}_\cap[F_\beta(Q_X)]$. In the following theorem, we show that β_{IB} and β_{PF} are given by the values of $\mathsf{IB}'(0)$ and $\mathsf{PF}'(0)$, respectively, given in Theorem 4. A similar formulae β_{IB} and β_{PF} were given in [86].

Proposition 1. *We have,*

$$\beta_{\mathsf{IB}} = \sup_{Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y \| P_Y)}{D_{\mathsf{KL}}(Q_X \| P_X)},$$

and

$$\beta_{\mathsf{PF}} = \inf_{Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y \| P_Y)}{D_{\mathsf{KL}}(Q_X \| P_X)}.$$

Kim et al. [80] have recently proposed an efficient algorithm to estimate β_{IB} from samples of P_{XY} involving a simple optimization problem. This algorithm can be readily adapted for estimating β_{PF}.

Proposition 1 implies that in optimizing the Lagrangians (17) and (18), we can restrict the Lagrange multiplier β, that is

$$\mathcal{L}_{\mathsf{IB}}(\beta) = H(Y) - \beta H(X) - \mathcal{K}_{\cup}[F_{\beta}(Q_X)]\big|_{P_X}, \qquad \text{for} \qquad \beta \in [0, \beta_{\mathsf{IB}}], \tag{19}$$

and

$$\mathcal{L}_{\mathsf{PF}}(\beta) = H(Y) - \beta H(X) - \mathcal{K}_{\cap}[F_{\beta}(Q_X)]\big|_{P_X}, \qquad \text{for} \qquad \beta \in [\beta_{\mathsf{PF}}, \infty). \tag{20}$$

Remark 1. *As demonstrated by Kolchinsky et al. [53], the boundary points* 0 *and* β_{IB} *are required for the computation of* $\mathcal{L}_{\mathsf{IB}}(\beta)$*. In fact, when* Y *is a deterministic function of* X*, then* only $\beta = 0$ *and* $\beta = \beta_{\mathsf{IB}}$ *are required to compute the* IB *and other values of* β *are vacuous. The same argument can also be used to justify the inclusion of* β_{PF} *in computing* $\mathcal{L}_{\mathsf{PF}}(\beta)$*. Note also that since* $F_{\beta}(Q_X)$ *becomes convex for* $\beta > \beta_{\mathsf{IB}}$*, computing* $\mathcal{K}_{\cap}[F_{\beta}(Q_X)]$ *becomes trivial for such values of* β*.*

Remark 2. *Observe that the lower convex envelope of any function* f *can be obtained by taking Legendre-Fenchel transformation (aka. convex conjugate) twice. Hence, one can use the existing linear-time algorithms for approximating Legendre-Fenchel transformation (e.g., [87,88]) for approximating* $\mathcal{K}_{\cup}[F_{\beta}(Q_X)]$*.*

Once $\mathcal{L}_{\mathsf{IB}}(\beta)$ and $\mathcal{L}_{\mathsf{PF}}(\beta)$ are computed, we can derive IB and PF via standard results in optimization (see [3] (Section IV) for more details):

$$\mathsf{IB}(R) = \inf_{\beta \in [0, \beta_{\mathsf{IB}}]} \beta R + \mathcal{L}_{\mathsf{IB}}(\beta), \tag{21}$$

and

$$\mathsf{PF}(r) = \sup_{\beta \in [\beta_{\mathsf{PF}}, \infty]} \beta r + \mathcal{L}_{\mathsf{PF}}(\beta). \tag{22}$$

Following the convex analysis approach outlined by Witsenhausen and Wyner [3], IB and PF can be directly computed from $\mathcal{L}_{\mathsf{IB}}(\beta)$ and $\mathcal{L}_{\mathsf{PF}}(\beta)$ by observing the following. Suppose for some β, $\mathcal{K}_{\cup}[F_{\beta}(Q_X)]$ (resp. $\mathcal{K}_{\cap}[F_{\beta}(Q_X)]$) at P_X is obtained by a convex combination of points $F_{\beta}(Q^i)$, $i \in [k]$ for some $Q^1, \ldots, Q^k$ in $\mathcal{P}(\mathcal{X})$, integer $k \geq 2$, and weights $\lambda_i \geq 0$ (with $\sum_i \lambda_i = 1$). Then $\sum_i \lambda_i Q^i = P_X$, and T^* with properties $P_{T^*}(i) = \lambda_i$ and $P_{X|T^*=i} = Q^i$ attains the minimum (resp. maximum) of $H(Y|T) - \beta H(X|T)$. Hence, $(I(X;T^*), I(Y;T^*))$ is a point on the upper (resp. lower) boundary of $\mathcal{M}$; implying that $\mathsf{IB}(R) = I(Y;T^*)$ for $R = I(X;T^*)$ (resp. $\mathsf{PF}(r) = I(Y;T^*)$ for $r = I(X;T^*)$). If for some β, $\mathcal{K}_{\cup}[F_{\beta}(Q_X)]$ at P_X coincides with $F_{\beta}[P_X]$, then this corresponds to $\mathcal{L}_{\mathsf{IB}}(\beta) = 0$. The same holds for $\mathcal{K}_{\cup}[F_{\beta}(Q_X)]$. Thus, all the information about the functional IB (resp. PF) is contained in the subset of the domain of $\mathcal{K}_{\cup}[F_{\beta}(Q_X)]$ (resp. $\mathcal{K}_{\cap}[F_{\beta}(Q_X)]$) over which it differs from $F_{\beta}(Q_X)$. We will revisit and generalize this approach later in Section 3.

We can now instantiate this for the binary symmetric case. Suppose X and Y are binary variables and $P_{Y|X}$ is binary symmetric channel with crossover probability δ, denoted by $\mathsf{BSC}(\delta)$ and defined as

$$\mathsf{BSC}(\delta) = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}, \tag{23}$$

for some $\delta \geq 0$. To describe the result in a compact fashion, we introduce the following notation: we let $h_{\mathsf{b}} : [0,1] \to [0,1]$ denote the binary entropy function, i.e., $h_{\mathsf{b}}(p) = -p \log p - (1-p)\log(1-p)$. Since this function is strictly increasing $[0, \frac{1}{2}]$, its inverse exists and is denoted by $h_{\mathsf{b}}^{-1} : [0,1] \to [0, \frac{1}{2}]$. Moreover, $a * b \coloneqq a(1-b) + b(1-a)$ for $a, b \in [0,1]$.

Lemma 4 (Mr. and Mrs. Gerber's Lemma)**.** *For* $X \sim \mathsf{Bernoulli}(p)$ *for* $p \leq \frac{1}{2}$ *and* $P_{Y|X} = \mathsf{BSC}(\delta)$ *for* $\delta \geq 0$*, we have*

$$\mathsf{IB}(R) = h_{\mathsf{b}}(p * \delta) - h_{\mathsf{b}}\left(\delta * h_{\mathsf{b}}^{-1}\big(h_{\mathsf{b}}(p) - R\big)\right), \tag{24}$$

and

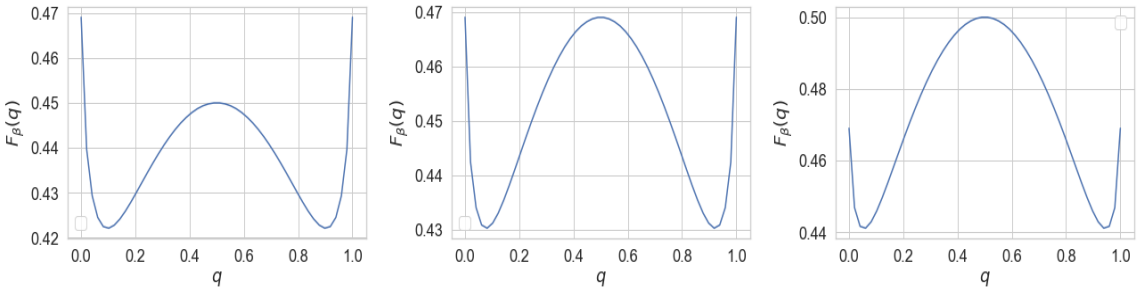
$$\mathsf{PF}(r) = h_{\mathsf{b}}(p * \delta) - \alpha h_{\mathsf{b}}\left(\delta * \frac{p}{z}\right) - \bar{\alpha} h_{\mathsf{b}}(\delta), \tag{25}$$

where $r = h_{\mathsf{b}}(p) - \alpha h_{\mathsf{b}}\left(\frac{p}{z}\right)$, $z = \max(\alpha, 2p)$, *and* $\alpha \in [0,1]$.

The result in (24) was proved by Wyner and Ziv [2] and is widely known as Mrs. Gerber's Lemma in information theory. Due to the similarity, we refer to (25) as Mr. Gerber's Lemma. As described above, to prove (24) and (25) it suffices to derive the convex and concave envelopes of the mapping $F_\beta : [0,1] \to \mathbb{R}$ given by

$$F_\beta(q) \coloneqq F_\beta(Q_X) = h_{\mathsf{b}}(q * \delta) - \beta h_{\mathsf{b}}(q), \tag{26}$$

where $q * \delta \coloneqq q\bar{\delta} + \delta\bar{q}$ is the output distribution of $\mathsf{BSC}(\delta)$ when the input distribution is $\mathsf{Bernoulli}(q)$ for some $q \in (0,1)$. It can be verified that $\beta_{\mathsf{IB}} \leq (1-2\delta)^2$. This function is depicted in Figure 4 depending of the values of $\beta \leq (1-2\delta)^2$.



(**a**) $\beta = 0.55$ (**b**) $\beta = 0.53$ (**c**) $\beta = 0.49$

Figure 4. The mapping $q \mapsto F_\beta(q) = H(Y') - \beta H(X')$ where $X' \sim \mathsf{Bernoulli}(q)$ and Y' is the result of passing X' through $\mathsf{BSC}(0.1)$, see (26).

2.3. *Operational Meaning of* IB *and* PF

In this section, we illustrate several information-theoretic settings which shed light on the operational interpretation of both IB and PF. The operational interpretation of IB has recently been extensively studied in information-theoretic settings in [29,30]. In particular, it was shown that IB specifies the rate-distortion region of noisy source coding problem [19,89] under the logarithmic loss as the distortion measure and also the rate region of the lossless source coding with side information at the decoder [90]. Here, we state the former setting (as it will be useful for our subsequent analysis of cardinality bound) and also provide a new information-theoretic setting in which IB appears as the solution. Then, we describe another setting, the so-called dependence dilution, whose achievable rate region has an extreme point specified by PF. This in fact delineate an important difference between IB and PF: while IB describes the entire rate-region of an information-theoretic setup, PF specifies only a corner point of a rate region. Other information-theoretic settings related to IB and PF include CEO problem [91] and source coding for the Gray-Wyner network [92].

2.3.1. Noisy Source Coding

Suppose Alice has access only to a noisy version X of a source of interest Y. She wishes to transmit a rate-constrained description from her observation (i.e., X) to Bob such that he can recover Y with small average distortion. More precisely, let (X^n, Y^n) be n i.i.d. samples of $(X,Y) \sim P_{XY}$. Alice encodes her observation X^n through an encoder $\phi : \mathcal{X}^n \to \{1, \ldots, K_n\}$ and sends $\phi(X^n)$ to Bob. Upon receiving $\phi(X^n)$, Bob reconstructs a "soft" estimate of Y^n via a decoder $\psi : \{1, \ldots, K_n\} \to \widehat{\mathcal{Y}}^n$ where $\widehat{\mathcal{Y}} = \mathcal{P}(\mathcal{Y})$.

That is, the reproduction sequence $\hat{y}^n$ consists of n probability measures on $\mathcal{Y}$. For any source and reproduction sequences y^n and $\hat{y}^n$, respectively, the distortion is defined as

$$d(y^n, \hat{y}^n) := \frac{1}{n}\sum_{i=1}^{n} d(y_i, \hat{y}_i),$$

where

$$d(y, \hat{y}) := \log \frac{1}{\hat{y}(y)}. \tag{27}$$

We say that a pair of rate-distortion (R, D) is achievable if there exists a pair (ϕ, ψ) of encoder and decoder such that

$$\limsup_{n\to\infty} \mathbb{E}[d(Y^n, \psi(\phi(X^n)))] \le \mathsf{D}, \qquad \text{and} \qquad \limsup_{n\to\infty} \frac{1}{n}\log K_n \le \mathsf{R}. \tag{28}$$

The noisy rate-distortion function $\mathsf{R}^{\mathsf{noisy}}(\mathsf{D})$ for a given $\mathsf{D} \ge 0$, is defined as the minimum rate R such that (R, D) is an achievable rate-distortion pair. This problem arises naturally in many data analytic problems. Some examples include feature selection of a high-dimensional dataset, clustering, and matrix completion. This problem was first studied by Dobrushin and Tsybakov [19], who showed that $\mathsf{R}^{\mathsf{noisy}}(\mathsf{D})$ is analogous to the classical rate-distortion function

$$\mathsf{R}^{\mathsf{noisy}}(\mathsf{D}) = \inf_{\substack{P_{\hat{Y}|X}:\mathbb{E}[d(Y,\hat{Y})]\le \mathsf{D},\\ Y-\circ-X-\circ-\hat{Y}}} I(X;\hat{Y}). \tag{29}$$

It can be easily verified that $\mathbb{E}[d(Y,\hat{Y})] = H(Y|\hat{Y})$ and hence (after relabeling $\hat{Y}$ as T)

$$\mathsf{R}^{\mathsf{noisy}}(\mathsf{D}) = \inf_{\substack{P_{T|X}:I(Y;T)\ge R,\\ Y-\circ-X-\circ-T}} I(X;T), \tag{30}$$

where $R = H(Y) - \mathsf{D}$, which is equal to $\widetilde{\mathsf{IB}}$ defined in (4). For more details in connection between noisy source coding and IB, the reader is referred to [29,30,91,93]. Notice that one can study an essentially identical problem where the distortion constraint (28) is replaced by

$$\lim_{n\to\infty} \frac{1}{n} I(Y^n; \psi(\phi(X^n)))] \ge R, \qquad \text{and} \qquad \limsup_{n\to\infty} \frac{1}{n}\log K_n \le \mathsf{R}.$$

This problem is addressed in [94] for discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ and extended recently in [95] for any general alphabets.

2.3.2. Test against Independence with Communication Constraint

As mentioned earlier, the connection between IB and noisy source coding, described above, was known and studied in [29,30]. Here, we provide a new information-theoretic setting which provides yet another operational meaning for IB. Given n i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from joint distribution Q, we wish to test whether X_i are independent of Y_i, that is, Q is a product distribution. This task is formulated by the following hypothesis test:

$$\begin{aligned} H_0 &: \; Q = P_{XY}, \\ H_1 &: \; Q = P_X P_Y, \end{aligned} \tag{31}$$

for a given joint distribution P_{XY} with marginals P_X and P_Y. Ahlswede and Csiszár [20] investigated this problem under a communication constraint: While Y observations (i.e., $Y_1, \ldots, Y_n$) are available,

the X observations need to be compressed at rate R, that is, instead of X^n, only $\phi(X^n)$ is present where $\phi : \mathcal{X}^n \to \{1, \ldots, K_n\}$ satisfies

$$\frac{1}{n} \log K_n \leq R.$$

For the type I error probability not exceeding a fixed $\varepsilon \in (0,1)$, Ahlswede and Csiszár [20] derived the smallest possible type 2 error probability, defined as

$$\beta_R(n,\varepsilon) = \min_{\substack{\phi:\mathcal{X}^n \to [K] \\ \frac{1}{n}\log K_n \leq R}} \min_{A \subset [K_n] \times \mathcal{Y}^n} \Big\{ (P_{\phi(X^n)} \times P_{Y^n})(A) : \ P_{\phi(X^n)\times Y^n}(A) \geq 1 - \varepsilon \Big\}.$$

The following gives the asymptotic expression of $\beta_R(n,\varepsilon)$ for every $\varepsilon \in (0,1)$. For the proof, refer to [20] (Theorem 3).

Theorem 7 ([20]). *For every $R \geq 0$ and $\varepsilon \in (0,1)$, we have*

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_R(n,\varepsilon) = \mathsf{IB}(R).$$

In light of this theorem, $\mathsf{IB}(R)$ specifies the exponential rate at which the type II error probability of the hypothesis test (31) decays as the number of samples increases.

2.3.3. Dependence Dilution

Inspired by the problems of information amplification [96] and state masking [97], Asoodeh et al. [21] proposed the dependence dilution setup as follows. Consider a source sequences X^n of n i.i.d. copies of $X \sim P_X$. Alice observes the source X^n and wishes to encode it via the encoder

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\},$$

for some $R > 0$. The goal is to ensure that any user observing $f_n(X^n)$ can construct a list, of fixed size, of sequences in $\mathcal{X}^n$ that contains likely candidates of the actual sequence X^n while revealing negligible information about a correlated source Y^n. To formulate this goal, consider the decoder

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to 2^{\mathcal{X}^n},$$

where $2^{\mathcal{X}^n}$ denotes the power set of $\mathcal{X}^n$. A *dependence dilution triple* $(R, \Gamma, \Delta) \in \mathbb{R}_+^3$ is said to be achievable if, for any $\delta > 0$, there exists a pair of encoder and decoder (f_n, g_n) such that for sufficiently large n

$$\Pr\left(X^n \notin g_n(J)\right) < \delta, \tag{32}$$

having fixed size $|g_n(J)| = 2^{n(H(X)-\Gamma)}$, where $J = f_n(X^n)$ and simultaneously

$$\frac{1}{n} I(Y^n; J) \leq \Delta + \delta. \tag{33}$$

Notice that without side information J, the decoder can only construct a list of size $2^{nH(X)}$ which contains X^n with probability close to one. However, after J is observed and the list $g_n(J)$ is formed, the decoder's list size can be reduced to $2^{n(H(X)-\Gamma)}$ and thus reducing the uncertainty about X^n by $n\Gamma \in [0, nH(X)]$. This observation can be formalized to show (see [96] for details) that the constraint (32) is equivalent to

$$\frac{1}{n} I(X^n; J) \geq \Gamma - \delta, \tag{34}$$

which lower bounds the amount of information J carries about X^n. Built on this equivalent formulation, Asoodeh et al. [21] (Corollary 15) derived a necessary condition for the achievable dependence dilution triple.

Theorem 8 ([21]). *Any achievable dependence dilution triple (R, Γ, Δ) satisfies*

$$\begin{cases} R & \geq \Gamma \\ \Gamma & \leq I(X;T) \\ \Delta & \geq I(Y;T) - I(X;T) + \Gamma, \end{cases}$$

for some auxiliary random variable T satisfying $Y \multimap X \multimap T$ and taking $|\mathcal{T}| \leq |\mathcal{X}| + 1$ values.

According to this theorem, $\mathsf{PF}(\Gamma)$ specifies the best privacy performance of the dependence dilution setup for the maximum amplification rate Γ. While this informs the operational interpretation of PF, Theorem 8 only provides an outer bound for the set of achievable dependence dilution triple (R, Γ, Δ). It is, however, not clear that PF characterizes the rate region of an information-theoretic setup.

The fact that IB fully characterizes the rate-region of an source coding setup has an important consequence: the cardinality of the auxiliary random variable T in IB can be improved to $|\mathcal{X}|$ instead of $|\mathcal{X}| + 1$.

2.4. Cardinality Bound

Recall that in the definition of IB in (4), no assumption was imposed on the auxiliary random variable T. A straightforward application of Carathéodory-Fenchel-Eggleston theorem (see e.g., [98] (Section III) or [79] (Lemma 15.4)) reveals that IB is attained for T taking values in a set $\mathcal{T}$ with cardinality $|\mathcal{T}| \leq |\mathcal{X}| + 1$. Here, we improve this bound and show $|\mathcal{T}| \leq |\mathcal{X}|$ is sufficient.

Theorem 9. *For any joint distribution P_{XY} and $R \in (0, H(X)]$, information bottleneck $\mathsf{IB}(R)$ is achieved by T taking at most $|\mathcal{X}|$ values.*

The proof of this theorem hinges on the operational characterization of IB as the lower boundary of the rate-distortion region of noisy source coding problem discussed in Section 2.3. Specifically, we first show that the extreme points of this region is achieved by T taking $|\mathcal{X}|$ values. We then make use of a property of the noisy source coding problem (namely, time-sharing) to argue that all points of this region (including the boundary points) can be attained by such T. It must be mentioned that this result was already claimed by Harremoës and Tishby in [99] without proof.

In many practical scenarios, feature X has a large alphabet. Hence, the bound $|\mathcal{T}| \leq |\mathcal{X}|$, albeit optimal, still can make the information bottleneck function computationally intractable over large alphabets. However, label Y usually has a significantly smaller alphabet. While it is in general impossible to have a cardinality bound for T in terms of $|\mathcal{Y}|$, one can consider approximating IB assuming T takes N values. The following result, recently proved by Hirche and Winter [100], is in this spirit.

Theorem 10 ([100]). *For any $(X, Y) \sim P_{XY}$, we have*

$$\mathsf{IB}(R, N) \leq \mathsf{IB}(R) \leq \mathsf{IB}(R, N) + \delta(N),$$

where $\delta(N) = 4N^{-\frac{1}{|\mathcal{Y}|}} \left[\log \frac{|\mathcal{Y}|}{4} + \frac{1}{|\mathcal{Y}|} \log N\right]$ and $\mathsf{IB}(R, N)$ denotes the information bottleneck functional (4) with the additional constraint that $|\mathcal{T}| \leq N$.

Recall that, unlike PF, the graph of IB characterizes the rate region of a Shannon-theoretic coding problem (as illustrated in Section 2.3), and hence any boundary points can be constructed

via time-sharing of extreme points of the rate region. This lack of operational characterization of PF translates into a worse cardinality bound than that of IB. In fact, for PF the cardinality bound $|\mathcal{T}| \leq |\mathcal{X}| + 1$ cannot be improved in general. To demonstrate this, we numerically solve the optimization in PF assuming that $|\mathcal{T}| = |\mathcal{X}|$ when both X and Y are binary. As illustrated in Figure 5, this optimization does not lead to a convex function, and hence, cannot be equal to PF.

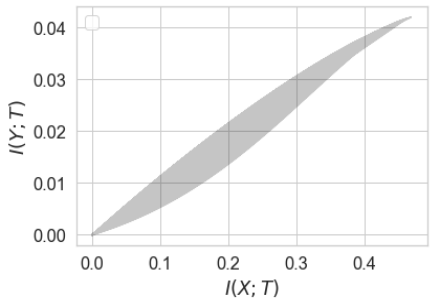


Figure 5. The set $\{(I(X;T), I(Y;T))\}$ with $P_X = \mathsf{Bernoulli}(0.9)$, $P_{Y|X=0} = [0.9, 0.1]$, $P_{Y|X=1} = [0.85, 0.15]$, and T restricted to be binary. While the upper boundary of this set is concave, the lower boundary is not convex. This implies that, unlike IB, $\mathsf{PF}(r)$ cannot be attained by binary variables T.

2.5. Deterministic Information Bottleneck

As mentioned earlier, IB formalizes an information-theoretic approach to clustering high-dimensional feature X into cluster labels T that preserve as much information about the label Y as possible. The clustering label is assigned by the soft operator $P_{T|X}$ that solves the IB formulation (4) according to the rule: $X = x$ is likely assigned label $T = t$ if $D_{\mathsf{KL}}(P_{Y|x} \| P_{Y|t})$ is small where $P_{Y|t} = \sum_x P_{Y|x} P_{X|t}$. That is, clustering is assigned based on the similarity of conditional distributions. As in many practical scenarios, a hard clustering operator is preferred, Strouse and Schwab [31] suggested the following variant of IB, termed as deterministic information bottleneck dIB

$$\mathsf{dIB}(P_{XY}, R) := \sup_{\substack{f:\mathcal{X}\to\mathcal{T}, \\ H(f(X))\leq R}} I(Y; f(X)), \tag{35}$$

where the maximization is taken over all deterministic functions f whose range is a finite set $\mathcal{T}$. Similarly, one can define

$$\mathsf{dPF}(P_{XY}, r) := \inf_{\substack{f:\mathcal{X}\to\mathcal{T}, \\ H(f(X))\geq r}} I(Y; f(X)). \tag{36}$$

One way to ensure that $H(f(X)) \leq R$ for a deterministic function f is to restrict the cardinality of the range of f: if $f : \mathcal{X} \to [e^R]$ then $H(f(X))$ is necessarily smaller than R. Using this insight, we derive a lower for $\mathsf{dIB}(P_{XY}, R)$ in the following lemma.

Lemma 5. *For any given P_{XY}, we have*

$$\mathsf{dIB}(P_{XY}, R) \geq \frac{e^R - 1}{|\mathcal{X}|} I(X;Y),$$

and

$$\mathsf{dPF}(P_{XY}, r) \leq \frac{e^r - 1}{|\mathcal{X}|} I(X;Y) + \Pr(X \geq e^r) \log \frac{1}{\Pr(X \geq e^r)}.$$

Note that both R and r are smaller than $H(X)$ and thus the multiplicative factors of $I(X;Y)$ in the lemma are smaller than one. In light of this lemma, we can obtain

$$\frac{e^R-1}{|\mathcal{X}|}I(X;Y) \le \mathsf{IB}(R) \le I(X;Y),$$

and

$$\mathsf{PF}(r) \le \frac{e^r-1}{|\mathcal{X}|}I(X;Y) + \Pr(X \ge e^r)\log\frac{1}{\Pr(X \ge e^r)}.$$

In most of practical setups, $|\mathcal{X}|$ might be very large, making the above lower bound for IB vacuous. In the following lemma, we partially address this issue by deriving a bound independent of $\mathcal{X}$ when Y is binary.

Lemma 6. *Let P_{XY} be a joint distribution of arbitrary X and binary $Y \sim \mathsf{Bernoulli}(q)$ for some $q \in (0,1)$. Then, for any $R \ge \log 5$ we have*

$$\mathsf{dIB}(P_{XY},R) \ge I(X;Y) - 2\alpha h_{\mathsf{b}}\Big(\frac{I(X;Y)}{2\alpha(e^R-4)}\Big),$$

where $\alpha = \max\{\log\frac{1}{q}, \log\frac{1}{1-q}\}$.

3. Family of Bottleneck Problems

In this section, we introduce a family of bottleneck problems by extending IB and PF to a large family of statistical measures. Similar to IB and PF, these bottleneck problems are defined in terms of boundaries of a two-dimensional convex set induced by a joint distribution P_{XY}. Recall that $R \mapsto \mathsf{IB}(P_{XY},R)$ and $r \mapsto \mathsf{PF}(P_{XY},r)$ are the upper and lower boundary of the set $\mathcal{M}$ defined in (6) and expressed here again for convenience

$$\mathcal{M} = \big\{(I(X;T), I(Y;T)) : Y \multimap X \multimap T, (X,Y) \sim P_{XY}\big\}. \tag{37}$$

Since P_{XY} is given, $H(X)$ and $H(Y)$ are fixed. Thus, in characterizing $\mathcal{M}$ it is sufficient to consider only $H(X|T)$ and $H(Y|T)$. To generalize IB and PF, we must therefore generalize $H(X|T)$ and $H(Y|T)$.

Given a joint distribution P_{XY} and two non-negative real-valued functions $\Phi : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^+$ and $\Psi : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}^+$, we define

$$\Phi(X|T) := \mathbb{E}\Big[\Phi(P_{X|T})\Big] = \sum_{t\in\mathcal{T}} P_T(t)\Phi(P_{X|T=t}), \tag{38}$$

and

$$\Psi(Y|T) := \mathbb{E}\Big[\Psi(P_{Y|T})\Big] = \sum_{t\in\mathcal{T}} P_T(t)\Psi(P_{Y|T=t}). \tag{39}$$

When $X \sim P_X$ and $Y \sim P_Y$, we interchangeably write $\Phi(X)$ for $\Phi(P_X)$ and $\Phi(Y)$ for $\Psi(P_Y)$.

These definitions provide natural generalizations for Shannon's entropy and mutual information. Moreover, as we discuss later in Sections 3.2 and 3.3, it also can be specialized to represent a large family of popular information-theoretic and statistical measures. Examples include information and estimation theoretic quantities such as Arimoto's conditional entropy of order α for $\Phi(Q_X) = ||Q_X||_\alpha$, probability of correctly guessing for $\Phi(Q_X) = ||Q_X||_\infty$, maximal correlation for binary case, and f-information for $\Phi(Q_X)$ given by f-divergence. We are able to generate a family of bottleneck problems using different instantiations of $\Phi(X|T)$ and $\Psi(Y|T)$ in place of mutual information in IB and PF. As we argue later, these problems better capture the essence of "informativeness" and "privacy"; thus providing analytical and interpretable guarantees similar in spirit to IB and PF.

Computing these bottleneck problems in general boils down to the following optimization problems

$$\mathsf{U}_{\Phi,\Psi}(\zeta) \coloneqq \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ \Phi(X|T) \le \zeta}} \Psi(Y|T), \tag{40}$$

and

$$\mathsf{L}_{\Phi,\Psi}(\zeta) \coloneqq \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ \Phi(X|T) \ge \zeta}} \Psi(Y|T). \tag{41}$$

Consider the set

$$\mathcal{M}_{\Phi,\Psi} \coloneqq \big\{(\Phi(X|T), \Psi(Y|T)) : Y \multimap X \multimap T, (X,Y) \sim P_{XY}\big\}. \tag{42}$$

Note that if both Φ and Ψ are continuous (with respect to the total variation distance), then $\mathcal{M}_{\Phi,\Psi}$ is compact. Moreover, it can be easily verified that $\mathcal{M}_{\Phi,\Psi}$ is convex. Hence, its upper and lower boundaries are well-defined and are characterized by the graphs of $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$, respectively. As mentioned earlier, these functional are instrumental for computing the general bottleneck problem later. Hence, before we delve into the examples of bottleneck problems, we extend the approach given in Section 2.2 to compute $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$.

3.1. Evaluation of $\mathsf{U}_{\Phi,\Psi}$ *and* $\mathsf{L}_{\Phi,\Psi}$

Analogous to Section 2.2, we first introduce the Lagrangians of $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$ as

$$\mathcal{L}^{\mathsf{U}}_{\Phi,\Psi}(\beta) \coloneqq \sup_{P_{T|X}} \Psi(Y|T) - \beta\Phi(X|T), \tag{43}$$

and

$$\mathcal{L}^{\mathsf{L}}_{\Phi,\Psi}(\beta) \coloneqq \inf_{P_{T|X}} \Psi(Y|T) - \beta\Phi(X|T), \tag{44}$$

where $\beta \ge 0$ is the Lagrange multiplier, respectively. Let (X', Y') be a pair of random variable with $X' \sim Q_X$ and Y' is the result of passing X' through the channel $P_{Y|X}$. Letting

$$F^{\Phi,\Psi}_{\beta}(Q_X) \coloneqq \Psi(Y') - \beta\Phi(X'), \tag{45}$$

we obtain that

$$\mathcal{L}^{\mathsf{U}}_{\Phi,\Psi}(\beta) = \mathcal{K}_{\cap}[F^{\Phi,\Psi}_{\beta}(Q_X)]\big|_{P_X} \qquad \text{and} \qquad \mathcal{L}^{\mathsf{L}}_{\Phi,\Psi}(\beta) = \mathcal{K}_{\cup}[F^{\Phi,\Psi}_{\beta}(Q_X)]\big|_{P_X}, \tag{46}$$

recalling that $\mathcal{K}_{\cap}$ and $\mathcal{K}_{\cup}$ are the upper concave and lower convex envelop operators. Once we compute $\mathcal{L}^{\mathsf{U}}_{\Phi,\Psi}$ and $\mathcal{L}^{\mathsf{L}}_{\Phi,\Psi}$ for all $\beta \ge 0$, we can use the standard results in optimizations theory (similar to (21) and (22)) to recover $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$. However, we can instead extend the approach Witsenhausen and Wyner [3] described in Section 2.2. Suppose for some β, $\mathcal{K}_{\cap}[F^{\Phi,\Psi}_{\beta}(Q_X)]$ (resp. $\mathcal{K}_{\cup}[F^{\Phi,\Psi}_{\beta}(Q_X)]$) at P_X is obtained by a convex combination of points $F^{\Phi,\Psi}_{\beta}(Q^i)$, $i \in [k]$ for some $Q^1, \dots, Q^k$ in $\mathcal{P}(\mathcal{X})$, integer $k \ge 2$, and weights $\lambda_i \ge 0$ (with $\sum_i \lambda_i = 1$). Then $\sum_i \lambda_i Q^i = P_X$, and T^* with properties $P_{T^*}(i) = \lambda_i$ and $P_{X|T^*=i} = Q^i$ attains the maximum (resp. minimum) of $\Psi(Y|T) - \beta\Phi(X|T)$, implying that $(\Phi(X|T^*), \Psi(Y|T^*))$ is a point on the upper (resp. lower) boundary of $M_{\Phi,\Psi}$. Consequently, such T^* satisfies $\mathsf{U}_{\Phi,\Psi}(\zeta) = \Psi(Y|T^*)$ for $\zeta = \Phi(X|T^*)$ (resp. $\mathsf{L}_{\Phi,\Psi}(\zeta) = \Psi(Y|T^*)$ for $\zeta = \Phi(X|T^*)$). The algorithm to compute $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$ is then summarized in the following three steps:

- Construct the functional $F^{\Phi,\Psi}_{\beta}(Q_X) \coloneqq \Psi(Y') - \beta\Phi(X')$ for $X' \sim Q_X$ and $Y' \sim Q_X P_{Y|X}$ and all $Q_X \in \mathcal{P}(\mathcal{X})$ and $\beta \ge 0$.
- Compute $\mathcal{K}_{\cap}[F^{\Phi,\Psi}(Q_X)]\big|_{P_X}$ and $\mathcal{K}_{\cup}[F^{\Phi,\Psi}(Q_X)]\big|_{P_X}$ evaluated at P_X.

- If for distributions $Q^1, \ldots, Q^k$ in $\mathcal{P}(X)$ for some $k \geq 1$, we have $\mathcal{K}_{\cap}[F^{\Phi,\Psi}(Q_X)]\big|_{P_X} = \sum_{i=1}^k \lambda_i F^{\Phi,\Psi}(Q^i)$ or $\mathcal{K}_{\cup}[F^{\Phi,\Psi}(Q_X)]\big|_{P_X} = \sum_{i=1}^k \lambda_i F^{\Phi,\Psi}(Q^i)$ for some $\lambda_i \geq 0$ satisfying $\sum_{i=1}^k \lambda_i = 1$, then then $P_{X|T=i} = Q_i$, $i \in [k]$ and $P_T(i) = \lambda_i$ give the optimal T^* in $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$, respectively.

We will apply this approach to analytically compute $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$ (and the corresponding bottleneck problems) for binary cases in the following sections.

3.2. Guessing Bottleneck Problems

Let P_{XY} be given with marginals P_X and P_Y and the corresponding channel $P_{Y|X}$. Let also $Q_X \in \mathcal{P}(\mathcal{X})$ be an arbitrary distribution on $\mathcal{X}$ and $Q_Y = Q_X P_{Y|X}$ be the output distribution of $P_{Y|X}$ when fed with Q_X. Any channel $P_{T|X}$, together with the Markov structure $Y \multimap X \multimap T$, generates unique $P_{X|T}$ and $P_{Y|T}$. We need the following basic definition from statistics.

Definition 1. *Let U be a discrete and V be an arbitrary random variables supported on $\mathcal{U}$ and $\mathcal{V}$ with $|\mathcal{U}| < \infty$, respectively. Then $\mathsf{P_c}(U)$ the probability of correctly guessing U and $\mathsf{P_c}(U|V)$ the probability of correctly guessing U given V are given by*

$$\mathsf{P_c}(U) \coloneqq \max_{u \in \mathcal{U}} P_U(u),$$

and

$$\mathsf{P_c}(U|V) \coloneqq \max_{g} \Pr(U = g(V)) = \mathbb{E}\left[\max_{u \in \mathcal{U}} P_{U|V}(u|V)\right].$$

Moreover, the multiplicative gain of the observation V in guessing U is defined as (the reason for ∞ in the notation becomes clear later)

$$I_\infty(U;V) \coloneqq \log \frac{\mathsf{P_c}(U|V)}{\mathsf{P_c}(U)}.$$

As the names suggest, $\mathsf{P_c}(U|V)$ and $\mathsf{P_c}(U)$ characterize the optimal efficiency of guessing U with or without the observation V, respectively. Intuitively, $I_\infty(U;V)$ quantifies how useful the observation V is in estimating U: If it is small, then it means it is nearly as hard for an adversary observing V to guess U as it is without V. This observation motivates the use of $I_\infty(Y;T)$ as a measure of privacy in lieu of $I(Y;T)$ in PF.

It is worth noting that $I_\infty(U;V)$ is not symmetric in general, i.e., $I_\infty(U;V) \neq I_\infty(V;U)$. Since observing T can only improve, we have $\mathsf{P_c}(Y|T) \geq \mathsf{P_c}(Y)$; thus $I_\infty(Y;T) \geq 0$. However, $I_\infty(Y;T) = 0$ does not necessarily imply independent of Y and T; instead, it means T is useless in estimating Y. As an example, consider $Y \sim \mathsf{Bernoulli}(p)$ and $P_{T|Y=0} = \mathsf{Bernoulli}(\delta)$ and $P_{T|Y=1} = \mathsf{Bernoulli}(\eta)$ with $\delta, \eta \leq \frac{1}{2} < p$. Then $\mathsf{P_c}(Y) = p$ and

$$\mathsf{P_c}(Y|T) = \max\{\bar{\delta}\bar{p}, \eta p\} + \bar{\eta} p.$$

Thus, if $\bar{\delta}\bar{p} \leq \eta p$, then $\mathsf{P_c}(Y|T) = \mathsf{P_c}(Y)$. This then implies that $I_\infty(Y;T) = 0$ whereas Y and T are clearly dependent; i.e, $I(Y;T) > 0$. While in general $I(Y;T)$ and $I_\infty(Y;T)$ are not related, it can be shown that $I(Y;T) \leq I_\infty(Y;T)$ if Y is uniform (see [65] (Proposition 1)). Hence, only with this uniformity assumption, $I_\infty(Y;T)$ implies the independence.

Consider $\Psi(Q_X) = -\sum_{x \in \mathcal{X}} Q_X(x) \log(Q_X(x))$ and $\Psi(Q_Y) = \|Q_Y\|_\infty$. Clearly, we have $\Phi(X|T) = H(X|T)$. Note that

$$\Psi(Y|T) = \sum_{t \in \mathcal{T}} P_T(t) ||P_{Y|T=t}||_\infty = \mathsf{P_c}(Y|T), \tag{47}$$

thus both measures $H(X|T)$ and $\mathsf{P_c}(Y|T)$ are special cases of the models described in the previous section. In particular, we can define the corresponding $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$. We will see later that $I(X;T)$ and $\mathsf{P_c}(Y|T)$ correspond to Arimoto's mutual information of orders 1 and ∞, respectively. Define

$$\mathsf{IB}^{(\infty,1)}(R) := \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) \le R}} I_\infty(Y;T). \tag{48}$$

This bottleneck functional formulated an interpretable guarantee:

$$\mathsf{IB}^{(\infty,1)}(R) \text{ characterizes the best error probability in recovering } Y \text{ among all } R\text{-bit summaries of } X$$

Recall that the functional $\mathsf{PF}(r)$ aims at extracting maximum information of X while protecting privacy with respect to Y. Measuring the privacy in terms of $\mathsf{P_c}(Y|T)$, this objective can be better formulated by

$$\mathsf{PF}^{(\infty,1)}(r) := \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I(X;T) \ge r}} I_\infty(Y;T), \tag{49}$$

with the interpretable privacy guarantee:

$$\mathsf{PF}^{(\infty,1)}(r) \text{ characterizes the smallest probability of revealing private feature } Y \text{ among all representations of } X \text{ preserving at least } r \text{ bits information of } X$$

Notice that the variable T in the formulations of $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$ takes values in a set $\mathcal{T}$ of arbitrary cardinality. However, a straightforward application of the Carathéodory-Fenchel-Eggleston theorem (see e.g., [79] (Lemma 15.4)) reveals that the cardinality of $\mathcal{T}$ can be restricted to $|\mathcal{X}|+1$ without loss of generality. In the following lemma, we prove more basic properties of $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$.

Lemma 7. *For any P_{XY} with Y supported on a finite set $\mathcal{Y}$, we have*

- $\mathsf{IB}^{(\infty,1)}(0) = \mathsf{PF}^{(\infty,1)}(0) = 0$.
- $\mathsf{IB}^{(\infty,1)}(R) = I_\infty(X;Y)$ *for any* $R \ge H(X)$ *and* $\mathsf{PF}^{(\infty,1)}(r) = I_\infty(X;Y)$ *for* $r \ge H(X)$.
- $R \mapsto \exp(\mathsf{IB}^{(\infty,1)}(R))$ *is strictly increasing and concave on the range* $(0, I_\infty(X;Y))$.
- $r \mapsto \exp(\mathsf{PF}^{(\infty,1)}(r))$ *is strictly increasing, and convex on the range* $(0, I_\infty(X;Y))$.

The proof follows the same lines as Theorem 1 and hence omitted. Lemma 7 in particular implies that inequalities $I(X;T) \le R$ and $I(X;T) \ge r$ in the definition of $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$ can be replaced by $I(X;T) = R$ and $I(X;T) = r$, respectively. It can be verified that I^∞ satisfies the data-processing inequality, i.e., $I^\infty(Y;T) \le I^\infty(Y;X)$ for the Markov chain $Y \multimap X \multimap T$. Hence, both $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$ must be smaller than $I_\infty(Y;X)$. The properties listed in Lemma 7 enable us to derive a slightly tighter upper bound for $\mathsf{PF}^{(\infty,1)}$ as demonstrated in the following.

Lemma 8. *For any P_{XY} with Y supported on a finite set $\mathcal{Y}$, we have*

$$\mathsf{PF}^{(\infty,1)}(r) \le \log\left[1 + \frac{r}{H(X)}\left(e^{I_\infty(Y;X)} - 1\right)\right],$$

and

$$\log\left[1 + \frac{R}{H(X)}\left(e^{I_\infty(Y;X)} - 1\right)\right] \le \mathsf{IB}^{(\infty,1)}(R) \le I_\infty(Y;X).$$

The proof of this lemma (and any other results in this section) is given in Appendix B. This lemma shows that the gap between $I_\infty(Y;X)$ and $\mathsf{IB}^{(\infty,1)}(R)$ when R is sufficiently close to $H(X)$ behaves like

$$I_\infty(Y;X) - \mathsf{IB}^{(\infty,1)}(R) \leq I_\infty(Y;X) - \log\left[1 + \frac{R}{H(X)}\left(e^{I_\infty(Y;X)} - 1\right)\right] \approx \left(1 - e^{-I_\infty(Y;X)}\right)\left(1 - \frac{R}{H(X)}\right).$$

Thus, $\mathsf{IB}^{(\infty,1)}(R)$ approaches $I_\infty(Y;X)$ as $R \to H(X)$ at least linearly.

In the following theorem, we apply the technique delineated in Section 3.1 to derive closed form expressions for $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$ for the binary symmetric case, thereby establishing similar results as Mr and Mrs. Gerber's Lemma.

Theorem 11. *For $X \sim \mathsf{Bernoulli}(p)$ and $P_{Y|X} = \mathsf{BSC}(\delta)$ with $p, \delta \leq \frac{1}{2}$, we have*

$$\mathsf{PF}^{(\infty,1)}(r) = \log\left[\frac{\bar{\delta} - (h_\mathsf{b}(p) - r)(\frac{1}{2} - \delta)}{1 - \delta * p}\right], \tag{50}$$

and

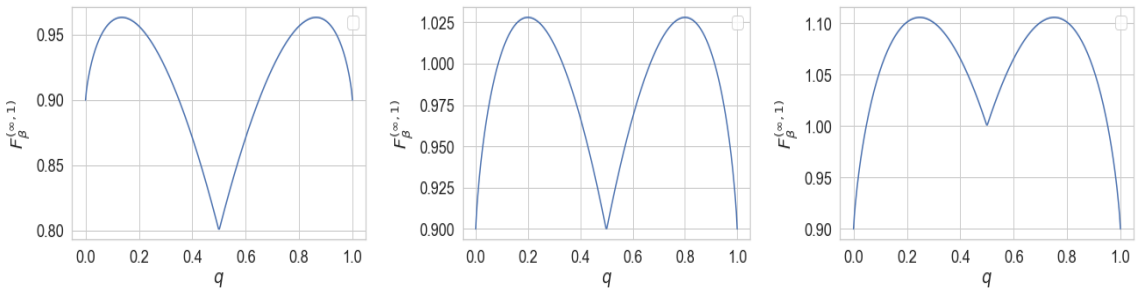
$$\mathsf{IB}^{(\infty,1)}(R) = \log\left[\frac{1 - \delta * h_\mathsf{b}^{-1}(h_\mathsf{b}(p) - R)}{1 - \delta * p}\right], \tag{51}$$

where $\bar{\delta} = 1 - \delta$.

As described in Section 3.1, to compute $\mathsf{IB}^{(\infty,1)}$ and $\mathsf{PF}^{(\infty,1)}$ it suffices to derive the convex and concave envelopes of the mapping $F_\beta^{(\infty,1)}(q) := \mathsf{P_c}(Y') + \beta H(X')$ where $X' \sim \mathsf{Bernoulli}(q)$ and Y' is the result of passing X' through $\mathsf{BSC}(\delta)$, i.e., $Y' \sim \mathsf{Bernoulli}(\delta * q)$. In this case, $\mathsf{P_c}(Y') = \max\{\delta * q, 1 - \delta * q\}$ and $F_\beta^{(\infty,1)}$ can be expressed as

$$q \mapsto F_\beta^{(\infty,1)}(q) = \max\{\delta * q, 1 - \delta * q\} + \beta h_\mathsf{b}(q). \tag{52}$$

This function is depicted in Figure 6.



(**a**) $\beta = 0.2$ (**b**) $\beta = 0.0.4$ (**c**) $\beta = 0.5$

Figure 6. The mapping $q \mapsto F_\beta^{(\infty,1)}(q) = \mathsf{P_c}(Y') + \beta H(X')$ where $X' \sim \mathsf{Bernoulli}(q)$ and $Y' \sim \mathsf{Bernoulli}(q)\mathsf{BSC}(0.1)$.

The detailed derivation of convex and concave envelope of $F_\beta^{(\infty,1)}$ is given in Appendix B. The proof of this theorem also reveals the following intuitive statements. If $X \sim \mathsf{Bernoulli}(p)$ and $P_{Y|X} = \mathsf{BSC}(\delta)$, then among all random variables T satisfying $Y \multimap X \multimap T$ and $H(X|T) \leq \lambda$, the minimum $\mathsf{P_c}(Y|T)$ is given by $\bar{\delta} - \lambda(0.5 - \delta)$. Notice that, without any information constraint (i.e., $\lambda = 0$), $\mathsf{P_c}(Y|T) = \mathsf{P_c}(Y|X) = \bar{\delta}$. Perhaps surprisingly, this shows that the mutual information constraint has a linear effect on the privacy of Y. Similarly, to prove (51), we show that among all R-bit representations T of X, the best achievable accuracy $\mathsf{P_c}(Y|T)$ is given by $1 - \delta * h_\mathsf{b}^{-1}(h_\mathsf{b}(p) - R)$. This can be proved by combining Mrs. Gerber's Lemma (cf. Lemma 4) and Fano's inequality as

follows. For all T such that $H(X|T) \geq \lambda$, the minimum of $H(Y|T)$ is given by $h_{\mathsf{b}}(\delta * h_{\mathsf{b}}^{-1}(\lambda))$. Since by Fano's inequality, $H(Y|T) \leq h_{\mathsf{b}}(1 - \mathsf{P}_{\mathsf{c}}(Y|T))$, we obtain $\delta * h_{\mathsf{b}}^{-1}(\lambda) \leq 1 - \mathsf{P}_{\mathsf{c}}(Y|T)$ which leads to the same result as above. Nevertheless, in Appendix B we give another proof based on the discussion of Section 3.1.

3.3. Arimoto Bottleneck Problems

The bottleneck framework proposed in the last section benefited from interpretable guarantees brought forth by the quantity I_∞. In this section, we define a parametric family of statistical quantities, the so-called Arimoto's mutual information, which includes both Shannon's mutual information and I_∞ as extreme cases.

Definition 2 ([22]). *Let $U \sim P_U$ and $V \sim P_V$ be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. Their Arimoto's mutual information of order $\alpha > 1$ is defined as*

$$I_\alpha(U;V) = H_\alpha(U) - H_\alpha(U|V), \tag{53}$$

where

$$H_\alpha(U) := \frac{\alpha}{1-\alpha} \log ||P_U||_\alpha, \tag{54}$$

is the Rényi entropy of order α and

$$H_\alpha(U|V) := \frac{\alpha}{1-\alpha} \log \sum_{v \in \mathcal{V}} P_V(v) ||P_{U|V=v}||_\alpha, \tag{55}$$

is the Arimoto's conditional entropy of order α.

By continuous extension, one can define $I - \alpha(U;V)$ for $\alpha = 1$ and $\alpha = \infty$ as $I(U;V)$ and $I_\infty(U;V)$, respectively. That is,

$$\lim_{\alpha \to 1^+} I_\alpha(U;V) = I(U;V), \qquad \text{and} \qquad \lim_{\alpha \to \infty} I_\alpha(U;V) = I_\infty(U;V). \tag{56}$$

Arimoto's mutual information was first introduced by Arimoto [22] and then later revisited by Liese and Vajda in [101] and more recently by Verdú in [102]. More in-depth analysis and properties of I_α can be found in [103]. It is shown in [71] (Lemma 1) that $I_\alpha(U;V)$ for $\alpha \in [1, \infty]$ quantifies the minimum loss in recovering U given V where the loss is measured in terms of the so-called α-loss. This loss function reduces to logarithmic loss (27) and $\mathsf{P}_{\mathsf{c}}(U|V)$ for $\alpha = 1$ and $\alpha = \infty$, respectively. This sheds light on the utility and/or privacy guarantee promised by a constraint on Arimoto's mutual information. It is now natural to use I_α for defining a family of bottleneck problems.

Definition 3. *Given a pair of random variables $(X,Y) \sim P_{XY}$ over finite sets $\mathcal{X}$ and $\mathcal{Y}$ and $\alpha, \gamma \in [1, \infty]$, we define $\mathsf{IB}^{(\alpha,\gamma)}$ and $\mathsf{PF}^{(\alpha,\gamma)}$ as*

$$\mathsf{IB}^{(\alpha,\gamma)}(R) := \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I_\gamma(X;T) \leq R}} I_\alpha(Y;T), \tag{57}$$

and

$$\mathsf{PF}^{(\alpha,\gamma)}(r) := \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I_\gamma(X;T) \geq r}} I_\alpha(Y;T), \tag{58}$$

Of course, $\mathsf{IB}^{(1,1)}(R) = \mathsf{IB}(R)$ and $\mathsf{PF}^{(1,1)}(r) = \mathsf{PF}(r)$. It is known that Arimoto's mutual information satisfies the data-processing inequality [103] (Corollary 1), i.e., $I_\alpha(Y;T) \leq I_\alpha(Y;X)$ for the Markov chain $Y \multimap X \multimap T$. On the other hand, $I_\gamma(X;T) \leq H_\gamma(X)$. Thus, both $\mathsf{IB}^{(\alpha,\gamma)}(R)$

and $\mathsf{PF}^{(\alpha,\gamma)}(r)$ equal $I_\alpha(Y;X)$ for $R, r \geq H_\gamma(X)$. Note also that $H_\alpha(Y|T) = \frac{\alpha}{1-\alpha}\log \Psi(Y|T)$ where $\Psi(Y|T)$ (see (39)) corresponding to the function $\Psi(Q_Y) = ||Q_Y||_\alpha$. Consequently, $\mathsf{IB}^{(\alpha,\gamma)}$ and $\mathsf{PF}^{(\alpha,\gamma)}$ are characterized by the lower and upper boundary of $\mathcal{M}_{\Phi,\Psi}$, defined in (42), with respect to $\Phi(Q_X) = ||Q_X||_\gamma$ and $\Psi(Q_Y) = ||Q_Y||_\alpha$. Specifically, we have

$$\mathsf{IB}^{(\alpha,\gamma)}(R) = H_\alpha(Y) + \frac{\alpha}{\alpha-1}\log \mathsf{U}_{\Phi,\Psi}(\zeta), \tag{59}$$

where $\zeta = e^{-(1-\frac{1}{\gamma})(H_\gamma(X)-R)}$, and

$$\mathsf{PF}^{(\alpha,\gamma)}(r) = H_\alpha(Y) + \frac{\alpha}{\alpha-1}\log \mathsf{L}_{\Phi,\Psi}(\zeta), \tag{60}$$

where $\zeta = e^{-(1-\frac{1}{\gamma})(H_\gamma(X)-r)}$ and $\Phi(Q_X) = \|Q_X\|_\gamma$ and $\Psi(Q_Y) = \|Q_Y\|_\alpha$. This paves the way to apply the technique described in Section 2.2 to compute $\mathsf{IB}^{(\alpha,\gamma)}$ and $\mathsf{PF}^{(\alpha,\gamma)}$. Doing so requires the upper concave and lower convex envelope of the mapping $Q_X \mapsto \|Q_Y\|_\alpha - \beta\|Q_X\|_\gamma$ for some $\beta \geq 0$, where $Q_Y \sim Q_X P_{Y|X}$. In the following theorem, we drive these envelopes and give closed form expressions for $\mathsf{IB}^{(\alpha,\gamma)}$ and $\mathsf{PF}^{(\alpha,\gamma)}$ for a special case where $\alpha = \gamma \geq 2$.

Theorem 12. *Let $X \sim \mathsf{Bernoulli}(p)$ and $P_{Y|X} = \mathsf{BSC}(\delta)$ with $p, \delta \leq \frac{1}{2}$. We have for $\alpha \geq 2$*

$$\mathsf{PF}^{(\alpha,\alpha)}(r) = \frac{\alpha}{1-\alpha}\log\left[\frac{\|p * \delta\|_\alpha}{\|q * \delta\|_\alpha}\right],$$

where $\|a\|_\alpha := \|[a, \bar{a}]\|_\alpha$ for $a \in [0,1]$ and $q \leq p$ solves

$$\frac{\alpha}{1-\alpha}\log\left[\frac{\|p\|_\alpha}{\|q\|_\alpha}\right] = r.$$

Moreover,

$$\mathsf{IB}^{(\alpha,\alpha)}(R) = \frac{\alpha}{\alpha-1}\log\left[\frac{\bar{\lambda}\|\delta\|_\alpha + \lambda\|\frac{q}{z} * \delta\|_\alpha}{\|p * \alpha\|_\alpha}\right],$$

where $z = \max\{2p, \lambda\}$ and $\lambda \in [0,1]$ solves

$$\frac{\alpha}{\alpha-1}\log\frac{\bar{\lambda} + \lambda\|\frac{p}{z}\|_\alpha}{\|p\|_\alpha} = R.$$

By letting $\alpha \to \infty$, this theorem indicates that for X and Y connected through $\mathsf{BSC}(\delta)$ and all variables T forming $Y \multimap X \multimap T$, we have

$$\mathsf{P_c}(X|T) \geq \lambda \quad \Longrightarrow \quad \mathsf{P_c}(Y|T) \geq \delta * \lambda, \tag{61}$$

which can be shown to be achieved T^* generated by the following channel (see Figure 7)

$$P_{T^*|X} = \begin{bmatrix} \frac{\lambda-p}{\bar{p}} & \frac{\bar{\lambda}}{\bar{p}} \\ 0 & 1 \end{bmatrix}. \tag{62}$$

Note that, by assumption, $p \leq \frac{1}{2}$, and hence the event $\{X = 1\}$ is less likely than $\{X = 0\}$. Therefore, (61) demonstrates that to ensure correct recoverability of X with probability at lest λ, the most private approach (with respect to Y) is to obfuscate the higher-likely event $\{X = 0\}$ with probability $\frac{\bar{\lambda}}{\bar{p}}$. As demonstrated in (61) the optimal privacy guarantee is linear in the utility parameter in the binary symmetric case. This is in fact a special case of the larger result recently proved in [65]

(Theorem 1): the infimum of $\mathsf{P_c}(Y|T)$ over all variables T such that $\mathsf{P_c}(X|T) \geq \lambda$ is piece-wise linear in λ, on equivalently, the mapping $e^r \mapsto \exp(\mathsf{PF}^{(\infty,\infty)}(r))$ is piece-wise linear.

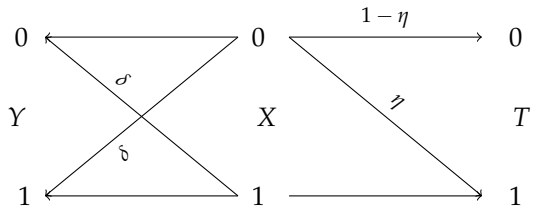


Figure 7. The structure of the optimal $P_{T|X}$ for $\mathsf{PF}^{(\infty,\infty)}$ when $P_{Y|X} = \mathsf{BSC}(\delta)$ and $X \sim \mathsf{Bernoulli}(p)$ with $\delta, p \in [0, \frac{1}{2}]$. If the accuracy constraint is $\mathsf{P_c}(X|T) \geq \lambda$ (or equivalently $I_\infty(X|T) \geq \log \frac{\lambda}{\bar{p}}$), then the parameter of optimal $P_{T|X}$ is given by $\eta = \frac{\bar{\lambda}}{\bar{p}}$, leading to $\mathsf{P_c}(Y|T) = \delta * \lambda$.

Computing $\mathsf{PF}^{(\alpha,\gamma)}$ analytically for every $\alpha, \gamma > 1$ seems to be challenging, however, the following lemma provides bounds for $\mathsf{PF}^{(\alpha,\gamma)}$ and $\mathsf{IB}^{(\alpha,\gamma)}$ in terms of $\mathsf{PF}^{(\infty,\infty)}$ and $\mathsf{IB}^{(\infty,\infty)}$, respectively.

Lemma 9. *For any pair of random variables (X, Y) over finite alphabets and $\alpha, \gamma > 1$, we have*

$$\frac{\alpha}{\alpha-1}\mathsf{PF}^{(\infty,\infty)}(f(r)) - \frac{\alpha}{\alpha-1}H_\infty(Y) + H_\alpha(Y) \leq \mathsf{PF}^{(\alpha,\gamma)}(r) \leq \mathsf{PF}^{(\infty,\infty)}(g(r)) + H_\alpha(Y) - H_\infty(Y),$$

and

$$\frac{\alpha}{\alpha-1}\mathsf{IB}^{(\infty,\infty)}(f(R)) - \frac{\alpha}{\alpha-1}H_\infty(Y) + H_\alpha(Y) \leq \mathsf{IB}^{(\alpha,\gamma)}(R) \leq \mathsf{IB}^{(\infty,\infty)}(g(R)) + H_\alpha(Y) - H_\infty(Y),$$

where $f(a) = \max\{a - H_\gamma(X) + H_\infty(X), 0\}$ and $g(b) = \frac{\gamma-1}{\gamma}b + H_\infty(X) - \frac{\gamma-1}{\gamma}H_\gamma(X)$.

The previous lemma can be directly applied to derive upper and lower bounds for $\mathsf{PF}^{(\alpha,\gamma)}$ and $\mathsf{IB}^{(\alpha,\gamma)}$ given $\mathsf{PF}^{(\infty,\infty)}$ and $\mathsf{IB}^{(\infty,\infty)}$.

3.4. f-Bottleneck Problems

In this section, we describe another instantiation of the general framework introduced in terms of functions Φ and Ψ that enjoys interpretable estimation-theoretic guarantee.

Definition 4. *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Furthermore, let U and V be two real-valued random variables supported over $\mathcal{U}$ and $\mathcal{V}$, respectively. Their f-information is defined by*

$$I_f(U;V) := D_f(P_{UV} \| P_U P_V), \tag{63}$$

where $D_f(\cdot\|\cdot)$ is the f-divergence [104] between distributions and defined as

$$D_f(P\|Q) := \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right].$$

Due to convexity of f, we have $D_f(P\|Q) \geq f(1) = 0$ and hence f-information is always non-negative. If, furthermore, f is strictly convex at 1, then equality holds if and only $P = Q$. Csiszár introduced f-divergence in [104] and applied it to several problems in statistics and information theory. More recent developments about the properties of f-divergence and f-information can be found in [23] and the references therein. Any convex function f with the property $f(1) = 0$ results in an f-information. Popular examples include $f(t) = t \log t$ corresponding to Shannon's mutual information, $f(t) = |t - 1|$ corresponding to T-information [83], and also $f(t) = t^2 - 1$ corresponding to χ^2-information [69] for. It is worth mentioning that if we allow α to be in $(0, 1)$ in Definition 2

(similar to [101]), then the resulting Arimoto's mutual information can be shown to be an f-information in the binary case for a certain function f, see [101] (Theorem 8).

Let $(X,Y) \sim P_{XY}$ be given with marginals P_X and P_Y. Consider functions Φ and Ψ on $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ defined as

$$\Phi(Q_X) := D_f(Q_X\|P_X) \qquad \text{and} \qquad \Psi(Q_Y) := D_f(Q_Y\|P_Y).$$

Given a conditional distribution $P_{T|X}$, it is easy to verify that $\Phi(X|T) = I_f(X;T)$ and $\Psi(Y|T) = I_f(Y;T)$. This in turn implies that f-information can be utilized in (40) and (41) to define general bottleneck: Let $f : (0,\infty) \to \mathbb{R}$ and $g : (0,\infty) \to \mathbb{R}$ be two convex functions satisfying $f(1) = g(1) = 0$. Then we define

$$\mathsf{IB}^{(f,g)}(R) := \sup_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I_g(X;T) \le R}} I_f(Y;T), \tag{64}$$

and

$$\mathsf{PF}^{(f,g)}(r) := \inf_{\substack{P_{T|X}: Y \multimap X \multimap T \\ I_g(X;T) \ge r}} I_f(Y;T). \tag{65}$$

In light of the discussion in Section 3.1, the optimization problems in $\mathsf{IB}^{(f,g)}$ and $\mathsf{IB}^{(f,g)}$ can be analytically solved by determining the upper concave and lower convex envelope of the mapping

$$Q_X \mapsto \mathcal{F}_\beta^{(f,g)} := D_f(Q_Y\|P_Y) - \beta D_g(Q_X\|P_X), \tag{66}$$

where $\beta \ge 0$ is the Lagrange multiplier and $Q_Y = Q_X P_{Y|X}$.

Consider the function $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}$ with $\alpha \in (1,\infty) \cup (1,\infty)$. The corresponding f-divergence is sometimes called Hellinger divergence of order α, see e.g., [105]. Note that Hellinger divergence of order 2 reduces to χ^2-divergence. Calmon et al. [68] and Asoodeh et al. [67] showed that if $I_{f_2}(Y;T) \le \varepsilon$ for some $\varepsilon \in (0,1)$, then the minimum mean-squared error (MMSE) of reconstructing any zero-mean unit-variance function of Y given T is lower bounded by $1-\varepsilon$, i.e., no function of Y can be reconstructed with small MMSE given an observation of T. This result serves a natural justification for I_{f_2} as an operational measure of both privacy and utility in a bottleneck problem.

Unfortunately, our approach described in Section 3.1 cannot be used to compute $\mathsf{IB}^{(f_2,f_2)}$ or $\mathsf{PF}^{(f_2,f_2)}$ in the binary symmetric case. The difficulty lies in the fact that the function $\mathcal{F}_\beta^{f_2,f_2}$, defined in (66), for the binary symmetric case is either convex or concave on its entire domain depending on the value of β. Nevertheless, one can consider Hellinger divergence of order α with $\alpha \ne 2$ and then apply our approach to compute $\mathsf{IB}^{(f_\alpha,f_\alpha)}$ or $\mathsf{PF}^{(f_\alpha,f_\alpha)}$. Since $D_{f_2}(P\|Q) \le (1+(\alpha-1)D_{f_\alpha}(P\|Q))^{1/(\alpha-1)} - 1$ (see [106] (Corollary 5.6)), one can justify I_{f_α} as a measure of privacy and utility in a similar way as I_{f_2}.

We end this section by a remark about estimating the measures studied in this section. While we consider information-theoretic regime where the underlying distribution P_{XY} is known, in practice only samples (x_i, y_i) are given. Consequently, the de facto guarantees of bottleneck problems might be considerably different from those shown in this work. It is therefore essential to asses the guarantees of bottleneck problems when accessing only samples. To do so, one must derive bounds on the discrepancy between $\mathsf{P_c}$, I_α, and I_f computed on the empirical distribution and the true (unknown) distribution. These bounds can then be used to shed light on the de facto guarantee of the bottleneck problems. Relying on [34] (Theorem 1), one can obtain that the gaps between the measures $\mathsf{P_c}$, I_α, and I_f computed on empirical distributions and the true one scale as $O(1/\sqrt{n})$ where n is the number of samples. This is in contrast with mutual information for which the similar upper bound scales as $O(\log n/\sqrt{n})$ as shown in [33]. Therefore, the above measures appear to be easier to estimate than mutual information.

4. Summary and Concluding Remarks

Following the recent surge in the use of information bottleneck (IB) and privacy funnel (PF) in developing and analyzing machine learning models, we investigated the functional properties of these two optimization problems. Specifically, we showed that IB and PF correspond to the upper and lower boundary of a two-dimensional convex set $\mathcal{M} = \{(I(X;T), I(Y;T)) : Y \multimap X \multimap T\}$ where $(X,Y) \sim P_{XY}$ represents the observable data X and target feature Y and the auxiliary random variable T varies over all possible choices satisfying the Markov relation $Y \multimap X \multimap T$. This unifying perspective on IB and PF allowed us to adapt the classical technique of Witsenhausen and Wyner [3] devised for computing IB to be applicable for PF as well. We illustrated this by deriving a closed form expression for PF in the binary case—a result reminiscent of the Mrs. Gerber's Lemma [2] in information theory literature. We then showed that both IB and PF are closely related to several information-theoretic coding problems such as noisy random coding, hypothesis testing against independence, and dependence dilution. While these connections were partially known in previous work (see e.g., [29,30]), we show that they lead to an improvement on the cardinality of T for computing IB. We then turned our attention to the continuous setting where X and Y are continuous random variables. Solving the optimization problems in IB and PF in this case without any further assumptions seems a difficult challenge in general and leads to theoretical results only when (X,Y) is jointly Gaussian. Invoking recent results on the entropy power inequality [25] and strong data processing inequality [27], we obtained tight bounds on IB in two different cases: (1) when Y is a Gaussian perturbation of X and (2) when X is a Gaussian perturbation of Y. We also utilized the celebrated I-MMSE relationship [107] to derive a second-order approximation of PF when T is considered to be a Gaussian perturbation of X.

In the second part of the paper, we argue that the choice of (Shannon's) mutual information in both IB and PF does not seem to carry specific operational significance. It does, however, have a desirable practical consequence: it leads to self-consistent equations [1] that can be solved iteratively (without any guarantee to convergence though). In fact, this property is unique to mutual information among other existing information measures [99]. Nevertheless, we argued that other information measures might lead to better interpretable guarantee for both IB and PF. For instance, statistical accuracy in IB and privacy leakage in PF can be shown to be precisely characterized by probability of correctly guessing (aka Bayes risk) or minimum mean-squared error (MMSE). Following this observation, we introduced a large family of optimization problems, which we call bottleneck problems, by replacing mutual information in IB and PF with Arimoto's mutual information[22] or f-information [23]. Invoking results from [33,34], we also demonstrated that these information measures are in general easier to estimate from data than mutual information. Similar to IB and PF, the bottleneck problems were shown to be fully characterized by boundaries of a two-dimensional convex set parameterized by two real-valued non-negative functions Φ and Ψ. This perspective enabled us to generalize the technique used to compute IB and PF for evaluating bottleneck problems. Applying this technique to the binary case, we derived closed form expressions for several bottleneck problems.

Author Contributions: All authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the National Science Foundation under Grant No. CIF 1900750 and CIF CAREER 1845852.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs from Section 2

Proof of Theorem 1.

- Note that $R = 0$ in optimization problem (4) implies that X and T are independent. Since Y, X and T form Markov chain $Y \multimap Y \multimap T$, independent of X and T implies independence of Y and T and thus $I(Y;T) = 0$. Similarly for $\mathsf{PF}(0)$.

- Since $I(X;T) \leq H(X)$ for any random variable T, we have $T = X$ satisfies the information constraint $I(X;T) \leq R$ for $R \geq H(X)$. Since $I(Y;T) \leq I(Y;X)$, this choice is optimal. Similarly for PF, the constraint $I(X;T) \geq r$ for $r \geq H(X)$ implies $T = X$. Hence, $\mathsf{PF}(r) = I(Y;X)$.
- The upper bound on IB follows from the data processing inequality: $I(Y;T) \leq \min\{I(X;T), I(X;Y)\}$ for all T satisfying the Markov condition $Y \multimap X \multimap T$.
- To prove the lower bound on PF, note that

$$I(Y;T) = I(X;T) - I(X;T|Y) \geq I(X;T) - H(X|Y).$$

- The concavity of $R \mapsto \mathsf{IB}(R)$ follows from the fact it is the upper boundary of the convex set $\mathcal{M}$, defined in (6). This in turn implies the continuity of $\mathsf{IB}(\cdot)$. Monotonicity of $R \mapsto \mathsf{IB}(R)$ follows from the definition. Strict monotonicity follows from the convexity and the fact that $\mathsf{IB}(H(X)) = I(X;Y)$.
- Similar as above.
- The differentiability of the map $R \mapsto \mathsf{IB}(R)$ follows from [94] (Lemma 6). This result in fact implies the differentiability of the map $r \mapsto \mathsf{PF}(r)$ as well. Continuity of the derivative of IB and PF on $(0, H(X))$ is a straightforward application of [108] (Theorem 25.5).
- Monotonicity of mappings $R \mapsto \frac{\mathsf{IB}(R)}{R}$ and $r \mapsto \frac{\mathsf{PF}(r)}{r}$ follows from the concavity and convexity of $\mathsf{IB}(\cdot)$ and $\mathsf{PF}(\cdot)$, respectively.
- Strict monotonicity of $\mathsf{IB}(\cdot)$ and $\mathsf{PF}(\cdot)$ imply that the optimization problems in (4) and (5) occur when the inequality in the constraints becomes equality.

□

Proof of Theorem 3. Recall that, according to Theorem 1, the mappings $R \mapsto \mathsf{IB}(R)$ and $r \mapsto \mathsf{PF}(r)$ are concave and convex, respectively. This implies that $\mathsf{IB}(R)$ (resp. $\mathsf{PF}(r)$) lies above (resp. below) the chord connecting $(0,0)$ and $(H(X), I(X;Y))$. This proves the lower bound (resp. upper bound) $\mathsf{IB}(R) \geq R\frac{I(X;Y)}{H(X)}$ (resp. $\mathsf{PF}(r) \leq r\frac{I(X;Y)}{H(X)}$).

In light of the convexity of PF and monotonicity of $r \mapsto \frac{\mathsf{PF}(r)}{r}$, we can write

$$\frac{\mathsf{PF}(r)}{r} \geq \lim_{r\to 0^+} \frac{\mathsf{PF}(r)}{r} = \inf_{r\neq 0} \frac{\mathsf{PF}(r)}{r} = \inf_{\substack{P_{T|X}: Y\multimap X\multimap T \\ I(X;T)\neq 0}} \frac{I(Y;T)}{I(X;T)} = \inf_{\mathcal{P}(\mathcal{X})\ni Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)},$$

where the last equality is due to [13] (Lemma 4) and Q_Y is the output distribution of the channel $P_{Y|X}$ when the input is distributed according to Q_X. Similarly, we can write

$$\frac{\mathsf{IB}(R)}{R} \leq \lim_{R\to 0^+} \frac{\mathsf{IB}(R)}{R} = \sup_{R\neq 0} \frac{\mathsf{IB}(R)}{R} = \sup_{\substack{P_{T|X}: Y\multimap X\multimap T \\ I(X;T)\neq 0}} \frac{I(Y;T)}{I(X;T)} = \sup_{\mathcal{P}(\mathcal{X})\ni Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)},$$

where the last equality is due to [82] (Theorem 4). □

Proof of Theorem 5. Let T_n be an optimal summeries of X^n, that is, it satisfies $T_n \multimap X^n \multimap Y^n$ and $I(X^n;T_n) = nR$. We can write

$$I(X^n,T_n) = H(X^n) - H(X^n|T_n) = \sum_{k=1}^n \Big[H(X_k) - H(X_k|X^{k-1},T_n)\Big] = \sum_{k=1}^n I(X_k;X^{k-1},T_n),$$

and hence, if $R_k := I(X_k;X^{k-1},T_n)$, then we have

$$R = \frac{1}{n}\sum_{k=1}^n R_k. \tag{A1}$$

We can similarly write

$$\begin{aligned} I(Y^n, T_n) &= H(Y^n) - H(Y^n|T_n) = \sum_{k=1}^{n} \Big[H(Y_k) - H(Y_k|Y^{k-1}, T_n)\Big] \\ &\le \sum_{k=1}^{n} \Big[H(Y_k) - H(Y_k|Y^{k-1}, X^{k-1}, T_n)\Big] \\ &= \sum_{k=1}^{n} \Big[H(Y_k) - H(Y_k|X^{k-1}, T_n)\Big] = \sum_{k=1}^{n} I(Y_k; X^{k-1}, T_n). \end{aligned}$$

Since we have $(T_n, X^{k-1}) \multimap X_k \multimap Y_k$ for every $k \in [n]$, we conclude from the above inequality that

$$I(Y^n, T_n) \le \sum_{k=1}^{n} I(Y_k; X^{k-1}, T_n) \le \sum_{k=1}^{n} \mathsf{IB}(P_{XY}, R_k) \le n\mathsf{IB}(P_{XY}, R), \tag{A2}$$

where the last inequality follows from concavity of the map $x \mapsto \mathsf{IB}(P_{XY}, x)$ and (A1). Consequently, we obtain

$$\mathsf{IB}(P_{X^nY^n}, nR) \le n\mathsf{IB}(P_{XY}, R). \tag{A3}$$

To prove the other direction, let $P_{T|X}$ be an optimal channel in the definition of IB, i.e., $I(X;T) = R$ and $\mathsf{IB}(P_{XY}, R) = I(Y;T)$. Then using this channel n times for each pair (X_i, Y_i), we obtain $T^n = (T_1, \ldots, T_n)$ satisfying $T^n \multimap X^n \multimap Y^n$. Since $I(X^n; T^n) = nI(X;T) = nR$ and $I(Y^n; T^n) = nI(Y;T)$, we have $\mathsf{IB}(P_{X^nY^n}, nR) \ge n\mathsf{IB}(P_{XY}, R)$. This, together with (A3), concludes the proof.

□

Proof of Theorem 4. First notice that

$$\lim_{R\to 0} \frac{\mathsf{IB}(R)}{R} = \sup_{R>0} \frac{\mathsf{IB}(R)}{R} = \sup_{\substack{P_{T|X}: \\ Y\multimap X\multimap T}} \frac{I(Y;T)}{I(X;T)} = \sup_{\substack{Q_X\in\mathcal{P}(\mathcal{X}) \\ Q_X\neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)},$$

where the last equality is due to [82] (Theorem 4). Similarly,

$$\lim_{r\to 0} \frac{\mathsf{PF}(r)}{r} = \inf_{r>0} \frac{\mathsf{PF}(r)}{r} = \inf_{\substack{P_{T|X}: \\ Y\multimap X\multimap T}} \frac{I(Y;T)}{I(X;T)} = \inf_{\substack{Q_X\in\mathcal{P}(\mathcal{X}) \\ Q_X\neq P_X}} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)},$$

where the last equality is due to [13] (Lemma 4).

Fix $x_0 \in \mathcal{X}$ with $P_X(x_0) > 0$ and let T be a Bernoulli random variable specified by the following channel

$$P_{T|X}(1|x) = \delta 1_{\{x=x_0\}},$$

for some $\delta > 0$. This channel induces $T \sim \mathsf{Bernoulli}(\delta P_X(x_0))$, $P_{Y|T}(y|1) = P_{Y|X}(y|x_0)$, and

$$P_{Y|T}(y|0) = \frac{P_Y(y) - P_{XY}(x_0, y)}{1 - \delta P_X(x_0)}.$$

It can be verified that

$$I(X;T) = -\delta P_X(x_0) \log P_X(x_0) + o(\delta),$$

and

$$I(Y;T) = \delta P_X(x_0) D_{\mathsf{KL}}(P_{Y|X}(\cdot|x_0)\|P_Y(\cdot)) + o(\delta).$$

Setting

$$\delta = \frac{r}{-P_X(x_0)\log P_X(x_0)},$$

we obtain

$$I(Y;T) = \frac{D_{\mathsf{KL}}(P_{Y|X}(\cdot|x_0)\|P_Y(\cdot))}{-\log P_X(x_0)} r + o(r),$$

and hence

$$\mathsf{PF}(r) \le \frac{D_{\mathsf{KL}}(P_{Y|X}(\cdot|x_0)\|P_Y(\cdot))}{-\log P_X(x_0)} r + o(r).$$

Since x_0 is arbitrary, the result follows. The proof for IB follows similarly. □

Proof of Lemma 1. When Y is an erasure of X, i.e., $\mathcal{Y} = \mathcal{X} \cup \{\perp\}$ with $P_{Y|X}(x|x) = 1-\delta$ and $P_{Y|X}(\perp|x) = \delta$, it is straightforward to verify that $D_{\mathsf{KL}}(Q_Y\|P_Y) = (1-\delta)D_{\mathsf{KL}}(Q_X\|P_X)$ for every P_X and Q_X in $\mathcal{P}(X)$. Consequently, we have

$$\inf_{Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)} = \sup_{Q_X \neq P_X} \frac{D_{\mathsf{KL}}(Q_Y\|P_Y)}{D_{\mathsf{KL}}(Q_X\|P_X)} = 1-\delta.$$

Hence, Theorem 3 gives the desired result.

To prove the second part, i.e., when X is an erasure of Y, we need an improved upper bound of PF. Notice that if perfect privacy occurs for a given P_{XY}, then the upper bound for $\mathsf{PF}(r)$ in Theorem 3 can be improved:

$$\mathsf{PF}(r) \le (r - r_0)\frac{I(X;Y)}{H(X) - r_0}, \tag{A4}$$

where r_0 is the largest $r \ge 0$ such that $\mathsf{PF}(r) = 0$. Here, we show that $r_0 = H(X|Y)$. This suffices to prove the result as (A4), together with Theorem 1, we have

$$\max\{r - H(X|Y), 0\} \le \mathsf{PF} \le (r - r_0)\frac{I(X;Y)}{H(X) - r_0} = (r - H(X|Y)).$$

To show that $\mathsf{PF}(H(X|Y)) = 0$, consider the channel $P_{T|X}(t|x) = \frac{1}{|\mathcal{Y}|}1_{\{t\neq\perp, x\neq\perp\}}$ and $P_{T|X}(\perp \mid \perp) = 1$. It can be verified that this channel induces T which is independent of Y and that

$$I(X;T) = H(T) - H(T|X) = H\left(\frac{1-\delta}{|\mathcal{Y}|}, \ldots, \frac{1-\delta}{|\mathcal{Y}|}, \delta\right) - (1-\delta)\log|\mathcal{Y}| = h_{\mathsf{b}}(\delta) = H(X|Y),$$

where $h_{\mathsf{b}}(\delta) := -\delta\log\delta - (1-\delta)\log(1-\delta)$ is the binary entropy function. □

Proof of Lemma 4. As mentioned earlier, equation (24) was proved in [2]. We thus give a proof only for (25).

Consider the problem of minimizing the Lagrangian $\mathcal{L}_{\mathsf{PF}}(\beta)$ (20) for $\beta \ge \beta_{\mathsf{PF}}$. Let $X' \sim Q_X = \text{Bernoulli}(q)$ for some $q \in (0,1)$ and Y' be the result of passing X' through $\mathsf{BSC}(\delta)$, i.e., $Y' \sim \text{Bernoulli}(q * \delta)$. Recall that $F_\beta(q) := F_\beta(Q_X) = h_{\mathsf{b}}(q * \delta) - \beta h_{\mathsf{b}}(q)$. It suffices to compute $\mathcal{K}_\cap[F_\beta(q)]$ the upper concave envelope of $q \mapsto F_\beta(q)$. It can be verified that $\beta_{\mathsf{IB}} \le (1-2\delta)^2$ and hence for all $\beta \ge (1-2\delta)^2$, $\mathcal{K}_\cap[F_\beta(q)] = F_\beta(0)$. A straightforward computation shows that $F_\beta(q)$ is symmetric around $q = \frac{1}{2}$ and is also concave in a region around $q = \frac{1}{2}$, where it reaches its local maximum. Hence, if β is such that

- $F_\beta(\frac{1}{2}) < F_\beta(0)$ (see Figure 4a), then $\mathcal{K}_\cap[F_\beta(q)]$ is given by the convex combination of $F_\beta(0)$ and $F_\beta(1)$.
- $F_\beta(\frac{1}{2}) = F_\beta(0)$ (see Figure 4b), then $\mathcal{K}_\cap[F_\beta(q)]$ is given by the convex combination of $F_\beta(0)$ and $F_\beta(\frac{1}{2})$ and $F_\beta(1)$.
- $F_\beta(\frac{1}{2}) > F_\beta(0)$ (see Figure 4c), then there exists $q_\beta \in [0, \frac{1}{2}]$ such that for $q \le q_\beta$, $\mathcal{K}_\cap[F_\beta(q)]$ is given by the convex combination of $F_\beta(0)$ and $F_\beta(q_\beta)$.

Hence, assuming $p \le \frac{1}{2}$, we can construct T^* that maximizes $H(Y|T) - \beta H(X|T)$ in three different cases corresponding three cases above:

- In the first case, T^* is binary and we have $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$ and $P_{X|T^*=1} = \mathsf{Bernoulli}(1)$ with $P_{T^*} = \mathsf{Bernoulli}(p)$.
- In the second case, T^* is ternary and we have $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$, $P_{X|T^*=1} = \mathsf{Bernoulli}(1)$, and $P_{X|T^*=2} = \mathsf{Bernoulli}(\frac{1}{2})$ with $P_{T^*} = (1-p-\frac{\alpha}{2}, p-\frac{\alpha}{2}, \alpha)$ for some $\alpha \in [0, 2p]$.
- In the third case, T^* is again binary and we have $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$ and $P_{X|T^*=1} = \mathsf{Bernoulli}(\frac{p}{\alpha})$ with $P_{T^*} = \mathsf{Bernoulli}(\alpha)$ for some $\alpha \in [2p, 1]$.

Combining these three cases, we obtain the result in (25). □

Proof of Lemma 2. Let $X = Y + \sigma N^{\mathsf{G}}$ where $\sigma > 0$ and $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of Y. According to the improved entropy power inequality proved in [25] (Theorem 1), we can write

$$e^{2(H(X)-I(Y;T))} \geq e^{2(H(Y)-I(X;T))} + 2\pi e\sigma^2,$$

for any random variable T forming $Y \multimap X \multimap T$. This, together with Theorem 5, implies the result. □

Proof of Corollary 1. Since (X, Y) are jointly Gaussian, we can write $X = Y + \sigma N^{\mathsf{G}}$ where $\sigma = \sigma_Y \frac{\sqrt{1-\rho^2}}{\rho}$ and σ_Y^2 is the variance of Y. Applying Lemma 2 and noticing that $H(X) = \frac{1}{2}\log(2\pi e(\sigma_Y^2 + \sigma^2))$, we obtain

$$I(Y;T) \leq \frac{1}{2}\log\frac{\sigma^2+\sigma_Y^2}{\sigma^2+\sigma_Y^2 e^{-2I(X;T)}} = \frac{1}{1-\rho^2+\rho^2 e^{-2I(X;T)}}, \tag{A5}$$

for all channels $P_{T|X}$ satisfying $Y \multimap X \multimap T$. This bound is attained by Gaussian $P_{T|X}$. Specifically, assuming $T = X + \tilde{\sigma}M^{\mathsf{G}}$ where $\tilde{\sigma}^2 = \sigma_Y^2 \frac{e^{-2R}}{\rho^2(1-e^{-2R})}$ for $R \geq 0$ and $M^{\mathsf{G}} \sim \mathcal{N}(0, \tilde{\sigma}^2)$ independent of X, it can be easily verified that $I(X;T) = R$ and $I(Y;T) = \frac{1}{1-\rho^2+\rho^2 e^{-2R}}$. This, together with (A5), implies $\mathsf{IB}(R) = \frac{1}{1-\rho^2+\rho^2 e^{-2R}}$. □

Next, we wish to prove Theorem 6. However, we need the following preliminary lemma before we delve into its proof.

Lemma A1. *Let X and Y be continuous correlated random variables with $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$. Then the mappings $\sigma \mapsto I(X;T_\sigma)$ and $\sigma \mapsto I(Y;T_\sigma)$ are continuous, strictly decreasing, and*

$$I(X;T_\sigma) \to 0, \quad \textit{and} \quad I(Y;T_\sigma) \to 0 \quad \textit{as} \quad \sigma \to \infty.$$

Proof. The finiteness of $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ imply that $H(X)$ and $H(Y)$ are finite. A straightforward application of the entropy power inequality (cf. [109] (Theorem 17.7.3)) implies that $H(T_\sigma)$ is also finite. Thus, $I(X;T_\sigma)$ and $I(Y;T_\sigma)$ are well-defined. According to the data processing inequality, we have $I(X;T_{\sigma+\delta}) < I(X;T_\sigma)$ for all $\delta > 0$ and also $I(Y;T_{\sigma+\delta}) \leq I(Y;T_\sigma)$ where the equality occurs if and only if X and Y are independent. Since, bu assumption X and Y correlated, it follows $I(Y;T_{\sigma+\delta}) < I(Y;T_\sigma)$. Thus, both $I(X;T_\sigma)$ and $I(Y;T_\sigma)$ are strictly decreasing.

For the proof of continuity, we consider two cases $\sigma = 0$ and $\sigma > 0$ separately. We first give the poof for $I(X;T_\sigma)$. Since $H(\sigma N^{\mathsf{G}}) = \frac{1}{2}\log(2\pi e\sigma^2)$, we have $\lim_{\sigma\to 0} H(\sigma N^{\mathsf{G}}) = \infty$ and thus $\lim_{\sigma\to 0} I(X;T_\sigma) = \infty$ that is equal to $I(X;T_0)$. For $\sigma > 0$, let σ_n be a sequence of positive numbers converging to σ. In light of de Bruijn's identity (cf. [109] (Theorem 17.7.2)), we have $H(T_{\sigma_n}) \to H(T_\sigma)$, implying the continuity of $\sigma \mapsto I(X;T_\sigma)$.

Next, we prove the continuity of $\sigma \mapsto I(Y;T_\sigma)$. For the sequence of positive numbers σ_n converging to $\sigma > 0$, we have $I(Y;T_{\sigma_n}) = H(T_{\sigma_n}) - H(T_{\sigma_n}|Y)$. We only need to show $H(T_{\sigma_n}|Y) \to H(T_\sigma|Y)$. Invoking again de Brujin's identity, we obtain $H(T_{\sigma_n}|Y=y) \to H(T_\sigma|Y=y)$ for each $y \in \mathcal{Y}$. The desired result follows from dominated convergence theorem. Finally, the The continuity of

$\sigma \mapsto I(Y;T_\sigma)$ when $\sigma = 0$ follows from [110] (p. 2028) stating that $H(T_{\sigma_n}|Y=y) \to H(X|Y=y)$ and then applying dominated convergence theorem.

Note that

$$0 \le I(Y;T_\sigma) \le I(X;T_\sigma) \le \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma^2}\right),$$

where σ_X^2 is the variance of X and the last inequality follows from the fact that $I(X;X+\sigma N^{\mathsf{G}})$ is maximized when X is Gaussian. Since by assumption $\sigma_X < \infty$, it follows that both $I(X;T_\sigma)$ and $I(Y;T_\sigma)$ converge to zero as $\sigma \to \infty$. □

In light of this lemma, there exists a unique $\sigma \ge 0$ such that $I(X;T_\sigma) = r$. Let σ_r denote such σ. Therefore, we have $\mathsf{PF}(r) = I(Y;T_{\sigma_r})$. This enables us to prove Theorem 6.

Prof of Theorem 6. The proof relies on the I-MMSE relation in information theory literature. We briefly describe it here for convenience. Given any pair of random variables U and V, the minimum mean-squared error (MMSE) of estimating U given V is given by

$$\mathsf{mmse}(U|V) := \inf_f \mathbb{E}[(U-f(V))^2] = \mathbb{E}[(U-\mathbb{E}[U|V])^2] = \mathbb{E}[\mathsf{var}(U|V)],$$

where the infimum is taken over all measurable functions f and $\mathsf{var}(U|V) = \mathbb{E}[(U-\mathbb{E}[U|V])^2|V]$. Guo et al. [107] proved the following identity, which is referred to as I-MMSE formula, relating the input-output mutual information of the additive Gaussian channel $T_\sigma = X + \sigma N^{\mathsf{G}}$, where $N^{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of X, with the MMSE of the input given the output:

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)} I(X;T_\sigma) = -\frac{1}{2\sigma^4}\mathsf{mmse}(X|T_\sigma). \tag{A6}$$

Since Y, X, and T_σ form the Markov chain $Y \multimap X \multimap T_\sigma$, it follows that $I(Y;T_\sigma) = I(X;T_\sigma) - I(X;T_\sigma|Y)$. Thus, two applications of (A6) yields

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)} I(Y;T_\sigma) = -\frac{1}{2\sigma^4}\left[\mathsf{mmse}(X|T_\sigma) - \mathsf{mmse}(X|T_\sigma,Y)\right]. \tag{A7}$$

The second derivative of $I(X;T_\sigma)$ and $I(Y;T_\sigma)$ are also known via the formula [111] (Proposition 9)

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)}\mathsf{mmse}(X|T_\sigma) = \frac{1}{\sigma^4}\mathbb{E}[\mathsf{var}^2(X|T_\sigma)] \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}(\sigma^2)}\mathsf{mmse}(X|T_\sigma,Y) = \frac{1}{\sigma^4}\mathbb{E}[\mathsf{var}^2(X|T_\sigma,Y)]. \tag{A8}$$

With these results in mind, we now begin the proof. Recall that σ_r is the unique σ such that $I(X;T_\sigma) = r$, thus implying $\mathsf{PF}^{\mathsf{G}}(r) = I(Y;T_{\sigma_r})$. We have

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathsf{PF}^{\mathsf{G}}(r) = \left[\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)} I(Y;T_\sigma)\right]_{\sigma=\sigma_r} \frac{\mathrm{d}}{\mathrm{d}r}\sigma_r^2. \tag{A9}$$

To compute the derivative of $\mathsf{PF}(r)$, we therefore need to compute the derivative of σ_r^2 with respect to r. To do so, notice that from the identity $I(X;T_{\sigma_r}) = r$ we can obtain

$$1 = \frac{\mathrm{d}}{\mathrm{d}r} I(X;T_{\sigma_r}) = \left[\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)} I(X;T_{\sigma_r})\right]_{\sigma=\sigma_r} \frac{\mathrm{d}}{\mathrm{d}r}\sigma_r^2 = -\frac{1}{2\sigma^4}\mathsf{mmse}(X|T_{\sigma_r})\frac{\mathrm{d}}{\mathrm{d}r}\sigma_r^2,$$

implying

$$\frac{\mathrm{d}}{\mathrm{d}r}\sigma_r^2 = \frac{-2\sigma^4}{\mathsf{mmse}(X|T_{\sigma_r})}.$$

Plugging this identity into (A9) and invoking (A7), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathsf{PF}^{\mathsf{G}}(r) = \frac{\mathsf{mmse}(X|T_{\sigma_r}) - \mathsf{mmse}(X|T_{\sigma_r}, Y)}{\mathsf{mmse}(X|T_{\sigma_r})}. \tag{A10}$$

The second derivative can be obtained via (A8)

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}\mathsf{PF}^{\mathsf{G}}(r) = 2\frac{\mathbb{E}[\mathsf{var}^2(X|T_{\sigma_r}, Y)]}{\mathsf{mmse}^2(X|T_{\sigma_r})} - 2\mathbb{E}[\mathsf{var}^2(X|T_{\sigma_r})]\frac{\mathsf{mmse}(X|T_{\sigma_r}, Y)}{\mathsf{mmse}^3(X|T_{\sigma_r})}.$$

Since $\sigma_r \to \infty$ as $r \to 0$, we can write

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathsf{PF}^{\mathsf{G}}(r)\Big|_{r=0} = \frac{\sigma_X^2 - \mathbb{E}[\mathsf{var}(X|Y)]}{\sigma_X} = \frac{\mathsf{var}(\mathbb{E}[X|Y])}{\sigma_X^2} = \eta(X, Y),$$

where $\mathsf{var}(\mathbb{E}[X|Y])$ is the variance of the conditional expectation X given Y and the last equality comes from the law of total variance. and

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}\mathsf{PF}^{\mathsf{G}}(r)\Big|_{r=0} = \frac{2}{\sigma_X^4}\left[\mathbb{E}[\mathsf{var}^2(X|Y)] - \sigma_X^2\mathbb{E}[\mathsf{var}(X|Y)]\right].$$

Taylor expansion of $\mathsf{PF}(r)$ around $r = 0$ gives the result. □

Proof of Theorem 9. The main ingredient of this proof is a result by Jana [112] (Lemma 2.2) which provides a tight cardinality bound for the auxiliary random variables in the canonical problems in network information theory (including noisy source coding problem described in Section 2.3). Consider a pair of random variables $(X, Y) \sim P_{XY}$ and let $d : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ be an arbitrary distortion measure defined for arbitrary reconstruction alphabet $\hat{\mathcal{Y}}$.

Theorem A1 ([112]). *Let $\mathcal{A}$ be the set of all pairs (R, D) satisfying*

$$I(X;T) \le \mathsf{R} \qquad \textit{and} \qquad \mathbb{E}[d(Y, \psi(T))] \le \mathsf{D},$$

for some mapping $\psi : \mathcal{T} \to \hat{\mathcal{Y}}$ and some joint distributions $P_{XYT} = P_{XY}P_{T|X}$. Then every extreme points of $\mathcal{A}$ corresponds to some choice of auxiliary variable T with alphabet size $|\mathcal{T}| \le |\mathcal{X}|$.

Measuring the distortion in the above theorem in terms of the logarithmic loss as in (27), we obtain that

$$\mathcal{A} = \{(\mathsf{R}, \mathsf{D}) \in \mathbb{R}_+^2 : \ \mathsf{R} \ge \mathsf{R}^{\mathsf{noisy}}(\mathsf{D})\},$$

where $R^{\mathsf{noisy}}(\mathsf{D})$ is given in (29). We observed in Section 2.3 that IB is fully characterized by the mapping $\mathsf{D} \mapsto R^{\mathsf{noisy}}(\mathsf{D})$ and thus by $\mathcal{A}$. In light of Theorem A1, all extreme points of $\mathcal{A}$ are achieved by a choice of T with cardinality size $|\mathcal{T}| \le |\mathcal{X}|$. Let $\{(\mathsf{R}_i, \mathsf{D}_i)\}$ be the set of extreme points of $\mathcal{A}$ each constructed by channel $P_{T_i|X}$ and mapping ψ_i. Due to the convexity of $\mathcal{A}$, each point $(\mathsf{R}, \mathsf{D}) \in \mathcal{A}$ is expressed as a convex combination of $\{(\mathsf{R}_i, \mathsf{D}_i)\}$ with coefficient $\{\lambda_i\}$; that is there exists a channel $P_{T|X} = \sum_i \lambda_i P_{T_i|X}$ and a mapping $\psi(T) = \sum_i \lambda_i \psi_i(T_i)$ such that $I(X;T) = \mathsf{R}$ and $\mathbb{E}[d(Y, \psi(T))] = \mathsf{D}$. This construction, often termed timesharing in information theory literature, implies that all points in $\mathcal{A}$ (including the boundary points) can be achieved with a variable T with $|\mathcal{T}| \le |\mathcal{X}|$. Since the boundary of $\mathcal{A}$ is specified by the mapping $R \mapsto \mathsf{IB}(R)$, we conclude that $\mathsf{IB}(R)$ is achieved by a variable T with cardinality $|\mathcal{T}| \le |\mathcal{X}|$ for very $R < H(X)$. □

Proof of Lemma 5. The following proof is inspired by [32] (Proposition 1). Let $\mathcal{X} = \{1, \dots, m\}$. We sort the elements in $\mathcal{X}$ such that

$$P_X(1)D_{\mathsf{KL}}(P_{Y|X=1}\|P_Y) \ge \cdots \ge P_X(m)D_{\mathsf{KL}}(P_{Y|X=m}\|P_Y).$$

Now consider the function $f:\mathcal{X}\to[M]$ given by $f(x)=x$ if $x<M$ and $f(x)=M$ if $x\geq M$ where $M=e^R$. Let $Z=f(X)$. We have $P_Z(i)=P_X(i)$ if $i<M$ and $P_Z(M)=\sum_{j\geq M}P_X(j)$. We can now write

$$\begin{aligned}
I(Y;Z) &= \sum_{i=1}^{M-1} P_X(i)D(P_{Y|X=i}\|P_Y) + P_Z(M)D(P_{Y|Z=M}\|P_Y)\\
&\geq \sum_{i=1}^{M-1} P_X(i)D(P_{Y|X=i}\|P_Y)\\
&\geq \frac{M-1}{|\mathcal{X}|}\sum_{i\in\mathcal{X}} P_X(i)D(P_{Y|X=i}\|P_Y)\\
&= \frac{M-1}{|\mathcal{X}|}I(X;Y).
\end{aligned}$$

Since $f(X)$ takes values in $[M]$, it follows that $H(f(X))\leq R$. Consequently, we have

$$\mathsf{dIB}(P_{XY},R)\geq \sup_{f:\mathcal{X}\to[M]} I(Y;f(X))\geq \frac{M-1}{|\mathcal{X}|}I(X;Y).$$

For the privacy funnel, the proof proceeds as follows. We sort the elements in $\mathcal{X}$ such that

$$P_X(1)D_{\mathsf{KL}}(P_{Y|X=1}\|P_Y)\leq\cdots\leq P_X(m)D_{\mathsf{KL}}(P_{Y|X=m}\|P_Y).$$

Consider now the function $f:\mathcal{X}\to[M]$ given by $f(x)=x$ if $x<M$ and $f(x)=M$ if $x\geq M$. As before, let $Z=f(X)$. Then, we can write,

$$\begin{aligned}
I(Y;Z) &= \sum_{i=1}^{M-1} P_X(i)D(P_{Y|X=i}\|P_Y) + P_Z(M)D(P_{Y|Z=M}\|P_Y)\\
&\leq \frac{M-1}{|\mathcal{X}|}\sum_{i\in\mathcal{X}} P_X(i)D(P_{Y|X=i}\|P_Y) + P_Z(M)D(P_{Y|Z=M}\|P_Y)\\
&= \frac{M-1}{|\mathcal{X}|}I(X;Y) + P_Z(M)\sum_{y\in\mathcal{Y}} P_{Y|Z}(y|M)\log\frac{P_{Y|Z}(y|M)}{P_Y(y)}\\
&\leq \frac{M-1}{|\mathcal{X}|}I(X;Y) + \Pr(X\geq M)\sum_{y\in\mathcal{Y}}\left[\sum_i P_{Y|X}(y|i)\frac{P_X(i)1_{\{i\geq M\}}}{\Pr(X\geq M)}\right]\log\frac{\sum_i P_{Y|X}(y|i)^{\frac{P_X(i)1_{\{i\geq M\}}}{\Pr(X\geq M)}}}{\sum_i P_{Y|X}(y|i)P_X(i)}\\
&\leq \frac{M-1}{|\mathcal{X}|}I(X;Y) + \sum_{y\in\mathcal{Y}}\sum_{i\geq M} P_{Y|X}(y|i)P_X(i)\log\frac{1}{\Pr(X\geq M)}\\
&= \frac{M-1}{|\mathcal{X}|}I(X;Y) + \Pr(X\geq M)\log\frac{1}{\Pr(X\geq M)}
\end{aligned}$$

where the last inequality is due to the log-sum inequality. □

Proof of Lemma 6. Employing the same argument as in the proof of [32] (Theorem 3), we obtain that there exists a function $f:\mathcal{X}\to[M]$ such that

$$I(Y;f(X))\geq\eta I(X;Y) \tag{A11}$$

for any $\eta\in(0,1)$ and

$$M\leq 4+\frac{4}{(1-\eta)\log 2}\log\frac{2\alpha}{(1-\eta)I(X;Y)}.$$

Since $h_{\mathsf{b}}^{-1}(x) \leq \frac{x \log 2}{\log \frac{1}{x}}$ for all $x \in (0,1]$, it follows from above that (noticing that $I(X;Y) \leq \alpha$)

$$M \leq 4 + \frac{I(X;Y)}{2\alpha} \frac{1}{h_{\mathsf{b}}^{-1}(\zeta)},$$

where $\zeta := \frac{(1-\eta)I(X;Y)}{2\alpha}$. Rearranging this, we obtain

$$h_{\mathsf{b}}^{-1}(\zeta) \leq \frac{I(X;Y)}{2\alpha(M-4)}.$$

Assuming $M \geq 5$, we have $\frac{I(X;Y)}{2\alpha(M-4)} \leq \frac{1}{2}$ and hence

$$\zeta \leq h_{\mathsf{b}}\Big(\frac{I(X;Y)}{2\alpha(M-4)}\Big),$$

implying

$$\eta \geq 1 - \frac{2\alpha}{I(X;Y)} h_{\mathsf{b}}\Big(\frac{I(X;Y)}{2\alpha(M-4)}\Big).$$

Plugging this into (A11), we obtain

$$I(Y;f(X)) \geq I(X;Y) - 2\alpha h_{\mathsf{b}}\Big(\frac{I(X;Y)}{2\alpha(M-4)}\Big).$$

As before, if $M = e^R$, then $H(f(X)) \leq R$. Hence,

$$\mathsf{dIB}(P_{XY}, R) \geq I(X;Y) - 2\alpha h_{\mathsf{b}}\Big(\frac{I(X;Y)}{2\alpha(e^R-4)}\Big),$$

for all $R \geq \log 5$. □

Appendix B. Proofs from Section 3

Proof of Lemma 8. To prove the upper bound on $\mathsf{PF}^{(\infty,1)}$, recall that $r \mapsto e^{\mathsf{PF}^{(\infty,1)}(r)}$ is convex. Thus, it lies below the chord connecting points $(0,0)$ and $(H(X), e^{I_\infty(X;Y)})$. The lower bound on $\mathsf{IB}^{(\infty,1)}$ is similarly obtained using the concavity of $R \mapsto e^{\mathsf{IB}^{(\infty,1)}(R)}$. This is achievable by an erasure channel. To see this consider the random variable T_δ taking values in $\mathcal{X} \cup \{\perp\}$ that is obtained by conditional distributions $P_{T_\delta|X}(t|x) = \bar{\delta} I_{t=x}$ and $P_{T_\delta|X}(\perp\ |x) = \delta$ for some $\delta \geq 0$. It can be verified that $I(X;T_\delta) = \bar{\delta} H(X)$ and $\mathsf{P_c}(Y|T_\delta) = \bar{\delta}\mathsf{P_c}(Y|X) + \delta \mathsf{P_c}(Y)$. By taking $\delta = 1 - \frac{R}{H(X)}$, this channel meets the constraint $I(X;T_\delta) = R$. Hence,

$$\mathsf{IB}^{(\infty,1)}(R) \geq \log\left[\frac{\mathsf{P_c}(Y|T_\delta)}{\mathsf{P_c}(Y)}\right] = \log\left[1 - \frac{R}{H(X)} + \frac{R}{H(X)}\frac{\mathsf{P_c}(Y|X)}{\mathsf{P_c}(Y)}\right].$$

□

Proof of Theorem 11. We begin by $\mathsf{PF}^{(\infty,1)}$. As described in Section 3.1, and similar to Mrs. Gerber's Lemma (Lemma 4), we need to construct the lower convex envelope $\mathcal{K}_\cup[F_\beta^{(\infty,1)}]$ of $F_\beta^{(\infty,1)}(q) = \mathsf{P_c}(Y') + \beta H(X')$ where $X' \sim \mathsf{Bernoulli}(q)$ and Y' is the result of passing X' through $\mathsf{BSC}(\delta)$, i.e., $Y' \sim \mathsf{Bernoulli}(\delta * q)$. In this case, $\mathsf{P_c}(Y') = \max\{\delta * q, 1 - \delta * q\}$. Hence, we need to determine the lower convex envelope of the map

$$q \mapsto F_\beta^{(\infty,1)}(q) = \max\{\delta * q, 1 - \delta * q\} + \beta h_{\mathsf{b}}(q). \tag{A12}$$

A straightforward computation shows that $F_\beta^{(\infty,1)}(q)$ is symmetric around $q = \frac{1}{2}$ and is also concave in q on $q \in [0, \frac{1}{2}]$ for any β. Hence, $\mathcal{K}_\cup[F_\beta^{(\infty,1)}]$ is obtained as follows depending on the values of β:

- $F_\beta^{(\infty,1)}(\frac{1}{2}) < F_\beta^{(\infty,1)}(0)$ (see Figure 6a), then $\mathcal{K}_\cup[F_\beta^{(\infty,1)}]$ is given by the convex combination of $F_\beta^{(\infty,1)}(0)$, $F_\beta^{(\infty,1)}(1)$, and $F_\beta^{(\infty,1)}(\frac{1}{2})$.
- $F_\beta^{(\infty,1)}(\frac{1}{2}) = F_\beta^{(\infty,1)}(0)$ (see Figure 6b), then $\mathcal{K}_\cup[F_\beta^{(\infty,1)}]$ is given by the convex combination of $F_\beta^{(\infty,1)}(0)$, $F_\beta^{(\infty,1)}(\frac{1}{2})$, and $F_\beta^{(\infty,1)}(1)$.
- $F_\beta^{(\infty,1)}(\frac{1}{2}) > F_\beta^{(\infty,1)}(0)$ (see Figure 6c), then $\mathcal{K}_\cup[F_\beta^{(\infty,1)}]$ is given by the convex combination of $F_\beta^{(\infty,1)}(0)$ and $F_\beta^{(\infty,1)}(1)$.

Hence, assuming $p \leq \frac{1}{2}$, we can construct T^* that minimizes $\mathsf{P_c}(Y|T) - \beta H(X|T)$. Considering the first two cases, we obtain that T^* is ternary with $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$, $P_{X|T^*=1} = \mathsf{Bernoulli}(1)$, and $P_{X|T^*=2} = \mathsf{Bernoulli}(\frac{1}{2})$ with marginal $P_{T^*} = [1 - p - \frac{\alpha}{2}, p - \frac{\alpha}{2}, \alpha]$ for some $\alpha \in [0, 2p]$. This leads to $\mathsf{P_c}(Y|T^*) = \bar{\alpha}\bar{\delta} + \frac{1}{2}\alpha$ and $I(X;T^*) = h_\mathsf{b}(p) - \alpha$. Note that $\mathsf{P_c}(Y|T^*)$ covers all possible domain $[\mathsf{P_c}(Y), \bar{\delta}]$ by varying α on $[0, 2p]$. Replacing $I(X;T^*)$ by r, we obtain $\alpha = h_\mathsf{b}(p) - r$ leading to $\mathsf{P_c}(Y|T^*) = \bar{\delta} - (h_\mathsf{b}(p) - r)(\frac{1}{2} - \delta)$. Since $\mathsf{P_c}(Y) = 1 - \delta * p$, the desired result follows.

To derive the expression for $\mathsf{IB}^{(\infty,1)}$, recall that we need to derive $\mathcal{K}_\cap[F_\beta^{(\infty,1)}]$ the upper concave envelope of $F_\beta^{(\infty,1)}$. It is clear from Figure 6 that $\mathcal{K}_\cap[F_\beta^{(\infty,1)}]$ is obtained by replacing $F_\beta^{(\infty,1)}(q)$ on the interval $[q_\beta, 1 - q_\beta]$ by its maximum value over q where

$$q_\beta \coloneqq \frac{1}{1 + e^{\frac{1-2\delta}{\beta}}},$$

is the maximizer of $F_\beta^{(\infty,1)}(q)$ on $[0, \frac{1}{2}]$. In other words,

$$\mathcal{K}_\cap[F_\beta^{(\infty,1)}(q)] = \begin{cases} F_\beta^{(\infty,1)}(q_\beta), & \text{for } q \in [q_\beta, 1 - q_\beta], \\ F_\beta^{(\infty,1)}(q), & \text{otherwise.} \end{cases}$$

Note that if $p < q_\beta$ then $\mathcal{K}_\cap[F_\beta^{(\infty,1)}]$ evaluated at p coincides with $F_\beta^{(\infty,1)}(p)$. This corresponds to all trivial $P_{T|X}$ such that $\mathsf{P_c}(Y|T) + \beta H(X|T) = \mathsf{P_c}(Y) + \beta H(X)$. If, on the other hand, $p \geq q_\beta$, then $\mathcal{K}_\cap[F_\beta^{(\infty,1)}$ is the convex combination of $F_\beta^{(\infty,1)}(q_\beta)$ and $F_\beta^{(\infty,1)}(1 - q_\beta)$. Hence, taking q_β as a parameter (say, α), the optimal binary T^* is constructed as follows: $P_{X|T^*=0} = \mathsf{Bernoulli}(\alpha)$ and $P_{X|T^*=1} = \mathsf{Bernoulli}(\bar{\alpha})$ for $\alpha \leq p$. Such channel induces

$$\mathsf{P_c}(Y|T^*) = \max\{\alpha * \delta, 1 - \alpha * \delta\} = 1 - \alpha * \delta,$$

as $\alpha \leq p \leq \frac{1}{2}$, and also

$$I(X;T^*) = h_\mathsf{b}(p) - h_\mathsf{b}(\alpha).$$

Combining these two, we obtain

$$\mathsf{P_c}(Y|T^*) = 1 - \delta * h_\mathsf{b}^{-1}(h_\mathsf{b}(p) - R).$$

□

Proof of Theorem 12. Let U_α and L_α denote the $\mathsf{U}_{\Phi,\Psi}$ and $\mathsf{L}_{\Phi,\Psi}$, respectively, when $\Psi(Q_X) = \Phi(Q_X) = \|Q_X\|_\alpha$. In light of (59) and (60), it is sufficient to compute L_α and U_α. To do so, we need to construct the lower convex envelope $\mathcal{K}_\cup[F_\beta^{(\alpha)}]$ and upper concave envelope $\mathcal{K}_\cap[F_\beta^{(\alpha)}]$ of the map $F_\beta^{(\alpha)}(q)$ given

by $q \mapsto \|Q_Y\|_\alpha - \beta\|Q_X\|_\alpha$ where $X' \sim \mathsf{Bernoulli}(q)$ and Y' is the result of passing X' through $\mathsf{BSC}(\delta)$, i.e., $Y' \sim \mathsf{Bernoulli}(\delta * q)$. In this case, we have

$$q \mapsto F_\beta^{(\alpha)}(q) = \|q * \delta\|_\alpha - \beta||q||_\alpha, \tag{A13}$$

where $\|a\|_\alpha$ is to mean $\|[a, \bar{a}]\|_\alpha$ for any $a \in [0, 1]$.

We begin by L_α for which we aim at obtaining $\mathcal{K}_\cup[F_\beta^{(\alpha)}]$. A straightforward computation shows that $F_\beta^{(\alpha)}(q)$ is convex for $\beta \leq (1-2\delta)^2$ and $\alpha \geq 2$. For $\beta > (1-2\delta)^2$ and $\alpha \geq 2$, it can be shown that $F_\beta^{(\alpha)}(q)$ is concave an interval $[q_\beta, 1-q_\beta]$ where q_β solves $\frac{d}{dq}F_\beta^{(\alpha)}(q) = 0$. (The shape of $q \mapsto F_\beta^{(\alpha)}(r)$ in is similar to what was depicted in Figure 4.) By symmetry, $\mathcal{K}_\cup[F_\beta^{(\alpha)}]$ is therefore obtained by replacing $F_\beta^{(\alpha)}(q)$ on this interval by $F_\beta^{(\alpha)}(q_\beta)$. Hence, if $p < q_\beta$, $\mathcal{K}_\cup[F_\beta^{(\alpha)}]$ at p coincides with $F_\beta^{(\alpha)}(p)$ which results in trivial $P_{T|X}$ (see the proof of Theorem 11 for more details). If, on the other hand, $p \geq q_\beta$, then $\mathcal{K}_\cup[F_\beta^{(\alpha)}]$ evaluated at p is given by a convex combination of $F_\beta^{(\alpha)}(q_\beta)$ and $F_\beta^{(\alpha)}(1-q_\beta)$. Relabeling q_β as a parameter (say, q), we can write an optimal binary T^* via the following: $P_{X|T*=0} = \mathsf{Bernoulli}(1-q)$ and $P_{X|T*=1} = \mathsf{Bernoulli}(q)$ for $q \leq p$. This channel induces $\Psi(Y|T^*) = \|q * \delta\|_\alpha$ and $\Phi(X|T^*) = \|q\|_\alpha$. Hence, the graph of L_α is given by

$$\big\{(\|q\|_\alpha, \|q * \delta\|_\alpha), 0 \leq q \leq p\big\}.$$

Therefore,

$$\sup_{\substack{P_{T|X} \\ H_\alpha(X|T) \geq \zeta}} H_\alpha(Y|T) = \frac{\alpha}{1-\alpha} \log \|q * \delta\|_\alpha,$$

where $q \leq p$ solves $\frac{\alpha}{1-\alpha} \log \|q\|_\alpha = \zeta$. Since the map $q \mapsto \|q\|_\alpha$ is strictly decreasing for $q \in [0, 0.5]$, this equation has a unique solution.

Next, we compute U_α or equivalently $\mathcal{K}_\cap[F_\beta^{(\alpha)}]$ the upper concave envelop of $F_\beta^{(\alpha)}$ defined in (A13). As mentioned earlier, $q \mapsto F_\beta^{(\alpha)}(q)$ is convex for $\beta \leq (1-2\delta)^2$ and $\alpha \geq 2$. For $\beta > (1-2\delta)^2$, we need to consider three cases: (1) $\mathcal{K}_\cap[F_\beta^{(\alpha)}]$ is given by the convex combination of $F_\beta^{(\alpha)}(0)$ and $F_\beta^{(\alpha)}(1)$, (2) $\mathcal{K}_\cap[F_\beta^{(\alpha)}]$ is given by the convex combination of $F_\beta^{(\alpha)}(0)$, $F_\beta^{(\alpha)}(\frac{1}{2})$, and $F_\beta^{(\alpha)}(1)$, (3) $\mathcal{K}_\cap[F_\beta^{(\alpha)}]$ is given by the convex combination of $F_\beta^{(\alpha)}(0)$ and $F_\beta^{(\alpha)}(q^\dagger)$ where $q^\dagger$ is a point $\in [0, \frac{1}{2}]$. Without loss of generality, we can ignore the first case. The other two cases correspond to the following solutions

- T^* is a ternary variable given by $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$, $P_{X|T^*=1} = \mathsf{Bernoulli}(1)$, and $P_{X|T^*=2} = \mathsf{Bernoulli}(\frac{1}{2})$ with marginal $T^* \sim \mathsf{Bernoulli}(1-p-\frac{\lambda}{2}, p-\frac{\lambda}{2}, \lambda)$ for some $\lambda \in [0, 2p]$. This produces

  $$\Psi(Y|T^*) = \bar{\lambda}\|\delta\|_\alpha + \lambda\|\frac{1}{2}\|_\alpha,$$

  and

  $$\Phi(X|T^*) = \bar{\lambda} + \lambda\|\frac{1}{2}\|_\alpha.$$

- T^* is a binary variable given by $P_{X|T^*=0} = \mathsf{Bernoulli}(0)$ and $P_{X|T^*=1} = \mathsf{Bernoulli}(\frac{p}{\lambda})$ with marginal $T^* \sim \mathsf{Bernoulli}(\lambda)$ for some $\lambda \in [2p, 1]$. This produces

  $$\Psi(Y|T^*) = \bar{\lambda}\|\delta\|_\alpha + \lambda\|\delta * \frac{p}{\lambda}\|_\alpha,$$

  and

  $$\Phi(X|T^*) = \bar{\lambda} + \lambda\|\frac{p}{\lambda}\|_\alpha.$$

Combining these two cases, can write

$$\mathsf{U}_\alpha(\zeta) = \bar{\lambda}\|\delta\|_\alpha + \lambda\|\frac{q}{z} * \delta\|_\alpha,$$

where

$$\zeta = \bar{\lambda} + \lambda\|\frac{q}{z}\|_\alpha,$$

and $z = \max\{2p, \lambda\}$. Plugging this into (59) completes the proof. □

Proof of Lemma 9. The facts that $\gamma \mapsto H_\gamma(X|T)$ is non-increasing on $[1, \infty]$ [103] (Proposition 5) and $(\sum_i |x_i|^\gamma)^{1/\gamma} \geq \max_i |x_i|$ for all $p \geq 0$ imply

$$\frac{\gamma - 1}{\gamma} H_\gamma(X|T) \leq H_\infty(X|T) \leq H_\gamma(X|T). \tag{A14}$$

Since $I_\infty(X;T) = H_\infty(X) - H_\infty(X|T)$, the above lower bound yields

$$I_\gamma(X;T) \geq \frac{\gamma}{\gamma - 1} I_\infty(X;T) - \frac{\gamma}{\gamma - 1} H_\infty(X) + H_\gamma(X), \tag{A15}$$

where the last inequality follows from the fact that $\gamma \mapsto H_\gamma(X)$ is non-increasing. The upper bound in (A14) (after replacing X with Y and γ with α) implies

$$I_\alpha(Y;T) \leq I_\infty(Y;T) + H_\alpha(Y) - H_\infty(Y). \tag{A16}$$

Combining (A15) and (A16), we obtain the desired upper bound for $\mathsf{PF}^{(\alpha,\gamma)}$. The other bounds can be proved similarly by interchanging X with Y and α with γ in (A15) and (A16). □

References

1. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 30 September–3 October 1999; pp. 368–377.
2. Wyner, A.; Ziv, J. A theorem on the entropy of certain binary sequences and applications: Part I. *IEEE Trans. Inf. Theory* **1973**, *19*, 769–772.
3. Witsenhausen, H.; Wyner, A. A conditional entropy bound for a pair of discrete random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 493–501.
4. Ahlswede, R.; Körner, J. On the connection between the entropies of input and output distributions of discrete memoryless channels. In Proceedings of the Fifth Conference on Probability Theory, Brasov, Romania, 1–6 September 1974.
5. Wyner, A. A theorem on the entropy of certain binary sequences and applications–II. *IEEE Trans. Inf. Theory* **1973**, *19*, 772–777.
6. Kim, Y.H.; El Gamal, A. *Network Information Theory*; Cambridge University Press: Cambridge, UK, 2012.
7. Slonim, N.; Tishby, N. Document clustering using word clusters via the information bottleneck method. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; pp. 208–215.
8. Still, S.; Bialek, W. How Many Clusters? An Information-Theoretic Perspective. *Neural Comput.* **2004**, *16*, 2483–2506.
9. Slonim, N.; Tishby, N. Agglomerative Information Bottleneck. In Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99, Denver, CO, USA, 29 November–4 December 1999; pp. 617–623.
10. Cardinal, J. Compression of side information. In Proceedings of the 2003 International Conference on Multimedia and Expo—Volume 1, Baltimore, MD, USA, 6–9 July 2003; Volume 2, pp. 569–572.

11. Zeitler, G.; Koetter, R.; Bauch, G.; Widmer, J. Design of network coding functions in multihop relay networks. In Proceedings of the 2008 5th International Symposium on Turbo Codes and Related Topics, Lausanne, Switzerland, 1–5 September 2008; pp. 249–254.
12. Makhdoumi, A.; Salamatian, S.; Fawaz, N.; Médard, M. From the Information Bottleneck to the Privacy Funnel. In Proceedings of the 2014 IEEE Information Theory Workshop (ITW 2014), Tasmania, Australia, 2–5 November 2014; pp. 501–505.
13. Calmon, F.P.; Makhdoumi, A.; Médard, M. Fundamental limits of perfect privacy. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 1796–1800.
14. Asoodeh, S.; Alajaji, F.; Linder, T. Notes on information-theoretic privacy. In Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 30 September–3 October 2014; pp. 1272–1278.
15. Ding, N.; Sadeghi, P. A Submodularity-based Clustering Algorithm for the Information Bottleneck and Privacy Funnel. In Proceedings of the 2019 IEEE Information Theory Workshop (ITW), Visby, Sweden, 25–28 August 2019; pp. 1–5.
16. Bertran, M.; Martinez, N.; Papadaki, A.; Qiu, Q.; Rodrigues, M.; Reeves, G.; Sapiro, G. Adversarially Learned Representations for Information Obfuscation and Inference. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 614–623.
17. Lopuhaä-Zwakenberg, M.; Tong, H.; Škorić, B. Data Sanitisation Protocols for the Privacy Funnel with Differential Privacy Guarantees. *arXiv* **2020**, arXiv:2008.13151.
18. Hsu, H.; Asoodeh, S.; Calmon, F. Obfuscation via Information Density Estimation. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Sicily, Italy, 26–28 August 2020; Volume 108, pp. 906–917.
19. Dobrushin, R.; Tsybakov, B. Information transmission with additional noise. *IRE Trans. Inf. Theory* **1962**, *8*, 293–304.
20. Ahlswede, R.; Csiszar, I. Hypothesis testing with communication constraints. *IEEE Trans. Inf. Theory* **1986**, *32*, 533–542. doi:10.1109/TIT.1986.1057194.
21. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Information extraction under privacy constraints. *Information* **2016**, *7*, 15.
22. Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. In *Topics in Information Theory, Coll. Math. Soc. J. Bolyai*; Csiszár, I., Elias, P. Eds.; North-Holland: Amsterdam, The Netherlands, 1977; Volume 16, pp. 41–52.
23. Raginsky, M. Strong Data Processing Inequalities andΦ-Sobolev Inequalities for Discrete Channels. *IEEE Trans. Inf. Theory* **2016**, *62*, 3355–3389. doi:10.1109/TIT.2016.2549542.
24. Hsu, H.; Asoodeh, S.; Salamatian, S.; Calmon, F.P. Generalizing Bottleneck Problems. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 531–535.
25. Courtade, T.A. Strengthening the entropy power inequality. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 2294–2298.
26. Globerson, A.; Tishby, N. *On the Optimality of the Gaussian Information Bottleneck Curve*; Technical Report; Hebrew University: Jerusalem, Israel, 2004.
27. Calmon, F.P.; Polyanskiy, Y.; Wu, Y. Strong data processing inequalities in power-constrained Gaussian channels. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Hongkong, China, 14–19 June 2015; pp. 2558–2562.
28. Rényi, A. On measures of dependence. *Acta Math. Acad. Sci. Hung.* **1959**, *10*, 441–451.
29. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38.
30. Zaidi, A.; Estella-Aguerri, I.; Shamai (Shitz), S. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy* **2020**, *22*, 151.
31. Strouse, D.; Schwab, D.J. The Deterministic Information Bottleneck. *Neural Comput.* **2017**, *29*, 1611–1630.
32. Bhatt, A.; Nazer, B.; Ordentlich, O.; Polyanskiy, Y. Information-Distilling Quantizers. *arXiv* **2018**, arXiv:1812.03031.
33. Shamir, O.; Sabato, S.; Tishby, N. Learning and Generalization with the Information Bottleneck. *Theor. Comput. Sci.* **2010**, *411*, 2696–2711.

34. Diaz, M.; Wang, H.; Calmon, F.P.; Sankar, L. On the Robustness of Information-Theoretic Privacy Measures and Mechanisms. *IEEE Trans. Inf. Theory* **2020**, *66*, 1949–1978.
35. El-Yaniv, R.; Souroujon, O. Iterative Double Clustering for Unsupervised and Semi-Supervised Learning. In Proceedings of the 12th European Conference on Machine Learning, Freiburg, Germany, 5–7 September 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 121–132.
36. Elidan, G.; Friedman, N. Learning Hidden Variable Networks: The Information Bottleneck Approach. *J. Mach. Learn. Res.* **2005**, *6*, 81–127.
37. Aguerri, I.E.; Zaidi, A. Distributed Information Bottleneck Method for Discrete and Gaussian Sources. *arXiv* **2017**, arXiv:1709.09082.
38. Aguerri, I.E.; Zaidi, A. Distributed Variational Representation Learning. *arXiv* **2019**. arXiv:1807.04193.
39. Strouse, D.; Schwab, D.J. Geometric Clustering with the Information Bottleneck. *Neural Comput.* **2019**, *31*, 596–612. doi:10.1162/neco_a_01136.
40. Cicalese, F.; Gargano, L.; Vaccaro, U. Bounds on the Entropy of a Function of a Random Variable and Their Applications. *IEEE Trans. Inf. Theory* **2018**, *64*, 2220–2230.
41. Koch, T.; Lapidoth, A. At Low SNR, Asymmetric Quantizers are Better. *IEEE Trans. Inf. Theory* **2013**, *59*, 5421–5445.
42. Pedarsani, R.; Hassani, S.H.; Tal, I.; Telatar, E. On the construction of polar codes. In Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings, St. Petersburg, Russia, 31 July–5 August 2011; pp. 11–15.
43. Tal, I.; Sharov, A.; Vardy, A. Constructing polar codes for non-binary alphabets and MACs. In Proceedings of the 2012 IEEE International Symposium on Information Theory Proceedings, Cambridge, MA, USA, 1–6 July 2012; pp. 2132–2136.
44. Kartowsky, A.; Tal, I. Greedy-Merge Degrading has Optimal Power-Law. *IEEE Trans. Inf. Theory* **2019**, *65*, 917–934.
45. Viterbi, A.J.; Omura, J.K. *Principles of Digital Communication and Coding*, 1st ed.; McGraw-Hill, Inc.: New York, NY, USA, 1979.
46. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the IEEE Information Theory Workshop (ITW), Jeju Island, Korea, 11–15 October 2015; pp. 1–5.
47. Shwartz-Ziv, R.; Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv* **2017**, arXiv:1703.00810.
48. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the Information Bottleneck Theory of Deep Learning. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
49. Goldfeld, Z.; Van Den Berg, E.; Greenewald, K.; Melnyk, I.; Nguyen, N.; Kingsbury, B.; Polyanskiy, Y. Estimating Information Flow in Deep Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 2299–2308.
50. Amjad, R.A.; Geiger, B.C. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2225–2239.
51. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. *arXiv* **2016**, arXiv:1612.00410.
52. Kolchinsky, A.; Tracey, B.D.; Wolpert, D.H. Nonlinear Information Bottleneck. *arXiv* **2017**, arXiv:1705.02436.
53. Kolchinsky, A.; Tracey, B.D.; Kuyk, S.V. Caveats for information bottleneck in deterministic scenarios. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
54. Chalk, M.; Marre, O.; Tkacik, G. Relevant Sparse Codes with Variational Information Bottleneck. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Barcelona, Spain, 9 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 1965–1973.
55. Wickstrøm, K.; Løkse, S.; Kampffmeyer, M.; Yu, S.; Principe, J.; Jenssen, R. *Information Plane Analysis of Deep Neural Networks via Matrix–Based Rényi's Entropy and Tensor Kernels*; *arXiv* **2019**, arXiv:1909.11396.
56. Matias, V.; Piantanida, P.; Rey Vega, L. The Role of the Information Bottleneck in Representation Learning. IEEE International Symposium on Information Theory (ISIT 2018), Vail, CO, USA, 17–22 June 2018. doi:10.1109/isit.2018.8437679.

57. Alemi, A.; Fischer, I.; Dillon, J. Uncertainty in the Variational Information Bottleneck. *arXiv* **2018**, arXiv:1807.00906.
58. Yu, S.; Jenssen, R.; Príncipe, J. Understanding Convolutional Neural Network Training with Information Theory. *arXiv* **2018**, arXiv:1804.06537.
59. Cheng, H.; Lian, D.; Gao, S.; Geng, Y. Evaluating Capability of Deep Neural Networks for Image Classification via Information Plane. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
60. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
61. Issa, I.; Wagner, A.B.; Kamath, S. An Operational Approach to Information Leakage. *IEEE Trans. Inf. Theory* **2020**, *66*, 1625–1657.
62. Cvitkovic, M.; Koliander, G. Minimal Achievable Sufficient Statistic Learning. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 1465–1474.
63. Asoodeh, S.; Alajaji, F.; Linder, T. On maximal correlation, mutual information and data privacy. In Proceedings of the IEEE 14th Canadian Workshop on Inf. Theory (CWIT), St. John's, NL, Canada, 6–9 July 2015; pp. 27–31.
64. Makhdoumi, A.; Fawaz, N. Privacy-utility tradeoff under statistical uncertainty. In Proceedings of the 51st Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2–4 October 2013; pp. 1627–1634. doi:10.1109/Allerton.2013.6736724.
65. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Estimation Efficiency Under Privacy Constraints. *IEEE Trans. Inf. Theory* **2019**, *65*, 1512–1534.
66. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Privacy-aware guessing efficiency. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
67. Asoodeh, S.; Alajaji, F.; Linder, T. Privacy-aware MMSE estimation. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1989–1993.
68. Calmon, F.P.; Makhdoumi, A.; Médard, M.; Varia, M.; Christiansen, M.; Duffy, K.R. Principal Inertia Components and Applications. *IEEE Trans. Inf. Theory* **2017**, *63*, 5011–5038.
69. Wang, H.; Vo, L.; Calmon, F.P.; Médard, M.; Duffy, K.R.; Varia, M. Privacy With Estimation Guarantees. *IEEE Trans. Inf. Theory* **2019**, *65*, 8025–8042.
70. Asoodeh, S. Information and Estimation Theoretic Approaches to Data Privacy. Ph.D. Thesis, Queen's University, Kingston, ON, Canada, 2017.
71. Liao, J.; Kosut, O.; Sankar, L.; du Pin Calmon, F. Tunable Measures for Information Leakage and Applications to Privacy-Utility Tradeoffs. *IEEE Trans. Inf. Theory* **2019**, *65*, 8043–8066.
72. Duchi, J.C.; Jordan, M.I.; Wainwright, M.J. Privacy aware learning. *J. Assoc. Comput. Mach. (ACM)* **2014**, *61*, 38.
73. Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; Tucker, G. On Variational Bounds of Mutual Information. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; Volume 97, pp. 5171–5180.
74. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 531–540.
75. Van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
76. Song, J.; Ermon, S. Understanding the Limitations of Variational Mutual Information Estimators. In Proceedings of the International Conference on Learning Representations, 26 April–1May 2020.
77. McAllester, D.; Stratos, K. Formal Limitations on the Measurement of Mutual Information. In Proceedings of the International Conference on Learning Representations, online, 26 April–1May 2020; Volume 108, pp. 875–884.
78. Rassouli, B.; Gunduz, D. On Perfect Privacy. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 2551–2555.
79. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*; Cambridge University Press: Cambridge, UK, 2011.

80. Kim, H.; Gao, W.; Kannan, S.; Oh, S.; Viswanath, P. Discovering Potential Correlations via Hypercontractivity. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4577–4587.
81. Ahlswede, R.; Gács, P. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.* **1976**, *4*, 925–939.
82. Anantharam, V.; Gohari, A.; Kamath, S.; Nair, C. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv* **2014**, arXiv:1304.6133v1
83. Polyanskiy, Y.; Wu, Y. Dissipation of Information in Channels With Input Constraints. *IEEE Trans. Inf. Theory* **2016**, *62*, 35–55.
84. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian Variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
85. Zaidi, A. Hypothesis Testing Against Independence Under Gaussian Noise. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 1289–1294. doi:10.1109/ISIT44484.2020.9173936.
86. Wu, T.; Fischer, I.; Chuang, I.L.; Tegmark, M. Learnability for the Information Bottleneck. *Entropy* **2019**, *21*, 924. doi:10.3390/e21100924.
87. Contento, L.; Ern, A.; Vermiglio, R. A linear-time approximate convex envelope algorithm using the double Legendre-Fenchel transform with application to phase separation. *Comput. Optim. Appl.* **2015**, *60*, 231–261.
88. Lucet, Y. Faster than the Fast Legendre Transform, the Linear-time Legendre Transform. *Numer. Algorithms* **1997**, *16*, 171–185.
89. Witsenhausen, H. Indirect rate distortion problems. *IEEE Trans. Inf. Theory* **1980**, *26*, 518–521.
90. Wyner, A. On source coding with side information at the decoder. *IEEE Trans. Inf. Theory* **1975**, *21*, 294–300.
91. Courtade, T.A.; Weissman, T. Multiterminal Source Coding Under Logarithmic Loss. *IEEE Trans. Inf. Theory* **2014**, *60*, 740–761.
92. Li, C.T.; El Gamal, A. Extended Gray-Wyner System With Complementary Causal Side Information. *IEEE Trans. Inf. Theory* **2018**, *64*, 5862–5878.
93. Vera, M.; Rey Vega, L.; Piantanida, P. Collaborative Information Bottleneck. *IEEE Trans. Inf. Theory* **2019**, *65*, 787–815.
94. Gilad-Bachrach, R.; Navot, A.; Tishby, N. An Information Theoretic Tradeoff between Complexity and Accuracy. In *Learning Theory and Kernel Machines*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 595–609.
95. Pichler, G.; Koliander, G. Information Bottleneck on General Alphabets. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 526–530. doi:10.1109/ISIT.2018.8437714.
96. Kim, Y.H.; Sutivong, A.; Cover, T. State mplification. *IEEE Trans. Inf. Theory* **2008**, *54*, 1850–1859.
97. Merhav, N.; Shamai, S. Information rates subject to state masking. *IEEE Trans. Inf. Theory* **2007**, *53*, 2254–2261.
98. Witsenhausen, H. Some aspects of convexity useful in information theory. *IEEE Trans. Inf. Theory* **1980**, *26*, 265–271. doi:10.1109/TIT.1980.1056173.
99. Harremoës, P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Nice, France, 24–29 June 2007; pp. 566–570.
100. Hirche, C.; Winter, A. An alphabet size bound for the information bottleneck function. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020.
101. Liese, F.; Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.
102. Verdú, S. α-mutual information. In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 1–6 February 2015; pp. 1–6.
103. Fehr, S.; Berens, S. On the Conditional Rényi Entropy. *IEEE Trans. Inf. Theory* **2014**, *60*, 6801–6810.
104. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
105. Sason, I.; Verdú, S. f -Divergence Inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006.
106. Guntuboyina, A.; Saha, S.; Schiebinger, G. Sharp Inequalities for f -Divergences. *IEEE Trans. Inf. Theory* **2014**, *60*, 104–121.

107. Guo, D.; Shamai, S.; Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 1261–1282.
108. Rockafellar, R.T. *Convex Analysis*; Princeton Univerity Press: Princeton, NJ, USA, 1997.
109. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
110. Linder, T.; Zamir, R. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Inf. Theory* **2008**, *40*, 2026–2031.
111. Guo, D.; Wu, Y.; Shitz, S.S.; Verdú, S. Estimation in Gaussian Noise: Properties of the Minimum Mean-Square Error. *IEEE Trans. Inf. Theory* **2011**, *57*, 2371–2385.
112. Jana, S. Alphabet sizes of auxiliary random variables in canonical inner bounds. In Proceedings of the 43rd Annual Conference on Information Sciences and Systems, Baltimore, MD, USA, 18–20 March 2009; pp. 67–71.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).