Generalizing Bottleneck Problems

Hsiang Hsu*, Shahab Asoodeh†, Salman Salamatian‡, and Flavio P. Calmon*
*Harvard University, {hsianghsu, fcalmon}@g.harvard.edu, †University of Chicago, shahab@uchicago.edu,
‡Massachusetts Institute of Technology, salmansa@mit.edu

***Abstract*—Given a pair of random variables $(X,Y)\sim P_{XY}$ and two convex functions f_1 and f_2, we introduce two bottleneck functionals as the lower and upper boundaries of the two-dimensional convex set that consists of the pairs $(I_{f_1}(W;X), I_{f_2}(W;Y))$, where I_f denotes f-information and W varies over the set of all discrete random variables satisfying the Markov condition $W\to X\to Y$. Applying Witsenhausen and Wyner's approach, we provide an algorithm for computing boundaries of this set for f_1, f_2, and discrete P_{XY}, . In the binary symmetric case, we fully characterize the set when (i) $f_1(t)=f_2(t)=t\log t$, (ii) $f_1(t)=f_2(t)=t^2-1$, and (iii) f_1 and f_2 are both ℓ^β norm function for $\beta>1$. We then argue that upper and lower boundaries in (i) correspond to Mrs. Gerber's Lemma and its inverse (which we call Mr. Gerber's Lemma), in (ii) correspond to estimation-theoretic variants of Information Bottleneck and Privacy Funnel, and in (iii) correspond to Arimoto Information Bottleneck and Privacy Funnel.**

I. Introduction

Few information-theoretic constructs have captured the attention of machine learning researchers and practitioners as the Information Bottleneck (IB) [1]. Given two correlated random variables X and Y with joint distribution P_{XY}, the goal of the IB is to determine a mapping $P_{W|X}$ that produces a new representation W of X such that (i) $W\to X\to Y$ and (ii) $I(W;Y)$ is maximized (information preserved) while minimizing $I(W;X)$ (compression). This tradeoff can be quantified by the Lagrangian functional $\mathsf{B}(P_{XY},\lambda)\triangleq \max_{P_{W|X}} I(W;Y)-\lambda I(W;X)$. The IB has proved useful in many machine learning problems, such as clustering [2] and natural language processing [3]. More recently, the IB framework has been used to analyze the training process of deep neural networks [4], [5].

In an inverse context, the Privacy Funnel (PF), introduced in [6], seeks to determine a mapping $P_{W|X}$ that minimizes $I(W;Y)$ (privacy leakage) while assuring $I(W;X)\geq x$ (revealing useful information). Analogously, the PF can be solved by considering the functional $\mathsf{F}(P_{XY},\lambda)\triangleq \min_{P_{W|X}} I(W;Y)+\lambda I(W;X)$. The privacy funnel (and its variants) has shown to be useful in information-theoretic privacy [6], [7].

The choice of mutual information in both the IB and the PF frameworks does not seem to carry any specific "operational" significance. It does, however, have a desirable practical consequence: it leads to self-consistent equations [1, Eq. 28] that can be solved iteratively in the IB case. In fact, this property is unique to mutual information among many other information metrics [8]. Nevertheless, at least in theory, one can replace the mutual information with a broader family of measures based on f-divergences[1].

In this paper, we consider a wider class of *bottleneck problems* which includes the IB and the PF. We define *f-information* between two random variables X and Y as $I_f(X;Y)\triangleq D_f(P_{XY}\|P_XP_Y)$, and introduce the following *bottleneck functional*

$$\mathsf{B}_{f_1,f_2}(P_{XY},x)\triangleq \max_{W\to X\to Y} I_{f_2}(W;Y)\ \mathsf{s.t.} I_{f_1}(W;X)\leq x, \tag{1}$$

and the *funnel functional*

$$\mathsf{F}_{f_1,f_2}(P_{XY},x)\triangleq \min_{W\to X\to Y} I_{f_2}(W;Y)\ \mathsf{s.t.} I_{f_1}(W;X)\geq x, \tag{2}$$

where f_1 and f_2 are convex functions. Different incarnations of f-information have already appeared, *e.g.*, T-information in [9] for $f(t)=|t-1|$. These metrics possess "operational" significance that are arguably more useful in statistical learning and privacy applications than mutual information. For instance, total variation and Hellinger distance play important roles in hypothesis testing [10] and χ^2-divergence in estimation problems [6]. Formulations (1) and (2) for a broader class of divergences can be potentially useful to emerging applications of information theory in machine learning.

Computing B and F reduces to characterizing the upper and lower boundaries, respectively, of the two-dimensional set

$$\Big\{\big(I_{f_1}(W;X),I_{f_2}(W;Y)\big):\ W\to X\to Y\Big\}. \tag{3}$$

It is worth mentioning that studying (3) is at the heart of the strong data processing inequalities [11] as well as fundamental limits of privacy [7]. Witsenhausen *et al.* [12] investigated the lower boundary of a related set $\Big\{(H(X|W),H(Y|W)):\ W\to X\to Y\Big\}$, where $H(\cdot)$ is the entropy function. In particular, they proposed an algorithm for analytically computing $H(Y|W)$ based on a dual formulation. When X is binary and $P_{Y|X}$ is a binary symmetric channel (BSC), the lower bound of $H(Y|W)$ corresponds to the well-known Mrs. Gerber's Lemma [13]. Related convex techniques have also been used to characterize some network information theoretic regions [14].

We generalize the approach in [12] to study boundaries of (3) for a broader class of f-information metrics, characterizing properties of new bottleneck problems of the form (1) and (2). In particular, we investigate the estimation-theoretic

[1]Given two probability distributions $P\ll Q$ and a convex function $f:(0,\infty)\to\mathbb{R}$ with $f(1)=0$, f-divergences is $D_f(P\|Q)\triangleq \mathbb{E}_Q\left[f(\frac{dP}{dQ})\right]$.

variants of information bottleneck and privacy funnel using χ^2-divergence, which we call *Estimation Bottleneck* and *Estimation Privacy Funnel*, respectively. In the binary symmetric case, the upper boundary corresponds to the inverse of Mrs. Gerber's Lemma, which we call Mr. Gerber's Lemma. We further extend these lemmas for Arimoto conditional entropy [15].

This paper is organized as follows. Section II introduces the geometry of bottleneck problems. In Section III, we formulate variational bottleneck problems and explore their use, and provide further applications on information inequalities in Section IV. Proofs of the results are available in [16].

II. Geometric Properties

A. Notation

Let X and Y be two random variables having joint distribution P_{XY} with supports $\mathcal{X} = [m] \triangleq \{1, \cdots, m\}$ and $\mathcal{Y} = [n]$, respectively. We denote by $P_X = \mathbf{q} \in \Delta_m$ the marginal probability vector with entries $[P_X(1), \cdots, P_X(m)]$, where $\Delta_m \triangleq \{\mathbf{x} \in \mathbb{R}^m : \sum_{i=1}^m x_i = 1, x_i \geq 0\}$ is a m-dimensional simplex. We denote by $\mathbf{T} \in \mathbb{R}^{n \times m}$ the stochastic matrix whose entries are the channel transformation $P_{Y|X}$, *i.e.* $[\mathbf{T}]_{i,j} = P_{Y|X}(i|j)$; thus, $P_Y = \mathbf{T}\mathbf{q} \in \Delta_n$. For a discrete random variable W with support $\mathcal{W}$, let $\mathbf{p}_w = \left[P_{X|W}(1|w), P_{X|W}(2|w), \ldots, P_{X|W}(m|w)\right]$, and let the marginal of W be $P_W(w) = \alpha_w$. We denote by h_m the entropy function, *i.e.* $h_m : \Delta_m \to \mathbb{R}$ with $h_m(\mathbf{q}) = -\sum_{i \in [m]} \mathbf{q}_i \log \mathbf{q}_i$ and $0 \log 0 \triangleq 0$. Finally, we denote the convex hull of a set $\mathcal{A}$ by $\mathsf{conv}\mathcal{A}$, and the boundary of a set by $\partial\mathcal{A}$.

B. Geometry of Bottleneck Problems

Let $f : \Delta_m \to \mathbb{R}$ and $g : \Delta_n \to \mathbb{R}$ be continuous and bounded mappings over simplices of dimension m and n, respectively. We study the set (3) by first considering a more general context, and then specialize it to different information metrics in following sections. We consider the set

$$\Big(\mathbb{E}\big[f(\mathbf{p}_w)\big], \mathbb{E}\big[g(\mathbf{T}\mathbf{p}_w)\big]\Big), \tag{4}$$

where $\mathbb{E}[f(\mathbf{p}_w)] = \sum_{w \in \mathcal{W}} \alpha_w f(\mathbf{p}_w)$, and $\mathbb{E}[g(\mathbf{T}\mathbf{p}_w)] = \sum_{w \in \mathcal{W}} \alpha_w g(\mathbf{T}\mathbf{p}_w)$. Recall that X, Y, and W form the Markov chain $W \to X \to Y$. Therefore, we are interested in the following set for a fixed channel $\mathbf{T}$:

$$\begin{aligned}\mathcal{C}(\mathbf{T}) \triangleq \Big\{ &(\mathbf{q}, \mathbb{E}[f(\mathbf{p}_w)], \mathbb{E}[g(\mathbf{T}\mathbf{p}_w)]) \,| \\ &\mathbf{p}_w \in \Delta_m, \sum_{w \in \mathcal{W}} \alpha_w \mathbf{p}_w = \mathbf{q}, \sum_{w \in \mathcal{W}} \alpha_w = 1 \Big\}. \end{aligned} \tag{5}$$

Moreover, we define $\mathcal{S}(\mathbf{T}) \triangleq \{(\mathbf{p}, f(\mathbf{p}), g(\mathbf{T}\mathbf{p})) \,|\, \mathbf{p} \in \Delta_m\}$. The next lemma is a direct generalization of [12, Lemma 2.1].

Lemma 1. *$\mathcal{C}(\mathbf{T})$ is convex and compact with $\mathcal{C}(\mathbf{T}) = \mathsf{conv}\mathcal{S}(\mathbf{T})$. In addition, all points in $\mathcal{C}(\mathbf{T})$ can be written as a convex combination of at most $m+1$ points of $\mathcal{S}(\mathbf{T})$; in other words, $|\mathcal{W}| \leq m+1$.*

Let the upper and lower boundaries of $\mathcal{C}$ be denoted by $U_{\mathbf{T}}$ and $L_{\mathbf{T}}$, respectively, i.e., we have

$$L_{\mathbf{T}}(\mathbf{q}, x) \triangleq \inf\{y | (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\}, \tag{6}$$

$$U_{\mathbf{T}}(\mathbf{q}, x) \triangleq \sup\{y | (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\}. \tag{7}$$

Under appropriate conditions on x (depending on the choice of f), $\{y | (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\}$ is non-empty, and hence the compactness of $\mathcal{C}(\mathbf{T})$ allows one to replace infimum and supremum in (6) and (7) with minimum and maximum, respectively. Moreover, it follows from the convexity of $\mathcal{C}(\mathbf{T})$ that $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ is convex and $U_{\mathbf{T}}(\mathbf{q}, \cdot)$ is concave.

C. Dual Formulations

Since $\mathcal{C}(\mathbf{T})$ is a convex set, its upper and lower boundaries are equivalently represented by its supporting hyperplanes. We use the dual approach introduced in [12] to evaluate $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ and $U_{\mathbf{T}}(\mathbf{q}, \cdot)$. For a given λ, define the conjugate function

$$L^*_{\mathbf{T}}(\mathbf{q}, \lambda) \triangleq \min\{-\lambda x + y | (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\}. \tag{8}$$

Note that the graph of $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ is the lower boundary of $\mathcal{C}(\mathbf{T})$. It then follows that the point (x, y) that achieves the minimum in (8) lies on the lower boundary of $\mathcal{C}(\mathbf{T})$ with supporting line of slope λ, and hence corresponds to a point $(x, L_{\mathbf{T}}(\mathbf{p}, x))$.

We now turn our attention to evaluating $L^*_{\mathbf{T}}(\mathbf{q}, \cdot)$. Let

$$\mathcal{S}_\lambda(\mathbf{T}) \triangleq \{(\mathbf{p}, y - \lambda x) \,|\, (\mathbf{p}, x, y) \in \mathcal{S}(\mathbf{T})\}. \tag{9}$$

We observe that $\mathcal{S}_\lambda(\mathbf{T})$ is the graph of the function $\phi(\cdot, \lambda)$ on Δ_m given by

$$\phi(\mathbf{p}, \lambda) = g(\mathbf{T}\mathbf{p}) - \lambda f(\mathbf{p}), \mathbf{p} \in \Delta_m. \tag{10}$$

Since the mapping $y - \lambda x$ preserves convexity, we have

$$\mathcal{C}_\lambda(\mathbf{T}) \triangleq \mathsf{conv}\mathcal{S}_\lambda(\mathbf{T}) = \{(\mathbf{q}, y - \lambda x) \,|\, (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\} \tag{11}$$

as the convex hull of the graph of $\phi(\cdot, \lambda)$. Thus, $L^*_{\mathbf{T}}(\mathbf{q}, \lambda)$ is given by the lower convex envelope of $\phi(\cdot, \lambda)$ at $\mathbf{q}$. The same would go, *mutatis mutandis*, for $U_{\mathbf{T}}(\mathbf{q}, x)$: its conjugate function $U^*_{\mathbf{T}}(\mathbf{q}, \cdot)$, defined as

$$U^*_{\mathbf{T}}(\mathbf{q}, \lambda) \triangleq \max\{-\lambda x + y | (\mathbf{q}, x, y) \in \mathcal{C}(\mathbf{T})\}$$

coincides with the upper concave envelope of $\phi(\cdot, \lambda)$ at $\mathbf{q}$.

These properties leads to a procedure for characterizing $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ and $U_{\mathbf{T}}(\mathbf{q}, \cdot)$. We illustrate $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ in details and $U_{\mathbf{T}}(\mathbf{q}, \cdot)$ will follow by using concave envelope instead. For $L^*_{\mathbf{T}}(\mathbf{q}, \lambda) = z^*$, there are two scenarios:

1) **Trivial case**: If $(\mathbf{q}, z^*)$ is in both $\mathcal{S}_\lambda$ and $\mathcal{C}_\lambda$, then $z^* = g(\mathbf{T}\mathbf{q}) - \lambda f(\mathbf{q})$. In this case, $(\mathbf{q}, x, L_{\mathbf{T}}(\mathbf{q}, x))$ simply reduces to $(\mathbf{q}, f(\mathbf{q}), g(\mathbf{T}\mathbf{q}))$, and the optimal W has $P_W(w) = 1$ for some w, independent of X.
2) **Non-trivial case**: If $z^* \neq \phi(\mathbf{q}, \lambda)$, then $(\mathbf{q}, z^*) \in \mathcal{C}_\lambda$ is the convex combination of points $(\mathbf{p}_i, \phi(\mathbf{p}_i, \lambda)) \in \mathcal{S}_\lambda$, with weights α_i where $i \in [k]$ for some $k \geq 2$, and $\sum_{i=1}^k \alpha_i = 1$. Then $(\mathbf{q}, x, L_{\mathbf{T}}(\mathbf{q}, x))$ is given by $\sum_{i=1}^k \alpha_i (\mathbf{p}_i, f(\mathbf{p}_i), g(\mathbf{T}\mathbf{p}_i))$. Moreover, an optimal W is attained by $P_W(i) = \alpha_i$ and $\mathbf{p}_w = \mathbf{p}_i$, $i \in [k]$.

Algorithm 1 Computing $(x, L_{\mathbf{T}}(\mathbf{q}, x))$ at slope λ

Input: $\lambda, \mathbf{q}$

Output: $(x, L_{\mathbf{T}}(\mathbf{q}, x))$

Compute $\phi(\mathbf{p}, \lambda)$, $\mathbf{p} \in \Delta_m$

$L^*_{\mathbf{T}}(\mathbf{p}, \lambda) \leftarrow$ convex envelope of $\phi(\mathbf{p}, \lambda)$

if $L^*_{\mathbf{T}}(\mathbf{q}, \lambda) = \phi(\mathbf{q}, \lambda)$ **then**

  return $(f(\mathbf{q}), g(\mathbf{Tq}))$

else

  $(\alpha_i, \mathbf{p}_i)_{i=1}^k \leftarrow \left\{ \left(a_i, \phi^{-1}(b_i, \lambda)\right)_{i=1}^k | b_i \in \phi(\cdot, \lambda), \sum_{i=1}^k a_i b_i = L^*_{\mathbf{T}}(\mathbf{q}, \lambda) \right\}$

  return $\left(\sum_{i=1}^k \alpha_i f(\mathbf{p}_i), \sum_{i=1}^k \alpha_i g(\mathbf{Tp}_i)\right)$

Hence, the points on the graph of $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ can be obtained by only considering the points of $\phi(\cdot, \lambda)$ which differs from its convex envelope $L^*_{\mathbf{T}}(\cdot, \lambda)$ since those are exactly the points where W is not induced from the trivial case. **Algorithm 1** summarizes our previous discussions.

D. Matched Channels

The geometry of bottleneck problems leads to intriguing properties of $\mathbf{p}_w$. Our previous discussions reveal that the points $\{\mathbf{p}_i\}_{i=1}^k$ used to form the convex envelope of $\phi(\cdot, \lambda)$ are special: they determine a channel $P_{X|W}$ such that for any distribution P_W, the resulting value of $(\mathbf{q}, \mathbb{E}[f(\mathbf{p}_w)], \mathbb{E}[g(\mathbf{Tp}_w)])$ is on the boundary of $\mathcal{C}(\mathbf{T})$ with supporting line of slope λ. In this case, we say that the points $\{\mathbf{p}_i\}_{i=1}^k$ form a *matched channel* for $\mathbf{T}$ and f, g.

Definition 1 (Matched Channel)**.** *For a fixed channel $P_{Y|X}$ and f, g, we say that $P_{X|W}$ is matched to $P_{Y|X}$ if there exists P_W such that $|\mathcal{W}| \geq 2$ and*

$$(\mathbb{E}[f(\mathbf{p}_w)], \mathbb{E}[g(\mathbf{Tp}_w)]) = (x, L_{\mathbf{T}}(\mathbf{q}, x)). \tag{12}$$

Using an elementary result in convex geometry (see Lemma 2 in Appendix.), we immediately have the following theorem.

Theorem 1. *Let $P_{X|W} = \mathbf{p}_w$ be a matched channel for $P_{Y|X}$. Then for any P_W, we have*

$$\left(\mathbb{E}[f(\mathbf{p}_w)], \mathbb{E}[g(\mathbf{Tp}_w)]\right) = (x, L_{\mathbf{T}}(\mathbf{q}, x)). \tag{13}$$

Proof. See the Appendix. □

From Theorem 1, we know that for any distribution P_X, matched channels $P_{X|W}$ are entirely determined by the points on the curve $\phi(\cdot, \lambda)$ whose convex combinations lead to the convex envelope of ϕ at P_X. It implies that as long as $\phi(\cdot, \lambda)$ meets its convex envelope at P_X, small perturbation around P_X does not change the matched channels $P_{X|W}$ but simply change the weight α_w. Thus, optimal mappings $P_{W|X}$ are surprisingly robust to small errors in estimation of P_X, which could potentially give pragmatic advantages when applying bottleneck problems to real data. However, if P_X changes, we can recover the matched channels by first solving α_i via $\mathbf{q} = \sum_{i=1}^{m+1} \alpha_i \mathbf{p}_w(i)$, where $\mathbf{p}_w(i) = P_{X|W=i}$, and then applying Bayes' rules.

Note that the properties above only hold when f and g do not depend on P_X. Specifically, matched channels do not exist for the cases studied in Section III-C and III-D.

III. Generalizing Bottleneck Problems

In this section, we demonstrate how the tools developed in Section II can be applied to new bottleneck problems of the form (1) and (2). We then revisit the IB and PF, and also study their estimation-theoretic variants.

Consider the Markov chain $W \to X \to Y$. Our goal is to describe the achievable pairs of f-divergences

$$(D_{f_1}(P_{WX} \| P_W P_X), D_{f_2}(P_{WY} \| P_W P_Y)). \tag{14}$$

Observe that for a given P_X we have

$$D_{f_1}(P_{WX} \| P_W P_X) \tag{15}$$

$$= \sum_{w \in \mathcal{W}} P_W(w) \left[\sum_{x \in \mathcal{X}} P_X(x) f_1\left(\frac{P_{X|W}(x|w)}{P_X(x)}\right) \right] \tag{16}$$

$$= \sum_{w \in \mathcal{W}} P_W(w) D_{f_1}\left(P_{X|W} \| P_X\right), \tag{17}$$

and hence $I_{f_1}(W; X)$ can be expressed as

$$I_{f_1}(W; X) = \sum_{w \in \mathcal{W}} \alpha_w f(\mathbf{p}_w) = \mathbb{E}\big[f(\mathbf{p}_w)\big], \tag{18}$$

for some function f. Similarly, define $D_{f_2}\left(P_{Y|W} \| P_Y\right) = g(P_{Y|W})$, we have

$$I_{f_2}(W; Y) = \sum_{w \in \mathcal{W}} \alpha_w g(\mathbf{Tp}_w) = \mathbb{E}\big[g(\mathbf{Tp}_w)\big]. \tag{19}$$

Hence the corresponding set $\{(\mathbf{q}, I_{f_1}(W; X), I_{f_2}(W; Y))\}$ for varying W has the same form as $\mathcal{C}(\mathbf{T})$. Letting

$$\phi(\mathbf{p}, \lambda) = D_{f_2}(\mathbf{Tp} \| P_Y) - \lambda D_{f_1}(\mathbf{p} \| P_X), \tag{20}$$

we can thus apply Algorithm 1 to characterize $\mathsf{B}_{f_1, f_2}(P_{XY}, \cdot)$ and $\mathsf{F}_{f_1, f_2}(P_{XY}, \cdot)$.

Next, we show that how the usual IB and PF fit in our formulation, and study their estimation-theoretic counterparts. We note, however, that the previous analysis does not require $f_1 = f_2$.

A. Information Bottleneck

Assuming $f_1(t) = f_2(t) = t \log t$ in the bottleneck functional (1), we have $D_{f_1}(P_{WX} \| P_W P_X) = I(W; X)$ and $D_{f_2}(P_{WY} \| P_W P_Y) = I(W; Y)$. Thus, the set of points $\{(x, U_{\mathbf{T}}(\mathbf{q}, x)) | 0 \leq x \leq H(X)\}$ corresponds to the set of solutions of the IB problem.

It is worth mentioning that the same geometric approach can also be applied directly to entropy functions which also leads to the IB formulation. In fact, this is exactly the setting studied in [12]. Specifically, choosing $f_1 = h_m$ and $f_2 = h_n$, the set of points

$$\{(h_m(\mathbf{q}) - x, h_n(\mathbf{Tq}) - L_{\mathbf{T}}(\mathbf{q}, x)) | 0 \leq x \leq h_m(\mathbf{q})\} \tag{21}$$

also corresponds to the set of solutions of the IB. The IB is closely related to strong data processing inequalities. See [11, Proposition 2] for more details in the case of the BSC (see also Fig. 1 (right)).

B. Privacy Funnel

Assuming $f_1(t) = f_2(t) = t\log t$ in the funnel functional (2), the set $\{(x, L_{\mathbf{T}}(\mathbf{q}, x)) \,|0 \le x \le H(X)\}$ corresponds to the set of solutions of the privacy funnel, introduced in [6]. Equivalently, using the entropy function as in Section III-A, the set of points

$$\{(h_m(\mathbf{q}) - x, h_n(\mathbf{T}\mathbf{q}) - U_{\mathbf{T}}(\mathbf{q}, x)) \,|0 \le x \le h_m(\mathbf{q})\} \tag{22}$$

also corresponds to the set of solutions, which follows from the fact that $U_{\mathbf{T}}(\mathbf{q}, \cdot)$ is monotonically non-decreasing; see Fig. 1 (right).

C. Estimation Bottleneck

One can move away from the usual IB and define new bottleneck problems by considering different functions f_1 and f_2. For instance, if $f_1(t) = f_2(t) = t^2 - 1$, then the corresponding f-information, called χ^2-information, is defined as

$$I_{f_1}(W;X) = \chi^2(W;X) \triangleq \mathbb{E}\left[\left(\frac{P_{W,X}(W,X)}{P_W(W)P_X(X)}\right)\right] - 1, \tag{23}$$

We simplify the notation in (1) for χ^2-information as

$$\mathsf{B}_{\chi^2}(P_{XY}, x) \triangleq \max_{W \to X \to Y} \chi^2(W;Y) \,\mathsf{s.t.}\, \chi^2(W;X) \le x, \tag{24}$$

The reason to specifically study χ^2- information are two-fold. First, it has been shown in [6] that $\chi^2(X;Y) = \sum_{i=1}^{d} \lambda_i(X;Y)$, where $\lambda_i(X;Y)$ is the i^{th} principal inertia component (PIC) of $P_{X,Y}$ and $d = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$. Moreover, if the PICs between X and Y are large, then the minimum mean square error (MMSE) $\mathsf{mmse}(X|Y)$ of estimating X given Y will be small [6, Theorem 1], thereby making reliable estimations. Hence, if the goal of an estimation problem is to minimize $\mathsf{mmse}(Y|W)$, we can equivalently consider maximizing $\chi^2(Y;W)$. Second, following the spirit of the IB, we also add the constraint $\chi^2(W;X) \le x$ for the new representation W, as χ^2-divergence serves as sharp bounds for any f-divergence [17].

Due to the above connection between $\mathsf{B}_{\chi^2}(P_{XY}, x)$ and estimation problems, we call $\mathsf{B}_{\chi^2}(P_{XY}, x)$ *Estimation Bottleneck (EB)* problem. Clearly, the set of points $\{(x, U_{\mathbf{T}}(\mathbf{q}, x)) \,|0 \le x \le m - 1\}$ corresponds to the set of solutions of (24); see Fig. 1 (left). The bound $x \le m - 1$ comes from the PIC analysis [6].

D. Estimation Privacy Funnel

Motivated by the connection between χ^2-information and estimation problems mentioned in Section III-C, we propose

$$\mathsf{F}_{\chi^2}(P_{XY}, x) \triangleq \min_{W \to X \to Y} \chi^2(W;Y) \,\mathsf{s.t.}\, \chi^2(W;X) \ge x, \tag{25}$$

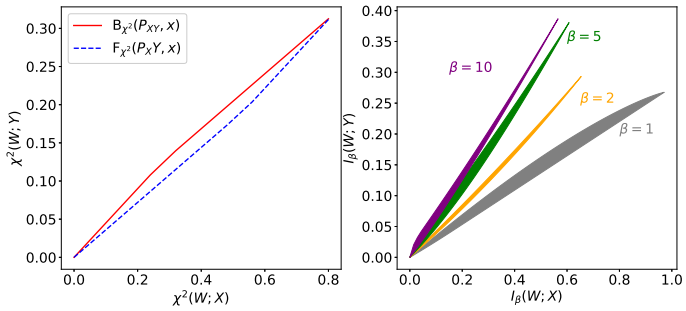

Fig. 1. $\mathbf{T}$ follows BSC with crossover probability $\delta = 0.1$ and $P\{X = 1\} = q = 0.1$. **Left**: The estimation bottleneck and privacy funnel. **Right**: Set of achievable pairs of Arimoto mutual information $\{I_\beta(W;X), I_\beta(W;Y)\}$ for BSC with $\delta = 0.2$ and $q = 0.4$. Note that when $\beta = 1$, the upper and lower boundaries correspond to the IB and the PF.

where the privacy is measured in terms of MMSE. The practical significance of (25) is justified as follows. Suppose Y represents private data (*e.g.* political preferences) and X (*e.g.* movie rating) is correlated with Y. The main objective, formulated by (25), is to construct a privacy-assuring mapping $P_{X|W}$ such that the information disclosed about Y by W is minimized, thus minimizing privacy leakage, while preserving the estimation efficiency that W provides about X. Similarly, the solutions of the estimation privacy funnel (25) correspond to the set of points $\{(x, L_{\mathbf{T}}(\mathbf{q}, x)) \,|0 \le x \le m - 1\}$ with $f_1(t) = f_2(t) = t^2 - 1$; see Fig. 1 (left).

IV. Mrs. and Mr. Gerber's Lemmas

The study of upper and lower boundaries of achievable mutual information pairs are essential in multi-user information theory [13]. In the binary symmetric case, we not only rephrase Mrs. Gerber's lemma [13], but also derive its counterpart for the PF. Furthermore, we discuss analogous results to Mrs. and Mr. Gerber's lemmas for Arimoto conditional entropy.

A. Mr. Gerber's Lemma

We apply the duality argument, developed in Section II-C, for $L_{\mathbf{T}}$ and $U_{\mathbf{T}}$ to characterize the IB and the PF the binary symmetric case. In particular, let $P_{Y|X}$ be the BSC with crossover probability δ and $q = \mathsf{Pr}(X = 1) \le \frac{1}{2}$. For $a \in [0,1]$, denote $\bar{a} = 1 - a$. We denote by $h_b(q)$ the binary entropy function $h_2([q, \bar{q}])$.

Let $f_1 = f_2 = h_2$. It was shown in [12] that

$$L_{\mathbf{T}}(q, x) = h_b\left(\delta \star h_b^{-1}(x)\right), \forall x \in [0, h_b(q)], \tag{26}$$

where $h_b^{-1} : [0,1] \to [0, \frac{1}{2}]$ is the inverse function of $h_b(\cdot)$ and $a \star b \triangleq (1-a)b + (1-b)a$, for $a, b \in [0,1]$. Eq. (26) is well-known as *Mrs. Gerber's Lemma (MGL)*. In this case, the matched channel is also a BSC with crossover probability $h_b^{-1}(x)$. Using the approach outlined in Section II, we derive a counterpart result for the upper boundary $U_{\mathbf{T}}$, and call it *Mr. Gerber's Lemma*.

Theorem 2 (Mr. Gerber's Lemma)**.** *For $0 \le q \le \frac{1}{2}$, we have*

$$U_{\mathbf{T}}(q, x) = \alpha h_b\left(\delta \star \frac{q}{z}\right) + \bar{\alpha} h_b(\delta), \tag{27}$$

where $x = \alpha h_b\left(\frac{q}{z}\right)$ *and* $z = \max(\alpha, 2q)$, $\alpha \in [0,1]$.

Proof. See the Appendix. □

In summary, in the binary symmetric case, the set of solutions for the IB follows from Mrs. Gerber's Lemma (26) and is given by $\{(h_b(q) - x, h_b(q \star \delta) - L_{\mathbf{T}}(q,x))\}$, and the set of solutions for the PF follows from Mr. Gerber's Lemma (27) and is given by $\{(h_b(q) - x, h_b(q \star \delta) - U_{\mathbf{T}}(q,x))\}$. The upper and lower boundaries of the achievable pairs $\{I(W;X), I(W;Y) : W \to X \to Y\}$ are therefore characterized by Mrs. and Mr. Gerber's Lemmas, respectively.

B. Achievable Pairs of Arimoto Conditional Entropy

Beside χ^2-divergence and the entropy functions, one can choose ℓ^β-norm $\|\cdot\|_\beta$ for f and g in (4), which results in Arimoto's version of conditional Rényi entropy (Arimoto conditional entropy) [15] of order $\beta > 1$:

$$H_\beta(X|W) \triangleq \frac{\beta}{1-\beta} \log \sum_{w \in \mathcal{W}} \alpha_w \|\mathbf{p}_w\|_\beta . \tag{28}$$

When $\beta = 1$, we define $H_1(X|W) = H(X|W)$. Hence, the set of achievable Arimoto conditional entropy pairs $(H_\beta(X|W), H_\beta(Y|W))$ can be obtained by the nonlinear mapping:

$$(x,y) \mapsto \left(\frac{\beta}{1-\beta}\log x, \frac{\beta}{1-\beta}\log y\right), (x,y) \in \mathcal{C}(\mathbf{T}) . \tag{29}$$

With (28) at hand, Arimoto mutual information [15] of order $\beta > 1$ can be defined as $I_\beta(X;W) \triangleq H_\beta(X) - H_\beta(X|W)$, where $H_\beta(X)$ is the Rényi entropy of order β. Arimoto conditional entropy has been proven useful in approximating the minimum error probability of Bayesian M-ary hypothesis testing [15].

C. Arimoto's Mr. and Mrs. Gerber's Lemmas

Due to the importance of Arimoto conditional entropy [15], we study the extensions of Mr. and Mrs. Gerber's Lemmas for Arimoto conditional entropy, naming them Arimoto's Mr. and Mrs. Gerber's Lemmas respectively.

Let $K_\beta(X|W) = \exp\left\{\frac{1-\beta}{\beta} H_\beta(X|W)\right\}$, and also $K_\beta(X) = \exp\left\{\frac{1-\beta}{\beta} H_\beta(X)\right\}$. Since $H_\beta(X|W) \le H_\beta(X)$ for $\beta > 1$ and the mapping $x \mapsto \exp\left\{\frac{1-\beta}{\beta} x\right\}$ is strictly decreasing, we have $K_\beta(X|W) \ge K_\beta(X)$. Define $L_{\mathbf{T}}(q,x)$ and $U_{\mathbf{T}}(q,x)$ respectively as the minimum and maximum of $K_\beta(Y|W)$ when $K_\beta(X|W) = x$ for $K_\beta(X) \le x \le 1$. For simplicity, denote $K_\beta(q) = K_\beta(X)$ if $X \sim$ Bernoulli(q). In this case, following section II-C, we have $\phi(p,\lambda) = K_\beta(\delta \star p) - \lambda K_\beta(p)$, which leads to the following theorem.

Theorem 3 (Arimoto's Mrs. Gerber's Lemma)**.** *For* $0 \le q \le 1/2$*, let* $\mathcal{L}^{(\beta)} \triangleq \{(x, L_{\mathbf{T}}(q,x)) \,|K_\beta(q) \le x \le 1\}$ *and for* $\beta > 1$*, we have*

$$\mathcal{L}^{(\beta)} = \{(K_\beta(p), K_\beta(p \star \delta)) \,|0 \le p \le q\} \tag{30}$$

In particular, $\frac{\beta}{1-\beta}\log y = \min_{W \to X \to Y} H_\beta(Y|W)$ *s.t.* $H_\beta(X|W) \ge \frac{\beta}{1-\beta}\log x$ *for* $(x,y) \in \mathcal{L}^{(\beta)}$.

Proof. See the Appendix. □

Analogous to this theorem, we also obtain the following generalization of Mr. Gerber's Lemma.

Theorem 4 (Arimoto's Mr. Gerber's Lemma)**.** *For* $0 \le q \le 1/2$*, let* $\mathcal{U}^{(\beta)} \triangleq \{(x, U_{\mathbf{T}}(q,x)) \,|K_\beta(q) \le x \le 1\}$ *and for* $\beta > 1$*, we have*

$$\mathcal{U}^{(\beta)} = \left\{\left(\bar\alpha + \alpha K_\beta\left(\frac{q}{z}\right), \alpha K_\beta\left(\frac{q}{z} \star \delta\right) + \bar\alpha K_\beta(\delta)\right)\right\}, \tag{31}$$

where $\alpha \in [0,1], z = \max\{\alpha, 2q\}$. *In particular, we have* $\frac{\beta}{1-\beta}\log y = \max_{W \to X \to Y} H_\beta(Y|W)$ *s.t.* $H_\beta(X|W) \le \frac{\beta}{1-\beta}\log x$ *for* $(x,y) \in \mathcal{U}^{(\beta)}$.

Consequently, for $\beta > 1$, Arimoto's Mrs. and Mr. Gerber's Lemmas jointly characterize the achievable sets $\{I_\beta(W;X), I_\beta(W;Y) : W \to X \to Y\}$; see Fig. 1 (right).

V. Final Remarks

In this paper, we study the geometric structure behind bottleneck problems, and generalize the IB and PF to a broader class of f-divergences. In particular, we consider estimation-theoretic variants of the IB and PF. Moreover, we show how bottleneck problems can be used to calculate the counterpart of Mrs. Gerber's lemma (called Mr. Gerber's Lemma), and derive versions of Mrs. and Mr. Gerber's lemmas for Arimoto conditional entropy. These results can be potentially useful for new applications of information theory in machine learning.

Appendix

A. Lemma 2

Lemma 2 ([18])**.** *Let* $\mathcal{A}$ *be a connected and non-empty subset of* $\mathbb{R}^n$*, and* $\mathcal{B} = \mathsf{conv}\mathcal{A}$*. Assume there exists* $\mathbf{x} \in \partial\mathcal{B}$ *such that* $x \notin \mathcal{A}$*, then there exists* $\{\mathbf{x}_i\}_{i=1}^m$*, where* $m \le n$ *with* $\mathbf{x} = \sum_{i\in[m]} \alpha_i \mathbf{x}_i$*,* $\mathbf{x}_i \in \partial\mathcal{A}$*,* $0 < \alpha_i < 1$*, and* $\sum_i \alpha_i = 1$*. Furthermore,* $\mathsf{conv}\{\mathbf{x}_i\}_{i=1}^m \subseteq \partial\mathcal{B}$.

B. Proof of Theorem 1 (Matched Channel)

Recall that $L_{\mathbf{T}}(\mathbf{q}, \cdot)$ is determined parametrically in λ by the points where $\phi(\cdot,\lambda)$ does not match its convex envelope $L^*_{\mathbf{T}}(\cdot,\lambda)$. Thus, the columns of the channel transformation matrix of a matched channel correspond to extreme points $P_{X|W}(\cdot|i) = \mathbf{p}_i$ where $\phi(\cdot,\lambda)$ matches $L^*_{\mathbf{T}}(\cdot,\lambda)$. However, there exists α_i where the convex combination $\sum_{i\in[k]} \alpha_i \mathbf{p}_i = \mathbf{q}$ corresponds to a point $\phi(\mathbf{q},\lambda) \ne L^*_{\mathbf{T}}(\mathbf{q},\lambda)$. Using lemma 2, any non-trivial convex combination of $\mathbf{p}_i$ will result in a point $\mathbf{q}$ which is on the convex envelope of $\phi(\cdot,\lambda)$ and determines a corresponding point on the curve $L_{\mathbf{T}}(\mathbf{q},\cdot)$.

C. Proof of Theorem 2 (Mr. Gerber's Lemma)

Take $f = g = h_2$, we have $\phi(p,\lambda) = h_b(p \star \delta) - \lambda h_b(p)$. For $\lambda \geq (1-2\delta)^2$, $\phi(\cdot,\lambda)$ is convex in p, and $U^*_{\mathbf{T}}(q,\lambda) = h_b(\delta)$. For $0 \leq \lambda < (1-2\delta)^2$, $\phi(\cdot,\lambda)$ is concave in a region centered at $p = \frac{1}{2}$, where it reaches a local maximum. Consequently, if $\phi\left(\frac{1}{2},\lambda\right) < h_b(\delta)$, the upper convex envelope of $\mathcal{S}_\lambda$ is the linear combination of $(0,\phi(0,\lambda))$ and $(1,\phi(1,\lambda))$ and $U_{\mathbf{T}}(q,x) = h_b(\delta)$. Assuming $p \leq \frac{1}{2}$, if $\phi\left(\frac{1}{2},\lambda\right) > h_b(\delta)$, then there exists $p_\lambda \in \left[0,\frac{1}{2}\right]$ such that for $p \leq p_\lambda$, $(p, U^*_{\mathbf{T}}(q,\lambda)) \in \mathcal{C}_\lambda$ is a convex combination of $(0,\phi(0,\lambda))$ and $(p_\lambda,\phi(p_\lambda,\lambda))$. Finally, if $\phi\left(\frac{1}{2},\lambda\right) = h_b(\delta)$, then any point on the upper convex envelope of $\mathcal{S}_\lambda$ also lies in $conv\left\{\phi(0,\lambda), \phi\left(\frac{1}{2},\lambda\right), \phi(1,\lambda)\right\}$.

Hence, assuming $p \leq \frac{1}{2}$, the distribution $P_{W,X}$ that achieves $U_{\mathbf{T}}(q,x)$ will be of two cases:

1) $Pr(X=1|W=0) = 0$, $Pr(X=1|W=1) = \frac{p}{\alpha}$ with $Pr(W=1) = \alpha$, $2p \leq \alpha \leq 1$.
2) W assuming values in $\{0,1,2\}$ with $Pr(X=1|W=0) = 0$, $Pr(X=1|W=1) = 1$, $Pr(X=1|W=2) = \frac{1}{2}$, with $Pr(W=0) = 1-p-\frac{\alpha}{2}$, $Pr(W=1) = p - \frac{\alpha}{2}$, $Pr(W=2) = \alpha$ and $0 \leq \alpha \leq 2p$.

Rearranging (1) and (2), the result in (27) follows.

D. Proof of Theorem 3 (Arimoto's Mr. Gerber's Lemma)

Since $\phi(\cdot,\lambda)$ is convex for $\lambda \leq (1-2\delta)^2$. For $\lambda > (1-2\delta)^2$, $\phi''(\cdot,\lambda)$ is negative on an interval $[p_\lambda, \bar{p_\lambda}]$, symmetric at $p = \frac{1}{2}$, and is positive elsewhere with local maximum at $p = \frac{1}{2}$. By symmetry, the lower convex envelope of the graph $\phi(\cdot,\lambda)$ is obtained by replacing $p = p_\lambda$ if $p \in [p_\lambda, \bar{p_\lambda}]$. Therefore, for a given $q \leq \frac{1}{2}$, if $p_\lambda \geq q$, then $(q, L_{\mathbf{T}}(q,\lambda))$ is a convex combination of $(p_\lambda,\phi(p_\lambda,\lambda))$ and $(\bar{p_\lambda},\phi(\bar{p_\lambda},\lambda))$. Hence, we have $L_{\mathbf{T}}(q,x) = K_\beta(p \star \delta)$ for $x = K_\beta(p)$ and $0 \leq p \leq q$.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of IEEE Allerton*, 2000.

[2] N. Tishby and N. Slonim, "Data clustering by markovian relaxation and the information bottleneck method," in *Proc. of NIPS*, 2001.

[3] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. of ACM SIGIR*, 2000.

[4] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. of IEEE ITW*, 2015.

[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[6] F. P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5011–5038, 2017.

[7] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Proc. of IEEE ISIT*, 2015.

[8] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. of IEEE ISIT*, 2007.

[9] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan 2016.

[10] ——, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC)*, vol. 6, pp. 2012–2016, 2014.

[11] F. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities in power-constrained gaussian channels," in *Proc. of IEEE ISIT*, 2015.

[12] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, 1975.

[13] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[14] C. Nair, "Upper concave envelopes and auxiliary random variables," *Int. J. Adv. Eng. Sci. Appl. Math.*, vol. 5, no. 1, pp. 12–20, 2013.

[15] I. Sason and S. Verdú, "Arimoto-Rényi conditional entropy and bayesian m-ary hypothesis testing," *arXiv preprint arXiv:1701.01974*, 2017.

[16] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems - extended version." [Online]. Available: https://github.com/HsiangHsu/ISIT-18-Extended-Version

[17] A. Makur and L. Zheng, "Bounds between contraction coefficients," in *Proc. of IEEE Allerton*, 2015.

[18] H. G. Eggleston, *Convexity*. Wiley Online Library, 1966.