Refined Pinsker’s and Reverse Pinsker’s Inequalities for Probability Distributions of Different Dimensions

MICHELE CAPRIO
PRECISE Center, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: caprio@seas.upenn.edu

This work was supported in part by the National Science Foundation under Grant CCF 1934964, and in part by the Army Research Office under Grant ARO MURI W911NF2010080.

ABSTRACT We provide optimal lower and upper bounds for the augmented Kullback-Leibler divergence in terms of the augmented total variation distance between two probability measures defined on two Euclidean spaces having different dimensions. We call them refined Pinsker’s and reverse Pinsker’s inequalities, respectively.

INDEX TERMS Kullback-Leibler divergence, total variation distance, optimal bounds, probability measures of different dimensions.

I. INTRODUCTION
Bounding the Kullback-Leibler (KL) divergence between probability measures (pm’s) defined on the same space in terms of their total variation (TV) distance is a well studied problem, of paramount importance in statistics and machine learning. Famous lower bounds are given by Pinsker’s inequality [5] and Vajda’s lower bound [12], while a famous upper bound is given by reverse Pinsker’s inequality [3], [11]. These results are particularly useful in Bayesian nonparametrics [2] and in the optimal quantization of pm’s [3].

In this note, we generalize results from [1] and [6] to find the optimal (defined below) lower and upper bounds for the KL divergence between pm’s defined on two Euclidean spaces having different dimensions in terms of their TV distance. The generalizations of KL divergence and TV distance to pm’s of different dimensions are called augmented KL divergence (AKL) and augmented total variation distance (ATV), respectively, and were first introduced in [4]. The AKL and the ATV could be used to measure the loss of information after projecting a probability measure $P$ down to a lower-dimensional subspace, e.g. via principal component analysis (PCA). That is, an interesting open research question is to determine whether the larger the AKL or the ATV between $P$ and its projection $\text{Proj}(P)$, the more likely it is to lose information in the projecting process, and if such loss depends on the projection we use. Another interesting information-theoretic application of AKL is the following: it can be used to calculate the divergence between two different dimensional distributions in the field of multi-target labeled probability distributions of a hybrid of continuous state and discrete label variables [8, Remark 3 and Equation (63)].

The main result of this paper, Theorem 10, states that for any given value $\delta$ of the ATV between two generic distributions defined on Euclidean spaces having different dimensions, we can give optimal lower and upper bounds to their augmented KL divergence.\footnote{As we shall see, the upper bound requires a mild assumption to hold.}

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.
When $P$ and $Q$ are defined on the same space, “optimality” should be understood as follows. For the refined Pinsker’s inequality, we mean the best lower bound on the KL divergence between $P$ and $Q$ given that their TV distance is some fixed value $\delta \geq 0$, that is, $\inf_{\delta} d_{TV}(P,Q) = \delta$ $D_{KL}(P\|Q)$. For the refined reverse Pinsker’s inequality, we mean the best upper bound on the KL divergence between $P$ and $Q$ over the class $A(\delta,m,M)$ of pm’s whose TV distance is equal to $\delta \geq 0$ and whose relative density $dP/dQ$ has finite lower and upper bounds $m$ and $M$, respectively, introduced in Definition 3.2. That is, $\sup_{P,Q\in A(\delta,m,M)} D_{KL}(P\|Q)$. As we can see, the meaning of “optimality” for the upper bound is slightly less general than that for the lower bound. As pointed out in [11, Section 1], this is due to the fact that for any $\varepsilon > 0$, there exists a pair $P,Q$ of pm’s such that $d_{TV}(P,Q) \leq \varepsilon$ while $D_{KL}(P\|Q) = \infty$. Consequently, a reverse Pinsker’s inequality which provides an upper bound on the KL divergence between $P$ and $Q$ when their TV distance is some fixed $\delta \geq 0$ may not exist in general, whence the necessity of working with $A(\delta,m,M)$. The generalizations of these “optimality” concepts to AKL and ATV are given in section III.

The note is divided as follows. Section II gives the needed background, and section III presents our main result. Section IV is a discussion.

II. PRELIMINARIES

A. PROBABILITY MEASURES ON THE SAME MEASURABLE SPACE

Pick two pm’s $P$, $Q$ defined on the same measurable space $(\Omega,\mathcal{F})$ and assume $P$ is absolutely continuous with respect to $Q$, written $P \ll Q$. This means that $Q(A) = 0$ implies $P(A) = 0$, $A \in \mathcal{F}$. Denote by $dP/dQ$ the relative density of $P$ with respect to $Q$, that is, $dP/dQ \equiv f$ is an $\mathcal{F}$-measurable functional on $\Omega$ such that for all $A \in \mathcal{F}$,

$$P(A) = \int_A f \, dQ.\tag{1}$$

Then, the KL divergence and the TV distance between $P$ and $Q$ are defined as

$$D_{KL}(P\|Q) := \int_\Omega \log \left(\frac{dP}{dQ}\right) \, dP$$

and $d_{TV}(P,Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|,\tag{1}$

respectively.3 Consider the following function, that – given some $\delta \geq 0$ – selects the smallest possible value of the KL divergence between pm’s whose TV distance is equal to $\delta$

$$\delta \mapsto L(\delta) := \inf_{d_{TV}(P,Q) = \delta} D_{KL}(P\|Q).\tag{2}$$

It is called the Vajda’s lower bound [12]. The following comes from [6, Theorem 1].

**Theorem 1 (Fedotov, Harremoës, and Topsøe):** Pick two probability measures $P$, $Q$ defined on a generic measurable space $(\Omega,\mathcal{F})$ and assume $d_{TV}(P,Q) = \delta \geq 0$. Then, curve $\gamma : \delta \mapsto (\delta, L(\delta))$ is a differentiable curve in the $(d_{TV},D_{KL})$-plane, symmetric around the $D_{KL}$-axes. In addition, using $t = \frac{d\delta}{d\gamma} \in \mathbb{R}_+$ as a parameter, $\gamma$ is parametrized by

$$\delta(t) = t \left(1 - \left(\coth(t) - \frac{1}{t}\right)^2\right)$$

$$L(\delta(t)) = \log \left(\frac{t}{\sinh(t)}\right) + t \coth(t) - \frac{t^2}{\sinh^2(t)} + \frac{2}{4}.\tag{3}$$

In [9, Corollary 1], the authors give an explicit value for $L(\delta)$, in contrast with (3) where the value is implicit.

**Corollary 2 (Reid and Williamson):** Pick two probability measures $P$, $Q$ defined on a generic measurable space $(\Omega,\mathcal{F})$ and assume $d_{TV}(P,Q) = \delta \geq 0$. Then,

$$L(\delta) = \min_{\gamma \in [\delta - 2,2 - \delta]} \left[\frac{\delta + 2 - \gamma}{4} \log \left(\frac{\gamma - 2 - \delta}{\gamma - 2 + \delta}\right)\right].\tag{4}$$

We now define $A(\delta,m,M)$, a set of pairs of probabilities that will be useful in the rest of the work. Before doing so, we need to introduce the concepts of essential infimum $\essinf f$ and essential supremum $\esssup f$ of $dP/dQ \equiv f$ with respect to $Q$. We have that

$$\essinf f := \inf\{b \in \mathbb{R} : Q(\{\omega \in \Omega : f(\omega) < b\}) = 0\},$$

$$\esssup f := \inf\{a \in \mathbb{R} : Q(\{\omega \in \Omega : f(\omega) > a\}) = 0\}.$$ 

**Definition 3:** Fix $\delta \geq 0$, $m > 0$, and $M < \infty$. We call $A(\delta,m,M)$ the set of all pm’s pairs $(P,Q)$ defined on a common measurable space $(\Omega,\mathcal{F})$ satisfying

1) $P \ll Q$,
2) $\essinf \frac{dP}{dQ} = m$,
3) $\esssup \frac{dP}{dQ} = M$,
4) $d_{TV}(P,Q) = \delta$.

The optimal upper bound for the KL divergence between (a pair of) pm’s belonging to $A(\delta,m,M)$ is defined as

$$U(A(\delta,m,M)) := \sup_{(P,Q)\in A(\delta,m,M)} D_{KL}(P\|Q).\tag{5}$$

We have the following important result.

**Theorem 4:** Pick any $\delta \geq 0$, $m > 0$, $M < \infty$, and assume $A(\delta,m,M) \neq \emptyset$. Then, for all $(P,Q) \in A(\delta,m,M)$, the following are optimal bounds

$$L(\delta) \leq D_{KL}(P\|Q) \leq U(A(\delta,m,M)).\tag{6}$$

**Proof:** The optimal upper bound for $D_{KL}(P\|Q)$ comes from equation (5). Its value, given in [1, Equation (9)], is

$$U(A(\delta,m,M)) = \delta \left(\frac{\log(M^{-1})}{1 - M^{-1}} + \frac{\log(m^{-1})}{m^{-1} - 1}\right).\tag{7}$$

The optimal lower bound comes from equation (3). An implicit parametric solution of the form of the graph of Vajda’s lower bound as $(V(t),L(t))_{t \in \mathbb{R}_+}$ is given in
Theorem 1, while an explicit value for $L(\delta)$ is given in Corollary 2.

Notice that in the case where $m = 1$ or $M = 1$, any $(P, Q) \in A(\delta, m, M)$ must be such that $\delta = d_{TV}(P, Q) = 0$. The right hand side of (7) is then understood as being equal to 0. In addition, the assumption that the pair $(P, Q)$ belongs to $A(\delta, m, M)$ is only needed to obtain the upper bound in (6), as pointed out in section I.

In [6, Theorem 7], the authors find a lower bound for $L(\delta)$ that makes computing a lower bound for the KL divergence in terms of the TV metric easier.

**Theorem 5 (Fedotov, Harremoës, and Topsøe):** Pick two probability measures $P, Q$ defined on a generic measurable space $(\Omega, \mathcal{F})$ and assume $d_{TV}(P, Q) = \delta \geq 0$. Then, the following is true

$$L(\delta) \geq \frac{1}{2} \delta^2 + \frac{1}{36} \delta^4 + \frac{1}{270} \delta^6 + \frac{221}{340200} \delta^8.$$ 

### B. PROBABILITY MEASURES ON TWO EUCLIDEAN SPACES WITH DIFFERENT DIMENSIONS

In this paper, we adopt the framework of [4] to prove a version of Theorems 4 and 5 for pm’s pairs $(P, Q)$ defined on two Euclidean spaces having different dimensions. Let $M(\Omega)$ denote the set of all Borel pm’s on $\Omega \subset \mathbb{R}^n$. For convenience, we restrict our attention to pm’s with densities so that we do not have to keep track of which measure is absolutely continuous to which other measure [4, Section III]; this is without loss of generality. Let $\lambda^n$ be the Lebesgue measure restricted to $\Omega \subset \mathbb{R}^n$. With respect to $\lambda^n$, we define

$$M_{\text{dens}}(\Omega) := \{\mu \in M(\Omega) : \mu \text{ has density}\}.$$ 

Notice that $\mu \in M_{\text{dens}}(\Omega)$ if and only if it is absolutely continuous with respect to $\lambda^n$. The Lebesgue measure is chosen because it is the most common measure; it can be substituted by any measure satisfying the condition that for any nonzero area, the measure of said area is positive. This requirement is needed to make $D_{KL}^{-}, D_{KL}^{+}, d_{TV}^{-},$ and $d_{TV}^{+}$ in Theorem 8 well defined.

We now introduce the machinery that we use to project a pm to a lower dimensional space and to embed a pm to a higher dimensional space. For any $d, n \in \mathbb{N}$, $d \leq n$, let

$$O(d, n) := \{V \in \mathbb{R}^{d \times n} : VV^T = I_d\},$$

that is, the Stiefel manifold of $d \times n$ matrices with orthonormal rows. For any $V \in O(d, n)$ and $b \in \mathbb{R}^d$, let

$$\varphi_{V, b} : \mathbb{R}^n \rightarrow \mathbb{R}^d, \quad x \mapsto \varphi_{V, b}(x) := Vx + b,$$

and for any $\mu \in M(\mathbb{R}^n)$, let $\varphi_{V, b}(\mu) \equiv \varphi_{V, b, \mu}$ be the pushforward of measure $\mu$ through function $\varphi_{V, b}$. That is, for every element $A$ of the sigma-algebra endowed to $\mathbb{R}^d$, $\varphi_{V, b}(\mu)(A) \equiv \varphi_{V, b, \mu}(A) := \mu(\varphi^{-1}_{V, b}(A)).$

**Definition 6:** Let $d, n \in \mathbb{N}$, $d \leq n$. For any $P \in M(\mathbb{R}^d)$ and $Q \in M(\mathbb{R}^n)$, the set of embeddings of $P$ into $\mathbb{R}^n$ is

$$\Phi^+(P, n) := \{\varphi \in M(\mathbb{R}^n) : \varphi_{V, b}(\alpha) = P, \text{ for some } V \in O(d, n), b \in \mathbb{R}^d\}$$

and the set of projections of $Q$ onto $\mathbb{R}^d$ is

$$\Phi^-(Q, d) := \{\beta \in M(\mathbb{R}^d) : \varphi_{V, b}(Q) = \beta, \text{ for some } V \in O(d, n), b \in \mathbb{R}^d\}.$$ 

**Remark 7:** Definition 6 is stating the following. The set of embeddings of a probability measure $P$ (defined on $\mathbb{R}^d$) onto $\mathbb{R}^n$, $n \geq d$, is given by those probabilities on $\mathbb{R}^n$ whose pushforward through function $\varphi_{V, b}$ recovers $P$, for some $V \in O(d, n)$ and $b \in \mathbb{R}^d$. The set of projections of a probability measure $Q$ (defined on $\mathbb{R}^n$) onto $\mathbb{R}^d$, $n \geq d$, is given by those probabilities on $\mathbb{R}^d$ that can be written as the pushforward of $Q$ through function $\varphi_{V, b}$, for some $V \in O(d, n)$ and $b \in \mathbb{R}^d$.

An important subset of $\Phi^+(P, n)$ is

$$\Phi_{\text{dens}}^+(P, n) := \{\alpha \in M_{\text{dens}}(\mathbb{R}^n) : \varphi_{V, b}(\alpha) = P, \text{ for some } V \in O(d, n), b \in \mathbb{R}^d\}.$$

The following relevant result comes from [4, Theorem III.4].

**Theorem 8 (Cai and Lim):** Let $d, n \in \mathbb{N}$, $d \leq n$. For any $P \in M(\mathbb{R}^d)$ and $Q \in M(\mathbb{R}^n)$, let

$$D_{KL}^{\lambda}(P || Q) := \inf_{\beta \in \Phi^-(Q, d)} D_{KL}(P || \beta),$$

$$D_{KL}^{+}(P || Q) := \inf_{\alpha \in \Phi^+(P, n)} D_{KL}(\alpha || Q),$$

$$d_{TV}^{-}(P, Q) := \inf_{\beta \in \Phi^-(Q, d)} d_{TV}(P, \beta),$$

$$d_{TV}^{+}(P, Q) := \inf_{\alpha \in \Phi^+(P, n)} d_{TV}(\alpha, Q).$$

Then,

$$D_{KL}^{\lambda}(P || Q) = D_{KL}^{+}(P || Q) \equiv \hat{D}_{KL}(P || Q)$$

and

$$d_{TV}^{-}(P, Q) = d_{TV}^{+}(P, Q) \equiv \hat{d}_{TV}(P, Q).$$

We call $\hat{D}_{KL}(P || Q)$ the augmented KL divergence ($AKL$), while $\hat{d}_{TV}(P, Q)$ the augmented TV distance ($ATV$). Notice that [4, Lemma III.2] guarantees the existence of quantities $D_{KL}^{\lambda}(P || Q), D_{KL}^{+}(P || Q), d_{TV}^{-}(P, Q),$ and $d_{TV}^{+}(P, Q)$.

### III. MAIN RESULT

Consider function

$$\delta \mapsto \hat{L}(\delta) := \inf_{\beta \in \Phi^-(Q, d)} \hat{D}_{KL}(P || Q).$$

Being the augmented counterpart of (2), we call it the augmented Vajda’s lower bound. Denote by $\alpha \in \Phi_{\text{dens}}^+(P, n)$ and $\beta \in \Phi^-(Q, d)$ the pm’s such that $\hat{D}_{KL}(P || Q) = D_{KL}(\alpha || Q) = D_{KL}(P || \beta)$, that is,

$$\alpha = \arg \inf_{\beta \in \Phi_{\text{dens}}^+(P, n)} D_{KL}(\alpha || Q)$$

and

$$\beta = \arg \inf_{\beta \in \Phi^-(Q, d)} D_{KL}(P || \beta).$$

Let then

$$\essinf \frac{d\alpha}{dQ} = m_1, \quad \esssup \frac{d\alpha}{dQ} = M_1.$$
\[
\text{ess inf } \frac{dP}{d\beta} = m_2, \quad \text{ess sup } \frac{dP}{d\beta} = M_2. \quad (11)
\]

Notice that (10) are taken with respect to \( P \), while (11) are taken with respect to \( \beta \). They correspond to (2) and (3) in Definition 3. We need to bound the relative densities \( d\mu/dQ \) and \( dP/d\beta \) otherwise we may have that \( \hat{d}_{TV}(P, Q) = \delta \), but \( \hat{D}_{KL}(P\|Q) = \infty \), similarly to what we pointed out in section I. We now define a set of pairs of probabilities that is the augmented counterpart of Definition 3.

**Definition 9:** Pick \( d, n \in \mathbb{N} \) such that \( d \leq n \). Fix \( \delta \geq 0 \), \( m_1, m_2 > 0 \) and \( M_1, M_2 < \infty \). \( \mathcal{A}(\delta, m_1, m_2, M_1, M_2) \) is the set of all \( pm \)'s belonging to the set \( \mathcal{A}(\delta, m_1, m_2, M_1, M_2) \) is the augmented Vajda's lower bound, then, is given by

\[
\hat{D}_{KL}(P\|Q). \quad (12)
\]

The following is our main result.

**Theorem 10:** Pick \( d, n \in \mathbb{N} \) such that \( d \leq n \). Fix \( \delta \geq 0 \), \( m_1, m_2 > 0 \), and \( M_1, M_2 < \infty \). Assume \( \mathcal{A}(\delta, m_1, m_2, M_1, M_2) \neq \emptyset \). Pick any \( (P, Q) \) in \( \mathcal{A}(\delta, m_1, m_2, M_1, M_2) \) and let

\[
\text{pol}_{\hat{d}_{TV}} := \frac{1}{2} \hat{d}_{TV}(P, Q)^2 + \frac{1}{36} \hat{d}_{TV}(P, Q)^4 + \frac{1}{270} \hat{d}_{TV}(P, Q)^6 + \frac{221}{340200} \hat{d}_{TV}(P, Q)^8.
\]

Then,

\[
\text{pol}_{\hat{d}_{TV}} \leq \hat{L}(\delta) \leq \hat{D}_{KL}(P\|Q) \leq \hat{U}(\mathcal{A}(\delta, m_1, m_2, M_1, M_2)). \quad (13)
\]

Before proving our result, let us remark that assuming \( (P, Q) \in \mathcal{A}(\delta, m_1, m_2, M_1, M_2) \) is only needed to upper bound \( \hat{D}_{KL}(P\|Q) \). The reason is that otherwise such upper bound may not exist, as pointed out earlier in this section. In addition, the second and the third inequalities in (13) are optimal. Finally, notice that there is an elegant relationship between Theorem 5 and the first inequality in (13). We can lower bound Vajda’s bound \( L(\delta) \) and the augmented Vajda’s bound \( \hat{L}(\delta) \) by the same polynomial, the first one in \( \delta = d_{TV}(P, Q) \) and the second one in \( \delta = \hat{d}_{TV}(P, Q) \).

**Proof:** The proof has four steps.

1. We first show that \( \text{pol}_{\hat{d}_{TV}} \leq \hat{D}_{KL}(P\|Q) \). We have that

\[
\hat{D}_{KL}(P\|Q) = \inf_{\beta \in \Phi^{-}(Q,d)} D_{KL}(P\|\beta)
\]

\[
\geq \inf_{\beta \in \Phi^{-}(Q,d)} \left[ \frac{1}{2} d_{TV}(P, \beta)^2 + \frac{1}{36} d_{TV}(P, \beta)^4 + \frac{1}{270} d_{TV}(P, \beta)^6 + \frac{221}{340200} d_{TV}(P, \beta)^8 \right]
\]

\[
\geq \frac{1}{2} d_{TV}(P, \beta)^2 + \frac{1}{36} d_{TV}(P, \beta)^4 + \frac{1}{270} d_{TV}(P, \beta)^6 + \frac{221}{340200} d_{TV}(P, \beta)^8.
\]

Here, the equality comes from Theorem 8, the first inequality is a consequence of Theorems 4 and 5, and the second inequality comes from Theorem 8 and the fact that the infimum of a sum is not smaller than the sum of the infima. Notice that if we substitute \( \inf_{\beta \in \Phi^{-}(Q,d)} \) with \( \inf_{\beta \in \Phi_{dens}(P,m)} \) the proof still holds thanks to Theorem 8.

2. The fact that \( \hat{D}_{KL}(P\|Q) \geq \hat{L}(\delta) \) comes from equation (8) and the assumption that \( \hat{d}_{TV}(P, Q) = \delta \). We also have the following result.

**Claim 11:** A version of parametrization (3) holds for \( \hat{L}(\delta) \). Let then

\[
\beta' = \arg \inf_{\beta \in \Phi^{-}(Q,d)} d_{TV}(P, \beta).
\]

If \( \beta = \beta' \), a version of equation (4) holds for \( \hat{L}(\delta) \).

**Proof:** To prove the first part of the claim, we begin by showing that \( \hat{D}_{KL} \) is convex, jointly in \( P \) and \( Q \). To see this, notice that, given two generic probability measures \( P, Q \) on the same measurable space \( \Omega, \mathcal{F} \), [6, Section II] points out that \( \hat{D}_{KL} \) is strictly convex, jointly in \( P \) and \( Q \). In our case, we have that \( \hat{D}_{KL}(P\|Q) = \inf_{\beta \in \Phi^{-}(Q,d)} D_{KL}(P\|\beta) \) because the infimum operator preserves convexity, we can conclude that \( \hat{D}_{KL} \) is convex, jointly in \( P \) and \( Q \). In addition, we have that \( \hat{L}(\delta) := \inf_{\hat{d}_{TV}(P, Q) = \delta} \hat{D}_{KL}(P\|Q) \).

Given the convexity of \( \hat{D}_{KL} \), and since the infimum operator preserves convexity, we can conclude that \( \hat{L} \) is convex as well. These convexity results entail that for any \( \delta \geq 0 \) for which \( \hat{d}_{TV}(P, Q) = \delta \), there exists a unique pair \( (P, Q) \in M(\mathbb{R}^d) \times M(\mathbb{R}^n) \) of probability measures such that \( \hat{D}_{KL}(P\|Q) \) is minimal among all distributions with augmented total variation equal to \( \delta \).

The augmented Vajda’s lower bound, then, is given by the function \( \delta \mapsto \hat{L}(\delta) = \hat{D}_{KL}(P_\delta\|Q_\delta) \). Let now \( \gamma \) denote the map \( \delta \mapsto (\delta, \hat{L}(\delta)) \). Parameter \( \delta \) cannot be used to give an explicit parametrization of \( \gamma \). Since both \( \hat{D}_{KL} \) and \( \hat{L} \) are convex functions, the convex conjugate [10] of both these functions can be explicitly calculated. To prove the statement, we follow the proof of [6, Theorem 1]. There, the authors use parameter \( t = d\hat{L}/d\delta \) from the convex conjugate of \( \hat{L} \) to parametrize \( \hat{L} \).

Before going on, we give two remarks. The first one is that in [6] the authors work with the so-called signed total variation metric \( \hat{d}_{TV}(P, Q) := 2 \sup_{\gamma \in \mathcal{F}} (P - Q(A)) \in [-2, 2] \) between probability measures defined on the same measurable space. This is merely a convenience choice (it is easier to obtain parametrization (3)), since

\[
d_{TV}(P, Q) = \begin{cases} \frac{1}{2} \hat{d}_{TV}(P, Q) & \text{if } \hat{d}_{TV}(P, Q) \geq 0 \\ -\frac{1}{2} \hat{d}_{TV}(P, Q) & \text{if } \hat{d}_{TV}(P, Q) < 0. \end{cases}
\]
Given that $d_{TV}^*$ is an f-divergence [4, Section I], Theorem 8 holds also if we use $d_{TV}^*$ in place of $d_{TV}$. In particular,

$$\hat{d}_{TV}(P, Q) = \inf_{\beta \in \Phi^-(Q, d)} d_{TV}(P, \beta)$$

$$= \inf_{\in \Phi^+(P, \alpha)} d_{TV}(\alpha, Q).$$

Notice that, because

$$\hat{d}_{TV}(P, Q) = \begin{cases} \frac{1}{2} \hat{d}_{TV}(P, Q) & \text{if } \hat{d}_{TV}(P, Q) \geq 0 \\ -\frac{1}{2} \hat{d}_{TV}(P, Q) & \text{if } \hat{d}_{TV}(P, Q) < 0, \end{cases}$$

in the proof that follows we abuse notation and denote by $\delta$ both the value of $\hat{d}_{TV}(P, Q)$ and that of $\hat{d}_{TV}^*(P, Q)$. The second remark is that in [6] the authors consider a two-elements state space on which $P$ and $Q$ are defined. As they highlight in [6, Section II], this simplification is without loss of generality since their results hold even in a continuous or a non-commutative setting.

In our more general case, we keep this simplification: we assume that $P$ is defined on the two-elements state space $\Omega = \{\omega_1, \omega_2\}$, so $P = (p_1, p_2 = 1 - p_1)$, $Q$ is defined on a higher-dimensional state space, and set $\Phi^-(Q, d)$ of projections of $Q$ onto $\Omega$ is a subset of $M(\Omega)$. This entails that $\beta = (p_1, p_2 = 1 - p_1)$.

Let $D_{KL}(p_1, p_2) := p_1 \log(p_1 / \beta_1)$. The convex conjugate of $D_{KL}$ is

$$\hat{D}^*(x, y) = \sup_{p_1, \beta_1} \left( \frac{x}{y} \right) \left( \frac{p_1}{\beta_1} - D_{KL}(p_1, \beta_1) \right).$$

We have

$$\frac{\partial}{\partial p_1} \left( \frac{x}{y} \right) \left( \frac{p_1}{\beta_1} - D_{KL}(p_1, \beta_1) \right) = x - \log \left( \frac{p_1}{\beta_1} \right) + \log \left( \frac{p_2}{\beta_2} \right)$$

$$= y + \frac{p_1}{\beta_1} - \frac{p_2}{\beta_2}.$$

To find the point where these partial derivatives are 0, we solve the simultaneous equations

$$x = \log \left( \frac{p_1}{\beta_1} \right) - \log \left( \frac{p_2}{\beta_2} \right)$$

$$y = \frac{p_2}{\beta_2} - \frac{p_1}{\beta_1},$$

whose solutions are

$$p_1 = e^t x + e^t y - 1$$

$$\beta_1 = \frac{1}{1 - e^t} - \frac{1}{y},$$

we obtain

$$\hat{D}^*(-2t, 2t) = \sup_{x, y} \left( \frac{x}{y} \right) \left( \frac{p_1}{\beta_1} \right) = t \delta.$$

Hence,

$$\hat{D}^*(-2t, 2t) = \sup_{x, y} \left( \frac{x}{y} \right) \left( \frac{p_1}{\beta_1} \right) = t \delta.$$

is the convex conjugate of $\hat{L}$, and $t$ must be the derivative of $\hat{L}$. We see that (14) and (15) solve our optimization problem. The parametrization of $\gamma$

$$\delta(t) = t \left( 1 - \left( \frac{\coth(t) - 1}{t} \right)^2 \right)$$

$$\hat{L}(\delta(t)) = \log \left( \frac{t}{\sinh(t)} \right) + t \coth(t) - \frac{t^2}{\sinh^2(t)}$$

is then obtained by direct evaluation of the quantities involved. A visual representation of $\hat{L}(\delta(t))$ is given in Figure 1.

Suppose now that $\beta = \beta' = \beta^*$. Then, this implies that we can write $\hat{L}(\delta)$ as $\inf_{d_{TV}(P, \beta^*) = \delta} D_{KL}(P, \beta^*)$.

Corollary 2 entails that for

$$\hat{d}_{TV}(P, Q) = d_{TV}(P, \beta^*) = \delta \geq 0,$$

we have that

$$\hat{L}(\delta) = \sup_{d_{TV}(P, Q) = \delta} \hat{D}_{KL}(P, Q)$$

$$= \sup_{d_{TV}(P, \beta^*) = \delta} D_{KL}(P, \beta^*)$$

$$= \min_{\gamma \in [\delta - 2, \delta]} \left( \frac{\delta + 2 - \gamma}{4} \log \left( \frac{\delta - 2 - \gamma}{\gamma - 2 + \delta} \right) \right.\left. + \left( \frac{\gamma + 2 - \delta}{4} \log \left( \frac{\gamma + 2 - \delta}{\gamma + 2 + \delta} \right) \right) \right).$$

Notice that in this case the parametrization in (3) holds too, but it is better to express $\hat{L}(\delta)$ explicitly as in (17).

(III) The fact that $\hat{D}_{KL}(P, Q) \leq \hat{U}(\mathcal{A}(\delta, m_1, m_2, M_1, M_2))$ comes from equation (12). We also have the following result.

Claim 12: A version of equation (7) holds for $\hat{U}(\mathcal{A}(\delta, m_1, m_2, M_1, M_2))$.

The use of the augmented signed total variation is clear here; had we used the augmented total variation (as defined in Theorem 8) instead, we would have equated $\gamma(y) \to \delta(y)$, since $\hat{d}_{TV}(P, Q) \in [0, 1]$, for all $(P, Q) \in M(\mathbb{R}^d) \times M(\mathbb{R}^d)$.
Once more, the equalities come from Theorem 8, and the second inequality comes from the fact that the infimum of a sum is not smaller than the sum of the infima. The last equality comes from our assumption that $\hat{d}_{TV}(P, Q) = \delta$. Notice that if we substitute $\inf_{\beta \in \Phi^+(Q, d)} \hat{d}_{TV}(\beta)$ with $\inf_{\alpha \in \Phi^+(Q, d)} \hat{d}_{TV}(\alpha)$, the proof still holds thanks to Theorem 8.

Theorem 10 is extremely important because for a given value $\delta$ of the augmented TV distance between two generic distributions, it gives us immediately a lower bound for the augmented KL divergence. In addition, if the essential suprema and essential infima in (10) and (11) are well defined, Theorem 10 also gives an upper bound for the augmented KL divergence. The next example gives another interesting byproduct of our main result.

Example 13: Consider a one-dimensional Gaussian distribution and write $\rho_1 = \mathcal{N}(\mu_1, \sigma^2)$, where $\mu_1 \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance. Consider then an $n$-dimensional Gaussian distribution and write $\rho_2 = \mathcal{N}(\nu, \Sigma)$, where $\nu \in \mathbb{R}^n$ is the mean vector and $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix. Call $\xi_1$ and $\xi_n$ the largest and smallest eigenvalues of $\Sigma$, respectively. Then, [4, Example VI.2] shows that

$$\hat{D}_{KL}(\rho_1 \| \rho_2) = \begin{cases} 
\frac{1}{2} \left[ \frac{\sigma^2}{\xi_n} - 1 + \log \left( \frac{\xi_n}{\sigma^2} \right) \right] & \text{if } \sigma < \sqrt{\xi_n} \\
\frac{1}{2} \left[ \frac{\sigma^2}{\xi_1} - 1 + \log \left( \frac{\xi_1}{\sigma^2} \right) \right] & \text{if } \sigma > \sqrt{\xi_n} \\
0 & \text{otherwise.}
\end{cases}$$

Call now $\xi$ the value taken by $\hat{D}_{KL}(\rho_1 \| \rho_2)$. Then, by Theorem 10, we find an upper bound to $\hat{d}_{TV}(\rho_1, \rho_2)$ by solving

$$\frac{1}{2} \delta^2 + \frac{1}{36} \delta^6 + \frac{1}{270} \delta^6 + \frac{221}{340200} \delta^8 \leq \xi,$$

where $\delta = \hat{d}_{TV}(\rho_1, \rho_2)$. 

Proof: We have that

$$\hat{D}_{KL}(P \| Q) = \inf_{\beta \in \Phi^+(Q, d)} D_{KL}(P \| \beta) \leq \frac{1}{\beta} (\log(M_2^{-1}) + \log(m_2^{-1})) 1 - M_2^{-1} + m_2^{-1} =: U_2.$$ 

Once more, the equalities come from Theorem 8, and the inequality comes from equation (7). We also have that

$$\hat{D}_{KL}(P \| Q) = \inf_{\alpha \in \Phi^+(P, n)} D_{KL}(\alpha \| Q) \leq \frac{1}{\alpha} (\log(M_2^{-1}) + \log(m_2^{-1})) 1 - M_2^{-1} + m_2^{-1} =: U_1.$$ 

Notice that in the case where $m_1 = 1$ or $M_1 = 1$, then $U_1$ is understood as being equal to 0. A similar reasoning holds for the case where $m_2 = 1$ or $M_2 = 1$, with $U_2$ in place of $U_1$. 

(IV) Finally, we show that pol$_{\hat{d}_{TV}} \leq \hat{L}(\delta)$. We have that

$$\hat{L}(\delta) := \inf_{\delta \in \Phi^-(Q, d)} \hat{D}_{KL}(P \| Q) = \inf_{\delta \in \Phi^-(Q, d)} \inf_{\beta \in \Phi^-(Q, d)} D_{KL}(P \| \beta) \geq \inf_{\delta \in \Phi^-(Q, d)} \inf_{\beta \in \Phi^-(Q, d)} \left[ \frac{1}{2} \hat{d}_{TV}(P, \beta)^2 + \frac{1}{36} \hat{d}_{TV}(P, \beta)^6 \right] \geq \frac{1}{2} \hat{d}_{TV}(P, Q)^2 + \frac{1}{36} \hat{d}_{TV}(P, Q)^6 + \frac{1}{270} \hat{d}_{TV}(P, Q)^6 + \frac{221}{340200} \hat{d}_{TV}(P, Q)^8.$$ 

Here, the first equality comes from definition (8), the second equality comes from Theorem 8, the first inequality comes from Theorem 5, and the second inequality comes from Theorem 10, and the last equality comes from our assumption that $\hat{d}_{TV}(P, Q) = \delta$. Notice that if we substitute $\inf_{\beta \in \Phi^+(Q, d)} \hat{d}_{TV}(\beta)$ with $\inf_{\alpha \in \Phi^+(P, n)} \hat{d}_{TV}(\alpha)$, the proof still holds thanks to Theorem 8. 

Theorem 10 is extremely important because for a given value $\delta$ of the augmented TV distance between two generic distributions, it gives us immediately a lower bound for the augmented KL divergence. In addition, if the essential suprema and essential infima in (10) and (11) are well defined, Theorem 10 also gives an upper bound for the augmented KL divergence. The next example gives another interesting byproduct of our main result.

Example 13: Consider a one-dimensional Gaussian distribution and write $\rho_1 = \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance. Consider then an $n$-dimensional Gaussian distribution and write $\rho_2 = \mathcal{N}(\nu, \Sigma)$, where $\nu \in \mathbb{R}^n$ is the mean vector and $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix. Call $\xi_1$ and $\xi_n$ the largest and smallest eigenvalues of $\Sigma$, respectively. Then, [4, Example VI.2] shows that

$$\hat{D}_{KL}(\rho_1 \| \rho_2) = \begin{cases} 
\frac{1}{2} \left[ \frac{\sigma^2}{\xi_n} - 1 + \log \left( \frac{\xi_n}{\sigma^2} \right) \right] & \text{if } \sigma < \sqrt{\xi_n} \\
\frac{1}{2} \left[ \frac{\sigma^2}{\xi_1} - 1 + \log \left( \frac{\xi_1}{\sigma^2} \right) \right] & \text{if } \sigma > \sqrt{\xi_n} \\
0 & \text{otherwise.}
\end{cases}$$

Call now $\xi$ the value taken by $\hat{D}_{KL}(\rho_1 \| \rho_2)$. Then, by Theorem 10, we find an upper bound to $\hat{d}_{TV}(\rho_1, \rho_2)$ by solving

$$\frac{1}{2} \delta^2 + \frac{1}{36} \delta^6 + \frac{1}{270} \delta^6 + \frac{221}{340200} \delta^8 \leq \xi,$$

where $\delta = \hat{d}_{TV}(\rho_1, \rho_2)$. 

FIGURE 1. Visual representation of $\hat{L}(\delta)$ in (16). As we can see, it is symmetric around the $D_{KL}(P \| Q)$ axis, which implies that using $\hat{d}_{TV}$ in place of $d_{TV}$ does not yield any loss of generality.
If instead we let \( \rho_1 \) and \( \rho_2 \) be a truncated one- and \( n \)-dimensional Gaussian, respectively, then we can use \( \hat{D}_{KL}(\rho_1 \| \rho_2) \leq \hat{U}(A(\delta, m_1, m_2, M_1, M_2)) \) from Theorem 10 and

\[
\hat{U}(A(\delta, m_1, m_2, M_1, M_2)) = \max \left\{ \hat{d}_{TV}(P, Q) \left( \frac{\log(M_1^{-1}) + \log(m_1^{-1})}{1 - M_1^{-1}} \right), \right. \\
\left. \hat{d}_{TV}(P, Q) \left( \frac{\log(M_2^{-1}) + \log(m_2^{-1})}{1 - M_2^{-1}} \right) \right\},
\]

from Claim 12 to find a lower bound for \( \delta = \hat{d}_{TV}(\rho_1, \rho_2) \). Notice that in this case we need the Gaussians to be truncated to ensure the essential suprema and essential infima in (10) and (11) are well defined.

Before concluding we point out that generalizing the proof that leads to equation (17) to the \( \beta \neq \beta' \) case is not easy; although we conjecture that a similar result holds, this will be the subject of future studies.

IV. CONCLUSION

In this note, we presented optimal upper and lower bounds for the augmented KL divergence in terms of the augmented TV distance. This is just the first step towards a deep study of augmented divergences that ideally should include structural properties, statistical analysis, duality, possible applications, and many more aspects. We plan to be at the forefront of this process.

More concretely, in the near future we plan to find bounds for more augmented divergences in terms of augmented metrics and vice versa, in the spirit of [7]. It would be especially interesting to generalize [9, Th. eorem 6] to the augmented framework of [4]. An encouraging result of this kind is presented in [4, Corollary III.6]: the authors give a bound for the augmented TV metric in terms of the augmented Hellinger squared divergence. We also plan to extend the second part of Claim 11 to the \( \beta \neq \beta' \) case.

ACKNOWLEDGMENT

The authors would like to thank Edric Tam, Yuhang Cai, and Vittorio Orlandi for their help with technical details, and Insup Lee, Oleg Sokolsky, Souradeep Dutta, Radoslav Ivanov, Kuk Jang, and Vivian Lin for inspiring this project and helpful discussions. Their deepest gratitude goes also to Sayan Mukherjee for covering the article processing charges and to two anonymous references for their generous suggestions regarding content and presentation.

REFERENCES

[1] O. Binette, “A note on reverse Pinsker inequalities,” IEEE Trans. Inf. Theory, vol. 65, no. 7, pp. 4094–4096, Jul. 2019.
[2] O. Binette and S. Guillotte, “Bayesian nonparametrics for directional statistics,” J. Stat. Planning Inference, vol. 216, pp. 118–134, Jan. 2022.
[3] G. Böcherer and B. C. Geiger, “Optimal quantization for distribution synthesis,” IEEE Trans. Inf. Theory, vol. 62, no. 11, pp. 6162–6172, Nov. 2016.
[4] Y. Cai and L.-H. Lim, “Distances between probability distributions of different dimensions,” IEEE Trans. Inf. Theory, vol. 68, no. 6, pp. 4020–4031, Jun. 2022.
[5] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[6] A. A. Fedotov, P. Harremoës, and F. Topsøe, “Refinements of Pinsker’s inequality,” IEEE Trans. Inf. Theory, vol. 49, no. 6, pp. 1491–1498, Jun. 2003.
[7] A. L. Gibbs and F. E. Su, “On choosing and bounding probability metrics,” Int. Statist. Rev., vol. 70, no. 3, pp. 419–435, Dec. 2002.
[8] T. Li, “A technical note on (labeled) RFS- AA fusion: Derivation from PHD consistency,” 2022, arXiv:2209.10433.
[9] D. M. Reid and C. R. Williamson, “Generalised Pinsker inequalities,” in Proc. 22nd Conf. Learn. Theory, Montreal, QC, Canada, 2009, pp. 1–21.
[10] R. T. Rockafellar, Convex Analysis. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[11] I. Sason and S. Verdú, “f-divergence inequalities,” IEEE Trans. Inf. Theory, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[12] I. Vajda, “Note on discrimination information and variation,” IEEE Trans. Inf. Theory, vol. IT-16, no. 6, pp. 771–773, Nov. 1970.

Michele Caprio received the B.Sc. and M.Sc. degrees in economics from Bocconi University, Milan, Italy, in 2015 and 2018, respectively, and the Ph.D. degree in statistics from Duke University, Durham, NC, USA, in 2022. He is currently a Postdoctoral Researcher at the PRECISE Center, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. His broad research interests include foundations of probability, mathematical statistics, and AI. More specifically, he is interested in imprecise probabilities and their applications to statistics and AI. Dr. Caprio was awarded the Aleane Webb Dissertation Research Fellowship and the IMS Hannan Travel Award, and he was the Finalist for the NESS Student Research Award, in 2022.