On the Universality of the Logistic Loss Function

Amichai Painsky and Gregory Wornell
EECS Department, MIT,
Cambridge, MA 02139,
Email: {amichai, gww}@mit.edu

***Abstract*—A loss function measures the discrepancy between the true values (observations) and their estimated fits, for a given instance of data. A loss function is said to be proper (unbiased, Fisher consistent) if the fits are defined over a unit simplex, and the minimizer of the expected loss is the true underlying probability of the data. Typical examples are the zero-one loss, the quadratic loss and the Bernoulli log-likelihood loss (log-loss). In this work we show that for binary classification problems, the divergence associated with smooth, proper and convex loss functions is bounded from above by the Kullback-Leibler (KL) divergence, up to a multiplicative normalization constant. It implies that by minimizing the log-loss (associated with the KL divergence), we minimize an upper bound to any choice of loss functions from this set. This property justifies the broad use of log-loss in regression, decision trees, deep neural networks and many other applications. In addition, we show that the KL divergence bounds from above any separable Bregman divergence that is convex in its second argument (up to a multiplicative normalization constant). This result introduces a new set of divergence inequalities, similar to the well-known Pinsker inequality.**

I. Introduction

Consider a weather forecaster that estimates the probability of rain on the following day. Its performance may be evaluated by different statistical measures. For example, we may count the number of times it assessed the chance of rain as greater than $t = 50\%$, while it eventually did not rain (and vice versa). This corresponds to a 0-1 loss (Table I). Alternatively, we may choose different threshold values, t, or completely different measures (quadratic loss, Bernoulli log-likelihood loss, etc.). Choosing a "good" measure is a well-studied problem, mostly in the context of *scoring rules* in decision theory [1]. Assuming that the desired measure is known in advance, the weather forecaster may be designed accordingly, to minimize that measure. In practice, different tasks entail inferring different information from the provided estimates. It means that the forecaster shall be designed according to a single measure that is "suitable" for a variety of possible purposes. This requirement is obviously quite challenging.

In this work we address this problem, as we show that for binary classification, the Bernoulli log-likelihood loss (log-loss) is a "universal" choice which dominates any alternative "analytically convenient" loss function (smooth, proper and convex). Specifically, we show that by minimizing the log-loss we minimize the regret (defined in Section II) associated with all possible alternatives in this set. This result justifies the use of log-loss in many learning applications, as it is the only measure that provides such universality guarantees.

Over the years, the log-loss was shown to have several favorable properties (for example, [2]–[4]). However, these properties are mostly motivated by information-theoretic principles. Here, we justify the use of log-loss directly from a decision theory perspective.

In addition, we show that our universality result may be viewed from a divergence analysis viewpoint, as we show that the divergence associated with the log-loss (KL divergence) bounds from above any separable Bregman divergence that is convex in its second argument, up to a multiplicative normalization constant. This result provides a new set of divergence inequalities, which have a similar nature as the well-known Pinsker inequality [2]. In that sense, our Bregman analysis may be viewed as a complementary set of results to the well known Pinsker-like f-divergence inequalities [5].

II. Basic Definitions

Let $Y \in \{0, 1\}$ be a Bernoulli distributed binary random variable with a parameter p. Let $\hat{Y}$ be an estimate of Y. A loss function $l(y, \hat{y})$ quantifies the difference between a realization of Y and its corresponding estimate. In this work we focus on probabilistic estimates, for which $\hat{y} \triangleq q \in [0, 1]$. In other words, $\hat{y} \triangleq q$ is a "soft" decision that corresponds to the probability of the event $y = 1$ (as opposed to a "hard decision" in which $\hat{y} \in \{0, 1\}$). A Binary loss function is defined as

$$l(y, q) = \mathbb{1}\{y = 0\}l_0(q) + \mathbb{1}\{y = 1\}l_1(q) \qquad (1)$$

where $\mathbb{1}\{\cdot\}$ is an indicator function and $l_k(q)$ is a loss function associated with the event $y = k$. Several examples of typical loss functions, such as 0-1 loss, quadratic loss and others are provided in Table I. Let

$$L(p, q) = E_Y l(Y, q) \qquad (2)$$

be the expected loss with respect to Y. Notice that $L(p, q)$ only depends on the Bernoulli parameter p and the estimate q. A *proper loss function* is a loss function for which the minimizer of the expected loss is the true underlying distribution of the random variable we are to estimate, $p = \arg\min_q L(p, q)$. This property is also known as Fisher-consistent or unbiased loss. A *strictly* proper loss function means that $q = p$ is a unique minimizer. In this work we require several regularity conditions for proper loss functions. We say that a proper loss function is *fair* if $l_0(0) = l_1(1) = 0$. This means that there is no loss incurred for perfect prediction. Further, We say that a proper loss function is *regular* if $\lim_{q \searrow 0} q l_1(q) =$

978-1-5386-4780-6/18/$31.00©2018 IEEE

$\lim_{q \nearrow 1}(1-q)l_0(q) = 0$. Intuitively, this condition ensures that making mistakes on events that never happen should not incur a penalty. In this paper we consider loss functions that are *fair* and *regular* unless stated otherwise.

Define the minimum of the expected proper loss as $G(p) \triangleq L(p,p)$. This term is also known as the *generalized entropy function* [1], *Bayes risk* [6] or *Bayesian envelope* [7]. For example, assuming $l(y,q)$ is the log-loss, then the generalized entropy function is Shannon entropy. Additional examples appear in Table I. The *regret* is defined as the difference between the expected loss and its minimum. For proper loss functions we have that $\Delta L(p,q) = L(p,q) - G(p)$. Savage [8] showed that a loss function $l(y,q)$ is proper and regular iff $G(p)$ is concave and for every $p,q \in [0,1]$ we have that

$$L(p,q) = G(q) + (p-q)G'(q).$$

This property allows us to draw an immediate connection between regret and Bregman divergence. Specifically, let $f : \mathbb{S} \to \mathbb{R}$ be a convex function over some convex set $\mathbb{S} \in \mathbb{R}^n$. Then its associated Bregman divergence is defined as

$$D_f(s||s_0) = f(s) - f(s_0) - \langle s - s_0, \nabla f(s_0)\rangle$$

for any $s, s_0 \in \mathbb{S}$, where $\nabla f(s_0)$ is the gradient of f at s_0. By setting $s = [0,1]$ we have that $\nabla f = f'$ and $\Delta L(p,q) = D_{-G}(p,q)$. This means that the regret of a proper loss function is uniquely associated with a Bregman divergence. An important example is the Kullback-Leibler (KL) divergence, $D_{\text{KL}}(p||q)$ associated with the log-loss. Additional examples appear in Table I.

Convex loss functions hold a special role in learning theory and optimization [6], [9]. Let $\underline{X}$ and Y be a set of explanatory variables (features) and an independent variable (target) respectively. Given a set of n i.i.d. samples of $\underline{X}$ and Y, the empirical risk minimization (ERM) criterion seeks to minimize $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i = 0)l_0(q_i) + \mathbb{1}(y_i = 1)l_1(q_i)$, where $q_i \triangleq q_i(\underline{x}_i)$. As the complexity of this problem increases, it is desirable for this minimization problem to be convex in the optimization parameter. Alternatively, assuming that p is known, minimizing the expected loss $L(p,q)$ has many desirable properties, both analytically and computationally, when the problem is convex. It is important to mention that convex proper loss functions correspond to Bregman divergences that are convex in their second parameter. This family of divergences are of a special interest in many applications [10], [11], and have an important role in our results.

III. Main Result

Our main result is as follows,

Theorem 1: Let $l(y,q)$ be a smooth and proper binary loss function with a corresponding generalized entropy function G. Assume that $l(y,q)$ is convex in q. Then for every $p,q \in [0,1]$,

$$D_{\text{KL}}(p||q) \geq \frac{1}{C(G)} D_{-G}(p||q)$$

where $C(G) > -\frac{1}{2}G''(p)|_{p=\frac{1}{2}}$ is a normalization constant (that does not depend on p or q).

A proof of this theorem is provided in Appendix A.

This result established that the KL divergence, associated with the log-loss, bounds from above the divergence of any smooth, proper and convex loss function, up to a multiplicative constant. In other words, by minimizing the log-loss we minimize an upper bound on any choice of such loss functions. The practical implications of this result are quite immediate. Assume that the performance measure according to which a learning algorithm is to be measured with is unknown a-priori to the experiment (for example, the weather forecaster, discussed in Section I). Then, minimizing the log-loss provides an upper bound to any possible choice of measure, associated with an "analytically convenient" loss function. This property makes the log-loss a universal choice for classification problems as it governs a large and significant class of measures.

We notice that the normalization constant $C(G)$ seems to provide a bound that is untight. However, it is unavoidable since proper loss functions are closed under affine transformations. It is important to mention that typical minimal values of $C(G)$ are relatively "small". For example, we have that $C(G) = 1$ for 0-1 loss and $C(G) = 2$ for both the quadratic loss and Boosting loss [9]. The corresponding quadratic bound, for example, is $D_{\text{KL}}(p||q) \geq (p-q)^2$.

In addition to the universality property, Theorem 1 allows us to analyze the local behavior of divergences associated with smooth, proper and convex loss functions, as demonstrated in Corollary 2.

Corollary 2: Let $l(y,q)$ be a smooth and proper binary loss function with a corresponding generalized entropy function G. Assume that $l(y,q)$ is convex in q. Then for every $p, p+dp \in [0,1]$,

$$\frac{1}{C(G)} D_{-G}(p||p+dp) \lesssim \frac{dp^2}{2}\mathcal{I}(p)$$

where $\mathcal{I}(p)$ is the Fisher information of a Bernoulli distributed random variable with a parameter p, and $\lesssim$ refers to inequality up to second order of Taylor expansion terms, $O(dp^2)$.

A proof of this theorem is provided in Appendix B.

Corollary 2 implies that when q is "close enough" to p, the divergence associated with the set of smooth, proper and convex binary loss functions is governed by the Fisher information of a Bernoulli random variable with a parameter p (up to the second order terms of the Taylor expansion). Since $\mathcal{I}(p)$ corresponds to the KL divergence (and henceforth, to log-loss) we conclude that the rate of convergence of any $D_{-G}(p||q)$ in this set is bounded from above by the rate of $D_{\text{KL}}(p||q)$, when $q = p + dp$. This provides an interesting trade-off between the universality of the log-loss, and its slower rate of convergence.

As stated in Section II, the divergence associated with a convex and proper loss function is a Bregman divergence that is convex in its second argument. This allows us to present our results from a divergence analysis perspective. Further, it allows us the extend our study to a greater alphabet size.

Define a *separable* Bregman divergence as

$$D_g(\underline{p}||\underline{q}) = \sum_{i=1}^{m} g(p_i) - g(q_i) - g'(q_i)(p_i - q_i)$$

for any $\underline{p}, \underline{q} \in [0,1]^m$ and a convex function $g : [0,1] \rightarrow [0,\infty]$. Notice that in the general case, the Bregman divergence is not restricted to the unit simplex. Separable Bregman divergences hold a fundamental role in divergence analysis, as shown in [4], [12].

Theorem 3: Let $D_g(\underline{p}||\underline{q})$ be a separable Bregman divergence, that is convex in $\underline{q}$. Then, for every $\underline{p}, \underline{q} \in [0,1]^m$,

$$D_{\text{KL}}(\underline{p}||\underline{q}) \geq \frac{1}{C(g)} D_g(\underline{p}||\underline{q})$$

where $D_{\text{KL}}(\underline{p}||\underline{q}) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} - \sum_{i=1}^{m} p_i + \sum_{i=1}^{m} q_i$ and $C(g) > g''(p)|_{p=1}$ is a normalization constant.
As a corollary, we further show that Theorem 3 holds for the case where $\underline{p}$ and $\underline{q}$ are constrained to the unit simplex. Proofs of Theorem 3 and its corollary are provided in Appendix C.

As in Theorem 1, we notice that the constant $C(g)$ is unavoidable since affine transformations preserve the convexity of $D_g(\underline{p}, \underline{q})$. For example, for the squared error case we have that

$$D_{\text{KL}}(\underline{p}||\underline{q}) \geq \frac{1}{2} \sum_{i=1}^{m} (p_i - q_i)^2. \tag{3}$$

Notice that the multiplicative constant in (3) is different than in the binary case ($C(g) = 1$). This is a result of the fundamental difference between the optimally conditions when minimizing some function $f(p_1, p_2)$ under the constraint that $p_1 + p_2 = 1$, as opposed to minimizing the same function using a single parameter, $f(p_1, 1 - p_1)$.

Further, it is important to mention that (3) resembles the well-known Pinsker inequality [2], that states

$$D_{\text{KL}}(\underline{p}||\underline{q}) \geq \frac{1}{2} \left(\sum_{i=1}^{m} |p_i - q_i| \right)^2. \tag{4}$$

Notice that the right-hand side of (4) is not a Bregman divergence (in fact it is a squared Csiszár divergence [2]) and therefore it is not considered in Theorem 3.

It is easy to verify that the Pinsker inequality is tighter than (3). However, the squared-error bound (3) is just a simple case of our broader result. In this sense, Theorem 3 may be viewed as an extension of Pinsker-like inequalities to the family of separable Bregman divergences that are convex in their second argument.

IV. Illustrations

We demonstrate our results in two illustrative experiments. In the first experiment we focus on ternary alphabet $y \in \{-1, 0, 1\}$ with a corresponding distribution $p(y) = [1/4, 1/2, 1/4]^T$. We are interested in q such that $E_q(Y)$ is a given fixed value. This implies a fixed expectation constraint on q. We examine smooth, proper and convex loss functions through their corresponding Bregman divergences (that are convex in their second moment), as discussed above. Notice that our constraint is linear in the q, so that our optimization problem is convex minimization over a convex set. Notice this problem may be easily generalized to larger alphabets and additional constraints on greater moments of Y. Fig. 1 demonstrates the results we achieve. The blue (upper) curve is $D_{\text{KL}}(\underline{p}||\underline{q})$ and the rest of the curves correspond to different Bregman divergences including quadratic loss and separable Mahalanobis distances [1]. We first notice that $E_p(Y) = 0$. This means that by allowing $E_q(Y) = 0$, we get an unbiased estimate and the divergence is zero. On the other hand, different values of $E_q(Y)$ result in bias. In this case, it is evident that the KL divergence bounds from above any choice of divergence, as expected.

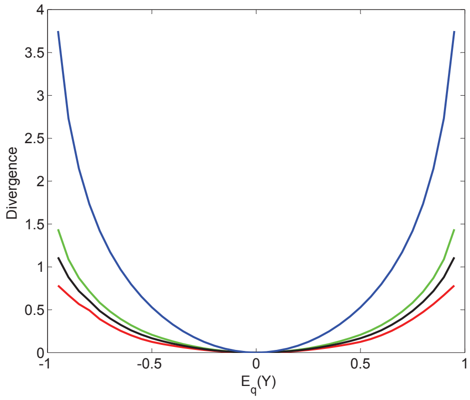

Fig. 1. Fixed expectation experiment: $Y \in \{-1, 0, 1\}$, $p = [1/4, 1/2, 1/4]^T$ and we optimize over q under a fixed expectation constraint, $E_q(Y)$. The blue (upper) curve is $D_{\text{KL}}(\underline{p}||\underline{q})$ while the rest of the curves are Bregman divergences, that are convex in $\underline{q}$.

In the second experiment we show that our bound holds for a broader range of practical problems. Assume that there exists some constraint that prevents $\underline{q}$ from converging to $\underline{p}$. This constraint may be statistical, computational or even algorithmic. We model this problem by stating that $D_\epsilon(\underline{p}||\underline{q}) \geq \epsilon$ for some (unknown) divergence measure D_ϵ and $\epsilon > 0$. In other words, we restrict $\underline{q}$ to be ϵ-far from $\underline{p}$ (in a $D_\epsilon(\underline{p}||\underline{q})$ sense). Fig. 2 demonstrates the results we achieve for $D_\epsilon(\underline{p}||\underline{q}) = \sum_{i=1}^{m} |p_i - q_i|$ (total variation, in the upper charts) and $D_\epsilon(\underline{p}||\underline{q}) = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{q_i}$ (Chi-Square, in the lower charts). The charts on the left show different divergence measures (in the same manner as in the previous experiment) for different ϵ values. The charts on the right demonstrate the KL divergence (blue curve on top), and the quadratic divergence, when we plug the minimizer of the KL divergence. Specifically, $\sum_{i=1}^{m} p_i \log \frac{p_i}{q_i^{\text{KL}}}$ and $\sum_{i=1}^{m} (p_i - q_i^{\text{KL}})^2$ where $\underline{q}^{\text{KL}} = \arg\min_{\underline{q}} D_{\text{KL}}(\underline{p}||\underline{q})$.

We first notice that our bound holds for the two choices of $D_\epsilon(\underline{p}||\underline{q})$, where greater ϵ values result in a greater bias than lower values, as expected. Second, notice that

$$D_{\text{KL}}(\underline{p}||\underline{q}) \geq D_{\text{KL}}(\underline{p}||\underline{q}^{\text{KL}}) \geq \frac{1}{C(g)} D_g(\underline{p}||\underline{q}^{\text{KL}}) \tag{5}$$

TABLE I
EXAMPLES OF BINARY LOSS FUNCTIONS

Loss function	$l(y,q)$	$G(p) = L(p,p)$	$D_{-G}(p\|\|q)$	$w(p)$
0-1 loss	$y\mathbb{1}\{q < \frac{1}{2}\}+$ $(1-y)\mathbb{1}\{q \geq \frac{1}{2}\}$	$p\mathbb{1}\{p < \frac{1}{2}\}+$ $(1-p)\mathbb{1}\{p \geq \frac{1}{2}\}$	$(1-2p)\mathbb{1}\{p < \frac{1}{2}, q \geq \frac{1}{2}\}+$ $(2p-1)\mathbb{1}\{p \geq \frac{1}{2}, q < \frac{1}{2}\}$	$2\delta(\frac{1}{2} - p)$
Quadratic loss	$y(1-q)^2 + (1-y)q^2$	$p(1-p)$	$(p-q)^2$	2
Log loss	$y\log\frac{1}{q}+$ $(1-y)\log\frac{1}{1-q}$	$p\log\frac{1}{p}+$ $(1-p)\log\frac{1}{1-p}$	$p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$	$\frac{1}{p(1-p)}$
Boosting loss	$2y\sqrt{\frac{1-q}{q}}+$ $2(1-y)\sqrt{\frac{q}{1-q}}$	$4\sqrt{p(1-p)}$	$2\left(p\sqrt{\frac{1-q}{q}} + (1-p)\sqrt{\frac{q}{1-q}}\right) -$ $4\sqrt{p(1-p)}$	$\frac{1}{(p(1-p))^{3/2}}$

for any separable Bregman divergence that is convex in $\underline{q}$. This means that we may use the minimizer of the KL divergence as an (untight) approximated "solution" for $D_g(\underline{p}||\underline{q})$. Indeed, the charts on the left demonstrate this inequality for quadratic divergence.

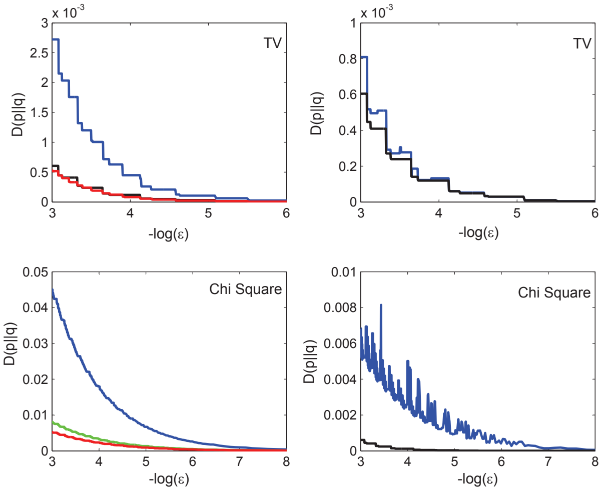

Fig. 2. Divergence constraint experiment: we minimize $D_g(\underline{p}||\underline{q})$ under the constraint that $D_\epsilon(\underline{p}||\underline{q}) \geq \epsilon$ for $D_\epsilon(\underline{p}||\underline{q})$ being total variation (upper charts) and Chi Square (lower charts). $D_g(\underline{p}||\underline{q})$ are the same as in Fig. 1. The charts on the left demonstrate Theorem 3 and the charts on the right demonstrate inequality (5), as described in the main text.

V. Conclusions and Discussion

In this work we introduce a fundamental inequality for divergence measures associated with smooth, proper and convex binary loss functions. We show that the KL divergence, associated with the Bernoulli log-likelihood loss function, bounds from above any divergence associated with this set of losses. This property makes the log-loss a universal choice, in the sense that it controls any "analytically convenient" alternative one may be interested in. The implications of this result span a broad variety of applications. In binary classification trees, the split criterion in each node is typically chosen between the Gini impurity (which corresponds to quadratic loss) and Information-Gain (corresponds to log-loss). The choice of a suitable splitting mechanism holds a long standing discussion with many statistical and computational implications (for example, [13]). In deep neural networks, the objective function is cross-entropy minimization (which again corresponds to log-loss) where several alternatives have (empirically) shown to be less successful over the years. Further, our result may extend the fundamental PAC-Bayes bound [14] to a universal setup which is independent in the choice of the loss.

As demonstrated in Theorem 3, our results may be viewed from a Bregman divergences perspective. Here, the applications of our results are universality guarantees for distributional clustering [15], clustering with Bregman divergences [16] and many others.

Appendix A: Sketch of proof for Theorem 1

A smooth and proper binary loss function satisfies

$$\frac{\partial}{\partial q} L(p,q)|_{q=p} = p l_0'(p) + (1-p) l_1'(p) = 0.$$

This means that

$$\frac{-l_1'(p)}{1-p} = \frac{l_0'(p)}{p} \triangleq w(p)$$

where $w(p)$ is defined as the *weight function*. Shuford et at. [17] showed that the converse is also true: a smooth binary loss function is proper only if the above holds, for $w(p)$ that satisfies $\int_\epsilon^{1-\epsilon} w(c)dc < \infty$, for all $\epsilon > 0$. Typical examples of weight functions for different losses appear in Table I. In addition, it is easy to verify that $\frac{d^2}{dp^2}G(p) = -w(p)$ for all proper binary loss functions. The convexity of the loss (with respect to q) implies that

$$\frac{\partial^2}{\partial q^2} L(p,q) = w(q) + (q-p)w'(q) \geq 0$$

for every fixed $p \in [0,1]$. Plugging $p=0$ and $p=1$ we get

$$-\frac{1}{q} \leq \frac{w'(q)}{w(q)} \leq \frac{1}{1-q}$$

for all $q \in (0,1)$. Integrating both sides achieves

$$\begin{aligned} &\text{For } \ 1 > q \geq \frac{1}{2}: \quad \frac{w\left(\frac{1}{2}\right)}{2q} \leq w(q) \leq \frac{w\left(\frac{1}{2}\right)}{2(1-q)} \\ &\text{For } \ \frac{1}{2} > q > 0: \quad \frac{w\left(\frac{1}{2}\right)}{2q} \geq w(q) \geq \frac{w\left(\frac{1}{2}\right)}{2(1-q)}. \end{aligned} \tag{6}$$

Similar results appear in Theorem 29 of [6]. Let us now look at $R(p,q) = C \cdot D_{\mathrm{KL}}(p||q) - D_{-G}(p||q)$ for a fixed p and find such C for which $R \geq 0$ for all p, q. We have

$$\begin{aligned} &\tfrac{\partial}{\partial q} R(p,q) = (q-p)\left(\tfrac{C}{q(1-q)} - w(q)\right) \\ &\tfrac{\partial^2}{\partial q^2} R(p,q) = C\left(\tfrac{p}{q^2} + \tfrac{1-p}{(1-q)^2}\right) - w(q) - (q-p)w'(q). \end{aligned}$$

We require that $q = p$ is a minimum. Notice that the second derivative condition, together with (6), yield that $C > \frac{1}{2}w\left(\frac{1}{2}\right) = -\frac{1}{2}G''(p)|_{q=p}$. This further implies that $\frac{C}{q(1-q)} - w(q) > 0$ so that $q = p$ is the global minimum (according to the first derivative condition), as desired. □

APPENDIX B: SKETCH OF PROOF FOR COROLLARY 2

Assume that $p, p+dp \in [0,1]$. Then, we may derive the Taylor expansion of $D_{-G}(p||p+dp)$ around p:

$$\begin{aligned} D_{-G}(p||p+dp) =& \frac{dp^2}{2}\frac{d^2}{dp^2}D_{-G}(p||q)|_{q=p} + O(dp^2) \simeq \\ & \frac{dp^2}{2}\frac{d^2}{dp^2}L(p,q)|_{q=p} = \frac{dp^2}{2}w(p) < \\ & \frac{dp^2}{2}\frac{C}{p(1-p)} = \frac{dp^2}{2} \cdot C \cdot \mathcal{I}(p). \end{aligned} \tag{7}$$

□

APPENDIX C: SKETCH OF PROOF FOR THEOREM 3

Let us begin by stating the derivatives of a separable Bregman divergence,

$$\begin{aligned} &\tfrac{\partial}{\partial q_i} D_g(\underline{p}||\underline{q}) = g''(q_i)(p_i - q_i) \\ &\tfrac{\partial^2}{\partial q_i^2} D_g(\underline{p}||\underline{q}) = g'''(q_i)(q_i - p_i) + g''(q_i). \end{aligned}$$

Assuming a fixed $\underline{\text{p}}$, the convexity of $D_g(\underline{p}||\underline{q})$ implies that its second derivative is non-negative for every $p_i \in [0,1]$. Specifically, for $p_i = \{0,1\}$ we get that

$$-\frac{1}{q_i} \leq \frac{g'''(q_i)}{g''(q_i)} \leq \frac{1}{1-q_i}.$$

Integrating the left inequality with respect to q_i attains

$$g''(q_i) \leq \frac{g''(1)}{q_i}. \tag{8}$$

As in Appendix A, we define $R(\underline{p},\underline{q}) = C \cdot D_{\mathrm{KL}}(\underline{p}||\underline{q}) - D_g(\underline{p}||\underline{q})$ and show that it is non-negative. Let us fix $\underline{q}$ and analyze $R(\underline{p},\underline{q})$ with respect to $\underline{p}$. We have that

$$\begin{aligned} &\tfrac{\partial}{\partial p_i} R(\underline{p},\underline{q}) = C\left(\log\tfrac{p_i}{q_i}\right) - g'(p_i) + g'(q_i) \\ &\tfrac{\partial^2}{\partial p_i^2} R(\underline{p},\underline{q}) = \tfrac{C}{p_i} - g''(p_i). \end{aligned}$$

Notice that the second derivative, together with (8) implies that $R(\underline{p},\underline{q})$ is convex (in $\underline{p}$) if $C \geq g''(1)$. This further means that $p_i = q_i$ is a unique minimizer, for this choice of C. Repeating the same derivation for any fixed $\underline{q}$ yields the desired property.

Assume now that $\underline{p}$ is given and it is over the unit simplex, while $\underline{q}$ is constrained to the same domain. Then, we may repeat the derivation above with corresponding Lagrange multipliers and reattain condition (8). Further, we may show that in this case, $R(\underline{p},\underline{q}) \geq 0$ for every $\underline{p},\underline{q} \in [0,1]^m$. This means that our inequality hold for any subset of $[0,1]^m$, such as the unit simplex, for example. □

REFERENCES

[1] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[3] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[4] J. Jiao, T. A. Courtade, A. No, K. Venkat, and T. Weissman, "Information measures: the curious case of the binary alphabet," *IEEE Trans. Inform. Theory*, vol. 60, no. 12, pp. 7616–7626, 2014.

[5] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. Inform. Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[6] M. D. Reid and R. C. Williamson, "Composite binary losses," *J. Machine Learning Res.*, vol. 11, no. Sep, pp. 2387–2422, 2010.

[7] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1280–1292, 1993.

[8] L. J. Savage, "Elicitation of personal probabilities and expectations," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 783–801, 1971.

[9] A. Buja, W. Stuetzle, and Y. Shen, "Loss functions for binary class probability estimation and classification: Structure and applications," *Working draft, November*, 2005.

[10] H. H. Bauschke and J. M. Borwein, "Joint and separate convexity of the bregman distance," *Studies in Computational Mathematics*, vol. 8, pp. 23–36, 2001.

[11] C. L. Byrne, *Iterative Optimization in Inverse Problems*. CRC Press, 2014.

[12] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. Int. Symp. Inform. Theory*. IEEE, 2007, pp. 566–570.

[13] A. Painsky and S. Rosset, "Cross-validated variable selection in tree-based methods improves predictive performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2142–2153, 2017.

[14] J. Langford, "Tutorial on practical prediction theory for classification," *J. Machine Learning Res.*, vol. 6, no. Mar, pp. 273–306, 2005.

[15] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of english words," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1993, pp. 183–190.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *J. Machine Learning Res.*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[17] E. H. Shuford, A. Albert, and H. E. Massengill, "Admissible probability measurement procedures," *Psychometrika*, vol. 31, no. 2, pp. 125–145, 1966.