1 Introduction

The goal of classification problem is to learn a discriminant function for classification of forthcoming data. Given data coming in i.i.d. pairs \((X_i, Y_i)\), for \(1 \leq i \leq n\), from some underlying distributions \(X \in \mathcal{X}\), and \(Y \in \mathcal{Y} = \{1, 2, 3, \ldots, k\}\), we aim to design a classifier, \(\gamma\), with smallest 0-1 loss, given by \(P(\gamma(X) \neq Y)\). However, minimizing 0-1 loss directly is often computationally intractable and practical algorithms are based on convex relaxations of 0-1 loss, say \(\Phi\), which is called the surrogate loss. It is natural to ask whether solving the minimization problem over \(\Phi\) leads to a minimizer of 0-1 loss; this is called consistency. In contrast to the binary classification, in the multi-class classification problem analysis of consistency of \(\Phi\) requires more sophisticated approaches.

The prior Bayes \(\Phi\)-risk is defined as the optimal achievable risk by having access only to probability of labels. Likewise, the posterior Bayes \(\Phi\)-risk is the optimal achievable risk once we further have access to posterior densities, given by \(p_c(x) = p(x \mid Y = c)\), for \(1 \leq c \leq k\). The difference is an important quantity as it indicates the amount of information that \(X\) has. In the binary classification problem, this gap was shown in [LV06] to be equal to an \(f\)-divergence between probability distributions \(P_1\) and \(P_2\), for a convex \(f : \mathbb{R} \to \mathbb{R}\).

In many applications, the covariates used for solving the minimization problem, are either received after passing through a dimension reducing quantizer, or are deliberately transformed using a feature selection stage in order to obtain a more meaningful interpretation. Decentralized detection problem, as described in [Tsi93], is an example of the first situation where we collect information from sensors located in different places and we aim to solve a hypothesis testing problem at a fusion center. Regarding to constraints on sensors power and the communication channels, we need to send quantized version of collected information to the fusion center. In experimental design problems, we often encounter the second situation, since we discard some information that are useless in our analysis.

For modeling purposes, let \(q : \mathcal{X} \to \mathcal{Z} = \{1, 2, \ldots, m\}\) be a quantizer. Our aim is jointly learning a classifier \(\gamma : \mathcal{Z} \to \mathcal{Y}\) and a quantizer \(q\), by minimizing the misclassification rate, given by \(P(\gamma(q(X)) \neq Y)\). Considering any convex surrogate loss \(\Phi\), we can find the optimal risk of quantizer \(q\), denoted as \(R_\Phi(q)\), by empirically minimizing \(\mathbb{E}[\Phi_Y(\gamma(q(X)))]\) over \(\gamma\). One fundamental question is finding all surrogate losses \(\Phi\) that are universally equivalent to the 0-1 loss functions, in the sense that, for all underlying probability distributions on \(\mathcal{X} \times \mathcal{Y}\),
the Φ loss and 0-1 loss induce the same ordering on the optimal risk of quantizers. By doing so, we are able to identify all convex surrogate losses that not only classification problem for them is easy to solve, but they also give us the same optimal classifier γ∗ and quantizer q∗ as the 0-1 loss, regardless of underlying distribution. For binary case, this problem can be written in terms of corresponding f-divergences and is studied in [NWJ09].

The main goal of this report is to define a generalized version of f-divergences, study its properties, and prove a general theorem for characterizing universally equivalent losses in the multi-class setting in terms of generalized f-divergences. We will also show that 0-1 loss is equivalent to the some multi-dimensional hinge loss that is widely-used in practice.

2 Generalized f-divergences

Let P1,...,Pk be probability distributions on a common σ-algebra F over a set X. Let f : Rk−1 → R be a convex function satisfying f(1) = 0. Let µ be any measure such that Pi ≪ µ for all i, and let πi = dPi/dµ denote the density of Pi with respect to µ. Then we define the generalized f-divergence between P1,...,Pk as

\[ D_f (P_1, \ldots, P_{k-1} \parallel P_k) := \int f \left( \frac{p_1(x)}{p_k(x)}, \ldots, \frac{p_{k-1}(x)}{p_k(x)} \right) p_k(x) d\mu(x). \]  

(1)

Lemma 2.1. The value of generalized f-divergence, defined in (1), does not depend on the base measure µ.

We now illustrate a few properties of these multi-way f-divergences, showing how they extend classical divergences. Given a measurable mapping q : X → N with a finite range, we define

\[ D_f (P_1, \ldots, P_{k-1} \parallel P_k \mid q) = \sum_{A \in q^{-1}(N)} f \left( \frac{P_1(A)}{P_k(A)}, \ldots, \frac{P_{k-1}(A)}{P_k(A)} \right) P_k(A). \]

This is the quantized version of the generalized f-divergences. When the function f is well-behaved, by choosing the quantizer wisely, we will not lose any information.

Lemma 2.2. Let f be a closed convex function. Then

\[ D_f (P_1, \ldots, P_{k-1} \parallel P_k) = \sup_q D_f (P_1, \ldots, P_{k-1} \parallel P_k \mid q), \]  

(2)

where the supremum is over quantizers of X.

For proving one of the most important properties of f-divergences, i.e., the data processing inequality, we first need the following lemma:

Lemma 2.3. Let f : Rk+ → R be convex. Let a : X → R+ and bi : X → R+, for 1 ≤ i ≤ k, be nonnegative measurable functions. Then for any finite measure µ defined on X, we have

\[ \left( \int a d\mu \right) f \left( \int b_1 d\mu \cdot \ldots \cdot \int b_k d\mu \right) \leq \int a(x) f \left( \frac{b_1(x)}{a(x)} \cdot \ldots \cdot \frac{b_k(x)}{a(x)} \right) d\mu(x). \]
Proposition 2.4. Let $f : \mathbb{R}_+^k \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Let $Q$ be a Markov kernel from $X$ to $Z$, that is, $Q(\cdot \mid x)$ is a probability distribution on $Z$ for each $x \in X$. Define the marginals $Q_{P_i}(A) := \int_X Q(A \mid x)dP_i(x)$ for each $i = 1, \ldots, k$. Then

$$D_f(Q_{P_1}, \ldots, Q_{P_k-1}\|Q_{P_k}) \leq D_f(P_1, \ldots, P_{k-1}\|P_k).$$

Proof. By Lemma 2.2, we have that

$$D_f(Q_{P_1}, \ldots, Q_{P_k-1}\|Q_{P_k}) = \sup \left\{ D_f(Q_{P_1}, \ldots, Q_{P_k-1}\|Q_{P_k} \mid q) : q \text{ is a quantizer of } Z \right\}.$$

It is consequently no loss of generality to assume that $Z$ is finite and $Z = \{1, 2, \ldots, m\}$. Let $\mu = \sum_{c=1}^k P_c$ be a dominating measure and let $p_c = dP_c/d\mu$. Then letting $q_{P_i}(i) = Q_{P_i}(\{i\})$, we obtain

$$D_f(Q_{P_1}, \ldots, Q_{P_k-1}\|Q_{P_k}) = \sum_{i=1}^m q_{P_k}(i)f\left(\frac{q_{P_1}(i)}{q_{P_1}(i)} \cdots \frac{q_{P_k-1}(i)}{q_{P_k}(i)}\right)$$

$$= \sum_{i=1}^m \left(\int_X q(i \mid x)p_k(x)d\mu(x)\right)f\left(\frac{\int_X q(i \mid x)p_1(x)d\mu(x)}{\int_X q(i \mid x)p_k(x)d\mu(x)}, \ldots, \frac{\int_X q(i \mid x)p_{k-1}(x)d\mu(x)}{\int_X q(i \mid x)p_k(x)d\mu(x)}\right)$$

$$\leq \sum_{i=1}^m \int_X q(i \mid x)f\left(\frac{p_1(x)}{p_k(x)}, \ldots, \frac{p_{k-1}(x)}{p_k(x)}\right)p_k(x)d\mu(x).$$

by Lemma 2.3. Noting that $\sum_{i=1}^m q(i \mid x) = 1$, we thus obtain our desired result. \qed

3 Correspondence between risks and $f$-divergences

In order to state the correspondence between risks and $f$-divergences, let us introduce some notations. Denote $\pi_c = P(Y = c)$ as the prior distribution on $Y$ and $p_c(x) = p(x \mid Y = c)$ as the posterior density, for $1 \leq c \leq k$. Each quantization method $q$, induces a conditional probability distribution $Q_c(z) = P(Y = c \mid q(x) = z)$ on $Z$, for almost all $y$. Let $\Psi_c(\alpha)$ be the loss associated with $Y = c$, where $\alpha \in \mathbb{R}_+^k$, with the condition $1^T\alpha = 0$, is the decision vector. For each discriminant function $\gamma : Z \to \mathbb{R}_+^k$, the risk, defined as the expectation of loss, is equal to

$$R_\Phi(\gamma \mid q) = E[\Phi_Y(\gamma(q(X)))] = \sum_{i=1}^m \Phi_Y(\gamma(i) \mid q(X) = i) P(q(X) = i)$$

$$= \sum_{i=1}^m \left[ \sum_{c=1}^k \left( \Phi_c(\gamma(i)) P(Y = c \mid q(X) = i) \right) \right] P(q(X) = i)$$

$$= \sum_{i=1}^m \left[ \sum_{c=1}^k \left( \pi_c \Phi_c(\gamma(i)) P_c(q(X) = i) \right) \right].$$

3
By taking the infimum over $\gamma$, we get the quantized posterior Bayes $\Phi$-risk:

$$R_\Phi(q) = \inf_{\gamma} R_\Phi(\gamma \mid q) = \sum_{i=1}^{m} \inf_{\alpha^T_1 = 0} \left( \sum_{c=1}^{k} \pi_c \Phi_c(\alpha) P_c(A_i) \right),$$

where $A_i = \{ x \in \mathcal{X} \mid q(x) = i \} = q^{-1}(i)$. Now we have the following proposition

**Proposition 3.1.** For any loss function $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_k)$ and any quantization rule $q : \mathcal{X} \to \{1, 2, \ldots, m\}$, define quantized $\Phi$-statistical information as the gap between the prior Bayes $\Phi$-risk and the posterior Bayes $\Phi$-risk. Then, $\Phi$-statistical information can be written as a generalized $f$-divergence. That is

$$I_\Phi(q) = R_\Phi - R_\Phi(q) = D_{f_\Phi}(P_1, P_2, \ldots, P_{k-1} \| P_k \mid q),$$

for some closed convex function $f_\Phi$. Further, $f_\Phi$ is given by

$$f_\Phi(t_1, t_2, \ldots, t_{k-1}) = \sup_{\alpha^T_1 = 0} \left\{ R_\Phi - \pi_k \Phi_k(\alpha) - \sum_{c=1}^{k-1} \pi_c \Phi_c(\alpha) t_c \right\}.$$

(3)

**Proof.** The proof is easy. By expanding the $\Phi$-statistical information

$$I_\Phi(q) = R_\Phi - R_\Phi(q) = \sum_{i=1}^{m} \sup_{\alpha^T_1 = 0} \left( R_\Phi - \sum_{c=1}^{k} \pi_c \Phi_c(\alpha) P_c(A_i) \right)$$

$$= \sum_{i=1}^{m} \sup_{\alpha^T_1 = 0} \left( R_\Phi - \sum_{c=1}^{k} \pi_c \Phi_c(\alpha) \frac{P_c(A_i)}{P_k(A_i)} \right) P_k(A_i) = D_{f_\Phi}(P_1, P_2, \ldots, P_{k-1} \| P_k \mid q).$$

It is easy to check that (3) satisfies above equality and hence we are done. \qed

## 4 Universal equivalent class characterization

As discussed in section §1, it is interesting to study whether there exist two surrogate loss functions that induce the same ordering on the optimal risk of quantizers. More precisely:

**Definition** Two surrogate loss functions $\Phi_1$ and $\Phi_2$ are *universally equivalent* if for any probability distribution $\mathbb{P}(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$, and any two quantization rules $q_1$ and $q_2$ on $\mathcal{X}$, there holds

$$R_{\Phi_1}(q_1) \leq R_{\Phi_1}(q_2) \Leftrightarrow R_{\Phi_2}(q_1) \leq R_{\Phi_2}(q_2).$$

It is easy to translate this definition in terms of $f_1$ and $f_2$ which are $f$-divergences associated to $\Phi_1$ and $\Phi_2$, respectively. This is true according to the Proposition 3.1.
Remark Let \( f_1 \) and \( f_2 \) be \( f \)-divergences associated with \( \Phi_1 \) and \( \Phi_2 \) respectively. Then \( \Phi_1 \) and \( \Phi_2 \) are universally equivalent if for any probability distribution \( P(X,Y) \) on \( \mathcal{X} \times \mathcal{Y} \), and any two quantization rules \( q_1 \) and \( q_2 \) on \( \mathcal{X} \), we have
\[
D_{f_1}(P_1, P_2, \ldots, P_{k-1} || P_k | q_1) \leq D_{f_1}(P_1, P_2, \ldots, P_{k-1} || P_k | q_2)
\]
\[
\Leftrightarrow D_{f_2}(P_1, P_2, \ldots, P_{k-1} || P_k | q_1) \leq D_{f_2}(P_1, P_2, \ldots, P_{k-1} || P_k | q_2),
\]
where for \( 1 \leq c \leq k \), \( P_c \) is the conditional distribution of \( X \) given \( Y = c \).

Observe that this definition is very stringent, in that it requires that the ordering between optimal \( \Phi_1 \) and \( \Phi_2 \) risks holds for all probability distributions \( P \) on \( \mathcal{X} \times \mathcal{Y} \). The following theorem provides necessary and sufficient conditions under which universal equivalency holds.

Theorem 4.1. Let \( f_1, f_2 : \mathbb{R}^K \to \mathbb{R} \) be closed convex \( f \)-divergences associated with \( \Phi_1 \) and \( \Phi_2 \), respectively. Then \( \Phi_1 \) and \( \Phi_2 \) are universally equivalent if and only if \( f_1(t) = af_2(t) + b^T t + c \) for some constants \( a > 0, b \in \mathbb{R}^K, c \in \mathbb{R} \).

Due to space constraints, the proof of the theorem is left to the appendix.

This theorem is the main result of the paper and we now discuss how using this theorem can help us to solve minimization of 0-1 loss problem. In the next section we compute the \( f \)-divergences associated with 0-1 and hinge loss to confirm their equivalence.

5 Universal equivalence of 0-1 and hinge loss

Example First consider the 0-1 loss, which is defined as the probability of misclassification. That is, the loss is zero if we choose the correct label and is 1 otherwise. This loss can be expressed in many ways as \( \Phi^{0-1}_c(\alpha) = (\Phi^{0-1}_1(\alpha), \ldots, \Phi^{0-1}_k(\alpha)) \), one such is
\[
\Phi^{0-1}_c(\alpha) = 1 - I(\alpha_c > \alpha_j \text{ for } 1 \leq j \leq c - 1 \text{ and } \alpha_c \geq \alpha_j \text{ for } c + 1 \leq j \leq k).
\]
This definition ensures that exactly one of \( \Phi^{0-1}_c \), for \( 1 \leq c \leq k \) is zero and the others are one. The prior Bayes risk is
\[
R_{\Phi^{0-1}} = \inf_{\alpha \in \mathcal{A}} \left( \sum_{c=1}^{k} \pi_c \Phi^{0-1}_c(\alpha) \right) = 1 - \max_{1 \leq c \leq k} \pi_c,
\]
While the posterior Bayes risk in the presence of the quantizer \( q \) is
\[
R_{\Phi^{0-1}}(q) = \sum_{i=1}^{m} \left( \sum_{c=1}^{k} \pi_c P_c(A_i) - \max_c \pi_c P_c(A_i) \right) = 1 - \sum_{i=1}^{m} \max_c \pi_c P_c(A_i).
\]
Thus, the \( f \)-divergence associated to 0-1 loss is equal to
\[
f_{\Phi^{0-1}}(t) = \max \left\{ \max_{1 \leq i \leq k-1} t_i \pi_i, \pi_k \right\} - \max_{1 \leq i \leq k} \pi_i.
\]
Example} Denote \( \phi^{\text{hinge}} \) as the hinge loss, defined by \( \phi^{\text{hinge}}(\alpha) = (1 - \alpha)_+ = \max\{1 - \alpha, 0\} \). There are many ways for expanding this loss function to the multi-dimensional setting. One general model for this purpose, introduced in [LLW04], is

\[
\Phi_c(\alpha) = \sum_{j \neq c} \phi^{\text{hinge}}(-\alpha_j),
\]

for \( 1 \leq c \leq k \). We also impose the constraint \( \alpha^T 1 = 0 \) in (5). This definition was shown in [Zha04, TB07], to be consistent whenever \( \phi \) is convex, decreasing, and \( \phi'(0) < 0 \). We here also adopt this model and for \( 1 \leq c \leq k \), define \( \Phi^{\text{hinge}}_c(\alpha) = \sum_{j \neq c} \phi^{\text{hinge}}(-\alpha_j) \). By computing prior Bayes \( \Phi^{\text{hinge}} \)-risk we get

\[
R_{\Phi^{\text{hinge}}} = \inf_{\alpha^T 1 = 0} \left( \sum_{c=1}^{k} \pi_c \sum_{j \neq c} (1 + \alpha_j)_+ \right) = \inf_{\alpha^T 1 = 0} \left( \sum_{c=1}^{k} (1 - \pi_c)(1 + \alpha_c)_+ \right).
\]

This convex optimization problem can be solved using KKT conditions. Define \( \alpha^* \) as

\[
\alpha_i^* = \begin{cases} k - 1 & \text{if } i = M, \\ -1 & \text{if } i \neq M. \end{cases}
\]

It is easy to check \( \alpha^* \) is a minimizer of infimum in (6). Therefore,

\[
R_{\Phi^{\text{hinge}}} = k(1 - \max_{1 \leq c \leq k} \pi_c).
\]

The posterior risk can also be calculated as

\[
R_{\Phi^{\text{hinge}}}(q) = k \left( 1 - \max_{1 \leq c \leq k} \pi_c P_c(A_i) \right).
\]

Hence, the \( f \)-divergence corresponding to the hinge loss is

\[
f_{\Phi^{\text{hinge}}}(t) = k \left( \max \left\{ \max_{1 \leq i \leq k - 1} t_i \pi_i, \pi_k \right\} - \max_{1 \leq i \leq k} \pi_i \right) = kf_{\pi, \psi_0 - 1}(t),
\]

where the last equality comes from (4). According to Theorem 4.1 we observe that 0-1 and the hinge loss are universally equivalent. More generally, using (4) we can state the following corollary of Theorem 4.1:

**Corollary 5.1.** All loss functions \( \Phi \) that are universally equivalent to the 0 – 1 loss, induce \( f \)-divergences of the form

\[
f_{\Phi}(t) = a \left( \max \left\{ \max_{1 \leq i \leq k - 1} t_i \pi_i, \pi_k \right\} - \max_{1 \leq i \leq k} \pi_i \right) + b^T t + c
\]

for some constants \( a > 0, b \in \mathbb{R}^{k-1}, c \in \mathbb{R} \).
Appendix A

Proof of Lemma 2.1 Let $\mu_1$ and $\mu_2$ be dominating measures; then $\mu = \mu_1 + \mu_2$ also dominates $P_1, \ldots, P_k$ as well as $\mu_1$ and $\mu_2$. We have for $\nu = \mu_1$ or $\nu = \mu_2$ that

$$
\frac{dP_i}{d\nu} = \frac{dP_i}{d\nu} \frac{d\nu}{d\mu} = \frac{dP_i}{d\mu} \quad \text{and} \quad \frac{dP_i}{d\nu} = \frac{dP_i}{d\mu},
$$

the latter two equalities holding $\mu$-almost surely by definition of the Radon-Nikodym derivative. Thus we obtain for $\nu = \mu_1$ or $\nu = \mu_2$ that

$$
\int f \left( \frac{dP_i}{d\nu}, \ldots, \frac{dP_{k-1}}{d\nu} \right) dP_k = \int f \left( \frac{dP_i}{d\nu}, \ldots, \frac{dP_{k-1}}{d\nu} \right) dP_k \frac{d\nu}{d\mu} = \int f \left( \frac{dP_i}{d\nu}, \ldots, \frac{dP_{k-1}}{d\nu} \right) dP_k \frac{d\nu}{d\mu}
$$

by definition of the Radon-Nikodym derivative. Moreover, we see that $\frac{d\nu}{d\mu} = \frac{dP_i}{dP_k/d\mu}$ a.s.-$\mu$, which shows that the base measure $\mu$ does not affect the integral.

For proving the lemma 2.2, we need some preliminaries definitions and lemmas. Given a measurable partition $\mathcal{P}$ of $\mathcal{X}$, define

$$
D_f(P_1, \ldots, P_{k-1} \| P_k \mid \mathcal{P}) = \sum_{A \in \mathcal{P}} f \left( \frac{P_i(A)}{P_k(A)} \right),
$$

In addition, given a sub-$\sigma$-algebra $\mathcal{G} \subset \mathcal{F}$, we let $P^\mathcal{G}$ denote the restriction of the measure $P$, defined on $\mathcal{F}$, to $\mathcal{G}$. Now we have the following preposition

**Proposition A.1.** Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ be a sequence of sub-$\sigma$-algebras of $\mathcal{F}$ and let $\mathcal{F}_\infty = \sigma(\cup_{n \geq 1} \mathcal{F}_n)$. Then

$$
D_f(P_1^{\mathcal{F}_1}, \ldots, P_{k-1}^{\mathcal{F}_1} \| P_k^{\mathcal{F}_1}) \leq D_f(P_1^{\mathcal{F}_2}, \ldots, P_{k-1}^{\mathcal{F}_2} \| P_k^{\mathcal{F}_2}) \leq \ldots \leq D_f(P_1^{\mathcal{F}_\infty}, \ldots, P_{k-1}^{\mathcal{F}_\infty} \| P_k^{\mathcal{F}_\infty})
$$

and moreover,

$$
\lim_{n \to \infty} D_f(P_1^{\mathcal{F}_n}, \ldots, P_{k-1}^{\mathcal{F}_n} \| P_k^{\mathcal{F}_n}) = D_f(P_1^{\mathcal{F}_\infty}, \ldots, P_{k-1}^{\mathcal{F}_\infty} \| P_k^{\mathcal{F}_\infty}).
$$

**Proof.** Define the measure $\nu = \frac{1}{k} \sum_{i=1}^k P_i$ and vectors $V^n$ via the Radon-Nikodym derivatives

$$
V^n = \frac{1}{k} \left( \frac{dP_1^{\mathcal{F}_n}}{d\nu^{\mathcal{F}_n}}, \ldots, \frac{dP_{k-1}^{\mathcal{F}_n}}{d\nu^{\mathcal{F}_n}} \right).
$$

Then $(1 - 1^T V^n) d\nu^{\mathcal{F}_n} = \frac{1}{k} dP_k^{\mathcal{F}_n}$, and $V^n$ is a martingale sequence adapted to the filtration $\mathcal{F}_n$ by standard properties of conditional expectation. Letting the set $C_k = \{ v \in \mathbb{R}_+^{k-1} \mid 1^T v \leq 1 \}$, we define $g : C_k \to \mathbb{R}$ by

$$
g(v) = f \left( \frac{v}{1 - 1^T v} \right) (1 - 1^T v).$$
Then $g$ is convex (it is a perspective function), and we have

$$
\mathbb{E}_\nu[g(V^n)] = \int f \left( \frac{dP_f^{F_n}}{dP_k^{F_n}}, \ldots, \frac{dP_{k-1}^{F_n}}{dP_k^{F_n}} \right) \left( 1 - \frac{1}{k} \sum_{i=1}^{k-1} \frac{dP_i^{F_n}}{dP_k^{F_n}} \right) d\nu = \frac{1}{k} D_f(P_{1}^{F_n} \ldots, P_{k-1}^{F_n} | P_k^{F_n}).
$$

Because $V^n \in C_k$ for all $n$, we see that $g(V^n)$ is a submartingale. This gives the first result of the proposition, that is, that the sequence $D_f(P_{1}^{F_n} \ldots, P_{k-1}^{F_n} | P_k^{F_n})$ is non-decreasing in $n$. Now, assume that the limit in the second statement is finite, as otherwise the result is trivial. Then using that $f(1) = 0$, we have by convexity that for any $v \in C_k$,

$$
g(v) = f \left( \frac{v}{1 - 1^T v} \right) (1 - 1^T v) + f(1) 1^T v \geq f(v + 1(1^T v)) \geq \inf_{v \in C_k} f(v + 1(1^T v)) > -\infty,
$$

the final inequality a consequence of the fact that $f$ is closed and hence attains its infimum. In particular, the sequence $g(V^n) - \inf_{v \in C_k} g(v)$ is a non-negative submartingale, and thus

$$
\sup_n \mathbb{E}_\nu \left[ |g(V^n) - \inf_{v \in C_k} g(v)| \right] = \lim_n \mathbb{E}_\nu \left[ |g(V^n) - \inf_{v \in C_k} g(v)| \right] < \infty.
$$

Coupled with this integrability guarantee, Doob’s second martingale convergence theorem thus yields the existence of a random vector $V^\infty \in \mathcal{F}_\infty$ such that

$$
0 \leq \lim_n \mathbb{E}_\nu \left[ |g(V^n) - \inf_{v \in C_k} g(v)| \right] = \mathbb{E}_\nu[|g(V^\infty)|] - \inf_{v \in C_k} g(v) < \infty.
$$

Because $\inf_{v \in C_k} g(v) > -\infty$, we have $\mathbb{E}_\nu[|g(V^\infty)|] < \infty$, giving the result of the proposition. \hfill \square

**Proof of Lemma 2.2**

Let the base measure $\mu = \frac{1}{k} \sum_{i=1}^{k} P_i$ and let $p_i = \frac{dp_i}{d\mu}$ be the associated densities of the $P_i$. Define the increasing sequence of partitions $\mathcal{P}^n$ of $\mathcal{X}$ by sets $A_{\alpha n}$ for vectors $\alpha = (\alpha_1, \ldots, \alpha_{k-1})$ and $B$ with

$$
A_{\alpha n} = \left\{ x \in \mathcal{X} \mid \frac{\alpha_j - 1}{2^n} \leq \frac{p_j(x)}{p_k(x)} < \frac{\alpha_j}{2^n} \text{ for } j = 1, \ldots, k-1 \right\}
$$

where we let each $\alpha_j$ range over $\{-n2^n, -n2^n+1, \ldots, n2^n\}$, and define and $B = (\cup_{\alpha} A_{\alpha n})^c = \mathcal{X} \setminus \cup_{\alpha} A_{\alpha n}$. Then we have

$$
\left( \frac{p_1(x)}{p_k(x)}, \ldots, \frac{p_{k-1}(x)}{p_k(x)} \right) = \lim_{n \to \infty} \sum_{\alpha \in \{-n2^n, \ldots, n2^n\}^{k-1}} \frac{\alpha}{2^n} \chi_{A_{\alpha n}}(x) + (n, \ldots, n) \chi_B(x),
$$

each term of which on the right-hand-side is $\mathcal{F}_n$-measurable, where $\mathcal{F}_n$ denotes the sub-$\sigma$-field generated by the partition $\mathcal{P}^n$. Defining $\mathcal{F}_\infty = \sigma(\cup_{n \geq 1} \mathcal{F}_n)$, we have

$$
D_f(P_{1}^{\mathcal{P}^n} \ldots, P_{k-1}^{\mathcal{P}^n} | P_k) \to D_f(P_{1}^{\mathcal{P}_\infty} \ldots, P_{k-1}^{\mathcal{P}_\infty} | P_k),
$$
where the limiting operation follows by Proposition A.1 and the final equality because of the containment \((\frac{p_1}{p_k}, \ldots, \frac{p_{k-1}}{p_k}) \in \mathcal{F}_\infty\).

\[\square\]

**Proof of Lemma 2.3** Recall that the perspective function \(g(y, t)\) defined by \(g(y, t) = tf(y/t)\) for \(t > 0\) is jointly convex in \(y\) and \(t\). The measure \(\nu = \mu/\mu(X)\) defines a probability measure, so that for \(X \sim \nu\), Jensen’s inequality implies

\[
\left(\int \nu f_1 \right) \left(\int \frac{b_1}{\nu} \right) \cdots \left(\int \frac{b_k}{\nu} \right) \leq E_{\nu} \left[ \frac{a(X)}{\nu} f_1 \left( \frac{b_1}{a(X)}, \ldots, \frac{b_k}{a(X)} \right) \right] = \frac{1}{\mu(X)} \int a(x) f_1 \left( \frac{b_1(x)}{a(x)}, \ldots, \frac{b_k(x)}{a(x)} \right) d\mu(x).
\]

Noting that \(\int b_i d\nu / \int \nu = \int b_i d\mu / \int d\mu\) gives the result. \(\square\)

For proving the main theorem we need the following definition

**Definition** Let \(f_1 : \mathbb{R}^K_+ \to \mathbb{R}\) and \(f_2 : \mathbb{R}^K_+ \to \mathbb{R}\) be continuous convex functions. Let \(m \in \mathbb{N}\) be arbitrary and the nonnegative sequences \(a_{ij} \geq 0, b_{ij} \geq 0, i = 1, \ldots, K\) and \(j = 1, \ldots, m\), satisfy \(\sum_{j=1}^m a_{ij} = \sum_{j=1}^m b_{ij}\) for each \(i = 1, \ldots, K\). Then we say that \(f_1\) and \(f_2\) are order-equivalent if for all \(m \in \mathbb{N}\) and all such sequences \(a_{ij}\) and \(b_{ij}\)

\[
\sum_{j=1}^m f_1(a_{1j}, a_{2j}, \ldots, a_{Kj}) \leq \sum_{j=1}^m f_1(b_{1j}, b_{2j}, \ldots, b_{Kj})
\]

if and only if

\[
\sum_{j=1}^m f_2(a_{1j}, a_{2j}, \ldots, a_{Kj}) \leq \sum_{j=1}^m f_2(b_{1j}, b_{2j}, \ldots, b_{Kj}).
\]

**Lemma A.2.** If for some nonnegative sequence \(a_k, b_k, 1 \leq k \leq M\) such that \(\sum_{k=1}^M a_k = \sum_{k=1}^M b_k\), then there exists nonnegative numbers \(z_{ij}\) for \(1 \leq i, j \leq M\) such that \(\sum_{i=1}^M z_{ij} = a_j\) \(\sum_{j=1}^M z_{ij} = b_j\) for \(1 \leq j \leq M\).

**Proof of Lemma** The proof is based on induction. When \(M = 1\), this is simple.

Now suppose the statement is true for all \(M - 1\), we are going to argue that the statement also holds for \(M\). Without loss of generality, we may assume that \(a_M = \min\{a_j \mid 1 \leq j \leq M\}\) and \(b_M = \min\{b_j \mid 1 \leq j \leq M\}\) (Otherwise, just reorder both sequence.). Now, W.L.O.G, let’s assume \(a_M \leq b_M\). (The case for \(b_M \leq a_M\) is similar.). Then we pick \(z_{M,M} = a_M\) and \(z_{i,M} = 0\) for all \(1 \leq i \leq M - 1\). Now assume that \(a_j^* = \max\{a_j \mid 1 \leq j \leq M - 1\}\), then we have \(a_j^* \geq b_M\) since \(a_j^* \geq \frac{1}{M} \sum_{k=1}^M a_k = \frac{1}{M} \sum_{k=1}^M b_k \geq b_M\). Then, let \(z_{j^*,M} = a_j^* - b_M\) and \(z_{i,M} = 0\) for all \(1 \leq i \leq K - 1\) and \(i \neq j^*\).

Now, if we denote \(a_j = a_j - z_{j,M}, b_j = b_j - z_{M,j}\) for \(1 \leq j \leq M - 1\), then we have
$\sum_{j=1}^{M-1} \tilde{a}_j = \sum_{j=1}^{M-1} \tilde{b}_j$. By induction hypotheses, there exists nonnegative numbers $z_{i,j}$ for $1 \leq i, j \leq M - 1$ such that $\sum_{i=1}^{M-1} z_{i,j} = a_j \sum_{j=1}^{M-1} z_{i,j} = b_j$ for $1 \leq j \leq M - 1$. It is easy to check now that the $z_{i,j}$ we constructed satisfy the requirement $\sum_{i=1}^{M} z_{i,j} = a_j$, $\sum_{j=1}^{M} z_{i,j} = b_j$ for $1 \leq j \leq M$.

\[\square\]

**Lemma A.3.** If two loss functions $\Phi_1$ and $\Phi_2$ are universally equivalent, then their corresponding $f$-divergences, denoted as $f_1$ and $f_2$, are order equivalent.

**Proof of Lemma** Given the nonnegative sequences $a_{ij} \geq 0$, $b_{ij} \geq 0$, $i = 1, \ldots, K - 1$ and $j = 1, \ldots, m$ such that $\sum_{j=1}^{m} a_{ij} = \sum_{j=1}^{m} b_{ij}$ for each $i = 1, \ldots, K - 1$, we may find some positive integer $M \geq m$ such that $M > \sum_{j=1}^{m} a_{ij}$ for all $i = 1, \ldots, K - 1$. Denote $\hat{a}_{i,j+1} = M - \sum_{j=1}^{m} a_{ij}$ and $\hat{b}_{i,j+1} = M - \sum_{j=1}^{m} b_{ij}$, for $i = 1, \ldots, K - 1$, and $\hat{a}_{i,l} = \hat{b}_{i,l} = 0$ for all $j + 1 < l \leq M$. Hence, we have $\hat{a}_{i,l} \geq 0$, $\hat{b}_{i,l} \geq 0$ for all $1 \leq i \leq K - 1, 1 \leq l \leq M$ and $\sum_{i=1}^{M} \hat{a}_{i,l} = \sum_{j=1}^{M} \hat{b}_{i,l} = M$.

Now, take $X$ as $\{(i,j) \mid 1 \leq i \leq M, 1 \leq j \leq M\}$. Let $Z = \{1, 2, \ldots, M\}$. Denote $q_1$ as the quantizer that maps $(i,j) \in X$ to $i \in Z$ and $q_2$ as the quantizer that maps $(i,j) \in X$ to $j \in Z$. Now, lemma A.2 shows that there exists some $z_{ij}^k$ for $1 \leq i \leq M$, $1 \leq j \leq M$, $1 \leq k \leq K$ such that $\sum_{i=1}^{M} z_{ij}^k = a_{kj}/M$ and $\sum_{j=1}^{M} z_{ij}^k = b_{kj}/M$. Denote $P_i$ as the probability measure that $P_i\{(i,j)\} = z_{ij}^k$ for $1 \leq i \leq M$, $1 \leq j \leq M$, $1 \leq k \leq K$.

Now, under this quantizer design, we have by order equivalence of $\Phi_1$ and $\Phi_2$,

$$\frac{1}{M} \sum_{j=1}^{M} f_1(a_{1j}, \ldots, a_{(K-1)j}) \leq \frac{1}{M} \sum_{j=1}^{M} f_1(b_{1j}, \ldots, b_{(K-1)j})$$

if and only if

$$\frac{1}{M} \sum_{j=1}^{M} f_2(a_{1j}, \ldots, a_{(K-1)j}) \leq \frac{1}{M} \sum_{j=1}^{M} f_2(b_{1j}, \ldots, b_{(K-1)j})$$

where $f_1$ and $f_2$ are the associated $f$-divergence. Note that, since $a_{ij} = b_{ij}$ for all $1 \leq i \leq K - 1$ and $j \geq m$. Therefore, we have

$$\sum_{j=1}^{m} f_1(a_{1j}, \ldots, a_{(K-1)j}) \leq \sum_{j=1}^{m} f_1(b_{1j}, \ldots, b_{(K-1)j})$$

if and only if

$$\sum_{j=1}^{m} f_2(a_{1j}, \ldots, a_{(K-1)j}) \leq \sum_{j=1}^{m} f_2(b_{1j}, \ldots, b_{(K-1)j})$$

Since this is for all nonnegative sequence of $a_{ij} \geq 0$, $b_{ij} \geq 0$, $i = 1, \ldots, K - 1$ and $j = 1, \ldots, m$ such that $\sum_{j=1}^{m} a_{ij} = \sum_{j=1}^{m} b_{ij}$ for each $i = 1, \ldots, K - 1$, $f_1$ and $f_2$ are thus order equivalent.

\[\square\]

We give one technical lemma that will be useful throughout our derivation.
Lemma A.4 (Karamata’s majorization inequality). Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. If $x_1 \geq x_2 \geq \cdots \geq x_n$, $y_1 \geq y_2 \geq \cdots \geq y_n$, and

$$x_1 + \cdots + x_k \geq y_1 + \cdots + y_k \text{ for } k \in \{1, \ldots, n\},$$

then

$$f(x_1) + \cdots + f(x_n) \geq f(y_1) + \cdots + f(y_n).$$

Lemma A.5. Let $f_1 : \mathbb{R}_+ \to \mathbb{R}$ and $f_2 : \mathbb{R}_+ \to \mathbb{R}$ be order equivalent. If further $f_1(i) = f_2(i)$ for $i = 0, 1, 2$, then we have $f_1(x) = f_2(x)$ for all $x \in [0, 2]$.

Proof. Let

$$S = \{x \geq 0 \mid f_1(x) = f_2(x)\}$$

denote the set of equal points of $f_1$ and $f_2$. The lemma will follow once we show the following three claims:

(a) First, we show that if $0 \in S, t \in S, 2t \in S$ for some $t > 0$, then $t/2 \in S$.

(b) If $0 \in S, t \in S, 2t \in S$, then $3t/2 \in S$.

(c) Based on these two observations, an inductive argument guarantees that $[0, 2] \subseteq S$.

We begin with claim (a). For $q \in [0, 1]$ and $i \in \{1, 2\}$, define the functions

$$\phi_i(q) = q f_i(2t) + 2 f_i \left( \frac{t}{2} \right) - (2q + 1) f_i(t) - (1 - q) f_i(0).$$

We claim that

$$\text{sign}(\phi_1(q)) = \text{sign}(\phi_2(q)) \text{ for all } q \in [0, 1]. \quad (7)$$

Indeed, let $q = \frac{r}{s}$ for some $r, s \in \mathbb{N}$ with $s \geq r$. Then

$$s \cdot \phi_i(q) = rf_i(2t) + 2sf_i \left( \frac{t}{2} \right) - (2r + s) f_i(t) - (s - r) f_i(0)$$

$$= f_i(2t) + \cdots + f_i(2t) + f_i \left( \frac{t}{2} \right) + \cdots + f_i \left( \frac{t}{2} \right) - f_i(t) + \cdots + f_i(t) - f_i(0) + \cdots + f_i(0).$$

Hence, we have $\phi_i(q) \geq 0$ if and only if

$$f_i(2t) + \cdots + f_i(2t) + f_i \left( \frac{t}{2} \right) + \cdots + f_i \left( \frac{t}{2} \right) \geq f_i(t) + \cdots + f_i(t) + f_i(0) + \cdots + f_i(0).$$

Using that the $f_i$ are ordering equivalent, we set $m = 2s + r$ and note that $2rt + 2s \cdot \frac{t}{2} = (2r + s)t + (s - r) \cdot 0$. Using the definition of order equivalence, we reach the conclusion (7).
for the rationals $\mathbb{Q}$, that is, that $\text{sign}(\phi_1(q)) = \text{sign}(\phi_2(q))$ for $q \in \mathbb{Q} \cap [0, 1]$. To extend the result to $[0, 1]$, we use the continuity of the $\phi_i$ and the fact that $\text{cl}(\mathbb{Q} \cap [0, 1]) = [0, 1]$, which yields the desired result (7).

By the convexity of $f_i$, Karamata’s majorization inequality (Lemma A.4) implies that

$$f_i(2t) + 2f_i \left( \frac{t}{2} \right) \geq 3f_i(t) \quad \text{and} \quad 2f_i \left( \frac{t}{2} \right) \leq f_i(t) + f_i(0).$$

That is, we have

$$\phi_1(1) = f_i(2t) + 2f_i \left( \frac{t}{2} \right) - 3f_i(t) \geq 0 \quad \text{and} \quad \phi_1(0) = 2f_i \left( \frac{t}{2} \right) - f_i(0) - f_i(t) \leq 0.$$

Using the intermediate value theorem, there exists $r \in [0, 1]$ such that $\phi_1(r) = 0$, and as $\text{sign}(\phi_1) = \text{sign}(\phi_2)$, we have $\phi_1(r) = \phi_2(r) = 0$. Rewriting this using the definition of $\phi_1, \phi_2$, we have

$$0 = \phi_i(r) = rf_i(2t) + 2f_i \left( \frac{t}{2} \right) - (2r + 1)f_i(t) - (1 - r)f_i(0),$$

which we rewrite as

$$f_i \left( \frac{t}{2} \right) = \frac{1}{2} ((2r + 1)f_i(t) + (1 - r)f_i(0) - rf_i(2t)).$$

As $f_1(x) = f_2(x)$ for $x \in \{0, t, 2t\}$, the above equation guarantees that $f_1(t/2) = f_2(t/2)$, completing our proof of claim (a).

A similar technique yields our claim (b), that is, that if $0, t, 2t \in S$, then $3t/2 \in S$. Consider now functions $\phi_i : [0, 1] \to \mathbb{R}$ defined by

$$\phi_i(q) = f_i \left( \frac{3t}{2} \right) + \frac{q}{2} f_i(0) - \frac{2q + 1}{2} f_i(t) - \frac{1 - q}{2} f_i(2t).$$

As in our argument for claim (a), we have $\text{sign}(\phi_1) = \text{sign}(\phi_2)$ on $\mathbb{Q} \cap [0, 1]$: if we assume $q = \frac{r}{s}$ for some $s, r \in \mathbb{N}$ with $s \geq r$, then $\phi_i(q) \geq 0$ if and only if

$$f_i \left( \frac{3t}{2} \right) + \cdots + f_i \left( \frac{3t}{2} \right) + \frac{q}{2} f_i(0) + \cdots + \frac{q}{2} f_i(0) \geq f_i(t) + \cdots + f_i(t) + f(2t) + \cdots f(2t).$$

Using the order-equivalence of $f_1$ and $f_2$ by taking $m = 2s + r$ (note that $2s \left( \frac{3t}{2} \right) + r \cdot 0 = (2r + s)t + 2(s - r)t$, we see that $\text{sign}(\phi_1(q)) = \text{sign}(\phi_2(q))$ for $q \in \mathbb{Q} \cap [0, 1]$. The obvious extension to $[0, 1]$ follows, so that $\text{sign}(\phi_1) = \text{sign}(\phi_2)$ on $[0, 1]$. Now, note that by convexity of the $f_i$, Karamata’s majorization inequality (Lemma A.4) implies

$$\phi_i(0) = \frac{1}{2} \left( 2f_i \left( \frac{3t}{2} \right) - f_i(2t) - f_i(t) \right) \leq 0 \quad \text{and} \quad \phi_i(1) = \frac{1}{2} \left( 2f_i \left( \frac{3t}{2} \right) + f_i(0) - 3f_i(t) \right) \leq 0.$$
The intermediate value theorem implies there exists \( r \in [0, 1] \) such that \( \phi_i(r) = 0 \) for \( i = 1, 2 \), as \( \text{sign}(\phi_1) = \text{sign}(\phi_2) \), whence

\[
\phi_i(r) = f_i \left( \frac{3t}{2} \right) + \frac{r}{2} f_i(0) - \frac{2r + 1}{2} f_i(t) - \frac{1 - r}{2} f_i(2t) = 0.
\]

This immediately implies

\[
f_i \left( \frac{3t}{2} \right) = \frac{2r + 1}{2} f_i(t) + \frac{1 - r}{2} f_i(2t) - \frac{r}{2} f_i(0).
\]

As \( f_1(x) = f_2(x) \) for \( x = 0, t, 2t \), the above formula implies \( f_1(x) = f_2(x) \) also at \( x = 3t/2 \), yielding our claim (b).

The final part of our argument to note that claims (a) and (b) can be generalized to any three numbers \( \{s, s+t, s+2t\} \in [0,2] \). That is, if \( \{s, s+t, s+2t\} \subset S \), then \( s + \frac{1}{2} t \) and \( s + \frac{3}{2} t \) belong to \( S \). This follows when we define the shifted functions \( \tilde{f}_i(x) = f_i(s + x) \), noting that \( \tilde{f}_1 \) and \( \tilde{f}_2 \) are order equivalent (Def. A). Applying claims (a) and (b) at \( 0, t, 2t \) to the \( \tilde{f}_i \), we see that \( \tilde{f}_1(t/2) = \tilde{f}_2(t/2) \) and \( \tilde{f}_1(3t/2) = \tilde{f}_2(3t/2) \). An inductive argument immediately implies that all the dyadic numbers, those rationals of the form \( q = m/2^n \), \( m = 0, \ldots, 2^{n+1} \), belong to the set \( S \). As \( f_1 \) and \( f_2 \) are continuous, we have that \( S = \{x \geq 0 \mid f_1(x) = f_2(x)\} \) is closed, and by the denseness of the dyadic numbers, it must be the case that \( S \supset [0,2] \). \( \square \)

One may use the argument we have shown above for proof of the last observation to get the following corollary,

**Corollary A.6.** Let \( f_1 : \mathbb{R}_+ \to \mathbb{R} \) and \( f_2 : \mathbb{R}_+ \to \mathbb{R} \) be order equivalent. If further \( f_1(a + ib) = f_2(a + ib) \) for \( i = 0, 1, 2 \), then we have \( f_1(x) = f_2(x) \) for all \( x \in [a, a + 2b] \).

**Lemma A.7.** Let \( f_1, f_2 : \mathbb{R}_+ \to \mathbb{R} \) be order equivalent. If \( f_1(i) = f_2(i) \) for \( i = 0, 1, 2 \), and further \( 2f_1(1) \neq f_1(2) + f_1(0) \), then \( f_1(t) = f_2(t) \) for all \( t \in \mathbb{R} \).

**Proof.** Convexity of \( f_1 \) together with the given condition implies \( 2f_1(1) - f_1(2) - f_1(0) < 0 \). Since \( f_1 \) and \( f_2 \) are order equivalent, we obtain \( 2f_2(1) - f_2(2) - f_2(0) < 0 \). Let \( n \geq 3 \) be an arbitrary integer and consider functions \( \phi_i : [0,1] \to \mathbb{R} \) for \( i \in \{1, 2\} \), defined by:

\[
\phi_i(q) = q f_i(n) + (-1 + nq) f_i(0) - ((n + 2)q - 2) f_i(1) - (1 - q) f_i(2).
\]

We claim that (7) still holds. Indeed, first for rational numbers such as \( q = \frac{r}{s} \), where \( r, s \in \mathbb{N} \) with \( s \geq r \) we have

\[
s \cdot \phi_i(q) = r f_i(n) + (-r + ns) f_i(0) - ((n + 2)r - 2s) f_i(1) - (s - r) f_i(2)
= f_i(n) + \cdots + f_i(n) + f_i(0) + \cdots + f_i(0) - f_i(1) + \cdots + f_i(1) - f_i(2) + \cdots + f_i(2).
\]

\( r \) times\hspace{1cm} \( \frac{r}{s} \) times\hspace{1cm} \( \frac{n}{s} \) times\hspace{1cm} \( \frac{n}{s} \) times\hspace{1cm} \( \frac{(n + 2) - 2s}{s} \) times\hspace{1cm} \( \frac{(n + 2) - 2s}{s} \) times\hspace{1cm} \( \frac{(s - r)}{s} \) times.
Hence, we have $\phi_i(q) \geq 0$ if and only if
\[
\underbrace{f_i(1) + \cdots + f_i(1)}_{(n+2)r-s \text{ times}} + \underbrace{f_i(0) + \cdots + f_i(0)}_{ns-r \text{ times}} \geq \underbrace{f_i(1) + \cdots + f_i(1)}_{(n+2)s-r \text{ times}} + \underbrace{f_i(2) + \cdots + f_i(2)}_{(s-r) \text{ times}}.
\]
If $ns - r < 0$ we can move the term containing $f(0)$ to the right hand side. Likewise, if $(n+2)r-s < 0$ we can move the term containing $f(1)$ to the left hand side. We set $m = ns = r+(ns-r) = ((n+2)r-s)+(s-r)$ and note that $n.r+0.(ns-r) = ((n+2)r-s).1+(s-r).2$.

Since $f_1$ and $f_2$ are order equivalent, $\phi_1$ and $\phi_2$ satisfy (7) for $q \in \mathbb{Q} \cap [0,1]$. Using the the continuity of $\phi_1$ and $\phi_2$ we can extend this result to all numbers in $[0,1]$, which proves the claim. By convexity of $f_i$, Karamata’s Majorization Inequality (Lemma A.4) implies that:
\[
\phi_i(1) = f_i(n) + (n-1)f_i(0) - nf_i(1) \geq 0 \quad \text{and} \quad \phi_i(0) = 2f_i(1) - f_i(2) - f_i(0) < 0.
\]

Using the intermediate value theorem, there exists $u \in (0,1]$ such that $\phi_1(u) = 0$ and since $\text{sign}(\phi_1) = \text{sign}(\phi_2)$ we obtain $\phi_2(u) = 0$. We can rewrite this as:
\[
\phi_i(u) = uf_i(n) + (-1+nu)f_i(0) - ((n+2)u-2)f_i(1) - (1-u)f_i(2) = 0,
\]
which can also be written as:
\[
f_i(n) = \frac{1}{u}((-((n+2)u-2)f_i(1) + (1-u)f_i(2) - (nu-1)f_i(0)).
\]
This shows that $f_1(n) = f_2(n)$. Since $n$ was chosen arbitrarily, we obtain $f_1(n) = f_2(n)$ for all integers $n \geq 0$. For completing the proof we only need to use Corollary A.6. Precisely, for any $t \in \mathbb{R}$, pick $a$ and $b$ as integers with the condition $a < t < a + 2b$. Since, $f_1(a+ib) = f_2(a+ib)$ for $i \in \{0,1,2\}$, $f_1(t) = f_2(t)$ according to the Corollary A.6.

Now, we have the tools required for proving the main result. Here we only show the difficult direction for $K = 1$.

**Proof of Theorem 4.1** One way is easy to prove. If, $f_i(t) = af_2(t) + b^T t + c$ then by computing the $f_1$ and $f_2$-divergences we get:
\[
D_{f_1}(P_1, P_2, \ldots, P_{k-1} \parallel P_k \mid q) = aD_{f_2}(P_1, P_2, \ldots, P_{k-1} \parallel P_k \mid q)
\]
Since $a > 0$ this means that $\Phi_1$ and $\Phi_2$ induce the same ordering on quantizers, as required. For proving the other way assuming $K = 1$, using the A.3 we observe that $f_1$ and $f_2$, are order equivalent. Denote the set $T_i = \{t \in \mathbb{R}_+ \mid f_i(2t) + f_i(0) \neq 2f_i(t)\}$, for $i \in \{1,2\}$. We have two different cases:

(a) $T_1 \neq \emptyset$. Pick any $t^* \in T_1$. Note that $t^* \in T_2$, according to the ordering equivalence of $f_1$ and $f_2$. Define $\tilde{f}_i(t) = f_i(t/t^*)$ for $i \in \{1,2\}$. It is easy to check that $\tilde{f}_1(t)$ and $\tilde{f}_2(t)$ are order equivalent. Choose numbers $a, b, c$ as:
\[
a = \frac{\tilde{f}_1(0) + \tilde{f}_1(2) - 2\tilde{f}_1(1)}{\tilde{f}_2(0) + \tilde{f}_2(2) - 2\tilde{f}_2(1)},
\]
\[
b = \tilde{f}_1(0) - a\tilde{f}_2(0),
\]
\[
c = \tilde{f}_1(1) - \tilde{f}_1(0).
\]
Construct a new function \( \tilde{f}_2(t) = a\tilde{f}_2(t) + ct + b \). According to the choice of \( a, b, \) and \( c \), we have \( \tilde{f}_1(t) = \tilde{f}_2(t) \) for \( t \in \{0, 1, 2\} \). Furthermore, \( 2\tilde{f}_1(1) \neq \tilde{f}_1(2) + \tilde{f}_1(0) \), since \( t^* \in T_1 \). Lemma A.7 implies that \( \tilde{f}_1(t) = \tilde{f}_2(t) \) for all \( t \in \mathbb{R} \). That is, \( \tilde{f}_1(t/t^*) = a\tilde{f}_2(t/t^*) + ct + b \) for all \( t \in \mathbb{R} \). Since, \( t^* > 0 \) is fixed we obtain \( f_1(t) = a\tilde{f}_2(t) + (ct^*)t + b \), which implies the result in this case.

(b) \( T_1 = \emptyset \). Then for all \( t \in \mathbb{R}_+ \), \( f_1(2t) + f_1(0) = 2f_1(t) \). This also implies that \( f_2(2t) + f_2(0) = 2f_2(t) \) for all \( t \in \mathbb{R}_+ \), regarding to the ordering equivalence. Using the convexity of \( f_i \), we get that \( f_i \) is an affine function on \( \mathbb{R} \). Therefore, there exists \( b, c \in \mathbb{R} \) such that \( f_1(t) = f_2(t) + bt + c \), as desired.

\[ \square \]

References


