Shannon meets Blackwell and Le Cam: channels, codes, and statistical experiments

Maxim Raginsky

***Abstract*—The Blackwell–Le Cam decision theory provides an approximation framework for statistical experiments in terms of expected risks of optimal decision procedures. The Blackwell partial order formalizes an intuitive notion of which experiment of a given pair is "more informative" for the purposes of inference. The Le Cam deficiency is an approximation measure for any two statistical experiments (with the same parameter space), and it tells us how much we will lose if we base our decisions on one experiment rather than another. In this paper, we develop an extension of the Blackwell–Le Cam theory, starting from a partial ordering for channels introduced by Shannon. In particular, we define a new approximation measure for channels, which we call the Shannon deficiency, and use it to prove an approximation theorem for channel codes that extends an earlier result of Shannon. We also construct a broad class of deficiency-like measures for channels based on generalized divergences, relate them to several alternative notions of capacity, and prove new upper and lower bounds on the Le Cam deficiency.**

I. Introduction

In two seminal papers written in the early 1950's [1], [2], David Blackwell has introduced the concept of "comparison of statistical experiments" based on expected losses of statistical decision procedures. This work has sparked a great deal of interest and led to a number of alternative comparison criteria, including information-theoretic ones [3], [4]. Le Cam [5] has extended Blackwell's theory to an *approximation* approach, which has since then been developed into a comprehensive theory of experiments (cf. [6], [7] for thorough expositions). This theory has led to deep results in mathematical statistics, such as the proof of asymptotic equivalence of nonparametric regression and function filtering in white Gaussian noise [8].

In information-theoretic terms, a statistical experiment is just a noisy communication channel with an uncoded input [9]; all the processing is done at the output. While this set-up is the right one for statistics, in information theory we are also interested in such things as coding or modulation, which are interposed between the source and the channel. A coding/decoding comparison criterion for channels has been introduced in a short paper of Shannon [10]. Although Shannon's comparison criterion bears certain similarities to Blackwell's, it appears to have been developed independently (in fact, there is no reference to Blackwell in [10]) from a random coding perspective.

The present paper makes the following contributions:

- Following Le Cam [5], we augment the Shannon partial order with a norm-based *approximation* criterion. Thus, instead of asking which of two channels is "less noisy" or "more informative," we ask how well one channel can be *approximated* by another. By allowing randomization at the input, as well as some shared randomness between the input and the output terminals, we define a new approximation measure for channels based on the Shannon ordering. In parallel to Le Cam's terminology, we call it the *Shannon deficiency* of one channel w.r.t. another.
- Using this new notion of deficiency, we extend the result of Shannon [10], which says that for any code on a more noisy channel we can find a code on a less noisy channel that does at least as well, to an *arbitrary* pair of channels, where the notion of "doing at least as well" is replaced by "doing at least as well up to ε."
- The Le Cam and Shannon deficiencies are based on the total variation distance. By replacing the total variation with any generalized divergence between probability distributions that obeys a data processing inequality (a *g-divergence* in the terminology of [11]), we obtain a new broad class of deficiency-like quantities, many of which can be related to various generalizations of the channel capacity and can be used to derive upper and lower bounds on the Le Cam and Shannon deficiencies.

This work was supported in part by DARPA under project KECoM.

The author is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA. E-mail: m.raginsky@duke.edu.

II. Notation and preliminaries

We take all alphabets to be *standard Borel* [12] (i.e., isomorphic to a Borel subspace of a Polish space). This covers virtually all settings of practical interest. Any such space X will always be endowed with its Borel σ-algebra $\mathcal{B}(\mathsf{X})$. The space of all probability measures on X will be denoted by $\mathcal{P}(\mathsf{X})$. The *total variation distance* between $P, Q \in \mathcal{P}(\mathsf{X})$ is

$$\|P - Q\| \triangleq \sup_{A \in \mathcal{B}(\mathsf{X})} |P(A) - Q(A)|. \tag{1}$$

A *Markov* (or *transition probability*) *kernel* between X and Y is a mapping $T : \mathcal{B}(\mathsf{Y}) \times \mathsf{X} \to [0, 1]$, such that $T(\cdot|x) \in \mathcal{P}(\mathsf{Y})$ for all $x \in \mathsf{X}$ and $T(B|\cdot)$ is a measurable function on X for any $B \in \mathcal{B}(\mathsf{Y})$. We will denote the space of all such T by $\mathcal{M}(\mathsf{Y}|\mathsf{X})$. If both X and Y are finite, then any $T \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$ is a stochastic matrix with elements $T(y|x)$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$.

Any $T \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$ induces a mapping $\mathcal{P}(\mathsf{X}) \longrightarrow \mathcal{P}(\mathsf{Y})$, which we also will denote with a slight abuse of notation by T; it maps any $P \in \mathcal{P}(\mathsf{X})$ to $Q = TP \in \mathcal{P}(\mathsf{Y})$, where

$$Q(B) = TP(B) \triangleq \int_{\mathsf{X}} T(B|x) P(\mathrm{d}x), \quad \forall B \in \mathcal{B}(\mathsf{Y})$$

We will denote the composition of Markov kernels by juxtaposition. That is, for $T \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$ and $S \in \mathcal{M}(\mathsf{U}|\mathsf{Y})$, their

978-1-4577-0594-6/11/$26.00 ©2011 IEEE

composition $ST \in \mathcal{M}(\mathsf{U}|\mathsf{X})$ is defined by

$$ST(C|x) \triangleq \int_{\mathsf{Y}} S(C|y)T(\mathrm{d}y|x), \quad \forall x \in \mathsf{X}, C \in \mathcal{B}(\mathsf{U}).$$

The total variation distance (1) induces a distance on $\mathcal{M}(\mathsf{Y}|\mathsf{X})$:

$$\|T - T'\|_\infty \triangleq \sup_{x \in \mathsf{X}} \|T(\cdot|x) - T'(\cdot|x)\|.$$

The following *contractivity properties* are easily proved:

$$\|S(T - T')\|_\infty \le \|T - T'\|_\infty, \|(S - S')T\|_\infty \le \|S - S'\|_\infty. \tag{2}$$

A *channel* with input alphabet X and output alphabet Y is simply a Markov kernel $W \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$ [13]. Whenever we need to specify the input and the output alphabets explicitly, we will represent the channel as a triple $(\mathsf{X}, \mathsf{Y}, W)$. An *$M$-code* for $(\mathsf{X}, \mathsf{Y}, W)$ is a pair (E, D), where $E \in \mathcal{M}(\mathsf{X}|\{1, \ldots, M\})$ is a (possibly randomized) encoder and $D \in \mathcal{M}(\{1, \ldots, M\}|\mathsf{Y})$ is a (possibly randomized) decoder. The *average probability of error* of (E, D) on W is

$$\overline{\varepsilon}_W(E, D) \triangleq \frac{1}{M} \sum_{j=1}^{M} \Big(1 - DWE(j|j)\Big).$$

We say that (E, D) is an (M, ε)-code if $\overline{\varepsilon}_W(E, D) \le \varepsilon$.

III. Comparison of channels

A. Blackwell ordering and Le Cam deficiency

Consider two channels $(\mathsf{X}, \mathsf{Y}, W)$ and $(\mathsf{X}, \mathsf{U}, W')$ with a common input alphabet X. When are we justified in saying that W is less noisy (or more informative) than W'? One natural answer is if we can simulate a single use of W' by post-processing a single use of W. This is precisely the comparison criterion proposed by Blackwell [1], [2]. Following the terminology of [7], we will say that W is *Blackwell sufficient* for W' and write $W \succeq_{\mathrm{B}} W'$ or $W' \preceq_{\mathrm{B}} W$ if $W' = TW$ for some Markov kernel $T \in \mathcal{M}(\mathsf{U}|\mathsf{Y})$. Another way to state this is that W' can be realized as a *stochastic degrading* of W. If $W \succeq_{\mathrm{B}} W'$ and $W' \succeq_{\mathrm{B}} W$, we will say that they are *Blackwell equivalent* and write $W \sim_{\mathrm{B}} W'$.

Blackwell sufficiency induces only a *partial order* on the family of all channels with the same input alphabet X. In fact, most channels are incomparable. As an example, let $\mathsf{X} = \mathsf{Y} = \mathsf{U} = \mathbb{R}$, and consider the additive channels

$$W: \ Y = X + Z \quad \text{and} \quad W': \ U = X + Z'$$

where Z and Z' are real-valued random variables independent of the input X. Then $W \succeq_{\mathrm{B}} W'$ if and only if the law of Z is a convolution factor of the law of Z' [14].

The scope of Blackwell's theory was expanded considerably by Le Cam [5], who introduced the notion of a *deficiency* of one statistical experiment w.r.t. another. We can do the same for channels, where in our information-theoretic context the deficiency of W w.r.t. W' will quantify how well any post-processing of W can be used to *approximate* W. The (*Le Cam*) *deficiency* of W w.r.t. W' is given by

$$\delta(W, W') \triangleq \inf_{T \in \mathcal{M}(\mathsf{U}|\mathsf{Y})} \|TW - W'\|_\infty. \tag{3}$$

According to the so-called *Le Cam randomization criterion* [6], [7], the infimum in (3) is actually achieved, i.e., there exists some $T^* \in \mathcal{M}(\mathsf{U}|\mathsf{Y})$, such that $\delta(W, W') = \|T^*W - W'\|_\infty$. Note that $\delta(W, W') = 0$ iff $W \succeq_{\mathrm{B}} W'$.

B. Shannon ordering and Shannon deficiency

The Shannon ordering criterion [10] also revolves around the simulation of one channel by another, but it allows for pre- and post-processing, as well as for shared randomness between the input and the output terminals. Consider two channels $(\mathsf{X}, \mathsf{Y}, W)$ and $(\mathsf{X}', \mathsf{Y}', W')$, where the input alphabets need not necessarily be the same. Adopting our earlier terminology and the definitions of [10], we say that W is *Shannon sufficient* for W', and write $W \succeq_{\mathrm{S}} W'$, if there exist some $k \in \mathbb{N}$, a probability vector $\pi = (\pi_1, \ldots, \pi_k)$, and k pairs $(T_i, S_i) \in \mathcal{M}(\mathsf{X}|\mathsf{X}') \times \mathcal{M}(\mathsf{Y}'|\mathsf{Y}), 1 \le i \le k$, such that

$$W' = \sum_{i=1}^{k} \pi_i S_i W T_i.$$

In the special case $\mathsf{X} = \mathsf{X}'$ it is easy to see that $W \succeq_{\mathrm{B}} W'$ implies $W \succeq_{\mathrm{S}} W'$, but not necessarily *vice versa*. Thus, the Shannon ordering criterion is weaker than that of Blackwell. However, it is also only a partial order. To be able to compare any channel to any other, we adopt Le Cam's strategy and define the corresponding *Shannon deficiency* as follows:

$$\delta_{\mathrm{S}}(W, W') \triangleq \inf_{k \in \mathbb{N}} \inf_{\pi;\, (T_i, S_i)_{i=1}^k} \left\| \sum_{i=1}^{k} \pi_i S_i W T_i - W' \right\|_\infty. \tag{4}$$

The following bound is immediate:

$$\delta_{\mathrm{S}}(W, W') \le \inf_{T \in \mathcal{M}(\mathsf{X}|\mathsf{X}')} \delta(WT, W').$$

When $\mathsf{X} = \mathsf{X}'$, it can be weakened to $\delta_{\mathrm{S}}(W, W') \le \delta(W, W')$, which suggests that the Le Cam deficiency is a more stringent measure than the Shannon deficiency. Again, note that $\delta_{\mathrm{S}}(W, W') = 0$ iff $W \succeq_{\mathrm{S}} W'$.

IV. Basic properties of channel deficiencies

A. Operational interpretation

In [10], Shannon gave an operational interpretation to the $\succeq_{\mathrm{S}}$ channel ordering: If $W \succeq_{\mathrm{S}} W'$, then for any (M, ε)-code for W' we can find an (M, ε)-code for W. The following theorem is an extension of Shannon's result to an arbitrary pair of channels in terms of the Shannon deficiency (4):

Theorem 1. *Consider two channels $(\mathsf{X}, \mathsf{Y}, W)$ and $(\mathsf{X}', \mathsf{Y}', W')$. If $\delta_{\mathrm{S}}(W, W') \le \varepsilon'$, then for any (M, ε)-code for W' we can find an $(M, \varepsilon + \varepsilon')$-code for W.*

Proof: If $\delta_{\mathrm{S}}(W, W') \le \varepsilon'$, we can find some $k \in \mathbb{N}$, a

probability vector $\pi = (\pi_1, \ldots, \pi_k)$, and $(T_i, S_i)_{i=1}^k$, such that

$$\left\| \sum_{i=1}^k \pi_i S_i W T_i - W' \right\|_\infty \leq \varepsilon'. \tag{5}$$

Now let (E, D) be an (M, ε)-code for W'. For each $1 \leq i \leq k$, define an M-code (E_i, D_i) for W by letting $E_i = T_i E$ and $D_i = D S_i$. Let I be a random variable taking values in $\{1, \ldots, k\}$ according to the distribution π. Then

$$\begin{aligned}
&\mathbb{E}\left[\overline{\varepsilon}_W(E_I, D_I)\right] - \overline{\varepsilon}_{W'}(E, D) \\
&\leq \max_{j \in \{1,\ldots,M\}} \sum_{\ell=1}^M \left| \mathbb{E}[D_I W E_I(\ell|j)] - D W' E(\ell|j) \right| \\
&= \left\| D \left(\mathbb{E}[S_I W T_I] - W' \right) E \right\|_\infty \\
&\leq \left\| \mathbb{E}[S_I W T_I] - W' \right\|_\infty \\
&\leq \varepsilon',
\end{aligned}$$

where the third step uses (2), and the last step is by (5). Since (E, D) is an (M, ε)-code for W', this gives

$$\mathbb{E}\left[\overline{\varepsilon}_W(E_I, D_I)\right] \leq \varepsilon + \varepsilon'.$$

Therefore, there must exist at least one value $i^* \in \{1, \ldots, k\}$, for which $\overline{\varepsilon}_W(E_{i^*}, D_{i^*}) \leq \varepsilon + \varepsilon'$. The theorem is proved. ∎

B. Monotonicity of Le Cam deficiency under data processing

The following result is known (cf. [6]), but we give a self-contained proof, since we will need it later:

Theorem 2. *Consider any three channels* $(\mathsf{X}, \mathsf{Y}, W)$, $(\mathsf{X}, \mathsf{U}, W')$ *and* $(\mathsf{X}, \mathsf{Z}, V)$. *Then:*

1) *If* $W \succeq_{\mathrm{B}} W'$, *then* $\delta(W, V) \leq \delta(W', V)$.
2) *If* $W' \succeq_{\mathrm{B}} V$, *then* $\delta(W, V) \leq \delta(W, W')$.

Equality holds in both cases if $\succeq_{\mathrm{B}}$ *is replaced by* $\sim_{\mathrm{B}}$.

Proof: (1) Since $W \succeq_{\mathrm{B}} W'$, there exists some $T \in \mathcal{M}(\mathsf{U}|\mathsf{Y})$ such that $W' = TW$. Hence,

$$\begin{aligned}
\delta(W', V) &= \inf_{S \in \mathcal{M}(\mathsf{Z}|\mathsf{U})} \|SW' - V\|_\infty \\
&= \inf_{S \in \mathcal{M}(\mathsf{Z}|\mathsf{U})} \|STW - V\|_\infty \\
&\geq \inf_{S \in \mathcal{M}(\mathsf{Z}|\mathsf{Y})} \|SW - V\|_\infty \\
&\equiv \delta(W, V).
\end{aligned}$$

(2) Again, since $W' \succeq_{\mathrm{B}} V$, there exists some $T \in \mathcal{M}(\mathsf{Z}|\mathsf{U})$ such that $V = TW'$. Then

$$\begin{aligned}
\delta(W, V) &= \inf_{S \in \mathcal{M}(\mathsf{Z}|\mathsf{Y})} \|SW - TW'\|_\infty \\
&\leq \inf_{S \in \mathcal{M}(\mathsf{U}|\mathsf{Y})} \|TSW - TW'\|_\infty \\
&\leq \inf_{S \in \mathcal{M}(\mathsf{U}|\mathsf{Y})} \|SW - W'\|_\infty \\
&\equiv \delta(W, W'),
\end{aligned}$$

where the third step is by (2). ∎

C. Comparison to extremal channels

A channel with a *finite* input alphabet and an arbitrary output alphabet is called *semicontinuous* (SC) [15, Ch. 5]. For a fixed finite X, we have two *extremal* channels, W_0 and W_1, such that $W_0 \preceq_{\mathrm{B}} W \preceq_{\mathrm{B}} W_1$ for any $(\mathsf{X}, \mathsf{Y}, W)$. These are unique up to Blackwell equivalence, and can be represented by $W_0(\cdot|x) = e_{x_0}$ for some fixed $x_0 \in \mathsf{X}$ and by $W_1(\cdot|x) = e_x$, where e_x is the probability distribution that puts all mass on $x \in \mathsf{X}$.

Given an arbitrary SC channel $(\mathsf{X}, \mathsf{Y}, W)$, let us define

$$\overline{\delta}(W) \triangleq \delta(W, W_1) \qquad \text{and} \qquad \underline{\delta}(W) \triangleq \delta(W_0, W).$$

Theorem 3 (Torgersen [16])**.**

$$\overline{\delta}(W) = \inf_{T \in \mathcal{M}(\mathsf{X}|\mathsf{Y})} \max_{x \in \mathsf{X}} \Big(1 - TW(x|x)\Big) \tag{6}$$

$$\underline{\delta}(W) = \inf_{Q \in \mathcal{P}(\mathsf{Y})} \max_{x \in \mathsf{X}} \|Q - W(\cdot|x)\| \tag{7}$$

In operational terms, $\overline{\delta}(W)$ is the best maximal probability of error achievable on W with an arbitrary decoder and an identity encoder, $E(x|x) = 1$ for all $x \in \mathsf{X}$. On the other hand, $\underline{\delta}(W)$ has a geometric interpretation as the "radius" of the set $\{W(\cdot|x)\}_{x \in \mathsf{X}}$ w.r.t. the total variation distance. The unique $Q^* \in \mathcal{P}(\mathsf{Y})$ that attains the infimum in (7) can be thought as the "center" of this set. We will come back to this interpretation later and extend it to more general notions of "information radius" [9], [17].

We can also use the Shannon deficiency to define $\overline{\delta}_{\mathrm{S}}(W)$ and $\underline{\delta}_{\mathrm{S}}(W)$. The former can be bounded as $\overline{\delta}_{\mathrm{S}}(W) \leq \overline{\delta}(W)$. For the latter, it is easy to show from definitions that it is equal to the Le Cam deficiency $\underline{\delta}(W)$.

V. Generalized deficiencies

By using other statistical or information-theoretic measures of (dis)similarity between probability distributions instead of the total variation distance, it is possible to construct a wide variety of deficiency-like quantities. The resulting alternative (or generalized) deficiencies are of interest in their own right, and they can also be used to bound the Le Cam deficiencies from above and from below. We start by defining a broad class of deficiencies based on a very minimalistic requirement: monotonicity under data processing. Following recent work of Polyanskiy and Verdú [11], we call any mapping $\mathcal{D} : \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \to \overline{\mathbb{R}}$ a *g-divergence* if

$$\mathcal{D}(TP \| TQ) \leq \mathcal{D}(P \| Q) \tag{8}$$

for any $P, Q \in \mathcal{P}(\mathsf{X})$ and any Markov kernel $T \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$. Given the channels $(\mathsf{X}, \mathsf{Y}, W)$ and $(\mathsf{X}, \mathsf{U}, W')$, let us define the corresponding *g-deficiency* of W w.r.t. W' as

$$\delta_{\mathcal{D}}(W, W') \triangleq \inf_{T \in \mathcal{M}(\mathsf{U}|\mathsf{Y})} \sup_{x \in \mathsf{X}} \mathcal{D}\big(W'(\cdot|x) \big\| TW(\cdot|x)\big).$$

The data-processing inequality (8) is enough to ensure the same monotonicity properties as the Le Cam deficiency:

Theorem 4. *The analog of Theorem 2 holds if we replace* δ *with* $\delta_{\mathcal{D}}$ *throughout.*

As we shall see next, particular choices of a g-divergence with additional properties (such as convexity) lead to many useful alternative notions of deficiency.

A. f-deficiencies

A wide class of g-divergences is formed by the so-called f-divergences of Csiszár (cf. [18] and references therein). Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let P and Q be two probability measures on a space X. Then the f-divergence between P and Q is defined as

$$D_f(P\|Q) \triangleq \int_{\mathsf{X}} \frac{\mathrm{d}Q}{\mathrm{d}\mu} f\left(\frac{\mathrm{d}P/\mathrm{d}\mu}{\mathrm{d}Q/\mathrm{d}\mu}\right) \mathrm{d}\mu, \tag{9}$$

where μ is any σ-finite measure on X that dominates both P and Q[1], and we use the conventions

$$0 \cdot f(0/0) = 0, \quad 0 \cdot f(t/0) = t \lim_{s \searrow 0} s f(1/s), \forall t > 0. \tag{10}$$

With proper choice of f, we recover many of the frequently used measures of divergence, for example:

- $f(t) = t \log t$: the relative entropy $D(P\|Q)$.
- $f(t) = |t-1|$: the total variation distance $\|P - Q\|$.

Thus, given two channels $(\mathsf{X}, \mathsf{Y}, W)$ and $(\mathsf{X}, \mathsf{U}, W')$, we can define the *$f$-deficiency* of W w.r.t. W' as

$$\delta_f(W, W') \triangleq \inf_{T \in \mathcal{M}(\mathsf{U}|\mathsf{Y})} \sup_{x \in \mathsf{X}} D_f(W'(\cdot|x) \| TW(\cdot|x))$$

For an SC channel $(\mathsf{X}, \mathsf{Y}, W)$, we have an analog of Theorem 3:

Theorem 5. *Suppose that* $\lim_{t \searrow 0} f(t) = 0$. *Then*

$$\overline{\delta}_f(W) = \inf_{T \in \mathcal{M}(\mathsf{X}|\mathsf{Y})} \max_{x \in \mathsf{X}} f^*\left(TW(x|x)\right) \tag{11}$$

$$\underline{\delta}_f(W) = \inf_{Q \in \mathcal{P}(\mathsf{Y})} \max_{x \in \mathsf{X}} D_f(W(\cdot|x)\|Q), \tag{12}$$

where $f^*(t) \triangleq tf(1/t)$ *is the* Csiszár conjugate *of* f *[18].*

Remark 1. The f-deficiency $\underline{\delta}_f(W)$ is the *absolute f-informativity* (or the *f-radius*) of W [9].

Proof: In (9), let μ be the counting measure on X. Then

$$\begin{aligned}
\overline{\delta}_f(W) &= \inf_T \max_{x \in \mathsf{X}} \sum_{x' \in \mathsf{X}} TW(x'|x) \cdot f\left(\frac{e_x(x')}{TW(x'|x)}\right) \\
&= \inf_T \max_{x \in \mathsf{X}} TW(x|x) \cdot f\left(\frac{1}{TW(x|x)}\right) \qquad (13) \\
&= \inf_T \max_{x \in \mathsf{X}} f^*\left(TW(x|x)\right), \qquad (14)
\end{aligned}$$

where (13) uses (10) and the fact that $f(t) \to 0$ as $t \searrow 0$, while (14) uses the definition of f^*. This proves (11); the simple proof of (12) is omitted. ∎

For the special case $f(t) = t \log t$ (natural logarithms), let us use the term *I-deficiency* and write δ_I.

[1]It is not hard to show that the value of $D_f(P\|Q)$ does not depend on the choice of the dominating measure μ.

Theorem 6. *For any two channels* $(\mathsf{X}, \mathsf{Y}, W)$ *and* $(\mathsf{X}, \mathsf{U}, W')$,

$$\delta(W, W') \le \sqrt{\delta_I(W, W')/2}. \tag{15}$$

For any SC channel $(\mathsf{X}, \mathsf{Y}, W)$,

$$\overline{\delta}_I(W) = -\log\left(1 - \overline{\delta}(W)\right) \quad \textit{and} \quad \underline{\delta}_I(W) = C(W),$$

where $C(W) = \max_{P \in \mathcal{P}(\mathsf{X})} I(P, W)$ *is the Shannon capacity of* W. *Moreover, if* $(\mathsf{X}, \mathsf{Y}, W)$ *and* $(\mathsf{X}, \mathsf{U}, W')$ *are SC, then* $\delta_I(W, W') \le \log|\mathsf{X}|$.

Proof: The bound (15) follows from Pinsker's inequality. The expression for $\overline{\delta}_I(W)$ follows from Theorems 5 and 3. Next, using (12) and the minimax theorem, we have

$$\begin{aligned}
\underline{\delta}_I(W) &= \max_{P \in \mathcal{P}(\mathsf{X})} \min_{Q \in \mathcal{P}(\mathsf{Y})} \sum_{x \in \mathsf{X}} P(x) D(W(\cdot|x)\|Q) \\
&= \max_{P \in \mathcal{P}(\mathsf{X})} I(P, W) \equiv C(W),
\end{aligned}$$

Finally, by data processing we have $\delta_I(W, W') \le \delta_I(W_0, W_1) = C(W_1) = \log|\mathsf{X}|$. ∎

B. Rényi deficiencies

The *Rényi divergence* of order $\lambda \in (0,\infty)\backslash\{1\}$ between P and Q is [19]

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \int \frac{\mathrm{d}Q}{\mathrm{d}\mu} \left(\frac{\mathrm{d}P/\mathrm{d}\mu}{\mathrm{d}Q/\mathrm{d}\mu}\right)^\lambda \mathrm{d}\mu.$$

Although it is not an f-divergence (but rather a monotone transformation of one, the so-called Hellinger divergence [18]), it is a g-divergence [11]. Hence we can define the *Rényi deficiency* of order λ, which we will denote by $\delta_\lambda(\cdot,\cdot)$.

Theorem 7. *Let* $(\mathsf{X}, \mathsf{Y}, W)$ *and* $(\mathsf{X}, \mathsf{U}, W')$ *be SC. Then*

$$\begin{aligned}
\underline{\delta}_\lambda(W) &= C_\lambda(W) \\
&= \max_{P \in \mathcal{P}(\mathsf{X})} \min_{Q \in \mathcal{P}(\mathsf{Y})} \sum_{x \in \mathsf{X}} P(x) D_\lambda(W(\cdot|x)\|Q),
\end{aligned}$$

the order-λ capacity of W [19], and $\delta_\lambda(W, W') \le \log|\mathsf{X}|$.

Proof: An argument similar to the proof of (12) leads to

$$\underline{\delta}_\lambda(W) = \min_{Q \in \mathcal{P}(\mathsf{Y})} \max_{x \in \mathsf{X}} D_\lambda(W(\cdot|x)\|Q).$$

From Proposition 1 in [19], the right-hand side is equal to $C_\lambda(W)$. By data processing, $\delta_\lambda(W, W') \le \delta_\lambda(W_0, W_1) = C_\lambda(W_1)$. From [19, p. 28], $C_\lambda(W_1) = \max_{P \in \mathcal{P}(\mathsf{X})} H(P) = \log|\mathsf{X}|$, where $H(P)$ is the Shannon entropy of P. ∎

If both X and Y are finite and $\lambda = 1/(1+\rho)$ for some $\rho > 0, \rho \ne 1$, then we can express $\underline{\delta}_\lambda(W)$ in terms of Gallager's function $E_0(\rho, P, W)$ [20] as

$$\underline{\delta}_\lambda(W) = \frac{1}{\rho} \max_{P \in \mathcal{P}(\mathsf{X})} E_0(\rho, P, W).$$

C. Neyman–Pearson deficiencies

An interesting result from [11] is that there exist g-divergences that are not representable as monotone functions of f-divergences. One example given there uses the *Neyman–Pearson functions* [6]. Given two probability measures P and

Q on a space X, let us define for any $0 < \alpha < 1$ the number

$$\beta_\alpha(P,Q) \triangleq \inf_{T \in \mathcal{M}(\{0,1\}|\mathsf{X})} \{TQ(1) : TP(1) \geq \alpha\}.$$

Then $-\beta_\alpha(TP,TQ) \leq -\beta_\alpha(P,Q)$ for any $T \in \mathcal{M}(\mathsf{Y}|\mathsf{X})$ [11]. Using this fact, let us define the *Neyman–Pearson α-divergence* between P and Q by

$$\mathcal{D}_\alpha(P\|Q) \triangleq \alpha - \beta_\alpha(P,Q);$$

we will denote the corresponding g-deficiency by $\delta_\alpha(\cdot,\cdot)$.

Theorem 8. *The Le Cam deficiency $\delta(W,W')$ between any two channels $(\mathsf{X},\mathsf{Y},W)$ and $(\mathsf{X},\mathsf{U},W')$ satisfies the bound*

$$\delta(W,W') \geq \delta_\alpha(W,W')$$

for any $0 < \alpha < 1$. Moreover, if $(\mathsf{X},\mathsf{Y},W)$ and $(\mathsf{X},\mathsf{U},W')$ are SC, then $\delta_\alpha(W,W') \leq \alpha(1-1/|\mathsf{X}|)$.

Proof: We start by noting that $\|P - Q\| = \sup_{0<\alpha<1} \mathcal{D}_\alpha(P\|Q)$ [6, p. 36]. Then

$$\begin{aligned}\delta(W,W') &= \inf_T \sup_{x\in\mathsf{X}} \|TW(\cdot|x) - W'(\cdot|x)\| \\ &= \inf_T \sup_{0<\alpha<1} \sup_{x\in\mathsf{X}} \mathcal{D}_\alpha\big(W'(\cdot|x)\big\|TW(\cdot|x)\big) \\ &\geq \sup_{0<\alpha<1} \mathcal{D}_\alpha(W,W').\end{aligned}$$

If $(\mathsf{X},\mathsf{Y},W)$ and $(\mathsf{X},\mathsf{U},W')$ are SC, then by data processing $\delta_\alpha(W,W') \leq \delta_\alpha(W_0,W_1)$, which in turn is equal to

$$\alpha - \sup_{Q\in\mathcal{P}(\mathsf{X})} \min_{x\in\mathsf{X}} \beta_\alpha(e_x,Q) = \alpha(1-1/|\mathsf{X}|)$$

(see Lemma 1 in the Appendix). ∎

VI. Conclusion

We have extended the notion of comparison between statistical experiments [1], [2], as well as an approximation theory for experiments [5], to coding/decoding scenarios common in information theory. Starting from a comparison criterion for channels due to Shannon [10], we have defined an approximation measure for channels, which we have termed the *Shannon deficiency*, mimicking the terminology of [5]. An interesting, though challenging, future direction would be to develop an approximation-based perspective on the coding theorems of information theory, similar to Le Cam's programme in asymptotic statistics.

References

[1] D. Blackwell, "Comparison of experiments," in *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, 1951, pp. 93–102.

[2] ——, "Equivalent comparisons of experiments," *Ann. Math. Statist.*, vol. 24, pp. 265–272, 1953.

[3] D. V. Lindley, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, vol. 27, pp. 986–1005, 1956.

[4] P. K. Goel and M. H. DeGroot, "Comparison of experiments and information measures," *Ann. Statist.*, vol. 7, no. 5, pp. 1066–1077, 1979.

[5] L. Le Cam, "Sufficiency and approximate sufficiency," *Ann. Math. Statist.*, vol. 35, pp. 1419–1455, 1964.

[6] E. Torgersen, *Comparison of Statistical Experiments*. Cambridge Univ. Press, 1991.

[7] A. N. Shiryaev and V. G. Spokoiny, *Statistical Experiments and Decisions: Asymptotic Theory*. Singapore: World Scientific, 2000.

[8] L. D. Brown and M. G. Low, "Asymptotic equivalence of nonparametric regression and white noise," *Ann. Statist.*, vol. 24, no. 6, pp. 2384–2398, 1996.

[9] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Math. Hungar.*, vol. 2, no. 1–4, pp. 191–213, 1972.

[10] C. E. Shannon, "A note on a partial ordering for communication channels," *Inform. Control*, vol. 1, pp. 390–397, 1958.

[11] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf. on Commun., Control, and Comput.*, 2010, pp. 1327–1333.

[12] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. New York: Academic Press, 1967.

[13] R. L. Dobrushin, "A general formulation of the basic Shannon theorem in information theory," *Uspekhi Math. Nauk*, vol. 14, no. 6, pp. 3–103, 1959.

[14] M. Stone, "Non-equivalent comparisons of experiments and their use for experiments involving location parameters," *Ann. Statist.*, vol. 32, pp. 326–332, 1961.

[15] A. Feinstein, *Information Theory*. New York: McGraw-Hill, 1958.

[16] E. N. Torgersen, "Measures of information based on comparison with total information and with total ignorance," *Ann. Statist.*, vol. 9, no. 3, pp. 638–657, 1981.

[17] R. Sibson, "Information radius," *Z. Wahrsch. verw. Geb.*, vol. 14, pp. 149–160, 1969.

[18] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[19] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

[20] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

Appendix

Lemma 1. *For a finite alphabet X,*

$$\sup_{Q\in\mathcal{P}(\mathsf{X})} \min_{x\in\mathsf{X}} \beta_\alpha(e_x,Q) = \alpha/|\mathsf{X}|.$$

Proof: Fix some $Q \in \mathcal{P}(\mathsf{X})$ and $x \in \mathsf{X}$. Then

$$\beta_\alpha(e_x,Q) = \min_{T\in\mathcal{M}(\{0,1\}|\mathsf{X})} \{TQ(1) : Te_x(1) \geq \alpha\}. \quad (16)$$

Since $Te_x(1) = \sum_{x'\in\mathsf{X}} T(1|x')e_x(x') = T(1|x)$, we can rewrite (16) as

$$\beta_\alpha(e_x,Q) = \min_{T\in\mathcal{M}(\{0,1\}|\mathsf{X})} \{TQ(1) : T(1|x) \geq \alpha\}.$$

For any feasible T, we must have

$$TQ(1) = \sum_{x'\in\mathsf{X}} T(1|x')Q(x') \geq T(1|x)Q(x) \geq \alpha Q(x),$$

which means that $\beta_\alpha(e_x,Q) \geq \alpha Q(x)$. On the other hand, choosing T^* with $T^*(1|x) = \alpha$ and $T^*(1|x') = 0$ for all $x' \neq x$, we have $T^*Q(1) = \alpha Q(x)$, so $\beta_\alpha(e_x,Q) \leq \alpha Q(x)$. Thus, $\beta_\alpha(e_x,Q) = \alpha Q(x)$. This and the fact that

$$\max_{Q\in\mathcal{P}(\mathsf{X})} \min_{x\in\mathsf{X}} Q(x) = \min_{x\in\mathsf{X}} \max_{Q\in\mathcal{P}(\mathsf{X})} Q(x) = 1/|\mathsf{X}|$$

completes the proof. ∎