Channel optimization for binary hypothesis testing [⋆]

Giancarlo Baldan [*] **Munther Dahleh** [**]

[*] *MIT, Laboratory for Information and Decision systems, Cambridge 02139 USA (email: gbaldan@mit.edu)*
[**] *MIT, Laboratory for Information and Decision systems, Cambridge 02139 USA (email: dahleh@mit.edu)*

Abstract: In this paper we consider the classical binary hypothesis testing problem where the iid samples are obtained through a channel. Our goal is to study the relationship between the channel capacity and the goodness of the estimation measured by the Chernoff information in order to get an upper bound on the estimation performances as well as some insight on the structure of the optimal channel.

Keywords: Optimal Estimation; Hypotheses; Capacity.

1. INTRODUCTION

The binary hypothesis testing problem is probably the simplest estimation problem one can consider. In the classical setup a sequence of samples x_i is drawn from an unknown n–dimensional probability distribution which can be either p_1 (Hypothesis H_1) with prior probability π_1 or p_2 (Hypothesis H_2) with prior probability π_2. The problem is to infer from the samples which of the two hypothesis is correct or, to be precise, the most likely.

This problem is very well known and is optimally solved using the likelihood ratio test (LRT) as shown, for example, in [1]. Furthermore, applying a large deviation principle, an asymptotic analysis can be performed to show that the probability of error in the estimation decays exponentially in the number of samples with a rate given by the so called Chernoff Information.

In this paper, we will consider an extension to this problem motivated by the fact that each collected sample is always obtained through a measuring system that can affect the estimation process. To model the effects due to the measuring system we will consider that the observations at the source are available only through a finite capacity, discrete, memoryless stochastic channel. Our goal is to address the question of designing such a channel to maximize goodness of the estimation as measured by the decay rate of the probability of error as well as obtaining a relationship between the capacity of the channel and the quality of the estimation.

2. BASIC DEFINITIONS

In this section we briefly review some basic quantities defined in Information theory. These quantities will be used throughout the whole paper and some approximations will be introduced to make their definitions more tractable.

The Kullback Leibler distance is one of the most important quantities in information theory and measure the distance between two probability distribution p and q defined over the same alphabet $\mathcal{X}$. The Kullback distance is defined as:

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}, \tag{1}$$

where the logarithm will be always considered in base e. Since from definition (1) it's clear that $D(p\|q)$ doesn't depend on the alphabet but just on the two distributions themselves we will usually adopt the notation:

$$D(p\|q) = \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}, \tag{2}$$

where n is the cardinality of $\mathcal{X}$ and we omit the dependence on the alphabet.

Another measure of the distance between two distributions p and q which is closely related to the Binary Hypothesis testing problem, is the so called Chernoff information whose definition is:

$$C(p,q) = D(p^{\lambda_*}\|p) = D(p^{\lambda_*}\|q), \tag{3}$$

where p^λ is a probability distribution defined as:

$$p_i^\lambda = \frac{p_i^\lambda q_i^{1-\lambda}}{\sum_{i=1}^n p_i^\lambda q_i^{1-\lambda}},$$

and λ_* is such that $D(p^{\lambda_*}\|p) = D(p^{\lambda_*}\|q)$.

We will often consider discrete, memoryless, stochastic channels mapping the alphabet $\mathcal{X}$ into a finite alphabet $\mathcal{Y}$ whose cardinality is m. This kind of channels is completely described by a conditional probability distribution:

$$W(y|x) = \mathbb{P}(Y = y|X = x), \tag{4}$$

which can be regarded as an $m \times n$ stochastic matrix and will be often denoted just by W.

To measure the capacity of such a channel we will use the standard "information" definition:

$$C = \max_{p_x} I(X;Y) \tag{5}$$

[⋆] This work was supported under NSF/EFRI grant 0735956 and under AFOSR/MURI grant R6756-G2

where X is a random variable such that $X \sim p_x$ and $Y \sim Wp_x$ is the corresponding random variable obtained through the channel. $I(X;Y)$ is called mutual information between X and Y and is defined as:

$$I(X;Y) = D(p_{xy} \| p_x p_y) = E_x[D(W(\cdot|x) \| p_y)]. \quad (6)$$

3. PROBLEM FORMULATION

The problem we are trying to face is, essentially, an optimization one and, therefore, to provide a correct formulation, we have to identify three major components: optimization variables, cost function and constraints. In this section we will define these components trying to motivate the choices made.

3.1 Optimization variables.

We model our sample source as a discrete random variable X over a finite alphabet $\mathcal{X}$ such that $|\mathcal{X}| = n$. The mass distribution of X depend on the unknown hypothesis:

$$X \sim \begin{cases} p_1 & \text{under } H_1 \\ p_2 & \text{under } H_2 \end{cases}, \quad (7)$$

where $p_1, p_2 \in R^n$ and $\mathbb{P}[H_1] = \pi_1$, $\mathbb{P}[H_2] = \pi_2$.

The channel through which we obtain the measurement is supposed to be a discrete, memoryless, stochastic channel W mapping the alphabet $\mathcal{X}$ into a finite alphabet $\mathcal{Y}$ whose cardinality is m.

Both the dimension of the output alphabet m and the channel W itself, will be regarded as optimization variables, thus allowing a complete flexibility in the choice of the most suitable channel.

3.2 Constraints.

Without any further assumption on the class of feasible channels, any optimization problem would be solved by the choice $m = n$ and $W = I$ that makes the random variable X perfectly measurable as if there was no channel at all. To make the scenario more realistic we decided to introduce a constraint in the capacity of the channel as measured by the usual mutual information between X and Y:

$$\max_{p_x} I(X;Y) \leq C. \quad (8)$$

We made this choice because the capacity of a channel is a reasonable abstraction of its quality and is often the most critical specification for a communication system.

3.3 Cost function.

Since the random process observable after the channel y_n is still i.i.d. with a distribution that can be either $q_1 = Wp_1$ or $q_2 = Wp_2$ depending on the true hypothesis, it is reasonable to measure the quality of the estimation using a standard technique for binary hypothesis testing applied to the process y_n.

We chose to optimize the asymptotic performance of the system in terms of the probability of error. Specifically it is well known that for a Binary hypothesis testing problem there exist a sequence of optimal estimators $\hat{H}_n : \mathcal{Y}^n \mapsto \{1,2\}$, designed using a log-likelihood ratio, such that they minimize the probability of error given n samples:

$$\begin{aligned} P_e(n) = \, & \mathbb{P}(\hat{H}_n(y_1, \ldots, y_n) = 2 | H = 1)\pi_1 + \\ & \mathbb{P}(\hat{H}_n(y_1, \ldots, y_n) = 1 | H = 2)\pi_2. \end{aligned}$$

Moreover it has been shown that $P_e(n)$ decays exponentially with n at a rate given by the Chernoff information $C(q_1, q_2)$, that is:

$$-\lim_{n \to \infty} \frac{1}{n} \log P_e(n) = C(q_1, q_2).$$

Our goal is then to maximize the Chernoff Information $C(q_1, q_2) = C(Wp_1, Wp_2)$ in order for the probability of error to decay as fast as possible. The complete formulation of the optimization problem can be written as:

$$\begin{cases} \max_{W,m} \; C(Wp_1, Wp_2) \\ \text{s.t.} \\ \qquad \max_{p_X} I(X;Y) \leq C \\ \qquad \mathbb{1}^T W = \mathbb{1}^T \\ \qquad W_{i,j} \geq 0 \end{cases} \quad (9)$$

In section 5 we will assume that the capacity C is small enough so that the cost function and the constraints in (9) can be approximated by more tractable expression thus leading to an approximating optimization problem valid for small capacities. By explicitly solving this problem we will gain some insight regarding the structure of the solutions of (9) in the small C regime. In the next section we will introduce some basic tools useful to perform the required approximations.

4. EUCLIDEAN APPROXIMATIONS

In this section we present some approximations to the quantities defined in section 2. To obtain these approximations we follow the idea known as Euclidean Information theory and presented in details in [2]. We start considering a simple Taylor expansion $\log(1+x) = x - \frac{x^2}{2} + \varphi(x)$, where $\varphi(x) = o(x^2)$ when x tends to 0. Applying this expansion to the definition of the Kullback distance (2) we get

$$\begin{aligned} D(p\|q) &= -\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \\ &= -\sum_{i=1}^{n} p_i \log \left(1 + \frac{q_i - p_i}{p_i}\right) \\ &= \frac{1}{2} \sum_{i=1}^{n} \frac{(q_i - p_i)^2}{p_i} - \sum_{i=1}^{n} p_i \, \varphi\left(\frac{q_i - p_i}{p_i}\right) \\ &= \frac{1}{2} \|p - q\|^2_{[p]^{-1}} - \sum_{i=1}^{n} p_i \, \varphi\left(\frac{q_i - p_i}{p_i}\right) \quad (10) \end{aligned}$$

where $[p]$ is a diagonal matrix whose diagonal elements are given by $p_i, \; i = 1, \ldots, n$.

We can simplify the expression in (10) by noticing that the last summation is an infinitesimal of a superior order with respect to $\|p-q\|^2_{[p]^{-1}}$ as proved by the following inequality:

$$\frac{\left|\sum_{i=1}^{n} p_i\,\varphi\left(\frac{q_i-p_i}{p_i}\right)\right|}{\|p-q\|^2_{[p]^{-1}}} \le \frac{\left|\varphi\left(\frac{q_j-p_j}{p_j}\right)\right|}{p_j\,\frac{(q_j-p_j)^2}{p_j^2}} \xrightarrow{p\to q} 0,$$

where j is the index such that $\left|\varphi\left(\frac{q_j-p_j}{p_j}\right)\right|$ is maximum and can be regarded as a function of p. Furthermore the quantities $\|p-q\|^2_{[p]^{-1}}$ and $\|p-q\|^2_{[q]^{-1}}$ are infinitesimal of the same order as $p \to q$ since from the inequalities[1]:

$$\min_i \frac{q_i}{p_i} \le \frac{\|p-q\|^2_{[p]^{-1}}}{\|p-q\|^2_{[q]^{-1}}} \le \max_i \frac{q_i}{p_i},$$

it follows that

$$\lim_{p\to q} \frac{\|p-q\|^2_{[p]^{-1}}}{\|p-q\|^2_{[q]^{-1}}} = 1 \qquad (11)$$

simply applying the squeeze theorem.

By virtue of the last two observations the expression in (10) can be finally written as

$$D(p\|q) = \frac{1}{2}\|p-q\|^2_{[q]^{-1}} + o\left(\|p-q\|^2_{[q]^{-1}}\right), \qquad (12)$$

and we can now use this expression to approximate both the definition of capacity (5) and Chernoff information.

Regarding the Chernoff information we can approximate it with an easier Kullback distance as stated in the following proposition:

Proposition 1. If two probability distributions p and q, defined on the same alphabet $\mathcal{X}$, are close enough then the following approximation holds:

$$C(p,q) \approx \frac{1}{4}D(p\|q).$$

More formally we have:

$$\lim_{p\to q} \frac{C(p,q)}{D(p\|q)} = \frac{1}{4}.$$

Proof: See appendix A.

Regarding the definition of channel capacity(5), it's well known (see [3]) that if $p*$ is the optimal input distribution achieving the capacity and p_0 is the corresponding output distribution, we have $D(W_i\|p_0) = C \quad \forall i : p_i^* > 0$ and $D(W_i\|p_0) < C \quad \forall i : p_i^* = 0$. By virtue of this consideration, under the assumption of a small C, all the conditional distributions W_i will be close to p_0 and the distances are well approximated by the expression (12) thus obtaining:

$$\frac{1}{2}\|W_i - p_0\|^2_{[p_0^{-1}]} \le C \quad \forall i. \qquad (13)$$

It's easy to see that the converse is true as well; if we fix a point p_0 on the simplex and choose n probability vector W_i satisfying the constraints (13), the resulting channel will have a capacity less than C. Therefore conditions (13) are an alternative formulation of the channel capacity constraint (8) and their only disadvantage is that they require a new arbitrary probability vector p_0.

5. NOISY CHANNEL SOLUTION

With the term "noisy channel" we mean a channel whose capacity C is small. In this section we aim to approximate, under the assumption $C << 1$, the general problem (9) with a more tractable optimization problem, whose solution can be computed explicitly and will allow us to understand the behavior of (9) in the noisy channel regime.

If $C << 1$ we can take advantage of the constraints (13) since they imply that Wp_1 and Wp_2 are close no matter what p_1 and p_2 are. If Wp_1 and Wp_2 are close, by virtue of proposition 1 ,we can use the approximation:

$$C(Wp_1, Wp_2) = \frac{1}{4}D(Wp_1\|Wp_2)$$

and therefore maximizing the Chernoff information turns out to be equivalent to maximizing the Kullback distance $D(Wp_1\|Wp_2)$. Finally, using again equation (12), we can approximate the Chernoff information via an Euclidean distance:

$$C(Wp_1, Wp_2) = \frac{1}{4}D(Wp_1\|Wp_2) = \frac{1}{8}\|Wp_1 - Wp_2\|^2_{[p_0^{-1}]}. \qquad (14)$$

Using the approximation for the capacity constraint (13) and the result in (14), the original problem (9) can be approximated by:

$$\begin{cases} \max\limits_{W,m,p_0} \frac{1}{8}\|W(p_1-p_2)\|^2_{[p_0^{-1}]} \\ \text{s.t.} \\ \qquad \frac{1}{2}\|W_i - p_0\|^2_{[p_0^{-1}]} \le C \quad \forall i \\ \qquad \mathbb{1}^T W = \mathbb{1}^T \\ \qquad W_{i,j} \ge 0 \end{cases}, \qquad (15)$$

and the advantage of this formulation is that it leads to an analytical solution as stated in the following proposition.

Proposition 2. Choose arbitrarily $m \ge 2$ and p_0 in the m-dimensional simplex and then consider an arbitrary probability vector w_A such that $\frac{1}{2}\|w_A - p_0\|^2_{[p_0^{-1}]} = C$ as well as the only other vector w_B whose distance from p_0 is C and is opposite to w_A with respect to p_0, that is $w_B = 2p_0 - w_A$. Next consider the following channel:

$$W_i^* = \begin{cases} w_A & \text{if } p_1^i \ge p_2^i \\ w_B & \text{if } p_1^i < p_2^i \end{cases} \qquad \forall i = 1, \ldots, n, \qquad (16)$$

then channel (16) is the optimal solution of (15) and the associated optimal cost is

$$\frac{1}{4}C\|p_1 - p_2\|_1^2 \qquad (17)$$

Proof:

Let's consider m and p_0 fixed. We will prove the statement showing first that the expression (17) is an upper bound to the optimal value and then that W^* achieves that bound.

To bound the cost we'll use the fact that $p_1 - p_2$ adds up to zero and therefore if A is a matrix with all the columns equal to each others then $A(p_1 - p_2) = 0$. Formally we obtain:

[1] For results on bounding a ratio of two quadratic form we refer to [4]

$$\begin{aligned}\|W(p_1 - p_2)\|_{[p_0^{-1}]} &= \|(W - [p_0|\cdots|p_0])\,(p_1 - p_2)\|_{[p_0^{-1}]}\\ &= \left\|\sum_{i=1}^{n}(W_i - p_0)\,(p_1^i - p_2^i)\right\|_{[p_0^{-1}]}\\ &\le \sum_{i=1}^{n}|p_1^i - p_2^i|\,\|W_i - p_0\|_{[p_0^{-1}]}\\ &\le \sum_{i=1}^{n}|p_1^i - p_2^i|\sqrt{2C}\\ &= \sqrt{2C}\|p_1 - p_2\|_1\end{aligned}$$

which is equivalent to $\frac{1}{8}\|W(p_1-p_2)\|^2_{[p_0^{-1}]} \le \frac{1}{4}C\|p_1-p_2\|_1^2$.

To prove that W^* achieves this bound let's start defining the quantity

$$\alpha = \sum_{i:p_1^i \ge p_2^i}(p_1^i - p_2^i),$$

and noticing that, since $p_1 - p_2$ adds up to zero, we also have:

$$\alpha = -\sum_{i:p_1^i < p_2^i}(p_1^i - p_2^i),$$

$$2\alpha = \|p_1 - p_2\|_1.$$

Now, with some algebra we get:

$$\begin{aligned}&\frac{1}{8}\|W^*(p_1 - p_2)\|^2_{[p_0^{-1}]} =\\ &= \frac{1}{8}\left\|w_A\sum_{i:p_1^i\ge p_2^i}(p_1^i - p_2^i) + w_B\sum_{i:p_1^i<p_2^i}(p_1^i - p_2^i)\right\|^2_{[p_0^{-1}]}\\ &= \frac{1}{8}\,\|\alpha w_A - \alpha w_B\|^2_{[p_0^{-1}]}\\ &= \frac{1}{8}\,\|2\alpha(w_A - p_0)\|^2_{[p_0^{-1}]}\\ &= \alpha^2 C\\ &= \frac{1}{4}C\|p_1 - p_2\|_1^2.\end{aligned}$$

Remarkably, the optimal value we obtained considering m and p_0 fixed turned out to be completely independent of m and p_0 and, therefore, problem (15) is solved by a triplet (W^*, m, p_0) where m and p_0 can be chosen arbitrary provided that the definition of W^* in (16) yields a well defined stochastic matrix [2].

□

A graphical depiction of w_A and w_B, used to construct the optimal channel, is reported in figure 1; we point out that, since C is considered small, it's always possible to determine such a pair of vectors inside the simplex.

The result just proven shows that for small capacity the behavior of the Chernoff bound is linear in C and is proportional to the l_1 distance between the two hypothesis. In the next section we will present some observations, based just on simulations, regarding the behavior for larger C.

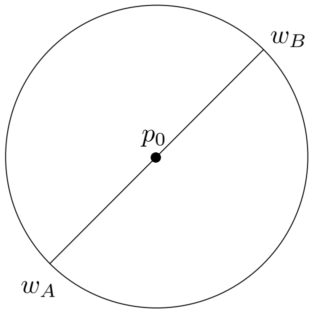

Fig. 1. Position of w_A and w_B in the simplex with respect to p_0

6. LARGE CAPACITY BEHAVIOR

As the capacity increases the problem (15) is no longer approximating the original optimization problem (9). In the general case finding an analytical solution to (9) is unrealistic but we can still make some remarks. In this section we will point out some of these interesting features and we will present a numerical result.

- For each m the solution of (9) as a function of C is monotone increasing and the optimal channel has always capacity C. This is true because the cost function can be shown to be convex and W belongs to a convex set by virtue of convexity of $I(Y, X)$ with respect to the channel.
- For each m the performances are not improving for $C \ge \log m$ because the maximum capacity achievable with an m dimensional output alphabet is always less than $\log m$. Moreover if $m = n$ then for $C \ge n$ we obtain exactly the Chernoff information since among the feasible channels there is the identity channel I which allows to measure the samples directly from the source. Finally the curves obtained for $m > n$ seem to be identical to the one obtained for $m = n$.
- Interestingly, for some choices of the Hypothesis p_1 and p_2, the Chernoff information is reached (with $m = n$) before the limit $C = \log n$

In figure 2 we show the solution of problem (9) where we kept m as a parameter. Only two different values of m have been taken into account but it's still possible to observe some of the behaviors just pointed out.

7. CONCLUSIONS AND FUTURE WORK

In this paper we considered a modified version of the binary hypothesis testing problem where the samples are measured through a channel. We looked for the best possible channel among those with a limited capacity and we showed that, if the channel has a small capacity, this optimization problem can be approximated by a quadratic one. The optimal solution for the approximating problem achieves an error exponent given by

$$\frac{1}{4}C\|p_1 - p_2\|_1^2$$

where C is the capacity of the channel while p_1 and p_2 are the two hypothesis. In the small C regime we were also able to provide an explicit formula for the optimal channel. It is not yet formally proved, although clearly supported by simulations, that the optimal solution of the

[2] Namely p_0 must be chosen far from the simplex borders so that w_A and w_B fall inside the simplex.

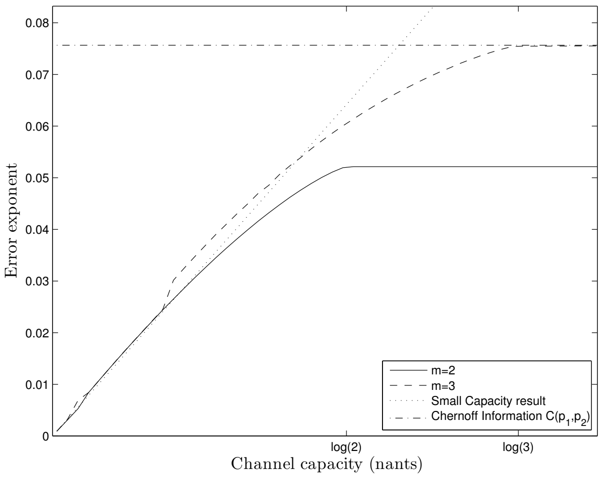

Fig. 2. Solution of problem (9) with $n = 3$, $p_1 = [0.53\ 0.13\ 0.34]'$ and $p_2 = [0.23\ 0.42\ 0.35]'$

approximating problem converges to the solution of the original problem as C tends to 0. We are currently working on some generalizations to the m-ary case as well as some non iid-based models like hidden Markov models.

8. ACKNOWLEDGMENT

The autors wish to thank Mesrob I. Ohannessian for his suggestions and help in proving proposition 1.

REFERENCES

[1] Cov:98 T. M. Cover, J. A. Thomas. Elements of Information Theory. Wiley–Interscience Publication, 1991.

[2] Euc:2008 S. Borade, L. Zheng. Euclidean Information Theory. *Communications, 2008 IEEE International Zurich Seminar on* pages 14–17, 2008.

[3] Gal:1968 R. Gallager Information Theory and Reliable Communication Wiley, 1968.

[4] QuadB:1999 F. Caliskan, C. Hajiyev Sensor fault detection in flight control systems based on the Kalman filter innovation sequence *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* volume 213, issue 3, pages 243–248, 1999.

[5] BhattB:1943 A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society* volume 35 pages 99-110, 1943.

Appendix A. PROOF OF PROPOSITION 1

We have to prove that:

$$\lim_{p\to q} \frac{C(p,q)}{D(p\|q)} = \frac{1}{4}.$$

First of all we want to point out that, if some of the components of q are equal to zero, then $D(p\|q)$ is not defined unless the same components of p are zero as well and the limit $p \to q$ is taken over the subspace in which $p_i = 0$ whenever $q_i = 0$. For this reason, without loss of generality, we can restrict our analysis to the case $q_i > 0\ \forall i$.

Since the definition of the Chernoff information in (3) is not in a closed form, in this section we'll provide an explicit expression to approximate it when p and q are close enough.

In order to easily deal with equation (3) let us introduce some notation conventions:

$$\begin{aligned}
D_q(\lambda) &= D(p^\lambda \| q),\\
D_p(\lambda) &= D(p^\lambda \| p),\\
\hat{D}_q(\lambda) &= \frac{1}{2}\|p^\lambda - q\|^2_{[q]^{-1}},\\
\hat{D}_p(\lambda) &= \frac{1}{2}\|p^\lambda - p\|^2_{[p]^{-1}}.
\end{aligned}$$

In [1] it's shown that the function $D_p(\lambda)$ is monotone decreasing in $\lambda \in [0, 1]$ while $D_q(\lambda)$ is increasing in the same interval; moreover there exist a unique $\lambda_* \in [0, 1]$ such that $D_p(\lambda_*) = D_q(\lambda_*)$.

It is also easy to show that $\hat{D}_p(\lambda)$ is monotone decreasing and $\hat{D}_q(\lambda)$ is monotone increasing in [0, 1] and that the unique value of λ satisfying the equation $\hat{D}_p(\lambda) = \hat{D}_q(\lambda)$ is $\lambda = 1/2$. In fact, if we denote with $\phi = \sum_{i=1}^n \sqrt{p_i q_i}$ the Bhattacharyya coefficient, we have:

$$\begin{aligned}
\hat{D}_q(1/2) &= \frac{1}{2}\sum_{i=1}^n \left[\frac{\sqrt{p_i q_i}}{\phi} - q_i\right]^2 \frac{1}{q_i}\\
&= \frac{1}{2}\sum_{i=1}^n \left[\frac{p_i q_i}{\phi^2} + q_i^2 - 2q_i\frac{\sqrt{p_i q_i}}{\phi}\right]\frac{1}{q_i}\\
&= \frac{1}{2}\sum_{i=1}^n \left[\frac{p_i}{\phi^2} + q_i - 2\frac{\sqrt{p_i q_i}}{\phi}\right]\\
&= \frac{1}{2}\left[\frac{1}{\left(\sum_{i=1}^n \sqrt{p_i q_i}\right)^2} - 1\right], \qquad \text{(A.1)}
\end{aligned}$$

which can be shown, with the same argumentation, to be equal to $\hat{D}_p(1/2)$.

We'll now show that the expression just found in (A.1) can be regarded as an approximation of the Chernoff information whose distance from the latter is infinitesimal of a superior order with respect to $\|p-q\|^2_{[q]^{-1}}$.

Let us start examining the difference $D_q - \hat{D}_q$ which, using the uniform bound $\|p^\lambda - q\|^2_{[q]^{-1}} \le \|p-q\|^2_{[q]^{-1}}\ \forall\lambda$ and by virtue of equation (12), turns out to be small as $p \to q$:

$$\begin{aligned}
D_q(\lambda) - \hat{D}_q(\lambda) &= \delta_q(\lambda)\\
&= o(\|p^\lambda - q\|^2_{[q]^{-1}})\\
&= o(\|p-q\|^2_{[q]^{-1}}) \quad \forall\lambda. \qquad \text{(A.2)}
\end{aligned}$$

Using the same argumentation and the result in (11) we can derive a similar result for the difference $D_p - \hat{D}_p$:

$$\begin{aligned}
D_p(\lambda) - \hat{D}_p(\lambda) &= \delta_p(\lambda) \\
&= o(\|p^\lambda - p\|^2_{[p]^{-1}}) \\
&= o(\|q - p\|^2_{[p]^{-1}}) \quad \forall\lambda \\
&= o(\|p - q\|^2_{[q]^{-1}}) \quad \forall\lambda. \qquad \text{(A.3)}
\end{aligned}$$

If we introduce now the two functions:

$$\begin{aligned}
f(\lambda) &= D_p(\lambda) - D_q(\lambda) \\
\hat{f}(\lambda) &= \hat{D}_p(\lambda) - \hat{D}_q(\lambda),
\end{aligned}$$

keeping in mind that $\hat{f}(1/2) = 0$, we can obtain the following bound on $f(1/2)$:

$$\begin{aligned}
|f\,(1/2)| &= \left|f\,(1/2) - \hat{f}\,(1/2)\right| \\
&= \left|D_p\,(1/2) - \hat{D}_p\,(1/2) + \hat{D}_q\,(1/2) - D_q\,(1/2)\right| \\
&\leq |\delta_p\,(1/2)| + |\delta_q\,(1/2)|\,. \qquad \text{(A.4)}
\end{aligned}$$

Using the fact that $\left|\frac{dD_q}{d\lambda}\right| \leq \left|\frac{df}{d\lambda}\right| \quad \forall\lambda$ and the results in (A.2), (A.3), (A.4), we can now show that the distance between $\hat{D}_q(1/2)$ and the Chernoff information $C(p,q) = D_q(\lambda_*)$ is small as $p \to q$:

$$\begin{aligned}
\left|\hat{D}_q(1/2) - D_q(\lambda_*)\right| &\leq |D_q(1/2) - D_q(\lambda_*)| + |\delta_q(1/2)| \\
&\leq |f(1/2) - f(\lambda_*)| + |\delta_q(1/2)| \\
&= |f(1/2)| + |\delta_q(1/2)| \\
&\leq |\delta_p\,(1/2)| + 2\,|\delta_q\,(1/2)| \\
&= o(\|p - q\|^2_{[q]^{-1}}) \qquad \text{(A.5)}
\end{aligned}$$

The result found in (A.5) allows us to write the Chernoff information in an explicit form suitable to our purposes; more precisely:

$$\begin{aligned}
C(p,q) &= \frac{1}{2}\left[\frac{1}{\left(\sum_{i=1}^n \sqrt{p_i q_i}\right)^2} - 1\right] + o\left(\|p - q\|^2_{[q]^{-1}}\right) \\
&= \hat{C}(p,q) + o\left(\|p - q\|^2_{[q]^{-1}}\right) \qquad \text{(A.6)}
\end{aligned}$$

In order to compute the limit $\lim_{p\to q} \frac{C(p,q)}{D(p\|q)}$ on the n-dimensional simplex, let's first reduce the dimension to an n-1 dimensional space where we get rid of the constraint $\sum_{i=1}^n p_i = 1$. In this lower dimensional space the approximate expression for the Kullback distance becomes:

$$\begin{aligned}
\frac{1}{2}\|p - q\|^2_{[q]^{-1}} =& \frac{1}{2}\left[\sum_{i=1}^{n-1} (p_i - q_i)^2 \frac{1}{q_i} + \left(1 - \sum_{i=1}^{n-1} p_i - q_n\right)^2 \frac{1}{q_n}\right] \\
=& \frac{1}{2}(\bar{p} - \bar{q})' \left([\bar{q}]^{-1} + \frac{1}{q_n}\mathbb{1}\mathbb{1}'\right)(\bar{p} - \bar{q}) \\
=& \frac{1}{2}(\bar{p} - \bar{q})' M_q (\bar{p} - \bar{q}), \qquad \text{(A.7)}
\end{aligned}$$

where $\bar{p}$ and $\bar{q}$ are $n-1$ dimensional vectors equal to the first $n-1$ elements of the vectors p and q while $\mathbb{1}$ is the $n-1$ dimensional vector of all 1. The approximate expression for the Chernoff information becomes:

$$\begin{aligned}
&\frac{1}{2}\left[\frac{1}{\left(\sum_{i=1}^n \sqrt{p_i q_i}\right)^2} - 1\right] = \\
&\qquad = \frac{1}{2\left(\sum_{i=1}^{n-1} \sqrt{p_i q_i} + \sqrt{q_n\left(1 - \sum_{i=1}^{n-1} p_i\right)}\right)^2} - \frac{1}{2} \\
&\qquad = F(\bar{p}, q). \qquad \text{(A.8)}
\end{aligned}$$

Before considering the limit let's compute a Taylor expansion of the function F around $\bar{q}$. After some straightforward computations we obtain:

$$\begin{aligned}
F|_{\bar{p}=\bar{q}} &= 0, \\
\left.\frac{\partial F}{\partial p_i}\right|_{\bar{p}=\bar{q}} &= 0 \quad \forall i = 1, \ldots, n-1, \\
\left.\frac{\partial F^2}{\partial p_i^2}\right|_{\bar{p}=\bar{q}} &= \frac{1}{4}\left(\frac{1}{q_i} + \frac{1}{q_n}\right) \quad \forall i = 1, \ldots, n-1, \\
\left.\frac{\partial F^2}{\partial p_i \partial p_j}\right|_{\bar{p}=\bar{q}} &= \frac{1}{4q_n} \quad \forall i \neq j,
\end{aligned}$$

therefore we have:

$$\begin{aligned}
\hat{C}(\bar{p}, q) =& \frac{1}{2!}\frac{1}{4}(\bar{p} - \bar{q})' \left([\bar{q}]^{-1} + \frac{1}{q_n}\mathbb{1}\mathbb{1}'\right)(\bar{p} - \bar{q}) + o(\|\bar{p} - \bar{q}\|^2) \\
=& \frac{1}{8}(\bar{p} - \bar{q})' M_q (\bar{p} - \bar{q}) + o(\|\bar{p} - \bar{q}\|^2)
\end{aligned}$$

Collecting the results obtained so far the limit we want to compute is trivial:

$$\begin{aligned}
\lim_{p\to q} \frac{C(p,q)}{D(p\|q)} &= \lim_{p\to q} \frac{\hat{C}(p,q)}{\frac{1}{2}\|p - q\|^2_{[q]^{-1}}} \\
&= \lim_{\bar{p}\to\bar{q}} \frac{\frac{1}{8}(\bar{p} - \bar{q})' M_q (\bar{p} - \bar{q}) + o(\|\bar{p} - \bar{q}\|^2)}{\frac{1}{2}(\bar{p} - \bar{q})' M_q (\bar{p} - \bar{q})} = \frac{1}{4}.
\end{aligned}$$