Notes on the capacity of the asymmetric binary channel

Steven Glenn Jackson

October 14, 2016

1 Unit conversion

Since this calculation involves derivatives, it is more convenient to measure information in nats. If desired, one can convert all results to bits at the end. One simple way to remember the conversion factor is to compute the entropy of a fair coin toss in both systems. Let $\mathcal{E}$ be the system of events (h, t), each occurring with probability $1/2$. In bits, we have

$$H(\mathcal{E}) = \frac{1}{2}\log_2\left(\frac{1}{1/2}\right) + \frac{1}{2}\log_2\left(\frac{1}{1/2}\right) = 1\,\mathrm{bit}$$

while in nats we have

$$H(\mathcal{E}) = \frac{1}{2}\ln\left(\frac{1}{1/2}\right) + \frac{1}{2}\ln\left(\frac{1}{1/2}\right) = \ln(2)\,\mathrm{nats}.$$

Therefore

$$1\,\mathrm{bit} = \ln(2)\,\mathrm{nats} \approx 0.693\,\mathrm{nats}.$$

Equivalently,

$$1\,\mathrm{nat} = \frac{1}{\ln(2)}\,\mathrm{bits} \approx 1.443\,\mathrm{bits}.$$

2 Simplification of the Lagrange multiplier equations

Certain simplifications of the Lagrange multiplier equations are easier to see in the context of a general channel than in the context of any particular channel; we begin with these. The material in this section closely parallels Subsection 3.4.2 and Theorem 3.4.3 in the text.

Consider a channel with input alphabet $A = \{a_1, \ldots, a_n\}$ and output alphabet $B = \{b_1, \ldots, b_k\}$. As usual, let $q_{i,j}$ denote the probability of the transition $a_i \to b_j$, i.e. the conditional probability that the output of the channel is b_j, given that the input is a_i. Also as usual, let $p_1, \ldots, p_n$ denote the input frequencies; in other words, p_i is the probability that the input is a_i. Finally, let $\mathcal{A}$ and $\mathcal{B}$ be the systems of events associated with the input alphabet and the output alphabet, respectively.

Let $F(p_1, \ldots, p_n)$ denote the mutual information $I(\mathcal{A}, \mathcal{B})$, regarded as a function of the input frequencies $p_1, \ldots, p_n$, with the transition probabilities treated as constants. We wish to maximize the value of F, subject to the constraint

$$g(p_1, \ldots, p_n) = p_1 + \cdots + p_n - 1 = 0.$$

Using the method of Lagrange multipliers, we look for points where the gradient of F is parallel to the gradient of g, that is, points where

$$\nabla F = \lambda \nabla g$$

for some scalar λ. Since $\frac{\partial g}{\partial p_s} = 1$ for all $s \in \{1, \ldots, n\}$, this condition is equivalent to

$$\frac{\partial F}{\partial p_s} = \lambda \qquad \forall s \in \{1, \ldots, n\}. \tag{1}$$

Now we compute the mutual information explicitly. By definition, the probability that the input is a_i is p_i. The probability that the input is a_i and the output is b_j is $p_i q_{ij}$. Finally, the probability that the output is b_j is $\sum_i p_i q_{ij}$. Thus, the mutual information (in nats) is

$$\begin{aligned} F(p_1, \ldots, p_n) &= \sum_{i,j} p_i q_{ij} \ln \left(\frac{p_i q_{ij}}{p_i \sum_t p_t q_{tj}} \right) \\ &= \sum_{i,j} p_i q_{ij} \left(\ln(q_{ij}) - \ln \left(\sum_t p_t q_{tj} \right) \right). \end{aligned}$$

Next let δ denote the Kronecker delta, i.e. δ_{ij} equals one if $i = j$ and zero otherwise. By the product rule, we have

$$\begin{aligned} \frac{\partial F}{\partial p_s} &= \sum_{i,j} \left(\delta_{is} q_{ij} \left(\ln(q_{ij}) - \ln \left(\sum_t p_t q_{tj} \right) \right) - p_i q_{ij} \frac{q_{sj}}{\sum_t p_t q_{tj}} \right) \\ &= \sum_j \left(q_{sj} \left(\ln(q_{sj}) - \ln \left(\sum_t p_t q_{tj} \right) \right) - \frac{\sum_i p_i q_{ij}}{\sum_t p_t q_{tj}} q_{sj} \right) \\ &= \sum_j \left(q_{sj} \ln \left(\frac{q_{sj}}{\sum_t p_t q_{tj}} \right) \right) - \sum_j q_{sj} \\ &= \sum_j \left(q_{sj} \ln \left(\frac{q_{sj}}{\sum_t p_t q_{tj}} \right) \right) - 1. \end{aligned}$$

Setting this equal to λ gives

$$\sum_j q_{sj} \ln \left(\frac{q_{sj}}{\sum_t p_t q_{tj}} \right) = \lambda + 1.$$

Putting $C = \lambda + 1$ and recalling the constraint $g(p_1, \ldots, p_n) = 0$ gives the *capacity equations*

$$\boxed{\begin{aligned} &\sum_j q_{sj} \ln\left(\frac{q_{sj}}{\sum_t p_t q_{tj}}\right) = C \qquad \forall s \in \{1, \ldots, n\} \\ &\sum_i p_i = 1 \end{aligned}} \tag{2}$$

which coincide with the equations appearing in Theorem 3.4.3 in the text.

Finally, suppose $(p_1, \ldots, p_n)$ is a solution of eq. (2). Then

$$\begin{aligned} F(p_1, \ldots, p_n) &= \sum_{i,j} p_i q_{ij} \ln\left(\frac{q_{ij}}{\sum_t p_t q_{tj}}\right) \\ &= \sum_i p_i C \\ &= C \end{aligned}$$

so that C is the channel capacity.

3 The asymmetric binary channel

Now consider a channel with $A = B = \{0, 1\}$ and the following transition probabilities:

$$\begin{aligned} P(0 \to 0) &= \alpha \\ P(0 \to 1) &= 1 - \alpha \\ P(1 \to 1) &= \beta \\ P(1 \to 0) &= 1 - \beta. \end{aligned}$$

For consistency with the notation of the previous section, we put $a_1 = b_1 = 0$ and $a_2 = b_2 = 1$; thus

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} \alpha & 1 - \alpha \\ 1 - \beta & \beta \end{bmatrix}.$$

The capacity equation (2) becomes

$$\begin{aligned} \alpha \ln\left(\frac{\alpha}{p_1\alpha + p_2(1-\beta)}\right) + (1-\alpha)\ln\left(\frac{1-\alpha}{p_1(1-\alpha) + p_2\beta}\right) &= C \\ (1-\beta)\ln\left(\frac{1-\beta}{p_1\alpha + p_2(1-\beta)}\right) + \beta \ln\left(\frac{\beta}{p_1(1-\alpha) + p_2\beta}\right) &= C \end{aligned} \tag{3}$$

and, of course, $p_1 + p_2 = 1$. Setting the left hand sides equal to each other, expanding the logarithms, and rearranging terms gives

$$\begin{aligned} &(1 - \alpha - \beta)(\ln(p_1\alpha + p_2(1-\beta)) - \ln(p_1(1-\alpha) + p_2\beta)) = \\ &\qquad -\alpha\ln(\alpha) - (1-\alpha)\ln(1-\alpha) + (1-\beta)\ln(1-\beta) + \beta\ln(\beta). \end{aligned} \tag{4}$$

Assume for the moment that $1 - \alpha - \beta \neq 0$. Then we have

$$\ln\left(\frac{p_1\alpha + p_2(1-\beta)}{p_1(1-\alpha) + p_2\beta}\right) = \frac{\alpha\ln(\alpha) + (1-\alpha)\ln(1-\alpha) + (1-\beta)\ln(1-\beta) + \beta\ln(\beta)}{1-\alpha-\beta}.$$

The expression on the right hand side is a constant that will appear frequently in the sequel; call it K. Exponentiating, we obtain

$$\frac{p_1\alpha + p_2(1-\beta)}{p_1(1-\alpha) + p_2\beta} = e^K. \tag{5}$$

Next, let

$$\begin{aligned} p_1 &= p \\ p_2 &= 1 - p. \end{aligned}$$

Substituting into eq. (5) gives

$$p\alpha + (1-p)(1-\beta) = e^K(p(1-\alpha) + (1-p)\beta).$$

After a little rearrangement, this becomes

$$p(\alpha + \beta - 1 - e^K(1-\alpha-\beta)) = e^K\beta + \beta - 1$$

whence the optimal input frequencies are

$$\begin{aligned} p_1 = p &= \frac{\beta(e^K+1) - 1}{(\alpha+\beta-1)(e^K+1)} \\ p_2 = 1 - p &= \frac{\alpha(e^K+1) - e^K}{(\alpha+\beta-1)(e^K+1)}. \end{aligned} \tag{6}$$

To compute the channel capacity, we first use the optimal input frequencies to compute the optimal output frequencies:

$$\begin{aligned} p_1\alpha + p_2(1-\beta) &= \frac{\alpha\beta(e^K+1) - \alpha + (1-\beta)\alpha(e^K+1) - (1-\beta)e^K}{(\alpha+\beta-1)(e^K+1)} \\ &= \frac{\alpha e^K + \beta e^K - e^K}{(\alpha+\beta-1)(e^K+1)} \\ &= \frac{e^K}{e^K+1} \\ p_1(1-\alpha) + p_2\beta &= \frac{(1-\alpha)\beta(e^K+1) - (1-\alpha) + \alpha\beta(e^K+1) - \beta e^K}{(\alpha+\beta-1)(e^K+1)} \\ &= \frac{1}{e^K+1}. \end{aligned}$$

Substituting these into eq. (3) gives the channel capacity:

$$\begin{aligned} C &= \alpha\ln\left(\frac{\alpha}{p_1\alpha + p_2(1-\beta)}\right) + (1-\alpha)\ln\left(\frac{1-\alpha}{p_1(1-\alpha) + p_2\beta}\right) \\ &= \alpha\ln(\alpha) - \alpha\ln\left(\frac{e^K}{e^K+1}\right) + (1-\alpha)\ln(1-\alpha) - (1-\alpha)\ln\left(\frac{1}{e^K+1}\right) \\ &= \alpha\ln(\alpha) + (1-\alpha)\ln(1-\alpha) - \alpha K + \ln(e^K+1). \end{aligned}$$

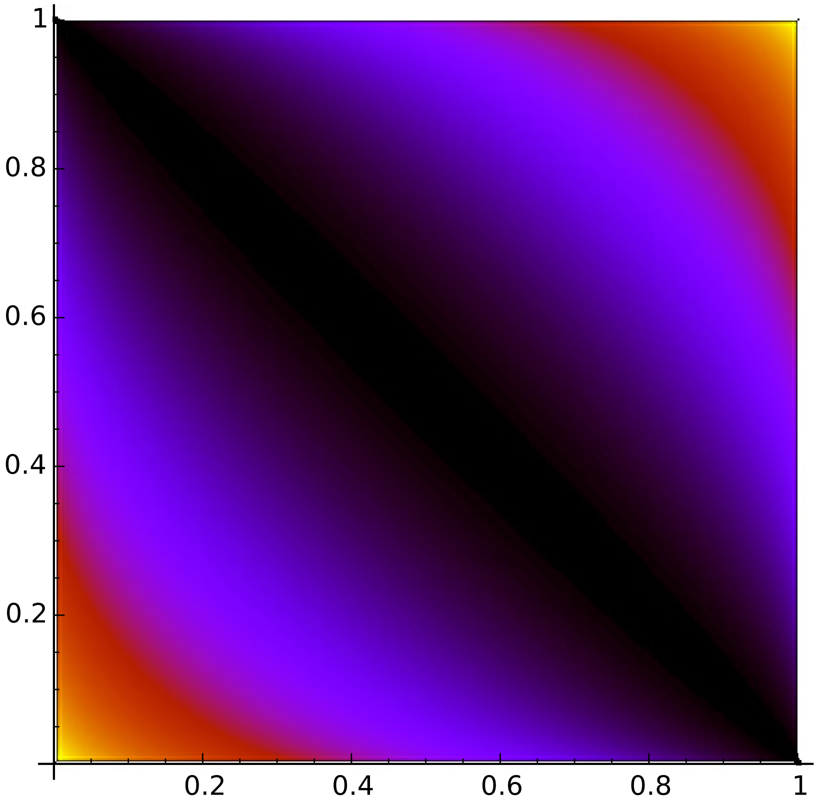

Figure 1: The capacity of the binary asymmetric channel

Finally, we consider the case where $1 - \alpha - \beta = 0$. In this case, the left hand side of eq. (4) collapses to zero, and after eliminating β via the identity $\beta = 1 - \alpha$, so does the right hand side. Thus, every (p_1, p_2) with $p_1 + p_2 = 1$ satisfies the capacity equation, and we may find the capacity by substituting any such pair we please into eq. (3). One convenient choice is $p_1 = 1$, $p_2 = 0$, which shows immediately that in this case the channel capacity is zero.

4 Visualization

Our formulas for the channel capacity and the optimal input frequencies are rather complicated, and it is not straightforward to visualize their behavior, so there is some advantage in using a computer algebra system to produce their graphs. This can be done in many different systems; here we use the free, open source system Sage (see http://www.sagemath.org). The script below

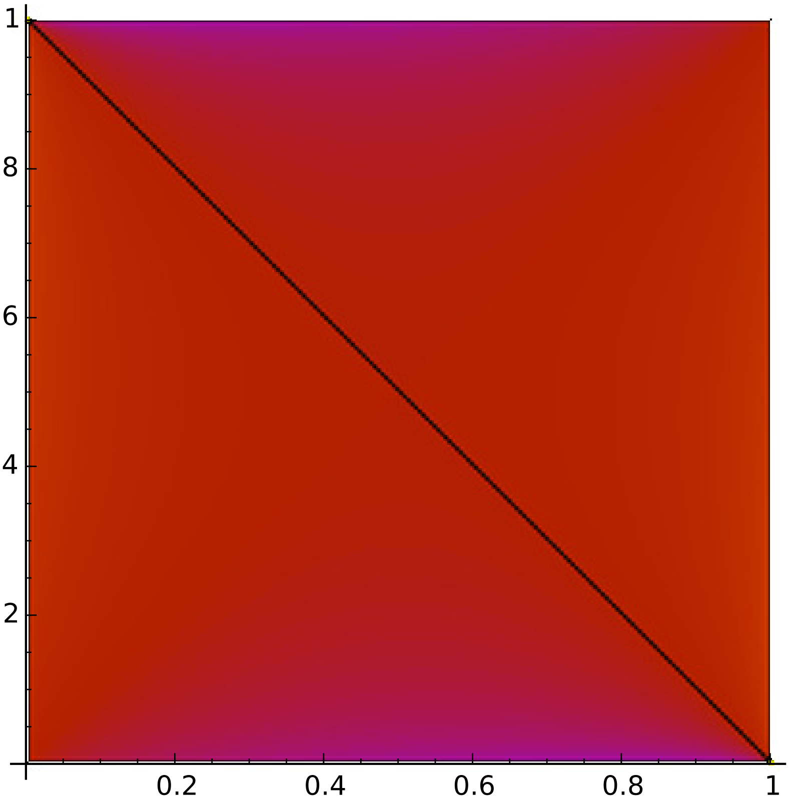

Figure 2: The optimal input frequency p_1

will produce the graph of the channel capacity, as well as the graphs of the optimal input frequencies.

```
# Graphs of the channel capacity and optimal input
    frequencies of the
# binary asymmetric channel.

def K(a,b):
    return(((1-b)*ln(1-b)+b*ln(b)-a*ln(a)-(1-a)*ln(1-a))
        / (1-a-b))

def C(a,b):
    if a+b == 1:
        return(0)
    k = K(a,b)
    return(RR((a*ln(a) + (1-a)*ln(1-a) - a*k + ln(1 +
        exp(k)))/ln(2)))

def P1(a,b):
```

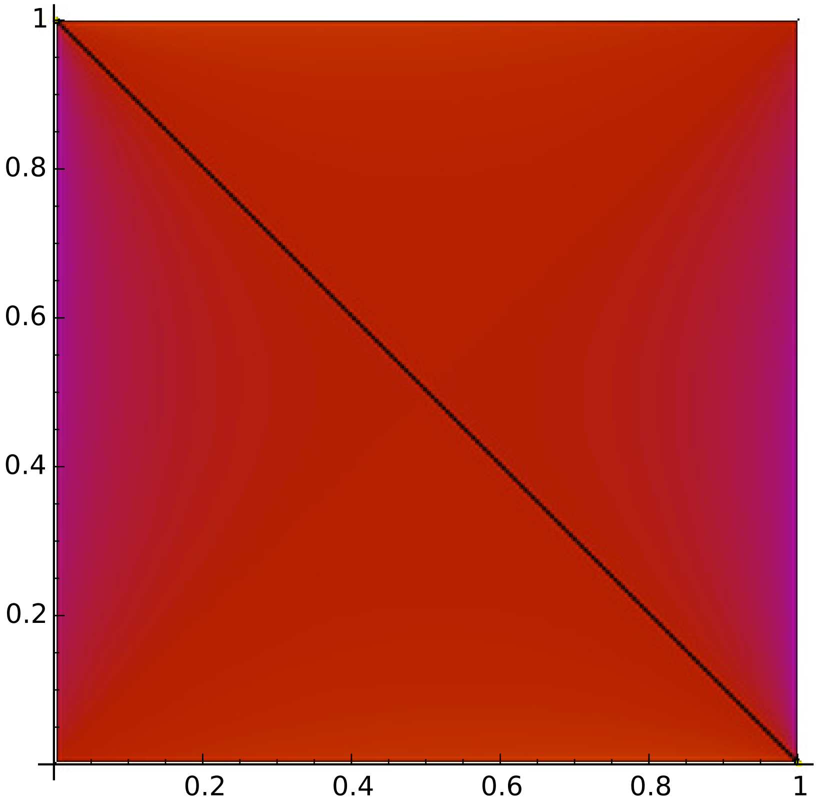

Figure 3: The optimal input frequency p_2

```
    # The initial code block avoids the ugliness near a
        + b = 1, where
    # the optimal frequencies are undefined.  Placing
        zeros along most
    # of this line creates a small black strip in the
        graph, and placing
    # ones at the corners we ensure that the graph uses
        the same color
    # scheme as the capacity graph.

    if abs(a+b-1) < 0.0001:
        if a==0 or a==1:
            return(1)
        return(0)
    k = K(a,b)
    return(RR((b*(exp(k) + 1) - 1)/((a+b-1)*(exp(k)+1))))

def P2(a,b):
    if abs(a+b-1) < 0.0001:
        if a==0 or a==1:
            return(1)
        return(0)
    k = K(a,b)
    return(RR((a*(exp(k)+1)-exp(k))/((a+b-1)*(exp(k)+1))))

c=density_plot(C, (0,1), (0,1), plot_points=200,
    cmap=’gnuplot’, aspect_ratio=1,
        figsize=6)
c.save(’c.eps’, aspect_ratio=1)

p1=density_plot(P1, (0,1), (0,1), plot_points=200,
    cmap=’gnuplot’, aspect_ratio=1,
        figsize=6)
p1.save(’p1.eps’, aspect_ratio=1)

p2=density_plot(P2, (0,1), (0,1), plot_points=200,
    cmap=’gnuplot’, aspect_ratio=1,
        figsize=6)
p2.save(’p2.eps’, aspect_ratio=1)
```

The graph of the channel capacity is shown in Figure 1. The parameter α is plotted on the horizontal axis, while β is on the vertical axis. Warmer colors indicate higher capacity; bright yellow corresponds to a capacity of 1 bit per transmission, visible at $\alpha = \beta = 1$ and also at $\alpha = \beta = 0$, while black corresponds to capacity zero, visible along the line $\alpha + \beta = 1$.

The graphs of the optimal frequencies are shown in Figures 2 and 3. The optimal frequencies are undefined along the line $\alpha + \beta = 1$. Note that the frequencies are rather insensitive to α and β, hovering near 1/2 except when one of these (but not the other) takes a really extreme value. This conforms to the intuition that one can’t really move information by sending, say, ones almost all the time—one must use a mixture of the two symbols to convey anything.