Preprint

Gilbert-Elliott channel & entropy rate of hidden Markov chains.

Yury Polyanskiy
(*Юрий Полянский*)

This preprint is about Gilbert-Elliott channel (GEC) w/o state knowledge. Most of the discussions however, reduce to questions regarding the binary hidden Markov chain, computation of its Blackwell measure and in particular its entropy rate.

Plan:

- Precise definitions of all processes involved.
- Overview of the convergence properties.
- α, ϕ and ψ-mixing coefficients.
- Markovity and proofs of convergence properties.
- Relation between T-operators on $[0,1]$ and $\mathbb{R}_+$.
- Properties of the K operator and the method of computation for the entropy rate.
- Iterations in projective space.
- Interpetation as a random walk in $PGL(2,\mathbb{R})$.
- Another method of computation of H.
- Attempt to prove othe expansion with $O(\log n)$ term.
- Proof of the expansion with $o(\sqrt{n})$ term.
- Bound for $|C_1 - C_0|$.
- Bound for $|V_0 - V_1|$.
- Entropy process in GEC may be not α-mixing.

1 GEC

1.1 Precise definitions

Let $\{S_j\}_{j=-\infty}^{\infty}$ be a homogeneous Markov process with states $\{1,2\}$, transition probabilities

$$\mathbb{P}[S_2 = 1|S_1 = 1] = \mathbb{P}[S_2 = 2|S_1 = 2] = 1 - \tau\,, \tag{1.1}$$
$$\mathbb{P}[S_2 = 2|S_1 = 1] = \mathbb{P}[S_2 = 1|S_1 = 2] = \tau\,, \tag{1.2}$$

and initial distribution

$$\mathbb{P}[S_1 = 1] = \mathbb{P}[S_1 = 2] = 1/2\,. \tag{1.3}$$

A quantity governing the dependence of the process S_j is given by

$$\mu = 1 - 2\tau\,.$$

Now for $\delta_1, \delta_2 \in [0,1]$ we define $\{Z_j\}_{j=-\infty}^{\infty}$ as conditionally independent given S_1^∞ and

$$\mathbb{P}[Z_j = 0|S_j = s] \;=\; 1 - \delta_s\,, \tag{1.4}$$
$$\mathbb{P}[Z_j = 1|S_j = s] \;=\; \delta_s\,. \tag{1.5}$$

NOTE: for the GEC w/o state known we always assume $1/2 \geq \delta_1 \geq \delta_2$.

The GEC channel acts on a binary vector X^n via binary addition:

$$Y^n = X^n + Z^n\,.$$

The capacity of the GEC channel is

$$C = \log 2 - H\,,$$

where H is the entropy rate of the process Z_j and is given by

$$H \;=\; \lim_{n\to\infty} \frac{1}{n} H(Z^n) \tag{1.6}$$
$$\;=\; \lim_{n\to\infty} \mathbb{E}\, h(\mathbb{P}[Z_n = 1|Z_1^{n-1}) \tag{1.7}$$
$$\;=\; \mathbb{E}\, h(\mathbb{P}[Z_0 = 1|Z_{-\infty}^{-1}])\,. \tag{1.8}$$

For convenience we define the following operation on $\mathbb{R} \times \{0,1\}$:

$$a^{\{z\}} = \begin{cases} a\,, & z = 1\,, \\ 1 - a\,, & z = 0 \end{cases}$$

We now define several related processes (the idea behind notation: a random variable with index j must be measurable with respect to $\sigma\{S^j, Z^j\}$):

$$\begin{aligned}
R_j &= \mathbb{P}[S_{j+1} = 1 | Z_1^j]\,, && (1.9)\\
Q_j &= \mathbb{P}[Z_{j+1} = 1 | Z_1^j] = \delta_1 R_j + \delta_2(1 - R_j)\,, && (1.10)\\
G_j &= -\log P_{Z_j|Z_1^{j-1}}(Z_j|Z_1^{j-1}) = -\log Q_{j-1}^{\{Z_j\}}\,, && (1.11)\\
R_j^* &= \mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0]\,, && (1.12)\\
Q_j^* &= \mathbb{P}[Z_{j+1} = 1 | Z_1^j, S_0] = \delta_1 R_j^* + \delta_2(1 - R_j^*)\,, && (1.13)\\
G_j^* &= -\log P_{Z_j|Z_1^{j-1},S_0}(Z_j|Z_1^{j-1}, S_0) = -\log {Q_{j-1}^*}^{\{Z_j\}}\,, && (1.14)\\
P_j &= \mathbb{P}[S_{j+1} = 1 | Z_{-\infty}^j]\,, && (1.15)\\
U_j &= \mathbb{P}[Z_{j+1} = 1 | Z_{-\infty}^j] = \delta_1 P_j + \delta_2(1 - P_j)\,, && (1.16)\\
F_j &= -\log P_{Z_j|Z_{-\infty}^{j-1}}(Z_j|Z_{-\infty}^{j-1}) = -\log U_j^{\{Z_j\}}\,. && (1.17)
\end{aligned}$$

Note that

- P_j, U_j and F_j are the only stationary processes (unless $\tau = 1/2$).
- All processes are bounded: for G_j and F_j this follows from (1.70).

1.2 Convergence properties

In Section 1.4 we show the following:

$$\left|\mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0 = 1] - \mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0 = 2]\right| \le \frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1}\,. \qquad (1.18)$$

From this and

$$\begin{aligned}
R_j &= \mathbb{P}[S_{j+1} = 1 | Z_1^j] && (1.19)\\
&= \sum_{a\in\{1,2\}} \mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0 = a]\,\mathbb{P}[S_0 = a | Z_1^j] && (1.20)
\end{aligned}$$

we find that[1]

$$\left|R_j - R_j^*\right| \leq \frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1} \qquad \text{a.s.}\,, \tag{1.21}$$

$$\left|Q_j - Q_j^*\right| \leq (\delta_1-\delta_2)\frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1} \qquad \text{a.s.}\,, \tag{1.22}$$

$$\left|R_j - P_j\right| \leq \frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1} \qquad \text{a.s.}\,, \tag{1.23}$$

$$\left|Q_j - U_j\right| \leq (\delta_1-\delta_2)\frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1} \qquad \text{a.s.}\,. \tag{1.24}$$

These bounds imply:

$$|R_j^* - R_j| \to 0 \quad \text{a.s.} \tag{1.25}$$

$$|P_j - R_j| \to 0 \quad \text{a.s.} \tag{1.26}$$

$$|Q_j^* - Q_j| \to 0 \quad \text{a.s.} \tag{1.27}$$

$$|U_j - Q_j| \to 0 \quad \text{a.s.} \tag{1.28}$$

For example, we show (1.23) as follows

$$|R_j - P_j| = \left|\sum_{a\in\{1,2\}} \mathbb{P}[S_{j+1}=1|Z_1^j, S_0=a](\mathbb{P}[S_0=a|Z_1^j] - \mathbb{P}[S_0=a|Z_{-\infty}^j])\right| \tag{1.29}$$

$$\leq \max_a \mathbb{P}[S_{j+1}=1|Z_1^j, S_0=a] - \min_a \mathbb{P}[S_{j+1}=1|Z_1^j, S_0=a] \tag{1.30}$$

$$\leq \frac{1}{2}\left|\ln\frac{\tau}{1-\tau}\right| |\mu|^{j-1}\,. \tag{1.31}$$

Alternatively, a martingale argument shows the following:

$$R_j, R_j^* \xrightarrow{d} P_0\,, \tag{1.32}$$

$$Q_j, Q_j^* \xrightarrow{d} U_0\,. \tag{1.33}$$

Introduce also T – a standard shift operator:

$$S_j \circ T = S_{j+1}\,, \quad Z_j \circ T = Z_{j+1}\,.$$

Then it is easy to derive (using stationarity of S_j, Z_j where necessary) that:

$$Q_j = \mathbb{E}\left[Q_j^* \middle| Z_1^j\right]\,, \tag{1.34}$$

$$Q_{j-1}\circ T = \mathbb{E}\left[Q_j \middle| Z_2^j\right]\,, \tag{1.35}$$

$$Q_j^* = \mathbb{E}\left[Q_{j-1}^*\circ T \middle| Z_1^j, S_0\right]\,. \tag{1.36}$$

[1]Notice that since R_0, R_0^*, P_0 are all in $[\tau, 1-\tau]$, direct argument shows that $|\mu|^{j-1}$ can be replaced with $|\mu|^j$.

On the other hand, we have:

$$\begin{aligned} H(Z_n|Z_1^{n-1}) &= \mathbb{E}\left[G_n\right] = \mathbb{E}\left[h(Q_n)\right], && (1.37)\\ H(Z_n|Z_1^{n-1}, S_0) &= \mathbb{E}\left[G_n^*\right] = \mathbb{E}\left[h(Q_n^*)\right], && (1.38)\\ H &= \mathbb{E}\left[F_0\right] = \mathbb{E}\left[h(U_0)\right], && (1.39) \end{aligned}$$

and so by using the Jensen's inequality and (1.33) we find that

$$\begin{aligned} \mathbb{E}\left[h(Q_1^*)\right] &\leq \mathbb{E}\left[h(Q_2^*)\right] \leq \cdots \leq \mathbb{E}\left[h(Q_n^*)\right] \nearrow H, && (1.40)\\ \mathbb{E}\left[h(Q_1)\right] &\geq \mathbb{E}\left[h(Q_2)\right] \geq \cdots \geq \mathbb{E}\left[h(Q_n)\right] \searrow H. && (1.41) \end{aligned}$$

1.3 α, ϕ and ψ-mixing coefficients of the process (S_j, Z_j).

Define σ-algebras

$$\mathcal{F}_b^a = \sigma\{S_b^a, Z_b^a\}.$$

Then the α, ϕ and ψ-mixing coefficients for this filtration are defined as

$$\begin{aligned} \alpha(n) &= \sup\left|\mathbb{P}[A,B] - \mathbb{P}[A]\mathbb{P}[B]\right|, && (1.42)\\ \phi(n) &= \sup\frac{|\mathbb{P}[A,B] - \mathbb{P}[A]\mathbb{P}[B]|}{\mathbb{P}[A]}, && (1.43)\\ \psi(n) &= \sup\frac{|\mathbb{P}[A,B] - \mathbb{P}[A]\mathbb{P}[B]|}{\mathbb{P}[A]\mathbb{P}[B]}, && (1.44) \end{aligned}$$

where the suprema are over $A \in \mathcal{F}_{-\infty}^0$, $B \in \mathcal{F}_n^\infty$ and $\mathbb{P}[A] > 0$, $\mathbb{P}[B] > 0$ where necessary.

Obviously, we have

$$\alpha(n) \leq \phi(n) \leq \psi(n). \qquad (1.45)$$

Here is a simple bound on $\psi(n)$. Notice that

$$\begin{aligned} \mathbb{P}[B|A] &= \mathbb{P}[B, S_n = 1|A] + \mathbb{P}[B, S_n = 2|A] && (1.46)\\ &= \mathbb{P}[B|S_n = 1]\mathbb{P}[S_n = 1|A] + \mathbb{P}[B|S_n = 2]\mathbb{P}[S_n = 2|A]. && (1.47) \end{aligned}$$

Now because S_j is such a simple Markov process, we can easily show that we have for any $a, b \in \{1, 2\}$

$$\mathbb{P}[S_n = a|S_0 = b] = \frac{1}{2} \pm \frac{1}{2}|\mu|^n, \qquad (1.48)$$

and hence

$$\begin{aligned} \left|\mathbb{P}[S_n = 1|A] - \frac{1}{2}\right| &= \left|\sum_{a\in\{1,2\}} \mathbb{P}[S_n = 1|S_0 = a]\mathbb{P}[S_0 = a|A] - \frac{1}{2}\right| && (1.49)\\ &\leq \sum_{a\in\{1,2\}} \left|\mathbb{P}[S_n = 1|S_0 = a] - \frac{1}{2}\right| \mathbb{P}[S_0 = a|A] && (1.50)\\ &\leq \frac{1}{2}|\mu|^n. && (1.51) \end{aligned}$$

Also notice that

$$\mathbb{P}[B] = \frac{1}{2}\mathbb{P}[B|S_n = 1] + \frac{1}{2}\mathbb{P}[B|S_n = 2]\,. \tag{1.52}$$

From (1.47) and (1.52) we derive

$$\begin{aligned}
\frac{|\mathbb{P}[AB] - \mathbb{P}[A]\mathbb{P}[B]|}{\mathbb{P}[A]\mathbb{P}[B]} &= \frac{|\mathbb{P}[B|A] - \mathbb{P}[B]|}{\mathbb{P}[B]} && (1.53)\\
&= \frac{\left|\sum_{a\in\{1,2\}} \mathbb{P}[B|S_n = a]\left(\mathbb{P}[S_n = a|A] - \frac{1}{2}\right)\right|}{\frac{1}{2}\sum_{a\in\{1,2\}} \mathbb{P}[B|S_n = a]} && (1.54)\\
&\leq |\mu|^n\,. && (1.55)
\end{aligned}$$

From (1.45) we conclude

$$\alpha(n) \leq \phi(n) \leq \psi(n) \leq |\mu|^n\,. \tag{1.56}$$

1.4 Markovity of R_j, R^*_j and P_j

First notice that R_j and Q_j, R^*_j and Q^*_j, and P_j and U_j determine each other (because $\delta_1 > \delta_2$). Since almost surely we have $Q_{j-1}, Q^*_{j-1}, U_{j-1} < 1/2$ we have that (Q_{j-1}, Z_j) and G_j, (Q^*_{j-1}, Z_j) and G^*_j, and (U_{j-1}, Z_j) and F_j determine each other.

Define two "R-propagation" functions $T_{0,1} : [0,1] \mapsto [\tau, 1-\tau]$:

$$\begin{aligned}
T_0(x) &= \frac{x(1-\tau)(1-\delta_1) + (1-x)\tau(1-\delta_2)}{x(1-\delta_1) + (1-x)(1-\delta_2)}\,, && (1.57)\\
T_1(x) &= \frac{x(1-\tau)\delta_1 + (1-x)\tau\delta_2}{x\delta_1 + (1-x)\delta_2}\,. && (1.58)
\end{aligned}$$

Note that $T_{0,1}$ are fractional-linear (Möbius) transformations, when viewed on complex plane.

Having defined these operators it is very easy to see that:

$$\begin{aligned}
R_{j+1} &= T_{Z_{j+1}}(R_j)\,, j \geq 0 && (1.59)\\
R^*_{j+1} &= T_{Z_{j+1}}(R^*_j)\,, j \geq -1 && (1.60)\\
P_{j+1} &= T_{Z_{j+1}}(P_j)\,, j \in \mathbb{Z}\,, && (1.61)
\end{aligned}$$

where for we start R_j and R^*_j as follows:

$$\begin{aligned}
R_0 &= 1/2\,, && (1.62)\\
R^*_{-1} &= 1\{S_0 = 1\}\,, && (1.63)\\
R^*_0 &= T_{0,1}(R^*_{-1}) = T_{Z_0}(R^*_{-1}) = (1-\tau)1\{S_0 = 1\} + \tau 1\{S_0 = 2\}\,. && (1.64)
\end{aligned}$$

Therefore, it can be shown that all of these processes are Markov and their transition kernel is given by

$$K(x,\cdot) = (\delta_1 x + \delta_2(1-x))\Delta_{T_1(x)}(\cdot) + ((1-\delta_1)x + (1-\delta_2)(1-x))\Delta_{T_0(x)}(\cdot)\,,$$

where

$$\Delta_{x_0}(A) = 1_A(x_0)$$

is a Dirac measure sitting at x_0.

Note that P_j is a stationary process, whose marginal distribution is equal to the unique stationary distribution of the kernel K. Unfortunatelly, this distribution is very hard to describe (in some cases it is concentrated on the set of measure 0).

So, processes R_j and R_j^* can be described as:

$$\begin{aligned} R_0 &\sim \Delta_{1/2}\,, && (1.65)\\ R_j &\sim K^j(1/2,\cdot)\,, && (1.66)\\ R_0^* &\sim \frac{1}{2}\Delta_\tau + \frac{1}{2}\Delta_{1-\tau}\,, && (1.67)\\ R_j^* &\sim \frac{1}{2}K^j(\tau,\cdot) + \frac{1}{2}K^j(1-\tau,\cdot)\,. && (1.68)\end{aligned}$$

Also notice that because of

$$\min(\tau, 1-\tau) \le T_0, T_1 \le \max(\tau, 1-\tau)$$

we have

$$\min(\tau, 1-\tau) \le R_j, R_j^*, P_j \le \max(\tau, 1-\tau) \tag{1.69}$$

and therefore also

$$\min(\delta_1\tau + \delta_2\bar\tau, \delta_1\bar\tau + \delta_2\tau) \le Q_j, Q_j^*, U_j \le \max(\delta_1\tau + \delta_2\bar\tau, \delta_1\bar\tau + \delta_2\tau)\,. \tag{1.70}$$

Finally, to prove (1.18) we first notice that

$$\begin{aligned} \mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0 = 1] &= T_{Z_j} \circ T_{Z_{j-1}} \cdot \circ T_{Z_1}(\mathbb{P}[S_1 = 1 | S_0 = 1])\,, && (1.71)\\ \mathbb{P}[S_{j+1} = 1 | Z_1^j, S_0 = 2] &= T_{Z_j} \circ T_{Z_{j-1}} \cdot \circ T_{Z_1}(\mathbb{P}[S_1 = 1 | S_0 = 2])\,. && (1.72)\end{aligned}$$

The (1.18) then follows from the following bound, valid for any $z \in \{0,1\}$ and $x, y \in (0,1)$:

$$|T_z(x) - T_z(y)| \le \frac{1}{4}|\mu| \left| \ln\frac{x}{1-x} - \ln\frac{y}{1-y} \right|\,. \tag{1.73}$$

We could directly prove (1.73) by establishing two facts:

1. There is a distance on $(0,1)$ defined by

$$d(x,y) = \left| \ln \frac{x}{1-x} - \ln \frac{y}{1-y} \right|$$

and T_z is μ-contracting in this distance.

2. The distance $d(x,y)$ and the usual $|x-y|$ are related via

$$|x-y| \leq \frac{1}{4} d(x,y)\,.$$

However, a more instructive proof follows if we discuss the relation with projective geometry. To treat T_0 and T_1 in one swoop, I will replace δ_1, δ_2 or $\bar{\delta}_1, \bar{\delta}_2$ with b_1, b_2 in the definition of T_z.

First notice that the operator T_z can be thought of as acting on a two-dimensional probability vector $[x, 1-x]'$:

$$\begin{pmatrix} x \\ 1-x \end{pmatrix} \mapsto \frac{1}{b_1 x + b_2(1-x)} \begin{pmatrix} b_1 \bar{\tau} x + b_2 \tau (1-x) \\ b_1 \tau x + b_2 \bar{\tau}(1-x) \end{pmatrix}\,.$$

Of course, this operator is not linear. However, we might notice that the denominator only serves the purpose of renormalizing the new vector so that it lies on the "probability simplex". But notice that the principal information that we need is only contained in the "direction" of the vector $\begin{pmatrix} x \\ 1-x \end{pmatrix}$. In mathematical terms this means that we must understand our vector as an element of the projective line $\mathbb{R}P^1$, which can be conveniently represented as a quotient of $\mathbb{R}^2$. Under this identification the positive quadrant $\mathbb{R}^2_+$ is identified with "positive" ray of $\mathbb{R}P^1$ which we will denote as $\mathbb{R}P^1_+ = [0, +\infty]$.

Having made this identification we may write that operator T_z works on the projective line element $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ as follows:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} b_1 \bar{\tau} & b_2 \tau \\ b_1 \tau & b_2 \bar{\tau} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \triangleq A_z \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\,.$$

In this way, we see that T_z might be thought as a linear transformation of $\mathbb{R}^2$ and therefore a projective transformation of $\mathbb{R}P^1$. Moreover, because all coefficients in the matrix are non-negative, this operator T_z maps positive-quadrant $\mathbb{R}^2_+$ into itself. Such transformation is notable because it contracts projective distance. The projective distance between the elements of $\mathbb{R}^2_+$ is defined as[1]

$$d_P(x,y) = \left| \ln \frac{x_1}{x_2} - \ln \frac{y_1}{y_2} \right| .$$

Note that this is a natural distance for a projective geometry because it is actually a log of the cross-ratio:

$$d_P(x,y) = |\ln(x,y;0,\infty)| = \ln \max[(x,y;0,\infty),(y,x;0,\infty)] .$$

Since cross-ratio is an invariant of a projective space (preserved under projective transformation), $d_P(x,y)$ is preserved under the stabilizer of the $[0,\infty]$ in the projective group $PGL(2,\mathbb{R})$; equivalently, this stabilizer is a subgroup of $GL(2,\mathbb{R})$ that maps $\mathbb{R}^2_+$ onto $\mathbb{R}^2_+$ (i.e. all non-negative $2x2$ matrices).

This distance is also natural in relation to hyperbolic geometry: if you represent rays by the points at which they intersect the line $x = 1$, then the (geodesic) distance between those points in the Poincare half-plane model is exactly $|\ln y_1 - \ln y_2|$.

If, however, T_z maps $\mathbb{R}^2_+$ *into* $\mathbb{R}^2_+$ then, as shown in Section 2 of [3] the mapping T_z is contracting:

$$d_P(T_z(x),T_z(y)) \le \tau_B(T_z) d_P(x,y) , \tag{1.74}$$

where $\tau_B(T_z)$ is a "projective norm" which has many different expressions, but for the $\mathbb{R}P^1$ case is given by

$$\tau_B \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{cases} \dfrac{\sqrt{\frac{ad}{bc}}-1}{\sqrt{\frac{ad}{bc}}+1} , & ad > bc , \\ \dfrac{\sqrt{\frac{bc}{ad}}-1}{\sqrt{\frac{bc}{ad}}+1} , & ad < bc , \end{cases}$$

In general,

$$\tau_B(T) = \frac{e^{\Delta/2}-1}{e^{\Delta/2}+1} = \tanh(\Delta/4) ,$$

where Δ is the d_P-diameter of the image of $\mathbb{R}^n_+ \leftrightarrow \mathbb{R}P^{n-1}_+$. In the one-dimensional case $\Delta = |\ln \frac{ad}{bc}|$. In general, (see the book of Seneta "Non-negative matrices and Markov Chains" and the paper of Hopf referenced there)

$$\Delta(T) \;\; = \;\; \ln \max_{ijkl} \frac{t_{ij}}{t_{ik}} \frac{t_{lk}}{t_{lj}} . \tag{1.75}$$

[1] In general, for other dimension > 2 we have:

$$d_P(x,y) = \ln \frac{\max_i x_i/y_i}{\min_j x_j/y_j} = \max_{i,j} \ln \frac{x_i y_j}{x_j y_i} .$$

Anyway, for our T we have

$$\tau_B(T_z) = |1 - 2\tau| = |\mu| \,.$$

This and the Birkhoff bound (1.74) imply

$$d_P(T_z(x), T_z(y)) \le |\mu| d_P(x, y) \,. \tag{1.76}$$

On the other hand, if we convert the projective line element $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ back to the probability vector:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1/(x_1 + x_2) \\ x_2/(x_1 + x_2) \end{pmatrix} \,,$$

then the projective distance can be written as

$$d_P(x, y) = d\left(\begin{pmatrix} x \\ 1 - x \end{pmatrix}, \begin{pmatrix} y \\ 1 - y \end{pmatrix}\right) = \left|\ln \frac{x}{1 - x} - \ln \frac{y}{1 - y}\right| \,.$$

Because the derivative of $\ln \frac{x}{1-x}$ is upper-bounded by 4 we have

$$|x - y| \le \frac{1}{4} d_P(x, y) \,. \tag{1.77}$$

Combining (1.76) and (1.77) we obtain (1.73).

Side note: It is always true (and easy to show) that for discrete distributions we have:

$$D(P||Q) \le d_P(P, Q) \,.$$

Therefore, all theorems on geometric-ergodicity that are derived using Birkhoff's contraction inequality also automatically imply a geoemtric convergence with respect to divergence! On the other hand, there is no corresponding lower-bound.

To see this we give a counter-example in the binary case. First notice that if $t = \frac{x}{1-x}$ and $\theta = \frac{y}{1-y}$ are projective-line elements corresponding to x and y then

$$d(x||y) = \frac{t}{1 + t} \ln \frac{t}{\theta} + \ln \frac{1 + \theta}{1 + t} \,.$$

Now if we set $t = 2\theta$ and send $\theta \to \infty$ we get:

$$d(x||y) \to 0 \,, \quad \text{but} \quad d_P(x, y) = \left|\ln \frac{t}{\theta}\right| = \ln 2 \not\to 0 \,.$$

1.5 Relation between T-operators on $[0,1]$ and $\mathbb{R}_+$

As explained above, we can think of operator T_b in two ways: first, as acting on $[0,1]$ and defined via (1.57), in which case T_b as an element of $PGL(2,\mathbb{R})$ is identified with

$$T_b \sim \begin{pmatrix} \bar{\tau}b_1 - \tau b_2 & \tau b_2 \\ b_1 - b_2 & b_2 \end{pmatrix}$$

On the other hand, T_b can be thought as acting on the the ratio $\frac{x}{1-x}$ and in this case it is an operator defined as

$$\tilde{T}_b(r) = \frac{\bar{\tau}b_1 r + \tau b_2}{\tau b_1 r + \bar{\tau} b_2} \sim \begin{pmatrix} \bar{\tau}b_1 & \tau b_2 \\ \tau b_1 & \bar{\tau} b_2 \end{pmatrix}$$

Both of them are elements of $PGL(2,\mathbb{R})$ and the relation between these two is given by

$$\tilde{T}_b = A T_b A^{-1} \tag{1.78}$$
$$T_b = A^{-1} \tilde{T}_b A \tag{1.79}$$
$$A \sim \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \tag{1.80}$$
$$A^{-1} \sim \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{1.81}$$

Operator A can be thought, geometrically, as a remapping of projective-elements from the coordinates based on $x_1 + x_2 = 1$ to coordinates based on $x_2 = 1$:

$$x \overset{A}{\mapsto} \frac{x}{1-x} \tag{1.82}$$
$$1 \overset{A}{\mapsto} \infty \tag{1.83}$$
$$\infty \overset{A}{\mapsto} -1 \tag{1.84}$$

Figure 1.1 gives an idea about both operators.

Finally, we can think of T_b as acting on $\mathbb{R}^2$ restricted to the simplex $x + y = 1$. In this case, it is easy to show that T_b is just a restriction of the following operator in $PSL(2,\mathbb{C})$:

$$L^{-1} T_b L = \begin{pmatrix} 1 & i \\ 1 & 1 \end{pmatrix} \circ \begin{pmatrix} \bar{\tau}b_1 & \tau b_2 \\ \tau b_1 & \bar{\tau} b_2 \end{pmatrix} \circ \begin{pmatrix} 1 & -i \\ -1 & 1 \end{pmatrix}$$

Here L is the transform that maps $\mathbb{R}^2_+ \cap \{x + y = 1\}$ to $\mathbb{R}_+$.

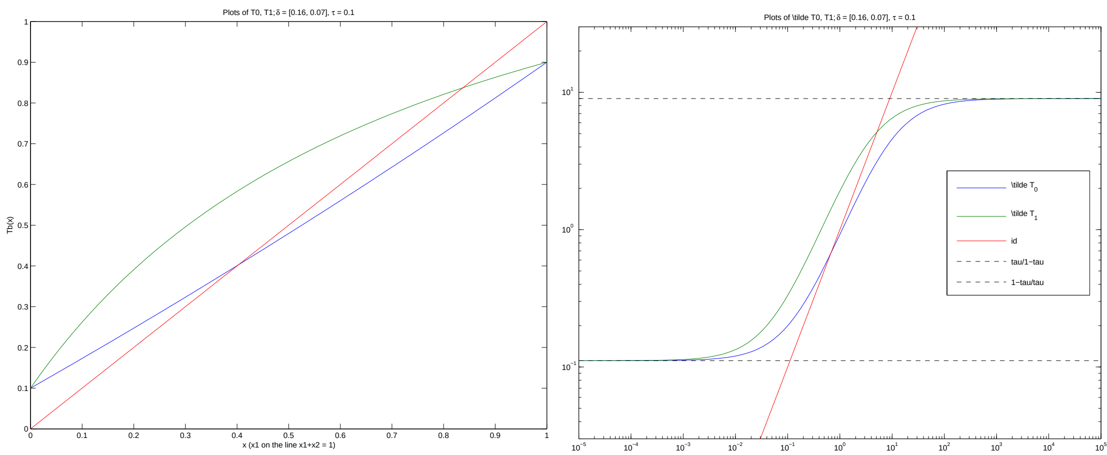

Figure 1.1: Operators T_0 and T_1 as acting on $[0,1]$ (left) and $\mathbb{R}_+$ (right)

1.6 Properties of K and method to compute H

K can be viewed as an operator acting on either measures ν or (by duality) as acting on functions f:

$$\nu K(\cdot) = \int_{[0,1]} \nu(dx) K(x, \cdot)\,, \tag{1.85}$$

$$Kf(x) = \int K(x, dy) f(y) \tag{1.86}$$

$$= (\delta_1 x + \delta_2(1-x)) f(T_1(x)) + ((1-\delta_1)x + (1-\delta_2)(1-x)) f(T_0(x)) \tag{1.87}$$

As an operator that acts on functions (defined on $[\tau, 1-\tau]$) it has the following properties:

- If f is monotone then Kf is monotone;
- If f is convex (concave) then Kf is convex (concave);
- $K^n f \to \mathbb{E}\left[f(P_0)\right]$.
- If f is $\cap$-concave then

$$\mathbb{E}\left[f(R_n)\right] = K^n f(1/2) \searrow \mathbb{E}\left[f(P_0)\right].$$

- If f is $\cap$-concave then

$$\mathbb{E}\left[f(R_n^*)\right] = \frac{1}{2} K^n f(\tau) + \frac{1}{2} K^n f(1-\tau) \nearrow \mathbb{E}\left[f(P_0)\right].$$

- Note that Kf only depends on the values of f on $[\tau, 1-\tau]$. But among all f continuous on $[\tau, 1-\tau]$ s.t. $\operatorname{supp} f \subseteq [\tau, 1-\tau]$ we have

$$Kf = 0 \Longrightarrow f = 0\,.$$

Proof outline: Show that $f(\tau) = f(1-\tau) = 0$. Then assume that there is some open interval U s.t. $f(U) > \epsilon$. Then $\exists x_n \to (1-\tau)$ s.t. $f(x_n) > \epsilon$, which contradicts continuity of f on $[\tau, 1-\tau]$.

To compute H notice that by (1.40) and (1.41) it is sufficient to compute $\mathbb{E}\left[h(Q_n^*)\right]$ and $\mathbb{E}\left[h(Q_n)\right]$. But

$$\mathbb{E}\left[h(Q_n^*)\right] = \mathbb{E}\left[h(\delta_1 R_n^* + \delta_2(1-R_n^*))\right] = P_{R_0^*} K^n \tilde{h}\,,$$

where

$$\tilde{h}(x) = h(\delta_1 x + \delta_2(1-x))$$

is a monotonically increasing (since $\delta_1, \delta_2 \le 1/2$), $\cap$-concave function. So the problem boils down to computing measures

$$\nu_n^* = \frac{1}{2}\Delta_\tau K^n + \frac{1}{2}\Delta_{1-\tau} K^n\,.$$

This can not be easily done because n-th measure is composed of 2^n Dirac measures and therefore requires too much memory to store. The obvious solution is to quantize the interval $[\tau, 1-\tau]$. This approach was proposed by Mushkin and Bar-David. It was criticized for not providing an accuracy estimate. However, notice that because of the monotonicity of $\tilde{h}$ we can "round-down" the location of each Δ-measure comprising ν_n^* towards the closest quantization level and still have a lower bound on $\mathbb{E}\left[h(Q_n^*)\right.$, Call the new (quantized) measure $\hat{\nu}_n^*$:

$$\hat{\nu}_n^* \tilde{h} \le \nu_n^* \tilde{h} \stackrel{\triangle}{=} \mathbb{E}\left[h(Q_n^*)\right].$$

Now because of the monotonicity of K this measure $\hat{\nu}_n^*$ can also be used insteado of the true ν_n^* as a source for computing ν_{n+1}^*. In other words, for any monotonically increasing f we have

$$(\hat{\nu}_n^* K) f \le (\nu_n^* K) f\,.$$

Therefore, such "rounding-down" at each stage leads to a lower bound to $\mathbb{E}\left[h(Q_n^*)\right]$ on each stage. Similar rounding-up can be used in computing $\mathbb{E}\left[h(Q_n)\right]$. The pair of these bounds is guaranteed to sandwich the true value of H:

$$\hat{\nu}_n^* \tilde{h} \le H \le \hat{\nu}_n \tilde{h}\,.$$

1.7 Definition of the Blackwell's measure

By construction we know that for any measurable and bounded f we have

$$\forall x_0 \in [0,1] : K^n f(x_0) \to \mathbb{E}\left[f(P_0)\right]. \tag{1.88}$$

Therefore, given the linearity of K we can define a measure μ (corresponding to P_0, of course) without a reference to P_0 as follows:

$$\mu f = \lim_{n\to\infty} K^n f(1/2).$$

It raises two questions: How to analytically (from the definition of K in (1.87)) prove that $\lim K^n f$ exists and is a constant? We know that when $f(x) = ax + b$ then $\mu f = a/2 + b$ (because $\mathbb{E}\left[P_0\right] = 1/2$); what about an analytical proof of this?

Finally, the main challenge is to be able to integrate any other (non-linear) functions f over μ with the ultimate goal of integrating a binary entropy function (see (1.88) also).

Answer to question 1: Note that since every continuous function on $[0,1]$ is uniformly continuous for any ϵ we can find a δ such that fluctuation of f on any δ-interval is below ϵ. Then according to representation (1.108) we see that for n large enough every operator $T_{b_1}\circ\cdots\circ T_{b_n}$ will map $[0,1]$ into an interval smaller than δ (because of Birkhoff contraction). Hence,

$$|f\circ T_{b_1}\circ\cdots\circ T_{b_n}(x) - f\circ T_{b_1}\circ\cdots\circ T_{b_n}(y)| \le \epsilon\,, \quad \forall x,y\in[0,1] \quad \forall b^n \in \{0,1\}^n\,.$$

Choosing $y = 1/2$ we get that $K^n f(x) \to K^n f(1/2)$. Note this is proof essentially depends on continuity of f (that is there is only a weak convergence of $\nu K^n \to \mu_{Blackwell}$).

Answer to question 2: Note that among linear polynomials $ax+b$ only $b = -a/2$ are eigenvalues of K:

$$K(ax+b) = \lambda(ax+b) \iff \lambda = 1-2\tau\,,\; b = -a/2\,.$$

Therefore,

$$K^n(x-1/2) = (1-2\tau)^n(x-1/2) \to 0$$

and therefore

$$K^n x \to K^n 1/2 = 1/2\,.$$

Here is another nice property of operator K:

$$\begin{aligned} Kf &= S(f\circ T_1)\,, && (1.89)\\ Sg &= p_1(x)g(x) + p_0(x) g\circ U\,, && (1.90)\\ U(x) &= T_1^{-1}\circ T_0 = \frac{\bar\delta_1\delta_2 x}{(\delta_2-\delta_1)x + \delta_1\bar\delta_2} \sim \begin{pmatrix} \bar\delta_1\delta_2 & 0\\ (\delta_2-\delta_1) & \delta_1\bar\delta_2\end{pmatrix}\,. && (1.91)\end{aligned}$$

Fantastically, U and S do not depend on τ!!!

Note: searching for an eigenvalue $f = \frac{ax+b}{cx+d}$ as in

$$Kf = uf + v(x - 1/2) \tag{1.92}$$

leads to $c = 0$, i.e. $f = ax + b$. I have got the same results for $f = \frac{ax^2+b}{cx+d}$, $f = ax^2 + bx + c$, $f = \frac{x^2+ax+b}{x^2+cx+d}$ and $f = \frac{x^3+ax+b}{x^3+cx+d}$. Some approach alternative to (1.92) is needed.

1.8 Easy case: $\tau = 0$

In the case $\tau = 0$ the operator K becomes much simpler (it also arises in the problem of binary hypothesis testing between Bernoulli(δ_1) and Bernoulli(δ_2)). We denote this operator by K_0:

$$K_0 f = p_0 \cdot f \circ T_0' + p_1 \cdot T_1',$$

where

$$T_0'(x) = \frac{x(1-\delta_1)}{x(1-\delta_1) + (1-x)(1-\delta_2)}, \tag{1.93}$$

$$T_1'(x) = \frac{x\delta_1}{x\delta_1 + (1-x)\delta_2}. \tag{1.94}$$

In this case the operator K_0 satisfies many nice properties. For example, there are algebraic relations:

$$K_0 \left(\frac{x}{1-x}\right)^n = (-1)^{n-1}(\alpha_n - \alpha_{n+1}) \sum_{k=0}^{n-1} (-1)^k \left(\frac{x}{1-x}\right)^k + \alpha_{n+1} \left(\frac{x}{1-x}\right),$$

where we denote for simplicity $\left(\frac{x}{1-x}\right)^0 = x$ (i.e. identity mapping). Constants α_j are merely

$$\alpha_j = \mathbb{E}_{P_1} \left(\frac{dP_1}{dP_2}\right)^j = \delta_1 \left(\frac{\delta_1}{\delta_2}\right)^j + \bar{\delta}_1 \left(\frac{\bar{\delta}_1}{\bar{\delta}_2}\right)^j$$

Similar relation can be derived for $K_0 \left(\frac{1-x}{x}\right)^n$ with α_j replaced by β_j where $\mathbb{E}_{P_1}$ is replaced with $\mathbb{E}_{P_2}$.

This allows (in principle) to decompose interesting functions f in Laurant series over $\left(\frac{x}{1-x}\right)^j, j \in (-\infty, \infty)$.

Interpretation with binary hypothesis testing is as follows:

$$K_0^k f(x) = \mathbb{E}^x f(\pi_k),$$

where $\pi_n = \mathbb{P}[\theta = 1 | \mathcal{F}_n]$ and $\pi_0 = x$, a.s. (a Markov process of conditional probabilities). I.e. K_0 is a Markov kernel of the π_n process.

So from here, since we know $\pi_n \to 1\{\theta = 1\}$ we get

$$K_0^n f \to x f(0) + (1-x) f(1) \quad n \to \infty\,. \tag{1.95}$$

Also, we know that

$$\frac{\pi_n}{1-\pi_n} = \frac{dP_1}{dP_2}\bigg|_{\mathcal{F}_n} \frac{\pi_0}{1-\pi_0}\,.$$

This demontstrates that functions $f = g(\ln x)$ are particularly easy to evaluate for interations, because $K_0^n f$ is then subject to law of large numbers, Chernoff bound, CLT etc.

What analytical properties of K_0 make it so simple and treatable? There are two:

1. Notice that K_0 is isomorphic to a simpler operator K_0' defined as

$$K_0' f = p_0(1) f \circ T_0' + p_1(1) f \circ T_1'\,. \tag{1.96}$$

   Indeed, a very simple argument shows that

$$K_0' \frac{f}{x} = \frac{1}{x} K_0 f\,,$$

   and therefore,

$$(K_0')^n f = \frac{1}{x} K_0^n f\,.$$

2. Another great property of K_0 is that T_0' and T_1' commute:

$$T_0' \circ T_1' = T_1' \circ T_0'$$

   This is possible iff matrices defining T_0 and T_1 are simultaneously diagonalizable (i.e. they have same eigenvectors, or equivalently T_0 and T_1 have two common fixed points: 0 and 1).[1]

   The row-eigenvectors when put together in a matrix become $\begin{matrix} 1 & 0 \\ -1 & 1 \end{matrix}$ which corresponds to a function $\frac{x}{1-x}$. So the change of variables $\frac{y=x}{1-x}$ establishes isomorphism of K_0' and

$$K_0'' g(y) = p_0(1) g(a_1 y) + p_1(1) g(a_2 y)\,.$$

3. Finally, if you want to reduce K_0'' to a convolutional operator, then we only need to change $z = \log y$ and then we get

$$K_0''' h(z) = p_0(1) h(z + a_1') p_1(1) h(z + a_2') = h * Q\,,$$

[1] Question: if K' is a fractional linear operator of the form (1.96) and T_0' and T_1' are fractional-linear and have two common fixed points. Is it true that all K' are isomorphic to each other? To some K_0'? (i.e. the one with fixed points 0 and 1).

for some discrete distribution Q. Properties of powers of convolutional operators are well known (see Feller) and they are simply an equivalent reformulation of the sums of iid r.v.'s. Note that typically asymptotic properties of convolutional-type operators are naturally studied via Fourier transform. It is interesting therefore that in K_0 we have commuting T_0', T_1' and our main tool is the Fourier transform. Perhaps to study general K we need to apply non-commutative Fourier analysis or something like that.

1.9 Isomorphism of iterations

Suppose that X is a topological space and $X*$ is a space of continous functions (in fact we can restrict to other subclasses, such as algebraic, or rational etc). Then operator $K : X^* \to X^*$ defines an iterative system. Our purpose is to evaluate (if it exists) the limit

$$g(x) = \lim_n K^n f\,.$$

We say that (X^*, K) is isomorphic to (Y^*, L) if there exists an isomorphism $U : X^* \to Y^*$ such that

$$K = U^{-1} \circ L \circ U$$

or diagrammatically:

$$K : X^* \xrightarrow{U} Y^* \xrightarrow{L} Y^* \xrightarrow{U^{-1}} X^*\,.$$

For example, in the previous subsection we showed that $([0, 1]^*, K_0)$ is isomorphic to a convolutional operator $(C(\mathbb{R}), (\cdot) * Q)$.

Not all operators are isomoprhic:

- $K = 1$ and $K = 0$ are not isomorphic.
- $K = 1$ and $K = \Delta_a$ $(f \mapsto f(a))$ are not isomorphic.
- K-ergodic markov kernel (e.g. K above for $\tau > 0$) and $K = K_0$.

In all cases, the proof is simply by considering the dimensionality of the space of limits $\lim K^n f$ (e.g. for ergodic K it is 1 dimensional, for K_0 it is 2-dimensional (1.95)).

1.10 Iterations in projective space

First, notice the following series for the binary entropy

$$h(x) = \ln 2 - \sum_{k=1}^{\infty} \frac{(2x-1)^{2k}}{2k(2k-1)} \tag{1.97}$$

converging for $0 < x < 1$. Then it is sufficient to study $\lim_{n\to\infty} K^n f(x)$ for polynomial functions $f(x)$.

We now define an operation $\hat{F}$ which maps functions on $\mathbb{RP}^1$ to functions on $\mathbb{R}^1$. First, function $F : \mathbb{RP}^1 \mapsto \mathbb{RP}^1$ is defined as

$$[s_1 : s_2] \overset{F}{\mapsto} [f_1(s_1, s_2) : f_2(s_1, s_2)] .$$

Then we set

$$\hat{F}(x) \triangleq \frac{f_1(x, 1-x)}{f_1(x, 1-x) + f_2(x, 1-x)} .$$

Finally, if we define pseudo-addition and pseudo-multiplication over $\mathbb{RP}^1$ as

$$[a_1 : a_2] \oplus [b_1 : b_2] \quad = \quad [a_1(a_2 + b_2) + a_2(a_1 + b_1) : b_1 b_2 - a_1 a_2] , \qquad (1.98)$$

$$[a_1 : a_2] \odot [b_1 : b_2] = [a_1 b_1 : a_2 b_2] .$$

(these operations are rational functions on $\mathbb{RP}^2 \mapsto \mathbb{RP}^1$; they are not defined everywhere, e.g. $[1 : 0] \odot [0 : 1]$ is undefined). It is easy to check that

$$\hat{F} \circ \hat{G} \quad = \quad \widehat{F \circ G} , \qquad (1.99)$$

$$\hat{F} \cdot \hat{G} \quad = \quad \widehat{F \odot G} , \qquad (1.100)$$

$$\hat{F} + \hat{G} \quad = \quad \widehat{F \oplus G} . \qquad (1.101)$$

Finally, we can show that if $\hat{F} = f$ then $\hat{G} = Kf$, where

$$G = \mathcal{K}F = \Big[\Pi \odot (F \circ T_1)\Big] \oplus \Big[\bar{\Pi} \odot (F \circ T_0)\Big] , \qquad (1.102)$$

where all functions are $\mathbb{RP}^1 \mapsto \mathbb{RP}^1$ and are defined as follows:

$$\Pi \quad : \quad [s_1 : s_2] \mapsto [\delta_1 s_1 + \delta_2 s_2 : \bar{\delta}_1 s_1 + \bar{\delta}_2 s_2] , \qquad (1.103)$$

$$\bar{\Pi} \quad : \quad [s_1 : s_2] \mapsto [\bar{\delta}_1 s_1 + \bar{\delta}_2 s_2 : \delta_1 s_1 + \delta_2 s_2] , \qquad (1.104)$$

$$T_1 \quad : \quad [s_1 : s_2] \mapsto [\delta_1 \bar{\tau} s_1 + \delta_2 \tau s_2 : \delta_1 \tau s_1 + \delta_2 \bar{\tau} s_2] , \qquad (1.105)$$

$$T_0 \quad : \quad [s_1 : s_2] \mapsto [\bar{\delta}_1 \bar{\tau} s_1 + \bar{\delta}_2 \tau s_2 : \bar{\delta}_1 \tau s_1 + \bar{\delta}_2 \bar{\tau} s_2] , \qquad (1.106)$$

Equation (1.102) replaces main equation (1.87) in the projective-space language. It might be more convenient since if treated as iteration for (homogeneous) functions $\mathbb{R}^2 \mapsto \mathbb{R}^2$ it maps polynomial F into polynomial G. So the main question is how to find $\lim \mathcal{K}^n F$ (NB: we know it must converge to a constant function).

1.11 Iterations on $\tau - x$ plane

We can also view iterations (1.87) as being done one the $\tau - x$ plane (or even $\mathbb{C}^2$?). Indeed we start with

$$F_0(x, \tau) = f(x) ,$$

for some polynomial $f(x)$ for which we want to compute $\mathbb{E}\left[f(P_0)\right]$. Then

$$F_{n+1}(x,\tau) = (\delta_1 x + \delta_2 \bar{x}) F_n\left(\frac{x\bar{\tau}\delta_1 + \bar{x}\tau\delta_2}{x\delta_1 + \bar{x}\delta_2}, \tau\right) + (\bar{\delta}_1 x + \bar{\delta}_2 \bar{x}) F_n\left(\frac{x\bar{\tau}\bar{\delta}_1 + \bar{x}\tau\bar{\delta}_2}{x\bar{\delta}_1 + \bar{x}\bar{\delta}_2}, \tau\right) \quad (1.107)$$

We know that $F_n(\cdot,\tau)$ must converge to a constant on $[0,1]$. Each $F_n(\cdot,\tau)$ has $2^{n+1}-1$ poles in x. This is because

$$F_n(x,\tau) = \sum_{b^n \in \{0,1\}^n} f \circ T_{b_1} \circ \cdots \circ T_{b_n} \cdot p_{b_n} \cdot p_{b_{n-1}} \circ T_{b_n} \cdots p_{b_1} \circ \cdots \circ T_{b_n}\,, \qquad (1.108)$$

where $p_0(x) = \bar{\delta}_1 x + \bar{\delta}_2 \bar{x}$, $p_1(x) = \delta_1 x + \delta_2 \bar{x}$.

Idea: can we establish some (integral? differential? combinatorial?) invariant that is preserved across the iterations for $F_n(x,\tau)$ and then use the fact that $F_n(x,1/2)$ is easily computable. E.g.,

$$\frac{\partial}{\partial\tau} F_n(x,\tau) = (\delta_2\bar{x} - \delta_1 x)\left(\frac{\partial F_{n-1}}{\partial x}\right)\left(\frac{x\bar{\tau}\delta_1 + \bar{x}\tau\delta_2}{x\delta_1 + \bar{x}\delta_2}, \tau\right) + (\bar{\delta}_2\bar{x} - \bar{\delta}_1 x)\left(\frac{\partial F_{n-1}}{\partial x}\right)\left(\frac{x\bar{\tau}\bar{\delta}_1 + \bar{x}\tau\bar{\delta}_2}{x\bar{\delta}_1 + \bar{x}\bar{\delta}_2}, \tau\right)$$

Another idea: look for functions F which are preserved by the iteration (1.107).

Finally, the curve $u = T_1(x), v = T_0(x)$ is an affine algebraic variety generated by

$$[(\bar{\tau} - u)\delta_1 + (u - \tau)\delta_2](v - \tau)\bar{\delta}_2 = (u - \tau)\delta_2[(\bar{\tau} - v)\bar{\delta}_1 + (v - \tau)\bar{\delta}_2]\,,$$

For some weird reason the $\delta_2 = 0$ case is special, for it the variety is defined as

$$v - \tau = 0\,.$$

1.12 Random walk on $PGL(2,\mathbb{R})$

The problem at hand can be interpreted as a random walk on $PGL(2,\mathbb{R})$. Consider the faithful (and sharply 2-transitive?) action of $PGL(2,\mathbb{R})$ on $\mathbb{RP}^1$ given as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} [x : y] = [ax + by : cx + dy]\,.$$

Consider the closed subset (a positive ray of $\mathbb{RP}^1$)

$$\mathbb{RP}^1_+ \triangleq \{[x : y] | x \geq 0, y \geq 0\} \subset \mathbb{RP}^1\,,$$

which is homeomorphic to $[0,1]$ under the map

$$\theta : [x : y] \mapsto \frac{x}{x + y}\,.$$

The closed subgroup of $PGL(2,\mathbb{R})$ fixing $\mathbb{RP}^1_+$ will be denoted $PGL_+(2,\mathbb{R})$:

$$PGL_+(2,\mathbb{R}) \stackrel{\triangle}{=} \{g \in PGL(2,\mathbb{R}) : g(\mathbb{RP}^1_+) \subset \mathbb{RP}^1_+\}$$

To define the random walk we take two elements $T_0, T_1 \in PGL_+(2,\mathbb{R})$ and a function $\pi : \mathbb{RP}^1_+ \to [0,1]$. The random walk is defined as follows:

$$g_0 = 1\,, \quad g_{t+1} = \begin{cases} T_1 \circ g_t\,, & \text{w.p. } \pi(g_t([1:1]))\,, \\ T_0 \circ g_t\,, & \text{w.p. } 1 - \pi(g_t([1:1]))\,. \end{cases}$$

The original problem:

$$T_0 = \begin{pmatrix} (1-\tau)(1-\delta_1) & \tau(1-\delta_2) \\ \tau(1-\delta_1) & (1-\tau)(1-\delta_2) \end{pmatrix}\,, \qquad T_1 = \begin{pmatrix} (1-\tau)\delta_1 & \tau\delta_2 \\ \tau\delta_1 & (1-\tau)\delta_2 \end{pmatrix}\,,$$

where $\delta_1, \delta_2, \tau \in [0,1]$ are some parameters and

$$\pi([x:y]) = \frac{\delta_1 x + \delta_2 y}{x+y}\,.$$

Note that in the interesting case $\tau < 1/2$ we have $T_0, T_1 \in PSL_+(2,\mathbb{R})$ and thus the random walk takes place in the $PSL_+(2,\mathbb{R})$ – this is safe to assume[1].

Let us denote by μ_t the measure on $PGL_+(2,\mathbb{R})$ corresponding to the distribution of random element g_t. Notice that as shown above we have for any probability measure ν on $\mathbb{RP}^1_+$:

$$a_*(\mu_t \times \nu) \to \nu_{Blackwell}\,, \quad t \to \infty \tag{1.109}$$

where $\nu_{Blackwell}$ is a certain (so called Blackwell) probability measure on $\mathbb{RP}^1_+$ and a_* corresponds to the map a defining the action:[2]

$$a : PGL(2,\mathbb{R}) \times \mathbb{RP}^1 \to \mathbb{RP}$$

To connect with the random processes defined above:

$$\begin{aligned} R_j &= \theta(g_j([1:1]))\,, && (1.110) \\ Q_j &= \pi(g_j([1:1]))\,, && (1.111) \\ P_0 &\stackrel{d}{=} \theta_*(\nu_{Blackwell}), && (1.112) \\ U_0 &\stackrel{d}{=} \pi_*(\nu_{Blackwell}). && (1.113) \end{aligned}$$

[1]Notice that above we frequently take $\delta_1 = \frac{1}{2}$ and $\delta_2 = 0$ to focus ideas. In this case this is not permitted since then operator $T_1 \notin PGL$ and we get a random walk on a *monoid of fractional linear functions* $\mathbb{RP}^1 \to \mathbb{RP}^1$ instead of the PGL (which is a group of invertible elements of that monoid).

[2]Corresponding to $\pi : \mathbb{RP}^1 \to \mathbb{R}$ we denote π_* the pushforward of measures on $\mathbb{RP}^1$ to measures on $\mathbb{R}$, etc.

The main goal: Learn how to integrate functions with respect to $\nu_{Blackwell}$. According to (1.109) we have

$$\int f d\nu_{Blackwell} = \lim_{t\to\infty} f(g_t([x_0 : y_0]))\,,$$

for any $[x_0 : y_0] \in \mathbb{RP}^1_+$. In particular the holy grail is to compute

$$\int h \circ \pi \, d\nu_{Blackwell}\,,$$

where $h(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy function. Note that according to (1.97) it is enough to learn how to integrate rational functions on $\mathbb{RP}^1$. In fact *nobody knows how to integrate even a single function* $f : \mathbb{RP}^1 \to \mathbb{R}$ *except the trivial case:*

$$f = c\theta + d$$

(Note: $\left|\theta \circ g_t - \frac{1}{2}\right| \le |1-2\tau|^t$ and thus $\int \theta d\nu_{Blackwell} = \frac{1}{2}$.) For example, can we get some expression in τ (for some fixed δ_1 and δ_2) for

$$\int \left(\frac{x}{x+y}\right)^2 d\nu_{Blackwell}?$$

Maybe any other polynomial in $\frac{x}{x+y}$?

Simplification: Analyze the simplified problem by taking T_0, T_1 as above but

$$\pi([x : y]) = \frac{1}{2}\,. \tag{1.114}$$

In this case distribution μ_t of g_t is just a t-fold convolution of the distribution μ_1 which is a sum of two delta-distributions on T_0 and T_1:

$$\mu_1 = \frac{1}{2}\Delta_{T_0} + \frac{1}{2}\Delta_{T_1}\,, \tag{1.115}$$

$$\mu_t = \mu_1 * \cdots * \mu_1\,. \tag{1.116}$$

It looks like all of the properties quoted above (in particular (1.109)) hold in this simplified setup. The idea is to apply Fourier transform to study this question: if $\hat{\mu}_1$ is the F.t. of μ_1 then

$$\hat{\mu}_t = (\hat{\mu}_1)^t\,.$$

Now, what is Fourier transform on $PGL(2,\mathbb{R})$ (or $PSL(2,\mathbb{R})$ when $\tau < 1/2$)?

Note: $\tau = 0$ is the only special case when we have nice closed-form expressions. What makes the most difference in is that when $\tau = 0$ we see that T_0 and T_1 commute. So if the non-commutativity is the key roadblock, then it is natural to attack it with the "non-commutative Fourier".

1.13 Monte Carlo method for computing H

An often used method of computation of H is based on the Shannon-McMillan-Breiman theorem, namely:

$$\frac{1}{n}\log P_{Z^n}[Z^n] \to -H \quad \text{a.s. and in } L_1\,. \tag{1.117}$$

The purpose of this note is to demonstrate numerically that a much faster convergence happens for the sequence:

$$\frac{1}{n}\sum_{j=1}^{n} h(Q_{j-1}) \to H \quad \text{a.s. and in } L_1\,. \tag{1.118}$$

Indeed, we have shown above that Q_k forms a Markov chain (i.e. easy to simulate) with

$$Q_0 = \frac{\delta_1+\delta_2}{2}$$

and

$$Q_j = \begin{cases} T_0'(Q_{j-1})\,, & \text{with probability } 1-Q_{j-1}\,,\\ T_1'(Q_{j-1})\,, & \text{with probability } Q_{j-1}\,, \end{cases}$$

where $T_0'(q)$ and $T_1'(q)$ are some fractional-linear functions (i.e. Mobius transforms), depending on τ, δ_1, δ_2. It can also easily be shown that

$$\tau \min(\delta_1,\delta_2) \le Q_j \le (1-\tau)\max(\delta_1,\delta_2) \qquad \text{a.s.} \tag{1.119}$$

Notice now that

$$\log P_{Z^n}[Z^n] = \sum_{k=1}^{n} \log P_{Z_k|Z^{k-1}}[Z_k|Z^{k-1}]\,, \tag{1.120}$$

and that

$$\begin{aligned} \mathbb{E}\left[\log P_{Z_k|Z^{k-1}}[Z_k|Z^{k-1}] \,\middle|\, Z^{k-1}\right] &= -h(\mathbb{P}[Z_k=1|Z_1^{k-1}]) && (1.121)\\ &= -h(Q_{k-1})\,. && (1.122) \end{aligned}$$

With this we can rewrite (1.120) as

$$-\log P_{Z^n}[Z^n] = \sum_{k=1}^{n}\left\{-h(Q_{k-1}) - \log P_{Z_k|Z^{k-1}}[Z_k|Z^{k-1}]\right\} + \sum_{k=1}^{n} h(Q_{k-1})\,.$$

Because of (1.119) the first term is a running sum of an a.s.-bounded martingale difference process and therefore by the Azuma inequality and (1.117) we prove (1.118).

This effect is demonstrated numerically for the case $\tau = 0.1, \delta_1 = 0.16, \delta_2 = 0.07$ and $H \approx 0.514$ bit on the Fig. 1.2.

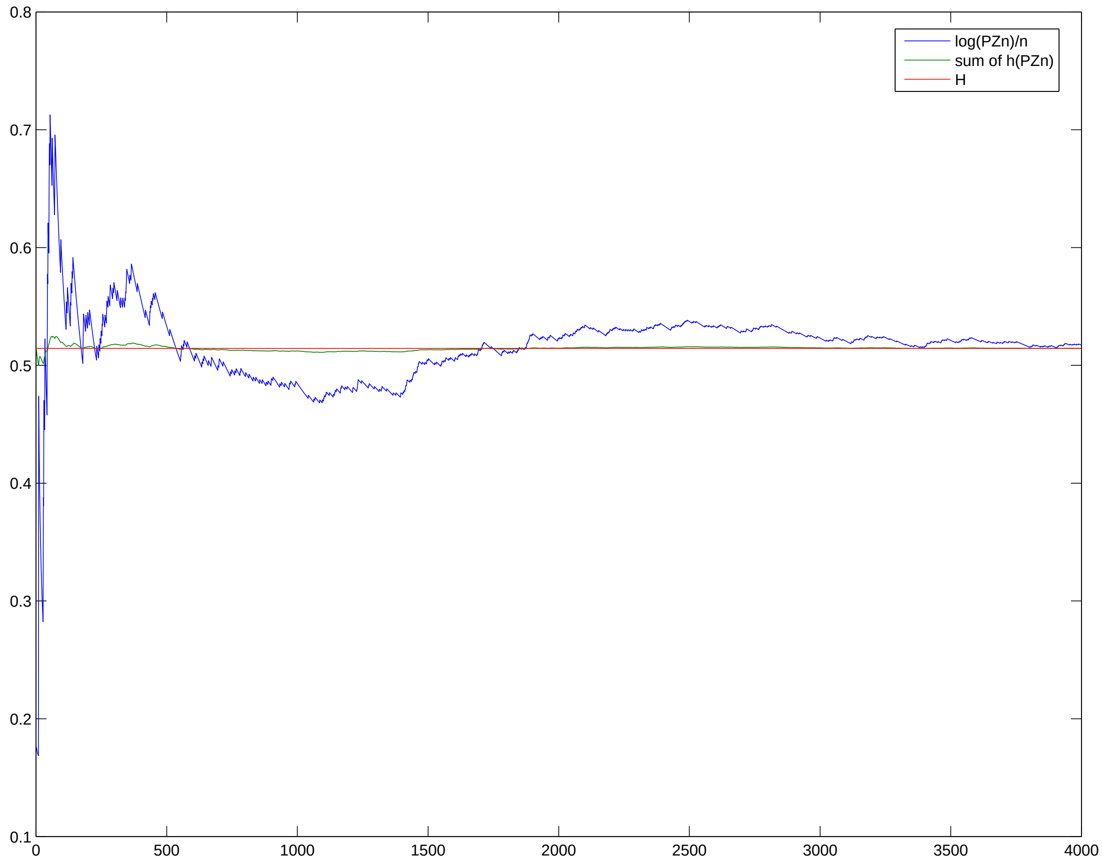

Figure 1.2: Comparing speed of convergence in (1.117) and (1.118). $\tau = 0.1, \delta_1 = 0.16, \delta_2 = 0.07$ and $H \approx 0.514$ bit.

1.14 Attempt to prove the $O(\log n)$ expansion

The capacity of the GEC without state knowledge is given by:

$$C_0 \quad = \quad \log 2 - H \,. \tag{1.123}$$

$$\tag{1.124}$$

Here is a theorem that we are willing to prove:

Theorem 1 *The dispersion of the Gilbert-Elliott channel with no state information and state transition probability $\tau \in (0,1)$ is*

$$V_0 = \mathrm{Var}\left[F_0\right] + 2\sum_{i=1}^{\infty} \mathbb{E}\left[(F_i - \mathbb{E}\left[F_i\right])(F_0 - \mathbb{E}\left[F_0\right])\right] \,. \tag{1.125}$$

Furthermore, provided that $V_0 > 0$ and regardless of whether ϵ is a maximal or average probability of error, we have

$$\log M^*(n, \epsilon) = nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon) + O(\log n)\,, \tag{1.126}$$

where C_0 is given by (1.123)*.*

Proof: The proof is a step-by-step repetition of the proof of Theorem for the CSIR case. In particular, the expression for the information density $i(X^n; Y^n)$ becomes

$$\begin{aligned} i(X^n; Y^n) &= \log \frac{P_{Y^n|X^n}}{P_{Y^n}} && (1.127) \\ &= n \log 2 - \sum_{j=1}^{n} G_j\,. && (1.128) \end{aligned}$$

For this $i(X^n; Y^n)$ we only need to establish that there exist a constant B_2 such that for any λ we have

$$\left| \mathbb{P}\left[i(X^n; Y^n) > nC_0 + \sqrt{nV_0}\lambda \right] - Q(\lambda) \right| \le \frac{B_2 \log n}{\sqrt{n}}\,. \tag{1.129}$$

The idea of proving (1.129) is to first approximate $i(X^n; Y^n)$ by a sum over a stationary process (following the proof of Theorem 2.6 in [2]) and then apply Tikhomirov's theorem [1].

Define

$$S_n = \log 2 - \sum_{j=1}^{n} F_j\,,$$

and notice that

$$\begin{aligned} \mathbb{E}\left[S_n\right] &= nC_0\,, && (1.130) \\ \mathrm{Var}[S_n] &= nV_0 + O(1)\,, && (1.131) \end{aligned}$$

where (1.131) follows from (1.56) by a usual argument (e.g., see [2]). So, similar to the proof of Theorem-CSIR, an application of Tikhomirov's theorem proves for some $B_2' > 0$:

$$\left| \mathbb{P}\left[S_n > nC_0 + \sqrt{nV_0}\lambda \right] - Q(\lambda) \right| \le \frac{B_2' \log n}{\sqrt{n}}\,. \tag{1.132}$$

We will show below that for some constant A_3 we have

$$|i(X^n; Y^n) - S_n| \le A_3 \quad \text{a.s.}\,. \tag{1.133}$$

Clearly (1.133) and (1.132) imply (1.129) after applying Taylor's expansion to $Q\left(\lambda \pm \frac{A_3}{\sqrt{nV_0}}\right)$.

To demonstrate (1.133) denote

$$q_{min} = \min(\delta_1\tau + \delta_2\bar{\tau}, \delta_1\bar{\tau} + \delta_2\tau)\,, \tag{1.134}$$

$$q_{max} = \max(\delta_1\tau + \delta_2\bar{\tau}, \delta_1\bar{\tau} + \delta_2\tau)\,. \tag{1.135}$$

By (1.70) we have

$$q_{min} \le Q_j, U_j \le q_{max}\,.$$

Then, by the definition of G_j and F_j we have

$$|G_j - F_j| \le \max_{[q_{min}, 1-q_{min}]} \left|\frac{d\log x}{dx}\right| \cdot |Q_j - U_j| \tag{1.136}$$

$$= \frac{\log e}{q_{min}}(\delta_1 - \delta_2)|\mu|^j\,. \tag{1.137}$$

where (1.150) is by (1.24).

But then we have

$$|i(X^n; Y^n) - S_n| = \left|\sum_{j=1}^n (G_j - F_j)\right| \tag{1.138}$$

$$\le \sum_{j=1}^n |G_j - F_j| \tag{1.139}$$

$$\le \frac{\log e}{q_{min}}(\delta_1 - \delta_2)\sum_{j=1}^n |\mu|^j \tag{1.140}$$

$$\le \frac{\log e}{q_{min}}(\delta_1 - \delta_2)\frac{|\mu|}{1 - |\mu|} = A_3\,. \tag{1.141}$$

This proves (1.133) with A_3 defined above.

It now remains to show that the process $\frac{1}{\sqrt{n}}\sum_{j=1}^n P_{Z_j|Z_{-\infty}^{j-1}}$ *satisfies Berry-Esseen type of bound. Note that this process is asymptotically normal with variance* V_0 *as shown by Theorem 2.6 in [2]. Unfortunally, I could not show that* F_j *is* α*-mixing and therefore application of Tikhomirov's theorem is not valid.*

1.15 Proof of the $o(\sqrt{n})$ expansion

Theorem 2 *The dispersion of the Gilbert-Elliott channel with no state information and state transition probability* $\tau \in (0, 1)$ *is*

$$V_0 = \mathrm{Var}\left[F_0\right] + 2\sum_{i=1}^\infty \mathbb{E}\left[(F_i - \mathbb{E}\left[F_i\right])(F_0 - \mathbb{E}\left[F_0\right])\right]\,. \tag{1.142}$$

Furthermore, provided that $V_0 > 0$ and regardless of whether ϵ is a maximal or average probability of error, we have

$$\log M^*(n,\epsilon) = nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon) + o(\sqrt{n})\,, \qquad (1.143)$$

where C_0 is given by (1.123).

Proof: *Achievability:* In this proof we will demonstrate how central-limit theorem result for the information density implies the $o(\sqrt{n})$ expansion. Otherwise, the proof is a step-by-step repetition of the proof of Theorem for the CSIR case. In particular, with equiprobable P_{X^n}, the expression for the information density $i(X^n;Y^n)$ becomes

$$\begin{aligned} i(X^n;Y^n) &= \log \frac{P_{Y^n|X^n}}{P_{Y^n}} && (1.144)\\ &= n\log 2 + \log P_{Z^n}(Z^n)\,. && (1.145) \end{aligned}$$

Provided that condition 2 of Theorem 2.6 in [2] holds, it yields the following:

$$\mathbb{P}\left[i(X^n;Y^n) > nC_0 + \sqrt{nV_0}\lambda\right] \to Q(\lambda)\,. \qquad (1.146)$$

We denote

$$\begin{aligned} q_{min} &= \min(\delta_1\tau + \delta_2\bar\tau, \delta_1\bar\tau + \delta_2\tau)\,, && (1.147)\\ q_{max} &= \max(\delta_1\tau + \delta_2\bar\tau, \delta_1\bar\tau + \delta_2\tau)\,. && (1.148) \end{aligned}$$

By (1.70) we have

$$q_{min} \le Q_j, U_j \le q_{max}\,.$$

Then, by the definition of G_j and F_j we have

$$\begin{aligned} |G_j - F_j| &\le \max_{[q_{min},1-q_{min}]} \left|\frac{d\log x}{dx}\right| \cdot |Q_j - U_j| && (1.149)\\ &= \frac{\log e}{q_{min}}(\delta_1 - \delta_2)|\mu|^j\,, && (1.150) \end{aligned}$$

where (1.150) is by (1.24). The bound (1.150) automatically proves that condition 2 of Theorem 2.6 in [2] is satisfied and therefore (1.146) holds.

Then, by Theorem **??** (DT bound) we know that there exists a code with M codewords and average probability of error p_e bounded by

$$\begin{aligned} p_e &\le \mathbb{E}\left[\exp\left\{-\left[i(X^n;Y^n) - \log\tfrac{M-1}{2}\right]^+\right\}\right] && (1.151)\\ &\le \mathbb{E}\left[\exp\left\{-\left[i(X^n;Y^n) - \log M\right]^+\right\}\right] && (1.152) \end{aligned}$$

where (1.152) is by monotonicity of $exp\{-[i(X^n;Y^n)-a]^+\}$ with respect to a. Furthermore, notice that for any random variable U and $a, b \in \mathbb{R}$ we have

$$\mathbb{E}\left[\exp\left\{-\left[U-a\right]^+\right\}\right] \leq \mathbb{P}[U \leq b] + \exp\{a-b\}\,. \tag{1.153}$$

To see (1.153), notice that $b < a$ is trivial and for $b \geq a$ we have

$$\begin{aligned}
\exp\left\{-\left[U-a\right]^+\right\} &= 1\{U \leq a\} + \exp\{a-U\} \cdot 1\{U > a\} && (1.154)\\
&\leq 1\{U \leq b\} + \exp\{a-U\} \cdot 1\{U > b\} && (1.155)\\
&\leq 1\{U \leq b\} + \exp\{a-b\}\,, && (1.156)
\end{aligned}$$

from which (1.153) follows by taking the expectation[1].

Fix some $\epsilon' > 0$ and set

$$\log \gamma_n = nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon - \epsilon')\,.$$

Then continuing from (1.152) we obtain

$$\begin{aligned}
p_e &\leq \mathbb{P}[i(X^n;Y^n) \leq \log\gamma_n] + \exp\{\log M - \log\gamma_n\} && (1.157)\\
&= \epsilon - \epsilon' + o(1) + \exp\{\log M - \log\gamma_n\}\,, && (1.158)
\end{aligned}$$

where (1.157) follows by applying (1.153) and (1.158) is by (1.146). If we set $\log M = \log\gamma_n - \log n$ then the right-hand side of (1.158) for large n falls below ϵ. Hence we conclude that for n large enough we have

$$\begin{aligned}
\log M^*(n,\epsilon) &\geq \log\gamma_n - \log n && (1.159)\\
&\geq nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon - \epsilon') - \log n\,, && (1.160)
\end{aligned}$$

or because ϵ' was arbitrary, this is equivalent to

$$\log M^*(n,\epsilon) \;\geq\; nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon) + o(\sqrt{n})\,. \tag{1.161}$$

Converse: To apply Theorem **??** (meta-converse) we choose the auxiliary channel $Q_{Y^n|X^n}$ which simply outputs an equiprobable Y^n independent of the input X^n:

$$Q_{Y^n|X^n}(y^n|x^n) = 2^{-n}\,. \tag{1.162}$$

Similar to the CSIR case we get

$$\beta_{1-\epsilon}\left(P_{X^nY^n}, Q_{X^nY^n}\right) \leq \frac{1}{M^*}\,, \tag{1.163}$$

[1]This upper-bound reduces (1.151) to the usual Feinstein's Lemma.

and also

$$\log \frac{P_{X^nY^n}}{Q_{X^nY^n}} = n\log 2 + \log P_{Z^n}(Z^n) \tag{1.164}$$

$$= i(X^n;Y^n)\,. \tag{1.165}$$

We pick any $\epsilon' > 0$ and set

$$\log \gamma_n = nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon+\epsilon')\,. \tag{1.166}$$

By (**??**) we have for $\alpha = 1-\epsilon$:

$$\beta_{1-\epsilon} \geq \frac{1}{\gamma_n}\left(1-\epsilon-\mathbb{P}\left[i(X^n;Y^n)\geq \log\gamma_n\right]\right) \tag{1.167}$$

$$= \frac{1}{\gamma_n}(\epsilon' + o(1))\,, \tag{1.168}$$

where (1.168) is from (1.146). Finally, from (1.163) we get

$$\log M^*(n,\epsilon) \leq -\log \beta_{1-\epsilon} \tag{1.169}$$

$$= \log\gamma_n - \log(\epsilon'+o(1)) \tag{1.170}$$

$$= nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon+\epsilon') + O(1) \tag{1.171}$$

$$= nC_0 - \sqrt{nV_0}Q^{-1}(\epsilon) + o(\sqrt{n})\,. \tag{1.172}$$

This concludes the proof.

1.16 Bounds on $|C_0 - C_1|$

First notice that

$$|C_0 - C_1| = C_1 - C_0$$

since $C_0 < C_1$ for all $\tau \in (0,1)$.

To estimate $(C_1 - C_0)$ we define

$$\xi_j = F_j + \log P_{Z_j|S_j}(Z_j|S_j) \tag{1.173}$$

and notice that

$$C_1 - C_0 = \mathbb{E}\left[\xi_0\right].$$

Two lemmas below will prove the following

Theorem 3 *Assuming $1/2 \geq \delta_1 \geq \delta_2 > 0$ and with $\tau \to 0$ we have*

$$C_1 - O(\sqrt{-\tau\ln\tau}) \leq C_0 \leq C_1 - O(\tau)\,.$$

Lemma 4 *Assuming* $1/2 \geq \delta_1 \geq \delta_2$ *and denoting*

$$p_{max} = \max\{\tau, 1-\tau\}\,, \tag{1.174}$$
$$p_{min} = \min\{\tau, 1-\tau\}\,, \tag{1.175}$$

we have the following:

$$\mathbb{E}\,[\xi_1] \geq h(\delta_1 p_{max} + \delta_2 p_{min})) - p_{max} h(\delta_1) - p_{min} h(\delta_2)\,. \tag{1.176}$$

In particular, with $\tau \to 0$ *we have*

$$\mathbb{E}\,[\xi_1] \geq O(\tau)\,.$$

Proof: Notice the following relation w.r.t. ξ_1:

$$\begin{aligned}
\mathbb{E}\,[\xi_1 | Z_{-\infty}^0] &= h(\delta_1 P_0 + \delta_2(1-P_0)) - P_0 h(\delta_1) - (1-P_0)h(\delta_2) && (1.177)\\
&= P_0 d(\delta_1 || \delta_1 P_0 + \delta_2(1-P_0)) && (1.178)\\
&\quad +(1-P_0) d(\delta_2 || \delta_1 P_0 + \delta_2(1-P_0)) && (1.179)\\
&\geq 0 && (1.180)
\end{aligned}$$

Looking at (1.179) we observe that the smallness of $\mathbb{E}\,[\xi_j] = C_1 - C_0$ implies that P_{j-1} is tightly concentrated around 0 and 1 (i.e. the state predictor is almost certain about the next state). This information about P_{j-1} will be used to prove the bound on $|V_1 - V_0|$.

We now aim to analyze the function:

$$f(x) = h(\delta_1 x + \delta_2(1-x)) - x h(\delta_1) - (1-x) h(\delta_2)\,.$$

This is a positive, $\cap$-concave function turning to 0 at the endpoints:

$$f(0) = f(1) = 0\,.$$

Therefore, since we know that P_0 always belongs to an interval between τ and $1-\tau$ we have

$$f(P_0) \geq \min(f(\tau), f(1-\tau))\,.$$

Once we check that the minimum is always attained at the point which is closer to 1 we get (1.176). This follows from an easily checkable inequality which is valid for $1/2 \geq \delta_1 \geq \delta_2$ and $x \in [0, 1/2]$:

$$h(\delta_1 x + \delta_2(1-x)) + (1-2x)[h(\delta_1) - h(\delta_2)] \geq h(\delta_1(1-x) + \delta_2 x)\,. \tag{1.181}$$

Note: this bound can be improved because actually P_0 always belongs to an interval between the two fixed points x_0 and x_1:

$$T_0(x_0) = x_0\,, \qquad T_1(x_1) = x_1\,.$$

For small τ this replaces p_{min} with something of the order $\tau \frac{1-\delta_2}{\delta_1 - \delta_2}$.

Lemma 5 *Assuming* $1/2 \geq \delta_1 \geq \delta_2 > 0$ *we have for* $\tau \to 0$

$$\mathbb{E}\left[\xi_1\right] \leq O\left(\sqrt{-\tau \log \tau}\right) . \tag{1.182}$$

Remark 1: The resulting bound is only used for estimating the rate of convergence of $C_0(\tau)$ to C_1 with $\tau \to 0$. Numerically this bound is not very tight and for this reason we do not provide all the constants.

Remark 2: A similar bound can be proved for $\delta_2 = 0$. For this case we must notice that

$$\mathbb{E}\left[(P_0 - 1\{S_1 = 1\})^2\right] = \mathbb{E}\left[P_0(1 - P_0)\right]$$

and then prove that if this expectation is small then P_0 is concentrated around 0 and 1 and therefore $\mathbb{E}\left[\xi_1\right]$ must also be small as follows from (1.179).

Proof: Because of the expression

$$\begin{aligned} H &= -\mathbb{E}\left[\log P_{Z_1|Z_{-\infty}^0}[Z_1|Z_{-\infty}^0]\right] && (1.183) \\ &= \mathbb{E}\, h(\delta_1 P_0 + \delta_2(1 - P_0)) && (1.184) \end{aligned}$$

we can write

$$\begin{aligned} C_1 - C_0 &= \mathbb{E}\left[\xi_1\right] && (1.185) \\ &= \mathbb{E}\left[h(\delta_1 P_0 + \delta_2(1 - P_0)) - h(\delta_1 1\{S_1 = 1\} + \delta_2 1\{S_1 = 2\})\right] . && (1.186) \end{aligned}$$

Because $\delta_2 > 0$ we can upper-bound the derivative of the function $h(\delta_1 x + \delta_2(1 - x))$ by

$$\left|\frac{d}{dx} h(\delta_1 x + \delta_2(1 - x))\right| \leq B_1 = (\delta_1 - \delta_2) \log \frac{1 - \delta_2}{\delta_2} .$$

So we have

$$\begin{aligned} \mathbb{E}\left[\xi_1\right] &\leq B_1 \mathbb{E}\left[|P_0 - 1\{S_1 = 1\}|\right] && (1.187) \\ &\leq B_1 \sqrt{\mathbb{E}\left[(P_0 - 1\{S_1 = 1\})^2\right]} , && (1.188) \end{aligned}$$

where (1.188) follows from the Lyapunov's inequality.

For any estimator $\hat{A}$ of $1\{S_1 = 1\}$ based on $Z_{-\infty}^0$ we have

$$\mathbb{E}\left[(P_0 - 1\{S_1 = 1\})^2\right] \leq \mathbb{E}\left[(\hat{A} - 1\{S_1 = 1\})^2\right] ,$$

because $P_0 = \mathbb{E}\left[1\{S_1 = 1\}|Z_{-\infty}^0\right]$ is an MMSE estimate.

We now take the following estimator

$$\hat{A}_n = 1\left\{\sum_{j=-n+1}^{0} Z_j \geq n\delta_a\right\} ,$$

where n is to be specified later and $\delta_a = \frac{\delta_1+\delta_2}{2}$. We then have the following simple estimate:

$$\begin{aligned}
\mathbb{E}\,[(\hat{A}_n - 1\{S_1 = 1\})^2] &= \mathbb{P}[1\{S_1 = 1\} \neq \hat{A}_n] && (1.189)\\
&\leq \mathbb{P}[\hat{A}_n \neq 1\{S_1 = 1\}, S_1 = \cdots = S_{-n+1}] && (1.190)\\
&\quad + 1 - \mathbb{P}[S_1 = \cdots = S_{-n+1}] && (1.191)\\
&= \frac{1}{2}\bar{\tau}^n \left(\mathbb{P}[B(n,\delta_1) < n\delta_a] + \mathbb{P}[B(n,\delta_2) \geq n\delta_a]\right) + 1 - \bar{\tau}^n && (1.192)
\end{aligned}$$

where $B(n, \delta)$ denotes the binomially distributed random variable. Using Chernoff bounds we can find that for some E_1 we have

$$\mathbb{P}[B(n,\delta_1) < n\delta_a] + \mathbb{P}[B(n,\delta_2) \geq n\delta_a] \leq 2e^{-nE_1}\,.$$

Then we get

$$\mathbb{E}\,[(\hat{A}_n - 1\{S_1 = 1\})^2] \leq 1 - \bar{\tau}^n(1 - e^{-nE_1})\,.$$

We now set denote for convenience:

$$\beta = -\ln\bar{\tau} = -\ln(1-\tau)\,.$$

and choose

$$n = \left\lceil -\frac{1}{E_1}\ln\frac{\beta}{E_1}\right\rceil\,.$$

Putting it all together we get

$$\mathbb{E}\,[(\hat{A}_n - 1\{S_1 = 1\})^2] \leq 1 - \bar{\tau}\cdot e^{-\frac{\beta}{E_1}\ln\frac{\beta}{E_1}}\left(1 - \frac{\beta}{E_1}\right)\,.$$

When $\tau \to 0$ we have $\beta = \tau + o(\tau)$ and then it is not hard to obtain

$$\mathbb{E}\,[(\hat{A}_n - 1\{S_1 = 1\})^2] \leq \frac{\tau}{E_1}\ln\frac{\tau}{E_1} + o(\tau\ln\tau)\,.$$

This proves (1.182).

1.17 Bound on $|V_0 - V_1|$

Expression (1.142) does not reveal the behavior of the V_0 with $\tau \to 0$. In the CSIR case we had an expression much easier to work with:

$$\begin{aligned}
V_1 &= \frac{1}{2}(v(\delta_1) + v(\delta_2))\\
&\quad + \left(\frac{h(\delta_1) - h(\delta_2)}{2}\right)^2\left(\frac{1}{\tau} - 1\right)\,. && (1.193)
\end{aligned}$$

This Section is devoted to proving the following result:

Theorem 6 *In the conditions of Theorem 2, assume that* $\delta_1 \geq \delta_2 > 0$*. Then, we have*

$$|V_0 - V_1| \leq 2\sqrt{V_1 \delta V} + \delta V \,, \tag{1.194}$$

where δV *satisfies*

$$\begin{aligned}
\delta V &\leq B_0 + \frac{B_0}{2(1-\sqrt{|\mu|})} \ln \frac{eB_1}{B_0} \,, && (1.195)\\
B_0 &= \frac{d_2(\delta_1||\delta_2)}{d(\delta_1||\delta_2)} |C_0 - C_1| \,, && (1.196)\\
B_1 &= \sqrt{\frac{B_0}{|\mu|}} \left(d(\delta_1||\delta_2) \left| \ln \frac{\tau}{1-\tau} \right| + \frac{h(\delta_1) - h(\delta_2)}{2|\mu|} \right) \,, && (1.197)\\
d_2(a||b) &= a \log^2 \frac{a}{b} + (1-a) \log^2 \frac{1-a}{1-b} \,. && (1.198)
\end{aligned}$$

Finally, with $\tau \to 0$ *this implies that*

$$V_0 = V_1 + o\left(\frac{1}{\tau}\right) = O\left(\frac{1}{\tau}\right) \,. \tag{1.199}$$

Proof: First define

$$\delta V = \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left[\sum_{j=1}^{n} \xi_j\right] \,, \tag{1.200}$$

where ξ_j was defined in (1.173). Also denote

$$\eta_j = -\log P_{Z_j|S_j}(Z_j|S_j) \,,$$

which then leads to

$$F_j = \eta_j + \xi_j \,.$$

Now notice that

$$\begin{aligned}
\mathbb{E}[\eta_j] &= \log 2 - C_1 \,, && (1.201)\\
\mathrm{Var}\left[\sum_{j=1}^{n} \eta_j\right] &= nV_1 + O(1) \,. && (1.202)
\end{aligned}$$

So that intuitively ξ_j is a correction term, compared to CSIR case, and allegedly for small τ it must be small.

As was discussed earlier, the (1.179) and the fact that $\mathbb{E}[\xi_j]$ tends to 0 with $\tau \to 0$ together imply that P_j's distribution is conecntrated around 0 and 1. Since $\mathbb{E}[\sigma_0^2]$ can be written as a function of P_{-1} this concentration will lead to an upper-bound on $\mathbb{E}[\sigma_0^2]$.

Now to the original question:

$$V_0 = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left[\sum_{j=1}^{n} F_j\right] \tag{1.203}$$

$$= \lim_{n\to\infty} \operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \eta_j + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \xi_j\right]. \tag{1.204}$$

By Cauchy-Schwartz we have

$$\operatorname{Var}[A+B] = \operatorname{Var}[A] + \operatorname{Var}[B] \pm 2\sqrt{\operatorname{Var}[A]\operatorname{Var}[B]}. \tag{1.205}$$

And we also have

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \eta_j\right] = V_1 + o(1), \tag{1.206}$$

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \xi_j\right] = \delta V + o(1). \tag{1.207}$$

Together from (1.204)-(1.207) we conclude (1.194).

We now move on to prove (1.195). Define a centered random variable:

$$\hat{\xi}_j = \xi_j - \mathbb{E}[\xi_j] = \xi_j - (C_1 - C_0).$$

Then, we have

$$\delta V = \mathbb{E}[\hat{\xi}_0^2] + 2\sum_{j=1}^{\infty} \mathbb{E}[\hat{\xi}_0\hat{\xi}_j].$$

Lemma 7 *In the conditions of Theorem 6, we have*

$$\mathbb{E}[\hat{\xi}_0]^2 \leq B_0. \tag{1.208}$$

Lemma 8 *In the conditions of Theorem 6, we have*

$$\mathbb{E}[\hat{\xi}_0\hat{\xi}_j] \leq B_1|\mu|^{j/2}. \tag{1.209}$$

Assuming these two lemmas we bound δV as follows. Set

$$N = \left\lceil \frac{2\ln\frac{B_0}{B_1}}{\ln|\mu|} \right\rceil.$$

Then we use Cauchy-Schwarz for $j < N$ and (1.209) for $j \geq N$:

$$\begin{aligned}\sum_{j=1}^{\infty} \mathbb{E}\,[\hat{\xi}_0 \hat{\xi}_j] &\leq (N-1)B_0 + B_1 \sum_{j \geq N} |\mu|^{j/2} && (1.210)\\ &\leq \frac{\ln \frac{B_0}{B_1}}{\ln \sqrt{|\mu|}} B_0 + \frac{B_0}{1-\sqrt{|\mu|}} && (1.211)\\ &\leq \frac{B_0}{1-\sqrt{|\mu|}} \ln \frac{eB_1}{B_0}\,. && (1.212)\end{aligned}$$

So we get, overall,

$$\delta V \leq B_0 + 2\frac{B_0}{1-\sqrt{|\mu|}} \ln \frac{eB_1}{B_0}\,,$$

which is exactly (1.195). Notice that we have actually demonstrated a very slightly stronger:

$$V_1 + \delta V - 2\sqrt{V_1 \cdot \delta V} \leq V_0 \leq V_1 + \delta V + 2\sqrt{V_1 \cdot \delta V}\,.$$

Finally, (1.199) follows once we notice that with $\tau \to 0$

$$\delta V \leq \frac{B_0 \ln \frac{1}{B_0} + o(B_0 \ln B_0)}{\tau}$$

and use Lemma 5.

Proof of Lemma 7: First notice that

$$\begin{aligned}\mathbb{E}\,[\xi_1^2 | Z_{-\infty}^0] &= P_0 d_2(\delta_1 || \delta_1 P_0 + \delta_2(1-P_0))\\ &\quad + (1-P_0) d_2(\delta_2 || \delta_1 P_0 + \delta_2(1-P_0))\,. && (1.213)\end{aligned}$$

We now state a Lemma:

Lemma 9 *Assume that $\delta_1 \geq \delta_2 > 0$ and $a, b \in [\delta_2, \delta_1]$, then*

$$\frac{d(a||b)}{d_2(a||b)} \geq \frac{d(\delta_1||\delta_2)}{d_2(\delta_1||\delta_2)}\,. \tag{1.214}$$

Applying this Lemma twice (with $a = \delta_1$ or δ_2 and $b = \delta_1 x + \delta_2 \bar{x}$) we obtain:

$$\begin{aligned}&x d_2(\delta_1 || \delta_1 x + \delta_2 \bar{x}) + \bar{x} d_2(\delta_2 || \delta_1 x + \delta_2 \bar{x})\\ &\quad \leq \frac{d_2(\delta_1||\delta_2)}{d(\delta_1||\delta_2)} \left(x d(\delta_1 || \delta_1 x + \delta_2 \bar{x}) + \bar{x} d(\delta_2 || \delta_1 x + \delta_2 \bar{x}) \right)\,. && (1.215)\end{aligned}$$

If we substitute $x = P_{-1}$ here, then from (1.213) and (1.179) we get that

$$\mathbb{E}\,[\xi_0^2 | Z_{-\infty}^{-1}] \leq \frac{d_2(\delta_1||\delta_2)}{d(\delta_1||\delta_2)} \mathbb{E}\,[\xi_0 | Z_{-\infty}^{-1}]\,.$$

Averaging this we obtain[1]

$$\mathbb{E}\,[\xi_0^2] \le \frac{d_2(\delta_1||\delta_2)}{d(\delta_1||\delta_2)}(C_1 - C_0)\,. \tag{1.217}$$

Proof of Lemma 9: We first notice that the base of the log cancels in (1.214) and so we replace log by ln below. Next observe that Lemma follows trivially if we proved the following two statements:

$$\forall \delta \in [0, 1/2] : \quad \frac{d(a||\delta)}{d_2(a||\delta)} \quad \text{is a non-increasing function of } a \in [0, 1/2]\,. \tag{1.218}$$

and

$$\frac{d(\delta_1||b)}{d_2(\delta_1||b)} \quad \text{is a non-decreasing function of } b \in [0, \delta_1]\,. \tag{1.219}$$

Steps for proving (1.218):

1. Take derivative of $\frac{d_2(a||\delta)}{d(a||\delta)}$ and require it to be non-negative. This leads to

$$\begin{cases} 2d(a||\delta) + \ln\frac{a}{\delta}\cdot\ln\frac{1-a}{1-\delta} \le 0\,, & \text{if } a \le \delta\,, \\ 2d(a||\delta) + \ln\frac{a}{\delta}\cdot\ln\frac{1-a}{1-\delta} \ge 0\,, & \text{if } a \ge \delta\,. \end{cases} \tag{1.220}$$

2. Now think of the expression in (1.220) as a function of δ:

$$f_a(\delta) = 2d(a||\delta) + \ln\frac{a}{\delta}\cdot\ln\frac{1-a}{1-\delta}\,.$$

   It is easy to check that

$$f_a(a) = 0\,, f_a'(a) = 0\,. \tag{1.221}$$

   So we need only to prove that

$$f_a(\delta) = \begin{cases} \text{a } \cup\text{-convex}\,, & \delta \in [0, a]\,, \\ \text{a } \cap\text{-concave}\,, & \delta \in [a, 1/2]\,. \end{cases} \tag{1.222}$$

   Indeed, if (1.222) holds then 0 will be an affine minorant for $f_a(\delta)$ on $[0, a]$ and an affine majorant on $[a, 1/2]$, which is exactly (1.220).

---

[1] Note that it can also be shown that analogous to (1.215) we have

$$\begin{aligned} &x d_2(\delta_1||\delta_1 x + \delta_2\bar{x}) + \bar{x} d_2(\delta_2||\delta_1 x + \delta_2\bar{x}) \\ &\quad\ge \frac{d_2(\delta_2||\delta_1)}{d(\delta_2||\delta_1)}\left(x d(\delta_1||\delta_1 x + \delta_2\bar{x}) + \bar{x} d(\delta_2||\delta_1 x + \delta_2\bar{x})\right)\,. \end{aligned} \tag{1.216}$$

which results in

$$\mathbb{E}\,[\xi_0^2] \ge \frac{d_2(\delta_2||\delta_1)}{d(\delta_2||\delta_1)}(C_1 - C_0)$$

and therefore shows that (1.217) can not be improved significantly.

3. To prove (1.222) we analyze the second derivative of f_a:

$$f''_a(\delta) = \frac{2a}{\delta^2} + \frac{2\bar{a}}{\bar{\delta}^2} - \frac{1}{\delta^2}\ln\frac{\bar{\delta}}{\bar{a}} - \frac{2}{\delta\bar{\delta}} - \frac{1}{\bar{\delta}^2}\ln\frac{\delta}{a}\,.$$

We now apply the following bounds to each of the log-terms:

$$1 - \frac{1}{x} \le \ln x \le x - 1 \tag{1.223}$$

(the LHS bound follows from the RHS bound applied to $\ln x = -\ln\frac{1}{x}$).

Application of the RHS bound of (1.223) yields

$$\begin{aligned} f''_a(\delta) &\le \frac{2a}{\delta^2} + \frac{2\bar{a}}{\bar{\delta}^2} - \frac{1}{\delta^2}\left(\frac{\bar{\delta}}{\bar{a}} - 1\right) - \frac{2}{\delta\bar{\delta}} - \frac{1}{\bar{\delta}^2}\left(\frac{\delta}{a} - 1\right), && (1.224)\\ &\le 0 \quad \text{whenever } \delta \ge a\,. && (1.225) \end{aligned}$$

And similarly application of the LHS bound of (1.223) yields

$$\begin{aligned} f''_a(\delta) &\ge \frac{2a}{\delta^2} + \frac{2\bar{a}}{\bar{\delta}^2} - \frac{1}{\delta^2}\left(1 - \frac{\bar{a}}{\bar{\delta}}\right) - \frac{2}{\delta\bar{\delta}} - \frac{1}{\bar{\delta}^2}\left(1 - \frac{a}{\delta}\right), && (1.226)\\ &\ge 0 \quad \text{whenever } \delta \le a\,. && (1.227) \end{aligned}$$

4. This finishes the proof of (1.218).

Proof of (1.219):

1. Taking derivative of $\frac{d(\delta_1||b)}{d_2(\delta_1||b)}$ with respect to b and requiring it to be non-negative is equivalent to

$$2(1-2b)\left(\delta\ln\frac{\delta}{b}\right)\left(\bar{\delta}\ln\frac{\bar{\delta}}{\bar{b}}\right) + (\delta\bar{b} + \bar{\delta}b)\left(\delta\ln^2\frac{\delta}{b} - \bar{\delta}\ln^2\frac{\bar{\delta}}{\bar{b}}\right) \ge 0\,. \tag{1.228}$$

2. It is convenient to introduce $x = b/\delta \in [0,1]$ and then we define

$$f_\delta(x) = 2(1-2\delta x)\delta\bar{\delta}\ln x \cdot \ln\frac{1-\delta x}{\bar{\delta}} + \delta(1 + x(1-2\delta))\left(\delta\ln^2 x - \bar{\delta}\ln^2\frac{1-\delta x}{\bar{\delta}}\right),$$

for which we must show

$$f_\delta(x) \ge 0\,. \tag{1.229}$$

If we think of $A = \ln x$ and $B = \ln\frac{1-\delta x}{\bar{\delta}}$ as independent variables, then (1.228) is equivalent to solving

$$2\gamma AB + \alpha A^2 - \beta B^2 \ge 0\,,$$

which after some manipulation (and observation that we naturally have a requirement $A \le 0 \le B$) reduces to

$$\frac{A}{B} \le -\frac{\gamma}{\alpha} - \frac{1}{\alpha}\sqrt{\gamma^2 + \alpha\beta}\,.$$

After substituting our values for A, B, α, β and γ we get that (1.228) will be shown if we can show

$$\frac{\ln x}{\ln \frac{1-\delta x}{\bar{\delta}}} \le -\frac{1-2\delta x}{1+x(1-2\delta)}\frac{\bar{\delta}}{\delta} - \left(\left(\frac{1-2\delta x}{1-2\delta x + x}\right)^2\left(\frac{\bar{\delta}}{\delta}\right)^2 + \frac{\bar{\delta}}{\delta}\right)^{1/2} \tag{1.230}$$

3. To show (1.230) we are allowed to upper-bound $\ln x$ and $\ln \frac{1-\delta x}{\bar{\delta}}$ (we need an upper-bound and not the lower-bound for the latter because $\ln x \le 0$). So we use the following two bounds correspondingly:

$$\begin{aligned} \ln x &\le (x-1) - (x-1)^2/2 + (x-1)^3/3 - (x-1)^4/4 + (x-1)^5/5\,, && (1.231)\\ \ln y &\le (y-1) - (y-1)^2/2 + (y-1)^3/3\,, && (1.232) \end{aligned}$$

both of which follow from the fact that the derivative of $\ln x$ of the corresponding order is always negative. Applying these bounds we find that we need to prove

$$\begin{aligned} &\frac{(x-1) - (x-1)^2/2 + (x-1)^3/3 - (x-1)^4/4 + (x-1)^5/5}{\delta(1-x) - \delta^2(1-x)^2/2 + \delta^3(1-x)^3/3} \\ &\le -\frac{1-2\delta x}{1+x(1-2\delta)}\frac{\bar{\delta}}{\delta} - \left(\left(\frac{1-2\delta x}{1-2\delta x + x}\right)^2\left(\frac{\bar{\delta}}{\delta}\right)^2 + \frac{\bar{\delta}}{\delta}\right)^{1/2} \end{aligned} \tag{1.233}$$

4. After a tedious algebra the (1.233) simplifies to

$$\frac{\delta^2(1-x)^3}{(1-\delta)^5} P_\delta(1-x) \ge 0\,, \tag{1.234}$$

where

$$\begin{aligned} P_\delta(x) \;=\; & -(4\delta^2 - 1)(1-\delta)^2/12 && (1.235)\\ & + (1-\delta)(4 - 5\delta + 4\delta^2 - 24\delta^3 + 24\delta^4)x/24 && (1.236)\\ & + (8 - 20\delta + 15\delta^2 + 20\delta^3 - 100\delta^4 + 72\delta^5)x^2/60 && (1.237)\\ & - (1-\delta)^3(11 - 28\delta + 12\delta^2)x^3/20 && (1.238)\\ & + (1-\delta)^3(1-2\delta)^2 x^4/5\,. && (1.239) \end{aligned}$$

5. Assume that $P_\delta(x_0) < 0$ for some x_0. For all $\delta \in (0, 1/2]$ we can easily check that $P_\delta(0) > 0$ and $P_\delta(1) > 0$. Therefore, there must be a root x_1 of P_δ in $(0, x_0)$ and a

root x_2 in $(x_0, 1)$ by continuity. It is also easily checked that $P'_\delta(0) > 0$ for all δ. Bu then we must have at least one root of P'_δ in $[0, x_1)$ and at least one root of P'_δ in $(x_2, 1]$.

Now, $P'_\delta(x)$ is a cubic polynomial such that $P'_\delta(0) > 0$. So it must have at least one root on the negative real axis and two roots on $[0, 1]$. But since $P''_\delta(0) > 0$, it must be that $P''_\delta(x)$ also has two roots on $[0, 1]$. But $P''_\delta(x)$ is a quadratic polynomial, so its roots are algebraic functions of δ, for which we can easily check that one of them is always larger than 1. So, $P'_\delta(x)$ has at most one root on $[0, 1]$. And therefore the assumption was incorrect and

$$P_\delta \geq 0 \quad \text{on } [0, 1]\,.$$

Proof of Lemma 8: We write first

$$\mathbb{E}\,[\hat{\xi}_j \hat{\xi}_0] = \mathbb{E}\,[\xi_j \xi_0] - (\mathbb{E}\,[\xi_0])^2 \ .$$

Now we can find from the definition of ξ_j that

$$\mathbb{E}\,[\xi_j | S^0_{-\infty}, Z^{j-1}_{-\infty}] = f(P_{j-1}, R^*_{j-1})\,,$$

where

$$f(x, y) = yd(\delta_1 || \delta_1 x + \delta_2(1 - x)) + (1 - y)d(\delta_2 || \delta_1 x + \delta_2(1 - x))\,.$$

Notice the following relation:

$$\frac{d}{d\lambda} H(\bar{\lambda} Q + \lambda P) = D(P || \bar{\lambda} Q + \lambda P) - D(Q || \bar{\lambda} Q + \lambda P) + H(P) - H(Q)\,.$$

This has two consequences. First it shows that a function

$$D(P || \bar{\lambda} Q + \lambda P) - D(Q || \bar{\lambda} Q + \lambda P)$$

is monotonically decaying with λ (since it is a derivative of a concave function). Second, we have the following general relation for the excess of the entropy above its affine approximation:

$$\left.\frac{d}{d\lambda}\right|_{\lambda=0} [H((1-\lambda)Q + \lambda P) - (1-\lambda)H(Q) - \lambda H(P)] \;=\; D(P||Q) \qquad (1.240)$$

$$\left.\frac{d}{d\lambda}\right|_{\lambda=1} [H((1-\lambda)Q + \lambda P) - (1-\lambda)H(Q) - \lambda H(P)] \;=\; -D(Q||P)\,. \qquad (1.241)$$

Also it is clear that for all other λ's the derivative is in between these two extreme values.

First we observe that

$$\max_{x,y\in[0,1]}\left|\frac{df(x,y)}{dy}\right| = \max_{x\in[0,1]}|d(\delta_1||\delta_1 x+\delta_2(1-x))-d(\delta_2||\delta_1 x+\delta_2(1-x))| \quad (1.242)$$

$$= \max(d(\delta_1||\delta_2), d(\delta_2||\delta_1)) \quad (1.243)$$

$$= d(\delta_1||\delta_2)\,, \quad (1.244)$$

where (1.243) is because the function under the absolute value is decreasing and (1.244) is because we are restricted to $\delta_2 \le \delta_1 \le \frac{1}{2}$. On the other hand, we see that

$$f(x,x) = h(\delta_1 x + \delta_2(1-x)) - xh(\delta_1) - (1-x)h(\delta_2) \ge 0\,. \quad (1.245)$$

Comparing with (1.240) and (1.241), we have:

$$\max_{x\in[0,1]}\left|\frac{df(x,x)}{dx}\right| = \max(d(\delta_1||\delta_2), d(\delta_2||\delta_1)) \quad (1.246)$$

$$= d(\delta_1||\delta_2)\,. \quad (1.247)$$

Together with (1.244) this also implies (though we don't need it):

$$||\nabla f||_2 \le \sqrt{5}d(\delta_1||\delta_2)\,. \quad (1.248)$$

We now return to the original question. First, we notice that:

$$\mathbb{E}\,[\xi_j|S^0_{-\infty}, Z^{j-1}_{-\infty}] = f(P_{j-1}, P_{j-1}) \pm B_2|\mu|^{j-1}\,, \quad (1.249)$$

where for convenience we denote

$$B_2 = \frac{1}{2}d(\delta_1||\delta_2)\left|\ln\frac{\tau}{1-\tau}\right|\,.$$

The (1.249) follows by observing that

$$P_{j-1} = T_{Z_{j-1}}\circ\cdots T_{Z_1}(P_0)\,, \quad (1.250)$$

$$R^*_{j-1} = T_{Z_{j-1}}\circ\cdots T_{Z_1}(R^*_0) \quad (1.251)$$

and applying (1.73).

An obvious generalization of (1.249) is the following:

$$\mathbb{E}\,[\xi_j|S^k_{-\infty}, Z^{j-1}_{-\infty}] = f(P_{j-1}, P_{j-1}) \pm B_2|\mu|^{j-1-k}\,. \quad (1.252)$$

Notice that by comparing the expression for $f(x,x)$ with (1.179) we have

$$\mathbb{E}\,[f(P_{j-1}, P_{j-1})] = \mathbb{E}\,[\xi_j]\,. \quad (1.253)$$

Next we show that

$$\mathbb{E}\left[\xi_j | S_{-\infty}^0, Z_{-\infty}^0\right] = \mathbb{E}\left[\xi_j\right] \pm |\mu|^{\frac{j-1}{2}} \left[2B_2 + B_3\right] , \tag{1.254}$$

where

$$B_3 = \frac{h(\delta_1) - h(\delta_2)}{2|\mu|} .$$

Denote

$$t(P_k, S_k) \stackrel{\triangle}{=} \mathbb{E}\left[f(P_{j-1}, P_{j-1}) | S_{-\infty}^k Z_{-\infty}^k\right] . \tag{1.255}$$

Then because of (1.247) and since P_k only affects the initial condition for P_{j-1} when written as (1.250), we have for arbitrary $x_0 \in [\tau, 1-\tau]$:

$$t(P_k, S_k) = t(x_0, S_k) \pm B_2 |\mu|^{j-k-1} . \tag{1.256}$$

On the other hand, as an average of $f(x,x)$ function $t(x_0, s)$ satisfies

$$0 \le t(x_0, S_k) \le \max_{x \in [0,1]} f(x,x) \le h(\delta_1) - h(\delta_2) .$$

From here and (1.48) we have

$$\mathbb{E}\left[t(x_0, S_k) | S_{-\infty}^0 Z_{-\infty}^0\right] = \frac{1}{2}[t(x_0, 1) + t(x_0, 2)] \pm \frac{h(\delta_1) - h(\delta_2)}{2} |\mu|^k ,$$

or together with (1.256):

$$\mathbb{E}\left[t(P_k, S_k) | S_{-\infty}^0 Z_{-\infty}^0\right] = \frac{1}{2}[t(x_0, 1) + t(x_0, 2)] \pm \left[\frac{h(\delta_1) - h(\delta_2)}{2} |\mu|^k + B_2 |\mu|^{j-k-1}\right] .$$

Since this is valid for any x_0 even x_0 depending on S_k we can average over x_0 distributed as $P_{P_k|S_k}$ and get

$$\mathbb{E}\left[t(P_k, S_k) | S_{-\infty}^0 Z_{-\infty}^0\right] = \mathbb{E}\left[t(P_k, S_k)\right] \pm \frac{h(\delta_1) - h(\delta_2)}{2} |\mu|^k . \tag{1.257}$$

Summing together (1.252), (1.253), (1.255), (1.256) and (1.257) we obtain that for arbitrary $0 \le k \le j-1$ we have

$$\mathbb{E}\left[\xi_j | S_{-\infty}^0 Z_{-\infty}^0\right] = \mathbb{E}\left[\xi_j\right] \pm \left[\frac{h(\delta_1) - h(\delta_2)}{2} |\mu|^k + 2B_2 |\mu|^{j-k-1}\right] .$$

Setting here $k = \lfloor j - 1/2 \rfloor$ we obtain (1.254).

With the help of (1.254) we obtain

$$\begin{aligned}
\mathbb{E}\left[\xi_0\xi_j\right] &= \mathbb{E}\left[\xi_0\mathbb{E}\left[\xi_j|S^0_{-\infty},Z^0_{-\infty}\right]\right] && (1.258)\\
&= \mathbb{E}\left[\xi_0\left(\mathbb{E}\left[\xi_j\right]\pm(2B_2+B_3)|\mu|^{\frac{j-1}{2}}\right)\right] && (1.259)\\
&= \left(\mathbb{E}\left[\xi_0\right]\right)^2\pm\mathbb{E}\left[|\xi_0|\right](2B_2+B_3)|\mu|^{\frac{j-1}{2}} && (1.260)\\
&= (C_0-C_1)^2\pm\sqrt{\mathbb{E}\left[(\xi_0)^2\right]}(2B_2+B_3)|\mu|^{\frac{j-1}{2}} && (1.261)\\
&= (C_0-C_1)^2\pm\sqrt{B_0}(2B_2+B_3)|\mu|^{\frac{j-1}{2}}\,, && (1.262)
\end{aligned}$$

where (1.261) is a Lyapunov's inequality and (1.262) is Lemma 7.

Finally, we have

$$|\mathbb{E}\left[\hat{\xi}_0\hat{\xi}_j\right]|\leq\sqrt{B_0}(2B_2+B_3)|\mu|^{\frac{j-1}{2}}\,.$$

1.18 Entropy process in GEC may be not α-mixing.

The purpose of this section is to show that the process

$$G_j=-\log P_{Z_j|Z_1^{j-1}}(Z_j|Z_1^{j-1})$$

at least for some δ_1,δ_2 and τ has the following property:

$$\sigma\{Z_1,Z_2,\dots Z_j\}=\sigma\{G_j\}\,,$$

or in other words single value of G_j determines all of the previous Z_1^j and in particular this means that the α-mixing coefficients for the G_j satisfy

$$\alpha_G(n)=1/4\,.$$

At the same time G_j's underlying time-shift transformation is a Bernoulli shift and therefore, G_j is ergodic and even (weakly) mixing.

Theorem 10 *Assume that $\tau<1/2$, $\delta_1>\delta_2$ and the following holds*

$$\begin{aligned}
&\delta_1^2(1-\delta_2)\sqrt{4(1-\delta_2)(1-\delta_1)\tau^2+(1-\tau)^2(\delta_1-\delta_2)^2}\\
&\qquad>(\delta_2-2\delta_1\delta_2+\delta_1^2)(\delta_1-\delta_2)(1-\tau)+(1-\delta_1)^2\delta_2\sqrt{4\delta_1\delta_2\tau^2+(1-\tau)^2(\delta_1-\delta_2)^2} \qquad (1.263)
\end{aligned}$$

Then, (P_{j-1},Z_j) can be computed from P_j and (R_{j-1},Z_j) can be computed from R_j. Consequently, $\{F_s,Z_s\}$ is a function of F_t for $s\leq t$ and similarly for the $\{G_s,Z_s\}$ and G_t. Finally, for the α-mixing coefficients this implies

$$\alpha_P(n)=\alpha_R(n)=\alpha_F(n)=\alpha_G(n)=1/4\qquad\forall n\geq 0\,.$$

Remark: Condition (1.263) is satisfied, for example, for $\delta_2 = 0$. In general, for any fixed δ_1 and τ, (1.263) holds for δ_2 sufficiently small. For example, for $\delta_1 = 1/2$ and $\tau = 0.1$ we have $\delta_2 \leq 0.025$.

Proof: We will only consider the case of (P_j, F_j, Z_j). First, let us show that if (P_{j-1}, Z_j) is computable from P_j then all the other claims follow. Indeed, as shown above, (P_j, Z_{j+1}) is a bijective function of F_{j+1}. Therefore, F_{j+1} determines P_j and, hence, by the claim of the theorem, determines (P_{j-1}, Z_j) which is F_j. So we have

$$\alpha_P(n) = \alpha_F(n) = 1/4.$$

By (1.61) we have

$$P_j = T_{Z_j}(P_{j-1})\,.$$

Define two fixed points (see Fig. 1.1 for an illustration):

$$T_1(x_1) = x_1\,, \tag{1.264}$$
$$T_0(x_0) = x_0\,. \tag{1.265}$$

Since almost surely there are infinitely many $Z_k = 1$ among $-\infty < k < j$ it is clear from recursion (1.61) that almost surely P_j belongs to the interval (x_0, x_1). At the same time both operators T_0 and T_1 map (x_0, x_1) into itself. Suppose that

$$T_0(x_1) < T_1(x_0)\,. \tag{1.266}$$

In this case it is easy to see that

$$T_0(x_0, x_1) \cap T_1(x_0, x_1) = \emptyset\,.$$

But then knowing the value of $T_{Z_j}(P_{j-1})$ we can exactly determine which operator was applied and what was the value of P_{j-1}. Finally, the condition (1.266) after some algebra reduces to (1.263).

To show the claim about (R_j, G_j, Z_j) we only need to notice that $R_0 = 1/2$ and $x_0 < 1/2 < x_1$. Then, R_j belongs to (x_0, x_1) and the same reasoning applies. **QED**.

References

[1] A. N. Tikhomirov, "On the convergence rate in the central limit theorem for weakly dependent random variables," *Theory of Probability and Its Applications*, Vol. XXV, No. 4, 1980.

[2] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theor. Prob. Appl.*, Vol. 7, No. 4, 1962.

[3] G. Birkhoff, "Extensions of Jentzsch's theorem.", *Transactions of the AMS*, 85:219-227, 1957.