Implementation Notes for entropy estimation based on NIST SP 800-90B non-IID track

June 23, 2023

1 Implementation notes for numerical computation of entropy estimates

1.1 Notes for the Markov estimate

In this section, we follow the convention using a column vector of probabilities $\boldsymbol{P}$ and a left stochastic matrix $\boldsymbol{T}$.

$$\boldsymbol{P} \equiv \begin{bmatrix} p_0 \\ p_1 \end{bmatrix} \tag{1}$$

$$\boldsymbol{T} \equiv \begin{bmatrix} t_{0,0} & t_{0,1} \\ t_{1,0} & t_{1,1} \end{bmatrix} \tag{2}$$

Considering a sequence of binary-valued samples of length λ, there are following four possible combinations of the first sample value and the last sample value:

a) $0 \to 0$
b) $0 \to 1$
c) $1 \to 0$
d) $1 \to 1$

For each combination, the probability of occurrence of the most likely sequence is expressed by the following equation, by using the parameters μ and ν:

a)

$$f_a(\boldsymbol{T}, \mu, \nu) \equiv p_0 t_{00}^{\nu} t_{11}^{\lambda-1-2\mu-\nu} t_{01}^{\mu} t_{10}^{\mu} \tag{3}$$

$$\begin{cases} 0 \leq 2\mu < \lambda - 1 \\ 0 \leq \nu \leq \lambda - 1 \\ 0 \leq \lambda - 1 - 2\mu - \nu \leq \lambda - 1 \end{cases} \tag{4}$$

b)

$$f_b(\boldsymbol{T}, \mu, \nu) \equiv p_0 t_{00}^{\nu} t_{11}^{\lambda-2-2\mu-\nu} t_{01}^{\mu} t_{10}^{\mu+1} \tag{5}$$

$$\begin{cases} 0 \leq 2\mu < \lambda - 1 \\ 0 \leq \nu < \lambda - 1 \\ 0 \leq \lambda - 2 - 2\mu - \nu \leq \lambda - 2 \end{cases} \tag{6}$$

c)

$$f_c(\boldsymbol{T}, \mu, \nu) \equiv p_1 t_{00}^{\nu} t_{11}^{\lambda-2-2\mu-\nu} t_{01}^{\mu+1} t_{10}^{\mu} \tag{7}$$

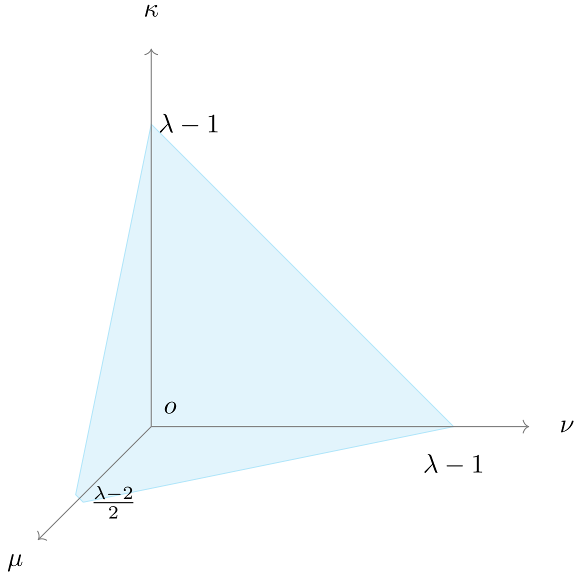
$$\begin{cases} 0 \leq 2\mu < \lambda - 1 \\ 0 \leq \nu < \lambda - 1 \\ 0 \leq \lambda - 2 - 2\mu - \nu \leq \lambda - 2 \end{cases} \tag{8}$$

d)

$$f_d(\boldsymbol{T}, \mu, \nu) \equiv p_1 t_{00}^{\nu} t_{11}^{\lambda-1-2\mu-\nu} t_{01}^{\mu} t_{10}^{\mu} \tag{9}$$

$$\begin{cases} 0 \leq 2\mu < \lambda - 1 \\ 0 \leq \nu \leq \lambda - 1 \\ 0 \leq \lambda - 1 - 2\mu - \nu \leq \lambda - 1 \end{cases} \tag{10}$$

If we denote κ as the exponent of t_{11}, then the possible values of μ, ν, and κ are expressed as integer coordinates in the approximately triangular region shown in the following figure:

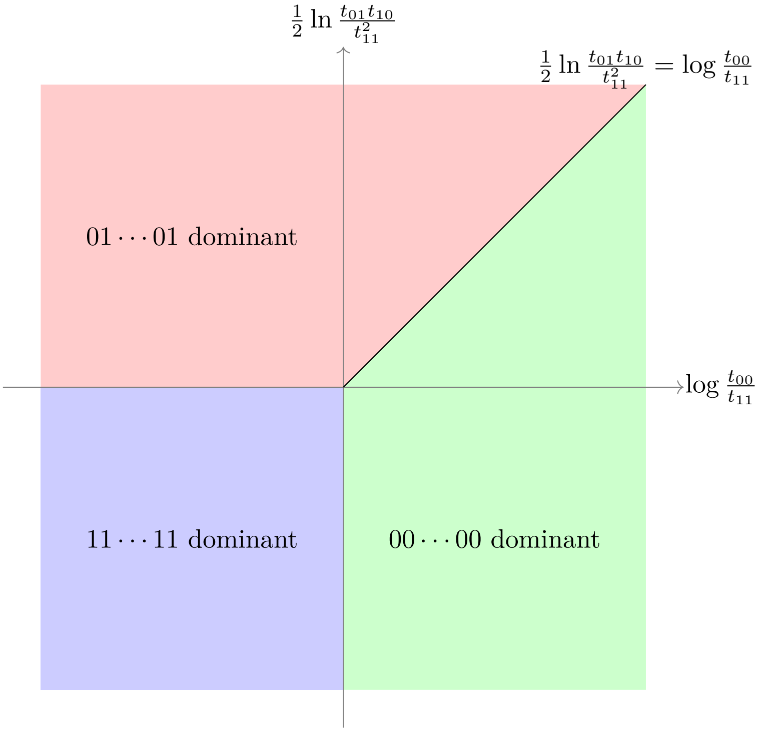


In order to calculate the extreme values of f_a, f_b, f_c, and f_d, we first consider the partial derivatives of $f_a(\boldsymbol{T}, \mu, \nu)$ with respect to μ, and ν.

$$\begin{aligned} \frac{\partial}{\partial \mu} f_a(\boldsymbol{T}, \mu, \nu) &= (-2 \ln t_{11} + \ln t_{01} + \ln t_{10}) f_a(\boldsymbol{T}, \mu, \nu) \\ &= \ln \frac{t_{01} t_{10}}{t_{11}^2} f_a(\boldsymbol{T}, \mu, \nu) \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial}{\partial \nu} f_a(\boldsymbol{T}, \mu, \nu) &= (\ln t_{00} - \ln t_{11}) f_a(\boldsymbol{T}, \mu, \nu) \\ &= \ln \frac{t_{00}}{t_{11}} f_a(\boldsymbol{T}, \mu, \nu) \end{aligned} \tag{12}$$

Taking into accout eq. (11) and eq. (12), and the property that the value $f_a(\boldsymbol{T}, \mu, \nu)$ is non-negative, the following rough phase diagram would be obtained, by taking the horizontal axis as $\log \frac{t_{00}}{t_{11}}$, and the vertical axis $\frac{1}{2} \log \frac{t_{01}t_{10}}{t_{11}^2}$.


$\frac{1}{2} \ln \frac{t_{01}t_{10}}{t_{11}^2}$
$\frac{1}{2} \ln \frac{t_{01}t_{10}}{t_{11}^2} = \log \frac{t_{00}}{t_{11}}$
$01 \cdots 01$ dominant
$\log \frac{t_{00}}{t_{11}}$
$11 \cdots 11$ dominant
$00 \cdots 00$ dominant

For example, if we consider the third quadrant, the conditional probability t_{11} is greater than or equal to t_{00} and $\sqrt{t_{01}t_{10}}$, then the sequence $11 \cdots 11$ will be dominant.

1.2 Notes for the collision estimate

In this section, we try to rewrite the following equation using elementary functions:

$$F(1/z) = \Gamma(3, z) z^{-3} \exp(z) \tag{13}$$

In NIST SP 800-90B, it is documented to evaluate incomplete gamma function using continued fraction. However, we try to obtain simpler expression using known properties of incomplete gamma function without using continued fraction.

First we try to use the following equations (see 8.4.8 and 8.4.11 in [2]).

$$\Gamma(n+1, z) = n! \exp(-z) e_n(z), \tag{14}$$

$$e_n(z) = \sum_{k=0}^{n} \frac{z^k}{k!} \tag{15}$$

With these two equations, $F(1/z)$ can be rewritten to as follows:

$$\begin{aligned}
F(1/z) &= \Gamma(3,z)z^{-3}\exp(z) \\
&= 2!\exp(-z)e_2(z)z^{-3}\exp(z) \\
&= 2z^{-3}e_2(z) \\
&= 2z^{-3}\sum_{k=0}^{2}\frac{z^k}{k!} \\
&= 2z^{-3}\left(1+z+\frac{z^2}{2}\right) \\
&= z^{-1}\left(2z^{-2}+2z^{-1}+1\right) \qquad (16)
\end{aligned}$$

By replacing the argument $1/z$ with q, the following expression can be obtained:

$$F(q) = q\left(2q^2+2q+1\right) \qquad (17)$$

From this expression, the range of $F(q)$ is $[0, \frac{5}{4}]$, with respect to the domain $q \in [0, 0.5]$.

If we denote $g(p)$ as the right hand side (RHS) of equation to be solved, in step 7 of 6.3.2 of NIST SP 800-90B[1], $g(p)$ can be rewritten by using Eq.(17).

$$\begin{aligned}
g(p) &\equiv pq^{-2}\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right]F(q) - pq^{-1}\frac{1}{2}(p^{-1}-q^{-1}) \\
&= pq^{-2}\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right]q\left(2q^2+2q+1\right) - pq^{-1}\frac{1}{2}(p^{-1}-q^{-1}) \\
&= pq^{-1}\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right]\left(2q^2+2q+1\right) - pq^{-1}\frac{1}{2}(p^{-1}-q^{-1}) \\
&= pq^{-1}\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right] - pq^{-1}\frac{1}{2}(p^{-1}-q^{-1}) \\
&+ pq^{-1}\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right]\left(2q^2+2q\right) \\
&= pq^{-1} + 2p\left[1+\frac{1}{2}(p^{-1}-q^{-1})\right](q+1) \\
&= pq^{-1} + 2p\left[(q+1)+\frac{1}{2}p^{-1}(q+1)-\frac{1}{2}(1+q^{-1})\right] \\
&= pq^{-1} + p\left[2(q+1)+p^{-1}(q+1)-(1+q^{-1})\right] \\
&= p\left[2(q+1)+p^{-1}(q+1)-1\right] \\
&= p\left[2q+1+p^{-1}(q+1)\right] \\
&= (2pq+p+q+1) \\
&= (2pq+2) \\
&= 2(pq+1) \\
&= 2\left[p(1-p)+1\right] \\
&= 2\left[-\left(p-\frac{1}{2}\right)^2+\frac{5}{4}\right] \qquad (18)
\end{aligned}$$

Figure 1 shows Eq.(18) graphically.

If $\bar{X}'$ is in range $[2, \frac{5}{2}]$, the solution p of step 7 of 6.3.2 of NIST SP 800-90B[1] can be expressed by the

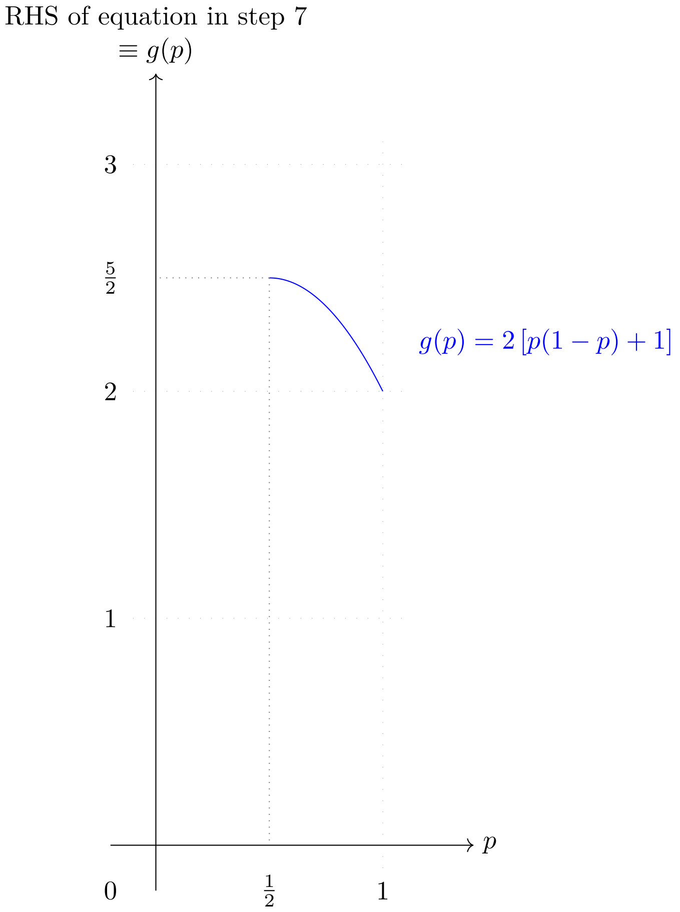


Figure 1 The right hand side of the equation in step 7 of 6.3.2 of NIST SP 800-90B

following equation:

$$p = \frac{1}{2} + \sqrt{\frac{5}{4} - \frac{\bar{X}'}{2}}. \tag{19}$$

1.3 Notes for the compression estimate

As documented in [3], Eq.(20) should be replaced by Eq.21.

$$G(z) = \frac{1}{\nu}\sum_{t=d+1}^{L}\sum_{u=1}^{t}\log_2(u)F(z,t,u) \tag{20}$$

$$G(z) = \frac{1}{\nu}\sum_{t=d+1}^{\lfloor L/b\rfloor}\sum_{u=2}^{t}\log_2(u)F(z,t,u) \tag{21}$$

$F(z,t,u)$ in Eq.(21) can be expressed by the following equation:

$$F(z,t,u) = \begin{cases} z^2(1-z)^{u-1} & \text{if} \quad u < t \\ z(1-z)^{t-1} & \text{if} \quad u = t \end{cases} \tag{22}$$

We have to evaluate Eq.(21), but this expression requires relatively large computational complexity due to its nested summation. First Figure 2 shows parameters u, t, where nested summation must be taken to compute $G(z)$.

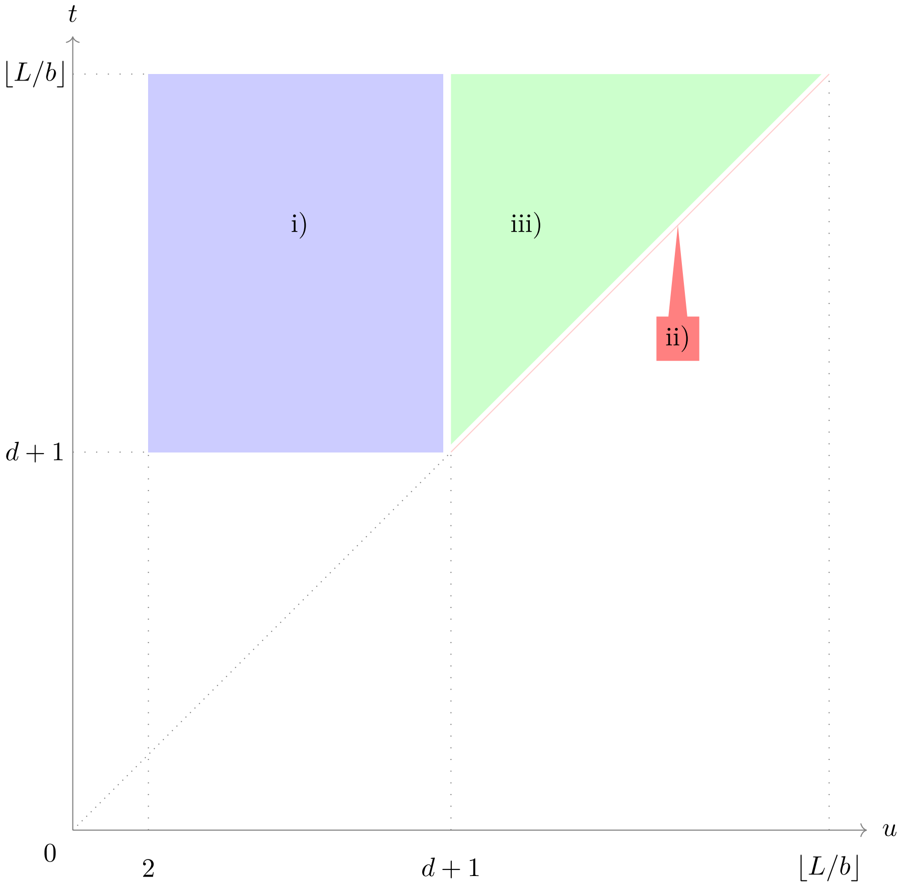


Figure 2 Parameters u, t, where nested summation must be taken

t-dependency of $F(z,t,u)$ must be considered when $u = t$. From the above, the parameters domain where summation must be taken can be divided into the following three groups:

i)

$$\begin{cases} 2 \leq u < d+1 \\ d+1 \leq t \leq \lfloor L/b \rfloor \end{cases} \tag{23}$$

ii)

$$\begin{cases} (u,t) = (\tau,\tau) \\ d+1 \leq \tau \leq \lfloor L/b \rfloor \end{cases} \tag{24}$$

iii)

$$\begin{cases} d+1 \leq u < t \\ d+1 < t \leq \lfloor L/b \rfloor \end{cases} \tag{25}$$

For groups i) and iii), the summation with respect to t can be taken first. By using the relation $\nu = \lfloor L/b \rfloor - d$, Eq.(21) can be rewritten to as follows:

$$\begin{aligned} G(z) = &\sum_{u=2}^{d} \log_2(u) z^2 (1-z)^{u-1} \\ &+ \frac{1}{\nu} \sum_{u=d+1}^{\lfloor L/b \rfloor} \log_2(u) z (1-z)^{u-1} \\ &+ \frac{1}{\nu} \sum_{u=d+1}^{\lfloor L/b \rfloor - 1} (\lfloor L/b \rfloor - u) \log_2(u) z^2 (1-z)^{u-1} \end{aligned} \tag{26}$$

With this expression, the computational complexity can be decreased from $\mathcal{O}(\nu^2)$ to $\mathcal{O}(\nu)$.

1.4 Notes for t-Tuple estimate, LRS estimate, MultiMMC Prediction and LZ78Y prediction estimate

In the following groups of entropy estimate, we have to handle t-tuples (pairs, triples, etc.)

a) t-Tuple estimate (NIST SP 800-90B 6.3.5)
b) Longest Repeated Substring (LRS) estimate (NIST SP 800-90B 6.3.6)
c) Multi Most Common in Window Prediction estimate (NIST SP 800-90B 6.3.7)
d) The MultiMMC Prediction estimate (NIST SP 800-90B 6.3.9)
e) The LZ78Y Prediction estimate (NIST SP 800-90B 6.3.10)

In order to express t-tuples, bitset or multiprecision integer is used without using array.

1.5 Notes for Multi Most Common in Window prediction estimate

In step 3-a-i of 6.3.7 of NIST SP 800-90B, it is required to compute the mode in the previous window of w_j before s_i. As the order of w_j is 1000, the computational complexity is estimated to be about $\mathcal{O}(1000n)$, and expected to be time-consuming. We attempt to reduce time-complexity by preparing histograms of certain lengths in advance, and in step 3-a-i, we use the histograms that fit within the target window, and compute the unavailable parts on-demand basis.

2 References

[1] Meltem Sönmez Turan, Elaine Barker, John Kelsey, Kerry A. McKay, Mary L. Baish, Mike Boyle *Recommendation for the Entropy Sources Used for Random Bit Generation*, NIST Special Publication 800-90B, Jan. 2018

[2] Franck W. J. Oliver, Daniel W. Lozier, Ronald F. Boisvert, Charles W. Clark, *NIST Handbook of Mathematical Functions*, National Institute of Standards and Technology, 2010

[3] G. Sakurai, *Proposed list of corrections for NIST SP 800-90B 6.3 Estimators*, Dec. 2022 `https://github.com/g-g-sakura/AnotherEntropyEstimationTool/blob/main/documentation/ProposedListOfCorrections_SP800-90B.pdf`