

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

March 4, 2023

1 Implementation notes for numerical computation of entropy estimates

1.1 Notes for the Markov estimate

To be documented.

1.2 Notes for the collision estimate

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

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

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]).

$$\Gamma(n+1, z) = n! \exp(-z) e_n(z), \quad (2)$$

$$e_n(z) = \sum_{k=0}^n \frac{z^k}{k!} \quad (3)$$

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

$$\begin{aligned} F(1/z) &= \Gamma(3, z)z^{-3} \exp(z) \\ &= 2! \exp(-z) e_2(z) z^{-3} \exp(z) \\ &= 2z^{-3} e_2(z) \\ &= 2z^{-3} \sum_{k=0}^2 \frac{z^k}{k!} \\ &= 2z^{-3} \left(1 + z + \frac{z^2}{2} \right) \\ &= z^{-1} (2z^{-2} + 2z^{-1} + 1) \end{aligned} \quad (4)$$

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

$$F(q) = q(2q^2 + 2q + 1) \quad (5)$$

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.(5).

$$\begin{aligned}
g(p) &\equiv pq^{-2} \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] F(q) - pq^{-1} \frac{1}{2}(p^{-1} - q^{-1}) \\
&= pq^{-2} \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] q(2q^2 + 2q + 1) - pq^{-1} \frac{1}{2}(p^{-1} - q^{-1}) \\
&= pq^{-1} \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] (2q^2 + 2q + 1) - pq^{-1} \frac{1}{2}(p^{-1} - q^{-1}) \\
&= pq^{-1} \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] - pq^{-1} \frac{1}{2}(p^{-1} - q^{-1}) \\
&\quad + pq^{-1} \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] (2q^2 + 2q) \\
&= pq^{-1} + 2p \left[1 + \frac{1}{2}(p^{-1} - q^{-1}) \right] (q + 1) \\
&= pq^{-1} + 2p \left[(q + 1) + \frac{1}{2}p^{-1}(q + 1) - \frac{1}{2}(1 + q^{-1}) \right] \\
&= pq^{-1} + p [2(q + 1) + p^{-1}(q + 1) - (1 + q^{-1})] \\
&= p [2(q + 1) + p^{-1}(q + 1) - 1] \\
&= p [2q + 1 + p^{-1}(q + 1)] \\
&= (2pq + p + q + 1) \\
&= (2pq + 2) \\
&= 2(pq + 1) \\
&= 2[p(1 - p) + 1] \\
&= 2 \left[-\left(p - \frac{1}{2}\right)^2 + \frac{5}{4} \right] \quad (6)
\end{aligned}$$

Figure 1 shows Eq.(6) 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 following equation:

$$p = \frac{1}{2} + \sqrt{\frac{5}{4} - \frac{\bar{X}'}{2}}. \quad (7)$$

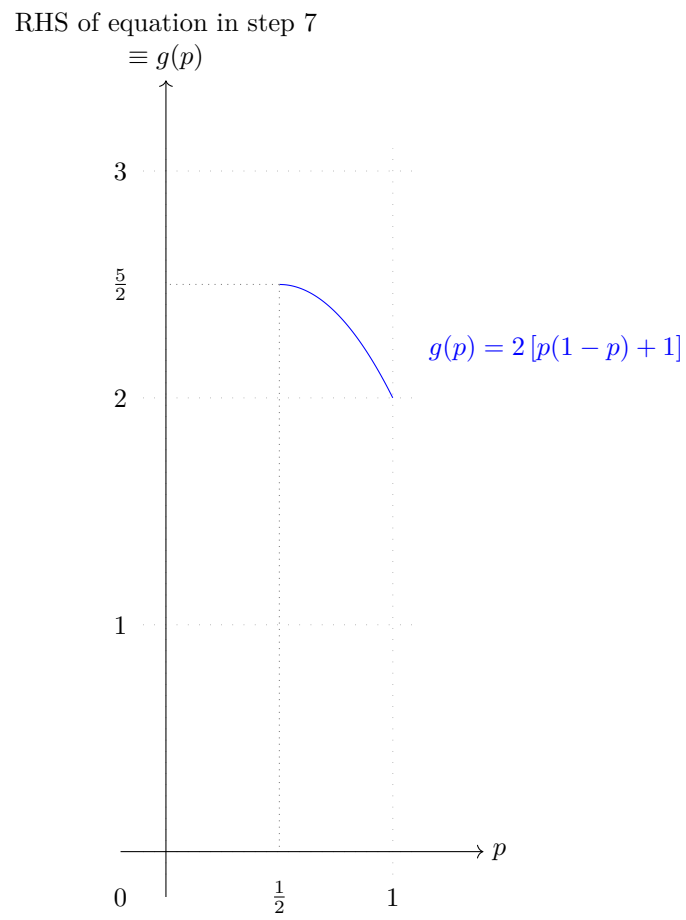


Figure 1 The right hand side of the equation in step 7 of 6.3.2 of NIST SP 800-90B

1.3 Notes for the compression estimate

As documented in [3], Eq.(8) should be replaced by Eq.9.

$$G(z) = \frac{1}{\nu} \sum_{t=d+1}^L \sum_{u=1}^t \log_2(u) F(z, t, u) \quad (8)$$

$$G(z) = \frac{1}{\nu} \sum_{t=d+1}^{\lfloor L/b \rfloor} \sum_{u=2}^t \log_2(u) F(z, t, u) \quad (9)$$

$F(z, t, u)$ in Eq.(9) can be expressed by the following equation:

$$F(z, t, u) = \begin{cases} z^2(1-z)^{u-1} & \text{if } u < t \\ z(1-z)^{t-1} & \text{if } u = t \end{cases} \quad (10)$$

We have to evaluate Eq.(9), 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 $G(z)$.

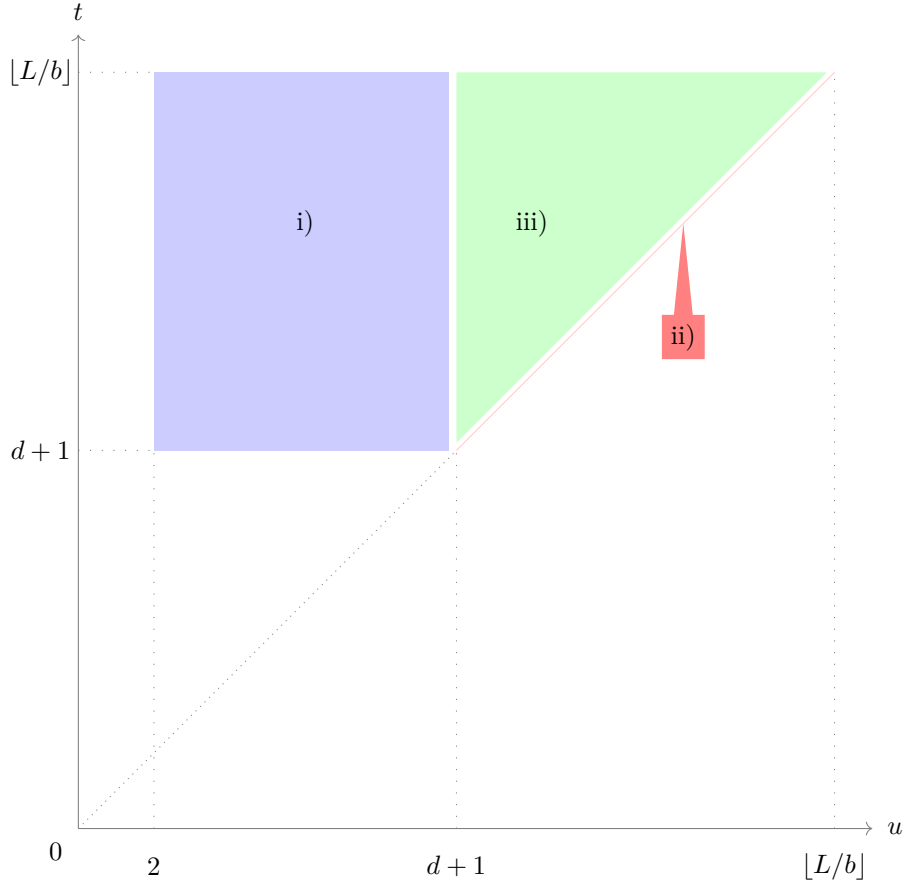


Figure 2 Parameters u, t , where nested summation must be taken

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)

$$\begin{cases} 2 \leq u < d+1 \\ d+1 \leq t \leq \lfloor L/b \rfloor \end{cases} \quad (11)$$

ii)

$$\begin{cases} (u, t) = (\tau, \tau) \\ d+1 \leq \tau \leq \lfloor L/b \rfloor \end{cases} \quad (12)$$

iii)

$$\begin{cases} d+1 \leq u < t \\ d+1 < t \leq \lfloor L/b \rfloor \end{cases} \quad (13)$$

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.(9) can be rewritten to as follows:

$$\begin{aligned} G(z) &= \sum_{u=2}^d \log_2(u) z^2 (1-z)^{u-1} \\ &\quad + \frac{1}{\nu} \sum_{u=d+1}^{\lfloor L/b \rfloor} \log_2(u) z (1-z)^{u-1} \\ &\quad + \frac{1}{\nu} \sum_{u=d+1}^{\lfloor L/b \rfloor - 1} (\lfloor L/b \rfloor - u) \log_2(u) z^2 (1-z)^{u-1} \end{aligned} \quad (14)$$

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.

- [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