Appendix

Proposition 2. For any solution (x_l, x_u) with $\beta \leq \gamma$, $x_l \in X$ and $x_u \in X$, there always exists another solution (x_l, x'_u) having $C(x_l, x'_u) \leq C(x_l, x_u)$, where $x'_u = \min X_c + 2^\beta$.

Proof. According to $\beta = \lceil \log(\max X_c - \min X_c + 1) \rceil$ in Formula 8, we have

$$\begin{aligned}\log(\max X_c - \min X_c + 1) &\leq \beta\\ \max X_c - \min X_c + 1 &\leq 2^\beta\\ \max X_c &\leq \min X_c + 2^\beta - 1\\ \max X_c &< x'_u.\end{aligned}$$

(1) For $x_u > x'_u$, it follows $\max X_c < x'_u < x_u = \min X_u$. Since there is no value between $\max X_c$ and $\min X_u$ in X, according to Definitions 2 and 4, we have $\min X'_u = \min X_u$.

(2) For $x_u \leq x'_u$, referring to Definition 4, we have $\max X'_u \geq \max X_u$.

Combining the above two cases, we can conclude that

$$\min X'_u \geq \min X_u.$$

For $n_u = |X_u|$ and $n'_u = |X'_u|$ introduced after Definition 4, it follows $n_u \geq n'_u$.

Let $n_\Delta = |X_u \setminus X'_u|$ be the size of the increment, having $n_\Delta = n_u - n'_u \geq 0$.

Given the same x_l and the corresponding identical X_l, n_l, we could get the difference C_Δ between $C(x_l, x'_u)$ and $C(x_l, x_u)$ defined in Formula 5,

$$\begin{aligned}C_\Delta &= C(x_l, x'_u) - C(x_l, x_u)\\ &= n_l(\lceil \log(\max X_l - x_{\min} + 1) \rceil + 1)\\ &\quad + n'_u(\lceil \log(x_{\max} - \min X'_u + 1) \rceil + 1)\\ &\quad + (n - n_l - n'_u)\lceil \log(\max X'_c - \min X'_c + 1) \rceil\\ &\quad - n_l(\lceil \log(\max X_l - x_{\min} + 1) \rceil + 1)\\ &\quad - n_u(\lceil \log(x_{\max} - \min X_u + 1) \rceil + 1)\\ &\quad - (n - n_l - n_u)\lceil \log(\max X_c - \min X_c + 1) \rceil. \quad (12)\end{aligned}$$

The same x_l also infers $\min X'_c = \min X_c$. Together with $n_u = n'_u + n_\Delta$, we have

$$C_\Delta = C_1 - C_2, \quad (13)$$

where

$$\begin{aligned}C_1 &= (n - n_l - n_u)\lceil \log(\max X'_c - \min X_c + 1) \rceil\\ &\quad + n_\Delta \lceil \log(\max X'_c - \min X_c + 1) \rceil\\ &\quad + n'_u(\lceil \log(x_{\max} - \min X'_u + 1) \rceil + 1)\end{aligned}$$

and

$$\begin{aligned}C_2 &= (n - n_l - n_u)\lceil \log(\max X_c - \min X_c + 1) \rceil\\ &\quad - n_\Delta(\lceil \log(x_{\max} - \min X_u + 1) \rceil + 1)\\ &\quad - n'_u(\lceil \log(x_{\max} - \min X_u + 1) \rceil + 1)\\ &= (n - n_l - n_u)\beta - n_\Delta(\gamma + 1) - n'_u(\gamma + 1).\end{aligned}$$

(i) Referring to Definition 2, we have $\max X'_c < x'_u = \min X_c + 2^\beta$. It follows

$$\begin{aligned}\log(\max X'_c - \min X_c + 1) &\leq \log(2^\beta)\\ \lceil \log(\max X'_c - \min X_c + 1) \rceil &\leq \beta.\end{aligned}$$

(ii) With the aforesaid proved $\min X'_u \geq \min X_u$, we infer

$$\lceil \log(x_{\max} - \min X'_u + 1) \rceil \leq \log(x_{\max} - \min X_u + 1) \rceil = \gamma.$$

Applying the above two conditions, we further derive

$$\begin{aligned}C_\Delta &\leq (n - n_l - n_u)\beta + n_\Delta \beta + n'_u(\gamma + 1)\\ &\quad - (n - n_l - n_u)\beta - n_\Delta(\gamma + 1) - n'_u(\gamma + 1)\\ &= n_\Delta(\beta - \gamma - 1) \leq 0.\end{aligned}$$

Given $\beta \leq \gamma$ and $n_\Delta \geq 0$, the conclusion is proved. □

Proposition 3. For any solution (x_l, x_u) with $\beta > \gamma$, $x_l \in X$ and $x_u \in X$, there always exists another solution (x_l, x'_u) having $C(x_l, x'_u) \leq C(x_l, x_u)$, where $x'_u = x_{\max} - 2^\gamma + 1$.

Proof. According to $\gamma = \lceil \log(x_{\max} - \min X_u + 1) \rceil$ in Formula 9, we have

$$\begin{aligned}\log(x_{\max} - \min X_u + 1) &\leq \gamma\\ x_{\max} - 2^\gamma + 1 &\leq \min X_u\\ x'_u &\leq \min X_u = x_u.\end{aligned}$$

(1) For $x'_u = x_u = \min X_u$, it is exactly the (x_l, x_u) solution, having $\min X'_u = x'_u = \min X_u, \max X'_c = \max X_c$.

(2) For $\max X_c < x'_u < x_u = \min X_u$, since there is no value between $\max X_c$ and $\min X_u$ in X, according to Definitions 2 and 4, we have $\min X'_u = \min X_u$, $\max X'_c = \max X_c$ as well.

(3) For $x'_u \leq \max X_c < x_u$, referring to Definitions 2 and 4, it follows $\max X'_c < \min X'_u \leq \max X_c < \min X_u$.

Combining the above three cases, we can infer that

$$\begin{aligned}\min X'_u &\leq \min X_u\\ \max X'_c &\leq \max X_c.\end{aligned}$$

For $n_u = |X_u|$ and $n_u = |X'_u|$ introduced after Definition 4, it follows $n'_u \geq n_u$. Let $n_\Delta = |X'_u \setminus X_u|$ be the size of the increment, having $n_\Delta = n'_u - n_u \geq 0$.

Given the same x_l and the corresponding identical X_l, n_l, we could get the difference C_Δ between $C(x_l, x'_u)$ and $C(x_l, x_u)$ defined in Formula 5,

$$\begin{aligned}C_\Delta &= C(x_l, x'_u) - C(x_l, x_u)\\ &= n_l(\lceil \log(\max X_l - x_{\min} + 1) \rceil + 1)\\ &\quad + n'_u(\lceil \log(x_{\max} - \min X'_u + 1) \rceil + 1)\\ &\quad + (n - n_l - n'_u)\lceil \log(\max X'_c - \min X'_c + 1) \rceil\\ &\quad - n_l(\lceil \log(\max X_l - x_{\min} + 1) \rceil + 1)\\ &\quad - n_u(\lceil \log(x_{\max} - \min X_u + 1) \rceil + 1)\\ &\quad - (n - n_l - n_u)\lceil \log(\max X_c - \min X_c + 1) \rceil.\end{aligned}$$

The same x_l also infers $\min X'_c = \min X_c$. Together with $n'_u = n_u + n_\Delta$, we have

$$C_\Delta = C_1 - C_2,$$

where

$$\begin{aligned} C_1 &= (n - n_l - n'_u)\lceil \log(\max X'_c - \min X_c + 1)\rceil \\ &\quad + n_\Delta(\lceil \log(x_{\max} - \min X'_u + 1)\rceil + 1) \\ &\quad + n_u(\lceil \log(x_{\max} - \min X'_u + 1)\rceil + 1) \end{aligned}$$

and

$$\begin{aligned} C_2 &= (n - n_l - n'_u)\lceil \log(\max X_c - \min X_c + 1)\rceil \\ &\quad - n_\Delta\lceil \log(\max X_c - \min X_c + 1)\rceil \\ &\quad - n_u(\lceil \log(x_{\max} - \min X_u + 1)\rceil + 1) \\ &= (n - n_l - n'_u)\beta - n_\Delta\beta - n_u(\gamma + 1). \end{aligned}$$

(i) Referring to Definition 2, we have $x'_u = x_{\max} - 2^\gamma + 1 \leq \min X'_u$. It follows

$$\begin{aligned} \log(x_{\max} - \min X'_u + 1) &\leq \log(2^\gamma) \\ \lceil \log(x_{\max} - \min X'_u + 1)\rceil &\leq \gamma. \end{aligned}$$

(ii) With the aforesaid proved $\max X'_c \leq \max X_c$, we infer

$$\lceil \log(\max X'_c - \min X_c + 1)\rceil \leq \log(\max X_c - \min X_c + 1)\rceil = \beta.$$

Applying the above two conditions, we further derive

$$\begin{aligned} C_\Delta &\leq (n - n_l - n'_u)\beta + n_\Delta(\gamma + 1) + n_u(\gamma + 1) \\ &\quad - (n - n_l - n'_u)\beta - n_\Delta\beta - n_u(\gamma + 1) \\ &= n_\Delta(\gamma + 1 - \beta) \leq 0. \end{aligned}$$

Given $\beta > \gamma$ and $n_\Delta \geq 0$, the conclusion is proved. ☐

Proposition 4. For normal distribution $X \sim N(\mu, \sigma^2)$, the approximation ratio ρ of BOS-M satisfies

$$\rho \leq \begin{cases} 2 & \text{if } \sigma \leq \frac{5}{3}, \\ \lceil \log(3\sigma - 1)\rceil & \text{otherwise.} \end{cases}$$

Proof. (1) Firstly, we prove the upper bound of the storage cost C_{approx} for BOS-M,

$$C_{\text{approx}} \leq \begin{cases} \lceil \log(6\sigma + 1)\rceil n & \text{if } \sigma < \frac{1}{2}, \\ 2n & \text{if } \frac{1}{2} \leq \sigma \leq \frac{5}{3}, \\ \lceil \log(3\sigma - 1)\rceil n & \text{otherwise.} \end{cases}$$

For normal distribution $X \sim N(\mu, \sigma^2)$, the median is μ, the maximum value $x_{\max}$ is approximately $\mu + 3\sigma$, and the minimum value $x_{\min}$ is approximately $\mu - 3\sigma$, which correspond to the 99.7% confidence interval under the empirical rule (within three standard deviations from the mean).

The storage cost C_β of BOS-M with bit-width β is

$$\begin{aligned} C_\beta &= C(\mu - 2^\beta, \mu + 2^\beta) \\ &= n_l\lceil \log(\mu - 2^\beta - x_{\min} + 1)\rceil \\ &\quad + n_u\lceil \log(x_{\max} - (\mu + 2^\beta) + 1)\rceil \\ &\quad + (n - n_l - n_u)\lceil \log((\mu + 2^\beta) - (\mu - 2^\beta) + 1)\rceil \\ &= n_l\lceil \log(3\sigma - 2^\beta + 1)\rceil + n_u\lceil \log(3\sigma - 2^\beta + 1)\rceil \\ &\quad + (n - n_l - n_u)(\beta + 1) \\ &= (n_l + n_u)\lceil \log(3\sigma - 2^\beta + 1)\rceil \\ &\quad + (n - n_l - n_u)(\beta + 1). \end{aligned}$$

With β increasing from 1 to $\lceil \log(6\sigma + 1)\rceil$, bit-widths of 3 parts of values decrease firstly, and then center values become smaller. Thus, C_β firstly decreases and then increases. The upper bound of C_{approx} is thus

$$C_{\text{approx}} \leq \min\{C_{\lceil \log(6\sigma+1)\rceil}, C_1\},$$

where

$$C_{\lceil \log(6\sigma+1)\rceil} = \lceil \log(6\sigma + 1)\rceil n,$$

and

$$C_1 = (n_l + n_u)\lceil \log(3\sigma - 1)\rceil + 2(n - n_l - n_u).$$

Considering 4 different cases below, we rewrite C_1 as

$$C_1 = 2n + (\lceil \log(3\sigma - 1)\rceil - 2)(n_l + n_u).$$

a) When $\sigma \leq \frac{1}{2}$ ($\mu - 2 < \mu - 3\sigma$), i.e., there are no upper and lower outliers with $\beta = 1$, we have $n_l + n_u = 0$ and $C_1 = 2n \geq \lceil \log(6\sigma + 1)\rceil n = C_{\lceil \log(6\sigma+1)\rceil}$.

b) When $\frac{1}{2} \leq \sigma < \frac{2}{3}$ ($\mu - 2 < \mu - 3\sigma$), i.e., there are no upper and lower outliers with $\beta = 1$, we have $n_l + n_u = 0$ and $C_1 = 2n < C_{\lceil \log(6\sigma+1)\rceil}$.

c) When $\frac{2}{3} \leq \sigma \leq \frac{5}{3}$, we have $0 \leq \lceil \log(3\sigma - 1)\rceil \leq 2$ and $2(n - n_l - n_u) \leq C_1 \leq 2n < C_{\lceil \log(6\sigma+1)\rceil}$.

d) When $\sigma > \frac{5}{3}$, we have $\lceil \log(3\sigma - 1)\rceil \geq 2$ and C_1 increases with σ growing. Then, when σ tends to positive infinity, we have there are no center values and $C_1 = \lceil \log(3\sigma - 1)\rceil n < C_{\lceil \log(6\sigma+1)\rceil}$.

Therefore, we conclude that

$$C_{\text{approx}} \leq \begin{cases} \lceil \log(6\sigma + 1)\rceil n & \text{if } \sigma < \frac{1}{2}, \\ 2n & \text{if } \frac{1}{2} \leq \sigma \leq \frac{5}{3}, \\ \lceil \log(3\sigma - 1)\rceil n & \text{otherwise.} \end{cases}$$

(2) Moreover, we can prove that $C_{\text{opt}} \geq n$.

The storage cost is larger than the sum of bit-width for each value, thus the optimal cost C_{opt} has

$$C_{\text{opt}} = \sum_{i=1}^{n} b_i,$$

where

$$b_i = \begin{cases} 1 & \text{if } x_i = x_{\min}, \\ \lceil \log(x_i - x_{\min} + 1)\rceil & \text{otherwise.} \end{cases}$$

Thus, we have $C_{\text{opt}} \geq n$, even when all values are the same.

(3) Finally, we derive that

$$\rho = \frac{C_{\text{approx}}}{C_{\text{opt}}} \leq \begin{cases} \lceil \log(6\sigma + 1) \rceil & \text{if } \sigma < \frac{1}{2}, \\ 2 & \text{if } \frac{1}{2} \leq \sigma \leq \frac{5}{3}, \\ \lceil \log(3\sigma - 1) \rceil & \text{otherwise.} \end{cases}$$

a) When $\sigma < \frac{1}{2}$, we have $\rho \leq \lceil \log(6\sigma + 1) \rceil \leq 2$.

b) When $\frac{1}{2} \leq \sigma \leq \frac{5}{3}$, we have $\rho \leq 2$.

c) When $\sigma > \frac{5}{3}$, we have $\rho \leq \lceil \log(3\sigma - 1) \rceil$.

To sum up, we conclude that

$$\rho = \frac{C_{\text{approx}}}{C_{\text{opt}}} \leq \begin{cases} 2 & \text{if } \sigma \leq \frac{5}{3}, \\ \lceil \log(3\sigma - 1) \rceil & \text{otherwise.} \end{cases}$$

$\square$