Analysis of the relationship between the number of views and samples

To uncover the relationship between the number of views and samples, we first explain the process of view generation. Then, we analyze and derive the relationship between the number of views and samples in theory.

I. View generation

For the generation of views, the proposed method begins with the original finest scale view, where each sample is considered as a granular ball. Then, it iteratively generates a coarser scale view by combining granular balls within the prior finer scale view until all samples are grouped into one granular ball, at which the coarsest scale view is obtained.

II. Relationship between the number of views and samples

In fact, the number of views varies with different datasets and is inherently determined by the number of samples. Across all datasets, although the number of views is different, it approximately increases with the number of samples. To uncover their relationship, theoretical analyses and experiments are performed.

Assume that the relationship between the number of granular balls in the k-th view and $(k+1)$-th view meets that

$$\lim_{|U|\to\infty} \mathbb{E}[\frac{|GBS_{k+1}|}{|GBS_k|}] = q, \tag{1}$$

where $|GBS_k|$ is the number of granular balls in the k-th view, $q(0 < q < 1)$ is a constant determined by the given data, and $|U|$ is the number of all samples.

In other words, for an infinite dataset, the ratio of the number of granular balls in the $(k+1)$-th view to that in the k-th view converges to a constant q, while this value may vary across different datasets and is determined by the specific dataset. This formula suggests that the number of granular balls decreases with the view generation process at a certain rate. Considering that the number of the original finest scale view $|GBS_1|$ is equal to the number of all samples $|U|$, some constraints can be introduced:

$$\begin{aligned} &\lim_{|U|\to\infty} \mathbb{E}[\frac{|GBS_{k+1}|}{|GBS_k|}] = q, (0 < q < 1) \\ &\text{s.t.} |GBS_k| \in N_+, 1 \le k \le K, \end{aligned} \tag{2}$$

where N_+ denotes the set of positive integers, and $K = |GBSV|$ is the number of generated views.

Further, the relationship between the number of views $K = |GBSV|$ and the number of sample $|U|$ can

be deduced as

$$\begin{aligned} & |GBS_{k+1}| = q|GBS_k|, (0 < q < 1), \\ & \text{s.t.} |GBS_k| \in N_+, 1 \leq k \leq K, \\ \Longleftrightarrow & |GBS_k| = q^{k-1}|GBS_1|, (0 < q < 1), \\ & \text{s.t.} |GBS_k| \in N_+, 1 \leq k \leq K, \\ \Longleftrightarrow & |GBS_k| = q^{k-1}|U|, (0 < q < 1), \\ & \text{s.t.} |GBS_k| \in N_+, 1 \leq k \leq K. \end{aligned} \tag{3}$$

Since there is only one granule ball in the last view, i.e., $|GBS_K| = 1$. Then, we have $1 \leq k \leq (-\log_q |U| + 1)(0 < q < 1)$ and $K = (-\log_q |U| + 1)(0 < q < 1)$. By substituting a base greater than 1, we have $K \sim \log |U|$, where the symbol $\sim$ means the similar order.

To examine the relationship between the number of views and sample size, we generated datasets with the number of samples from 50 to 20,000 at a step size of 50 and recorded the number of granular balls and views, respectively. The experimental results are shown in Table 1.

From Table 1, it can be seen that the number of generated views varies with different sizes of samples, but approximately increases with the sample size. To further validate their relationship, we conducted a statistical analysis on the data listed in Table 1, and the results are shown in Table 2.

In Table 2, the first and second columns denote the number of generated views and the average number of samples under the same number of views, respectively. The third and fourth columns represent the logarithmic value of the average sample size and the ratio of the logarithmic value of the number of generated views to the average sample size. Note that the results for the numbers of views 3, 4, 9, and 10 are presented due to insufficient data to reflect the tendency. By observing the results in the last column, the ratio values are approximately equal and converge to a constant (≈ 0.5), indicating that the number of views and the number of samples satisfy that:

$$\begin{aligned} & \frac{\log_{10} K}{|U|} \approx 0.5, \\ \Longleftrightarrow & |U| \approx 2 \log_{10} K, \\ \Longleftrightarrow & |U| \approx 2 \frac{\log_{\sqrt{10}} K}{\log_{\sqrt{10}} 10}, \\ \Longleftrightarrow & |U| \approx \log_{\sqrt{10}} K, \\ \Longleftrightarrow & |U| \approx -\log_{\frac{1}{\sqrt{10}}} K. \end{aligned} \tag{4}$$

In such case, the constant q is equal to $\frac{1}{\sqrt{10}}$. These results are consistent with the conclusion.

TABLE 1: The number of generated views with different numbers of samples.

$\|U\|$	K	$\|U\|$	K	$\|U\|$	K	$\|U\|$	K	$\|U\|$	K	$\|U\|$	K	$\|U\|$	K	$\|U\|$	K
50	3	2550	7	5050	8	7550	8	10050	8	12550	8	15050	9	17550	8
100	5	2600	7	5100	8	7600	8	10100	9	12600	8	15100	8	17600	8
150	4	2650	7	5150	7	7650	8	10150	8	12650	8	15150	8	17650	9
200	5	2700	7	5200	8	7700	7	10200	8	12700	8	15200	8	17700	9
250	6	2750	7	5250	7	7750	8	10250	8	12750	8	15250	8	17750	9
300	5	2800	7	5300	8	7800	8	10300	8	12800	8	15300	8	17800	8
350	5	2850	7	5350	8	7850	8	10350	8	12850	7	15350	7	17850	8
400	6	2900	7	5400	8	7900	8	10400	8	12900	8	15400	8	17900	9
450	5	2950	7	5450	8	7950	8	10450	8	12950	8	15450	8	17950	8
500	6	3000	7	5500	7	8000	8	10500	8	13000	8	15500	8	18000	8
550	5	3050	7	5550	8	8050	8	10550	8	13050	8	15550	9	18050	9
600	6	3100	7	5600	8	8100	8	10600	8	13100	8	15600	9	18100	8
650	6	3150	7	5650	7	8150	8	10650	8	13150	8	15650	8	18150	8
700	6	3200	8	5700	8	8200	8	10700	8	13200	8	15700	8	18200	8
750	6	3250	7	5750	7	8250	8	10750	8	13250	8	15750	8	18250	9
800	6	3300	7	5800	8	8300	8	10800	8	13300	8	15800	8	18300	9
850	7	3350	7	5850	8	8350	8	10850	8	13350	8	15850	8	18350	9
900	6	3400	7	5900	8	8400	8	10900	8	13400	8	15900	8	18400	8
950	6	3450	7	5950	7	8450	8	10950	8	13450	8	15950	8	18450	8
1000	7	3500	7	6000	8	8500	8	11000	8	13500	8	16000	8	18500	9
1050	6	3550	8	6050	8	8550	8	11050	9	13550	8	16050	9	18550	8
1100	7	3600	7	6100	8	8600	8	11100	8	13600	9	16100	8	18600	8
1150	7	3650	7	6150	8	8650	8	11150	8	13650	8	16150	8	18650	9
1200	7	3700	8	6200	8	8700	8	11200	8	13700	8	16200	8	18700	9
1250	6	3750	7	6250	7	8750	8	11250	8	13750	8	16250	8	18750	9
1300	7	3800	7	6300	8	8800	8	11300	8	13800	8	16300	9	18800	9
1350	6	3850	7	6350	8	8850	8	11350	8	13850	8	16350	8	18850	9
1400	7	3900	8	6400	8	8900	8	11400	8	13900	8	16400	8	18900	9
1450	7	3950	7	6450	8	8950	8	11450	8	13950	8	16450	9	18950	9
1500	6	4000	7	6500	8	9000	8	11500	9	14000	9	16500	8	19000	8
1550	7	4050	7	6550	8	9050	8	11550	8	14050	8	16550	8	19050	8
1600	6	4100	7	6600	8	9100	8	11600	8	14100	8	16600	9	19100	9
1650	6	4150	7	6650	7	9150	7	11650	8	14150	8	16650	8	19150	9
1700	7	4200	8	6700	8	9200	7	11700	8	14200	9	16700	9	19200	9
1750	7	4250	7	6750	7	9250	8	11750	8	14250	9	16750	8	19250	9
1800	7	4300	7	6800	8	9300	8	11800	8	14300	8	16800	8	19300	9
1850	7	4350	7	6850	8	9350	8	11850	8	14350	8	16850	9	19350	9
1900	7	4400	7	6900	8	9400	8	11900	8	14400	8	16900	9	19400	9
1950	7	4450	7	6950	8	9450	8	11950	8	14450	8	16950	8	19450	9
2000	7	4500	7	7000	7	9500	8	12000	8	14500	8	17000	9	19500	8
2050	7	4550	8	7050	7	9550	8	12050	8	14550	8	17050	9	19550	8
2100	7	4600	7	7100	8	9600	8	12100	8	14600	9	17100	8	19600	9
2150	7	4650	8	7150	7	9650	8	12150	8	14650	9	17150	9	19650	8
2200	7	4700	8	7200	8	9700	8	12200	8	14700	8	17200	9	19700	9
2250	7	4750	8	7250	8	9750	8	12250	8	14750	9	17250	8	19750	9
2300	7	4800	8	7300	7	9800	8	12300	8	14800	8	17300	8	19800	9
2350	7	4850	8	7350	8	9850	8	12350	8	14850	8	17350	9	19850	10
2400	7	4900	7	7400	8	9900	8	12400	8	14900	8	17400	9	19900	8
2450	7	4950	8	7450	8	9950	8	12450	8	14950	8	17450	8	19950	9
2500	7	5000	8	7500	8	10000	9	12500	8	15000	8	17500	9	20000	9

TABLE 2: The results of statistical analysis.

K	$\|U\|$	$\log_{10}\|U\|$	$(\log_{10} K)/\|U\|$
5	325.0000	2.5119	0.5024
6	931.2500	2.9691	0.4948
7	3621.5190	3.5589	0.5084
8	11294.0928	4.0529	0.5066