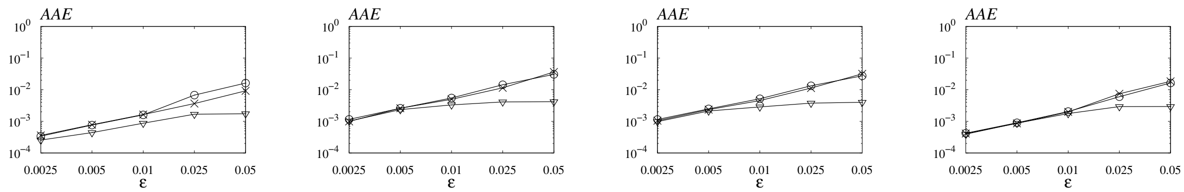
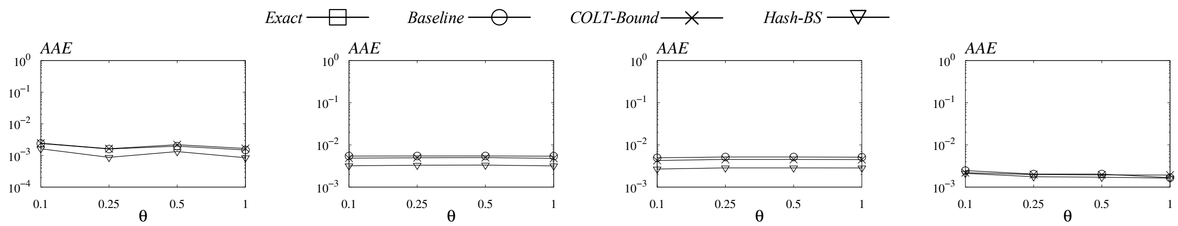


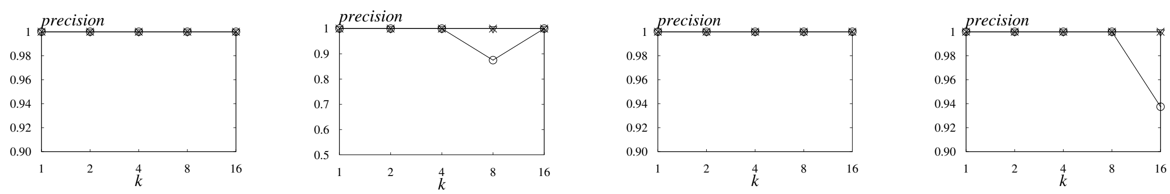
(a) enem (b) hus (c) pus (d) airline

Figure 5: Varying ϵ: Average absolute error of ϵ-approximate empirical variance algorithms.



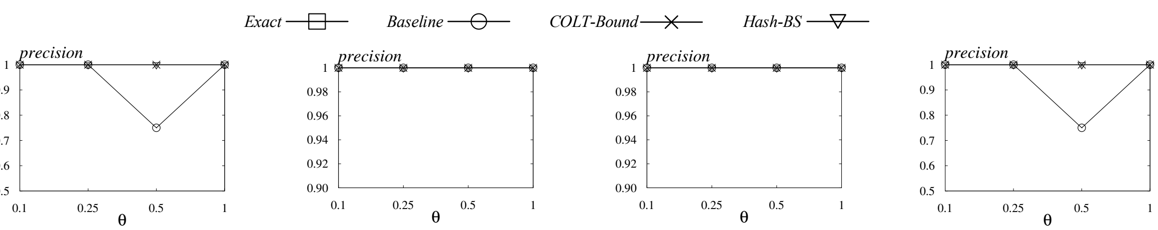
(a) enem (b) hus (c) pus (d) airline

Figure 6: Varying θ: Average absolute error of ϵ-approximate empirical variance algorithms.



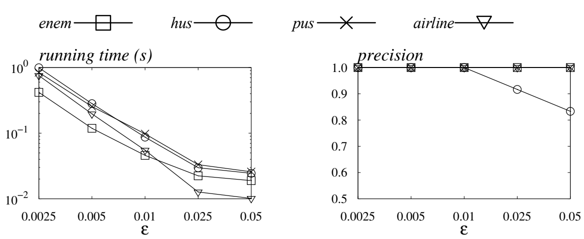
(a) enem (b) hus (c) pus (d) airline

Figure 7: Varying k: Query Precision of empirical variance top-k algorithms.



(a) enem (b) hus (c) pus (d) airline

Figure 8: Varying θ: Query Precision of empirical variance top-k algorithms.



(a) Running Time (b) Precision

Figure 9: Tuning ϵ: approximate top-k.

In the algorithm, the sample size m will double in each iteration and check whether m is large enough to satisfy the termination condition. So the algorithm terminates with $m \leq 2m^*$ with at least $1-p_f$ probability. The total number M^* of records we have ever added to the blocks for an attribute is at most $4m^*$ since we resample and double the number of records for each iteration, which is $O\left(\frac{r}{d\epsilon^2}\right)$. According to the previous analysis, $r = \lceil 4.5a \rceil$ and d is a constant with this setting where $a = \ln(1/p'_f)$. In Algorithm 1, $i_{\max} = \log_2\lceil n/b_0 \rceil$ and $p'_f = p_f/i_{\max}$. So M^* is $O(\log(\log n/p_f)/\epsilon^2)$.

The second stopping condition indicates that the number of records we have visited is no larger than the total number n' of records. Recall that θ is the ratio of the number of records satisfying the predicates to the total number of records. Since all records are hashed and we use the block sampling strategy, the ratio of M^* to the total number of sampled records is also θ in expectation. Then the expected running time of the ϵ-approximate algorithm is

$$O\left(\min\left\{n', \frac{\log(\log n/p_f)}{\epsilon^2\theta}\right\}\right),$$

which completes the proof of the theorem. □