# Quantile Estimation Based on the Principles of the Search on the Line

Anis Yazidi and Hugo Lewi Hammer

Department of Computer Science, Oslo Metropolitan University, Olso, Norway

**Abstract.** The goal of our research is to estimate the quantiles of a distribution from a large set of samples that arrive sequentially. We propose a novel quantile estimator that requires a *finite memory* and is simple to implement. Furthermore, the estimator falls under the family of incremental estimators, i.e., it utilizes the previously-computed estimates and *only* resorts to the last sample for updating these estimates. The estimator estimates the quantile on a set of discrete values. Choosing a low resolution results in fast convergence and low precision of the current estimate after convergence, while a high resolution results in slower convergence, but higher precision. The convergence results are based on the theory of Stochastic Point Location (SPL). The reader should note that the aim of the paper is to demonstrate its salient properties as a novel quantile estimator that uses only *finite* memory.

**Keywords:** *Discretized Estimation, Learning Automata, Stochastic Point Location, Quantile Estimation.*

## 1 Introduction

An incremental estimator, by definition, resorts to the last observation(s) in order to update its estimate. This is especially true of quantile estimators because they work with the samples as they come from a stochastic distribution. The research on developing incremental quantile estimators is sparse. Probably, one of the outstanding early and unique examples of incremental quantile estimators is due to Tierney, proposed in 1983 [20], and which resorted to the theory of stochastic approximation. Applications of Tierney's algorithm to network monitoring can be found in [4]. A very limited number of studies have devised *incremental* quantile estimators including the Frugal estimator due to Ma et al. [10], the higher fidelity Frugal due to Yazidi et al [26] and the DUMIQE estimator due to Yazidi and Hammer [25].

In order to appreciate the qualities of our estimator, we will present the estimator scheme proposed by Tierney [20]. Let $X$ be a random variable. Let $x(n)$ be a concrete realization of $X$ at time '$n$'. $x(n)$ is drawn from the distribution of $X$, $f_X(x)$. The intention of the exercise is to estimate the $q$-th quantile, the number $Q_q$ such that $F_X(Q_q) = q$. Tierney [20] achieved this by maintaining a running estimate $\widehat{Q_q}(n)$ at time $'n'$.

$$\widehat{Q_q}(n+1) = \widehat{Q_q}(n) + \frac{d_n}{n+1}(q - I(x(n) \leq \widehat{Q_q}(n))) \tag{1}$$

where $d_n = min(\frac{1}{\widehat{f_n(Q_q)}}, d_0 n^a)$. Here $0 < a < 1/2$, $d_o > 0$, and $f_n(Q_q)$ is an estimator of $f(Q_q)$ defined in [20]. The reason for invoking the $min$ operation in the above expression of $d(n)$ is the fact that the estimated density must be bounded to prevent the correction factor from "exploding". In other words, $f_n$ is current estimate of the density of $X$ at the $q$-th quantile. This is usually done based on maintaining a histogram structure. However, requiring the incremental constructions of local approximations of the distribution function in the neighborhood of the quantiles increases the complexity of the algorithm. Our goal is to present an algorithm that does not involve any local approximations of the distribution function. Recently, a generalization of the Tierney's [20] algorithm was proposed by [5] where the authors proposed a batch update of the quantile, where the quantile is updated every $M \geq 1$ observations.

A body of research has been focused on quantile estimation from data streams without making any specific assumption on the distribution of the data samples. We shall first review some the related work on estimating quantiles from data streams. However, as we will explain later, these related works require some memory restrictions which renders our work to be radically distinct from them. In fact, our approach requires storing only one sample value in order to update the estimate. The most representative work for this type of "streaming" quantile estimator is due to the seminal work of Munro and Paterson [11]. In [11], Munro and Paterson described a $p$-pass algorithm for selection using $O(n^{1/(2p)})$ space for any $p \geq 2$. Cormode and Muthukrishnan [6] proposed a more space-efficient data structure, called the Count-Min sketch, which is inspired by Bloom filters, where one estimates the quantiles of a stream as the quantiles of a random sample of the input. The key idea is to maintain a random sample of an appropriate size to estimate the quantile, where the premise is to select a subset of elements whose quantile approximates the true quantile. From this perspective, the latter body of research requires a certain amount of memory that increases as the required accuracy of the estimator increases [21]. Examples of these works are [2,7,11,21].

In [5], the authors proposed a modification of the stochastic approximation algorithm [20] in order to allow an update similar to the well-known Exponentially Weighted Moving Averages form for updates. This modification is particularly helpful in the case of non-stationary environments in order to cope with non-stationary data. Thus, the quantile estimate is a weighted combination of the new data that has arrived and the previously-computed estimate. Indeed, a "weighted" update scheme is applied to to incrementally build local approximations of the distribution function in the neighborhood of the quantiles. In the experiments results that we report, we consider the cases when data are generated randomly from stationary and non-stationary distributions.

This paper introduces a novel discretized quantile estimator. While we have earlier solved the binomial estimation problem using disretized estimators [23], this is the first solution to the quantile estimation.

We submit that the entire phenomenon of utilizing the concepts of disretization in quantile estimation is unexplored, and it is precisely here that we have our primary contributions.

### 1.1 Legacy SPL Solutions

To place our work in the right perspective, we briefly review the state of the art of the SPL problem, whose formulation and solution is central to our approach. The SPL problem, in its most elementary formulation, assumes that there is a Learning Mechanism (LM) whose task is to determine the optimal value of some variable (or parameter), $x$. We assume that there is an optimal choice for $x$ – an unknown value, say $x^* \in [0, 1)$. The SPL involves inferring the value $x^*$. Although the mechanism does not know the value of $x^*$, it was assumed that it has responses from an intelligent "Environment" (synonymously, referred to as the "Oracle"), $\varXi$, that is capable of informing it whether any value of $x$ is too small or too big. To render the problem both meaningful and distinct from its deterministic version, we would like to emphasize that the response from this Environment is assumed "faulty." Thus, $\varXi$ may tell us to increase $x$ when it should be decreased, and *vice versa*. However, to render the problem tangible, in [14] the probability of receiving an intelligent response was assumed to be $p > 0.5$, in which case $\varXi$ was said to be *Informative*. Note that the quantity "$p$" reflects on the "effectiveness" of the Environment. Thus, whenever the current $x < x^*$, the Environment correctly suggests that we increase $x$ with probability $p$. It simultaneously could have incorrectly recommended that we decrease $x$ with probability $(1 - p)$. The converse is true for $x \geq x^*$.

We can summarize the existing SPL-related literature as follows:

- Oommen [14] pioneered the study of the SPL when he proposed and analyzed an algorithm that operates on a discretized search space[1] while interacting with an informative Environment (i.e., $p > 0.5$). The search space was first sliced into $N$ sub-intervals at the positions $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$, where a larger value of $N$ ultimately implied a more accurate convergence to $x^*$. The algorithm then did a controlled random walk on this space by "obediently" following the Environment's advice in the discretized space. In spite of the Oracle's erroneous feedback, this discretized solution was proven to be $\epsilon$-optimal.
- An novel alternate *parallel* strategy that combined LA and pruning was used in [17] to solve the SPL. By utilizing the response from the environment, the authors of [17] partitioned the interval of search into three disjoint subintervals, eliminating at least one of the subintervals from fur-

[1] Some of the existing results about discretized automata are found in [1, 9, 12, 13, 15, 16, 19]. Indeed, the fastest reported LAs are the discretized pursuit, and discretized maximum likelihood and Bayesian estimator algorithms [1, 12, 16].

ther search, and by recursively searching the remaining interval(s) until the search interval was at least as small as the required resolution[2].

- In a subsequent work [18], Oommen *et al.* introduced the Continuous Point Location with Adaptive d-ARY Search (CPL-AdS) which was a generalization of the work in [17]. In CPL-AdS, the given search interval was subdivided into $d$ partitions representing $d$ disjoint subintervals, where $d > 3$. In each interval, initially, the midpoint of the given interval was considered to be the estimate of the unknown $x^*$. Each of the $d$ partitions of the interval was independently explored using an $\epsilon$-optimal two-action LA, where the two actions were those of selecting a point from the left or right half of the partition under consideration. Thereafter, the scheme proposed in [18] eliminated at least one of the subintervals from being searched further, and recursively searched the remaining pruned contiguous interval until the search interval was at least as small as the required resolution of estimation. Again, this elimination process essentially utilized the $\epsilon$-optimality property of the underlying LA and the monotonicity of the intervals to guarantee the convergence. By virtue of this property, at each epoch consisting of a certain number, $N_\infty$, of iterations, the algorithm could "$(1-\epsilon)$-confidently" discard regions of the search space.
- The authors of [8] proposed a rather straightforward modification of the latter CPL-AdS so as to also track changes in $x^*$. Indeed, to achieve the latter, the authors of [8] proposed to perform an *additional* parallel d-ARY search at each epoch on the original search interval. The limitation of this work is that the strategy proposed in [8] can only track $x^*$ under certain conditions relative to the frequency of change in $x^*$ and the length of an epoch. However, more importantly, the interesting facet of the solution presented in [18] is that it converges with an arbitrarily high accuracy even if the Oracle is a *stochastic compulsive liar* who is attempting to stochastically deceive the LM.
- Recently Yazidi *et al.* [22] proposed a *hierarchical* searching scheme for solving the SPL problem. The solution involves partitioning the line in a hierarchical tree-like manner, and of moving to relatively distant points, as characterized by those along the path of the tree. With regard to its advantages, this solution is an order of magnitude faster than the classical SPL solution [14]. The marginal drawback, however, is that it works under the premise that $p$ is a constant whose value is larger than the golden ratio conjugate.

## 2 Discretized Estimator

Let $Q_i = a + i.\frac{(b-a)}{N}$ and suppose that we are estimating the quantile in interval $[a, b]$, where $a > 0$. Note $Q_0 = a$ and $Q_N = b$. There is an implicit assumption that the true quantile lies in $[a, b]$. However, this is not a limitation of our

[2] The logic behind this is explained in the next item, when the authors generalized this scenario for the case when the number of partitions was $d > 3$.

scheme, the proof is valid for any bounded and probably non bounded function.

We suppose that the estimate at each time instant $\widehat{Q}(n)$ takes values from the $N+1$ values $Q_i = a + i.\frac{(b-a)}{N}$, where $0 \leq i \leq N$.

According to whether $q$ is lager or smaller than $1/2$, we will have two different estimation schemes.

- **Case 1**: $q < \frac{1}{2}$
  If $\widehat{Q}(n) \leq x(n)$ and $rand() \leq \frac{1}{2(1-q)}$
  $$\widehat{Q}(n+1) \leftarrow Min(\widehat{Q}(n) + 1/N, b) \tag{2}$$
  Else
  $$\widehat{Q}(n+1) \leftarrow Max(\widehat{Q}(n) - 1/N, a) \tag{3}$$
- **Case 2**: $q > \frac{1}{2}$
  If $\widehat{Q}(n) > x(n)$ and $rand() \leq \frac{1}{2q}$
  $$\widehat{Q}(n+1) \leftarrow Max(\widehat{Q}(n) - 1/N, a) \tag{4}$$
  Else
  $$\widehat{Q}(n+1) \leftarrow Min(\widehat{Q}(n) + 1/N, b) \tag{5}$$

where $Max(.,.)$ and $Min(.,.)$ denote the max and min operator of two real numbers while $rand()$ is a is a random number generated in $[0, 1]$.

**Theorem 1.** *We would like to estimate the q-th quantile to be estimated, i.e,* $Q^* = {F_X}^{-1}(q)$. *Applying the updating rules (2), (3), (4) and (5), we obtain:* $\lim_{n\to\infty} \lim_{N\to\infty} E(\widehat{Q}(n)) = Q^*$.

The proof of the theorem is quite involved and has been omitted here for the sake of space limitation. The full proof can be found in [24].

**Remarks**: A few remarks regarding our method for updating the estimates are not out of place. Indeed:

- First of all, it is pertinent to mention that although the rationale for updating is similar to that of the SSL algorithm [14], there are some fundamental differences. Unlike the latter, which explicitly assumes the existence of an "Oracle", in this case, our scheme simulates such an entity.
- Secondly, at this juncture, we emphasize that unlike the work of [14], the probability that the Oracle suggests the move in the correct direction, is not constant over the states of the estimator's state space. This is quite a significant difference, which basically reduces our model to a Markov Chain with state-dependent transition probabilities.

## 3 Experimental results

We compare our estimator to the the EWSA (Exponential weighted Stochastic Approximation) due to Chen et al. [5], which is a direct extension of the Stochastic Approximation based quantile estimator proposed by Tierney [20]. Note that both, our estimator and the EWSA are incremental estimators.

In this section, we compare both estimators for different distributions, under different resolution parameter and in stationary environments. The results are conclusive and demonstrate that the convergence of the algorithms conform to the theoretical results. We use different distributions namely:

- Uniform in $[0, 1]$
- Normal $N(0, 1)$
- Exponential distribution with mean 1 and variance 1
- Chi-square distribution with mean 1 and variance 2.

In all the experiments, we chose $a$ to be $-4$ and $b$ is fixed to $4$. Note that whenever the resolution is $N$, the estimate is moving with either additive or subtractive step size equal to $\frac{b-a}{N}$. A larger value of the resolution parameter $N$ implies a small step size, while a low value of the resolution parameter $N$ results in a smaller step size.

Initially, at time $0$, the estimates for the Discretized Quantile Estimator (DQE) are set to $Q_{N/2}$, while for the EWSA the initial value of the estimate is $(a + b)/2$.

The reader should note that the aim of the paper is to demonstrate its salient properties as a novel quantile estimator that uses only *finite* memory.

### 3.1 Comparison in Stationary Environments for Different Distributions

In this set of experiments, we examine a stationary environment. We used different resolutions namely $N = 30$, $N = 100$ and $N = 1000$ and as previously mentioned $[a, b] = [-4, 4]$. Given that the step size is obtained by the formula $\frac{b-a}{N}$, please note that the resulting step sizes of the resolutions $N = 30$, $N = 100$ and $N = 1000$ are $8/30$, $8/100$ and $8/1000$, respectively.

While for the EWSA, we use three different typical values, namely $\lambda = 0.01$, $\lambda = 0.05$, $\lambda = 0.1$, which illustrates the overall properties of this estimator. In fact, a low value of $\lambda$ permits slow updates of the estimates and is convenient for a stationary environment, while a high value of $\lambda$ for dynamic environment as it allows faster updates of the estimates.

Figure 1 depicts the case of estimating the $80\%$ quantile for the four different distributions: uniform, normal, exponential and Chi-square. We report the estimation error from an ensemble of $1000$ experiments. We observe that our DQE approaches the true value for all the four distributions asymptotically over time. We observe that given a low resolution ($N = 30$), the error drops very fast but stabilizes on a value between $0.05$ and $0.15$. Using a higher resolution, the error drops slower, but asymptotically the error becomes much smaller compared to the low resolution alternative. A very intriguing characteristic of our estimator is that by choosing a sufficiently high resolution, we are able to estimate the quantile as precise as we want. This is not possible for traditional off line quantile estimators without using an infinite amount of memory. Comparing the results of EWSA and DQE, it seems that the performance of the EWSA is highly dependent of $\lambda$ which can be hard to choose. E.g. for the normal distribution EWSA ends up with high errors for all the three values chosen for $\lambda$. In comparison the convergence properties of the DQE is far more consistent.

A very intriguing characteristic of our estimator, as the resolution increases, the estimation error diminishes (asymptotically). In fact, the limited memory of the estimator does not permit to achieve zero error, i.e, $100\%$ accuracy. As noted in the theoretical results, the convergence will be centred around the smallest interval $[Q_z, Q_{z+1}]$ containing the true quantile. Loosely speaking, a higher resolution increases the accuracy while a low resolution decreases the accuracy.

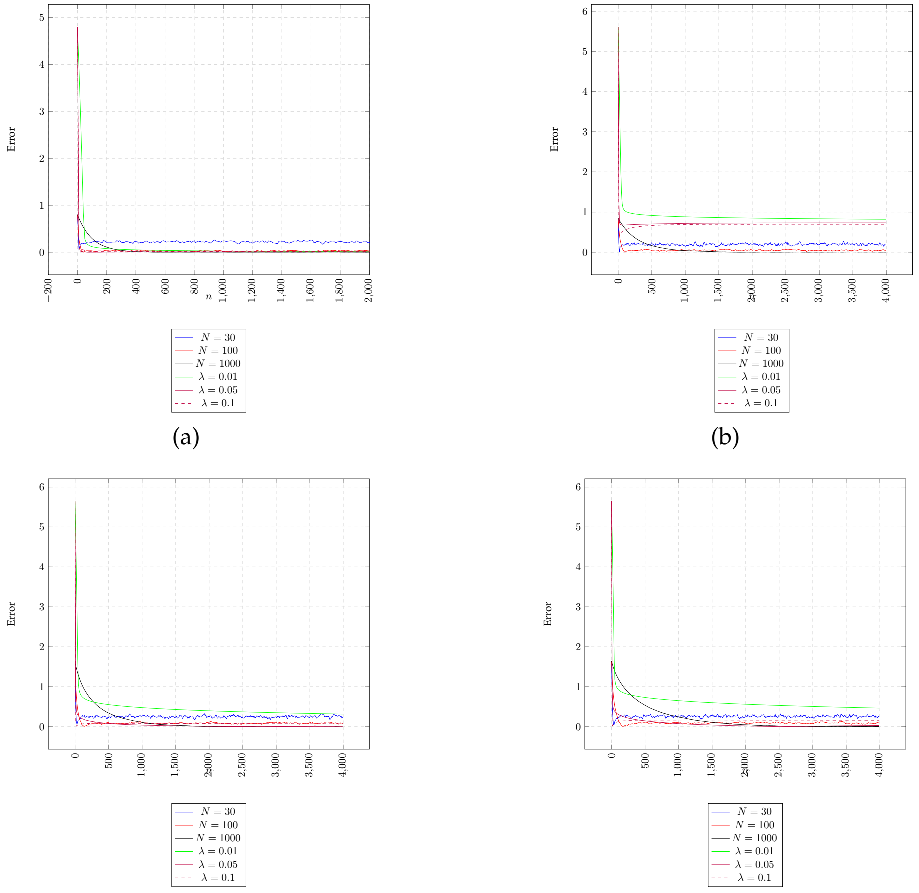
Figures 2, 3 and 4 depict the cases of estimating the $70\%$, $90\%$, $95\%$ quantiles respectively for the same four different distributions. In the comparison of the results for the different values of $N$ and for different values of $\lambda$ for the EWSA scheme, the conclusions are as for the 80% quantile case described above.

## 4 Conclusion

In this paper, we have designed a novel incremental discretized quantile estimator based on the theory of stochastic search on the line. We emphasize that the estimator can be easily implemented and that it is much simpler than the state of the art incremental estimator proposed by Tierney [20] as it does not require estimation of the density at the quantile. We have also extended our estimator in order to handle data arriving in a batch mode.

There are different extensions that can be envisaged for future work:

- We worked in finite Markov chain domain, and suppose that the true quantile lies in the interval $[a, b]$. As a future work, we plan to extend the proof to infinite state Markov chain.
- The existing algorithm for quantile estimation is designed for data elements that are added one by one. A possible extension is to generalize our algorithm to handle not only data insertions, but also dynamic data operations such as deletions and updates such as in [3].
- An interesting research direction is to simultaneously estimate more than a single quantile value. To achieve this, our present scheme will have to be modified so as to guarantee the monotonicity property of the quantiles, i.e, maintaining multiple quantile estimates while simultaneously ensuring that the estimates do not violate the monotonicity property.



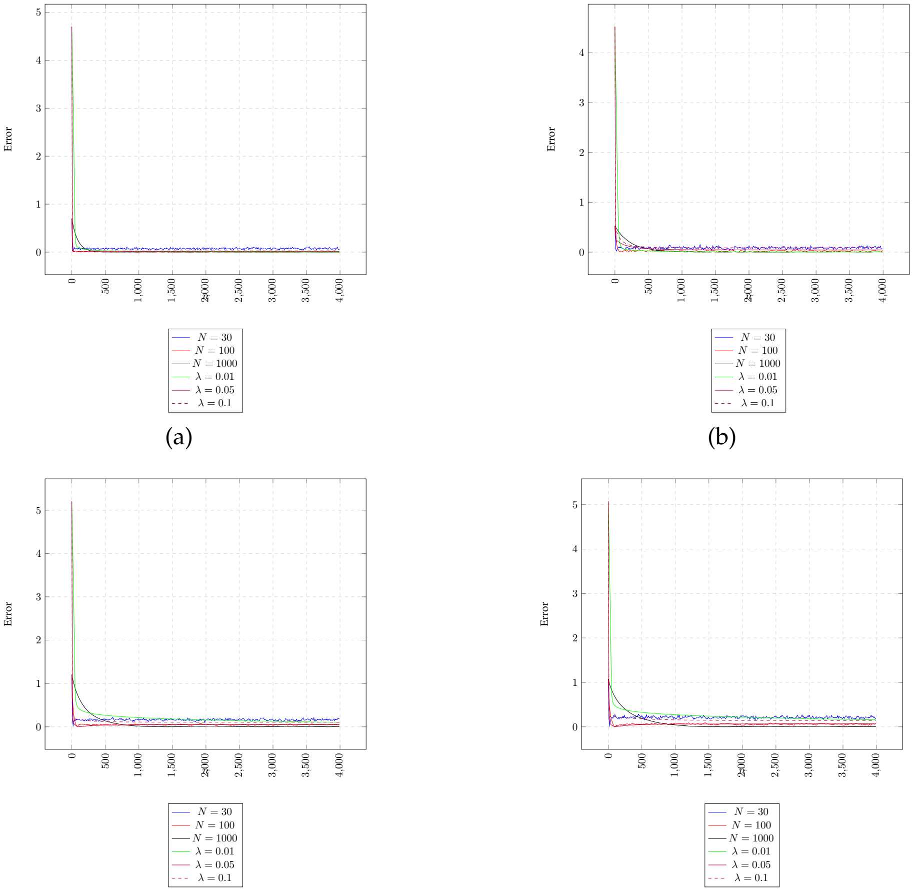
(a)

(b)

(c)

(d)

**Fig. 1.** This figure depicts the variation of the estimation error with time $n$ for the quantile of 80% for the DQE ($N = 30$, $N = 100$ and $N = 1000$) and for the EWSA ($\lambda = 0.01$, $\lambda = 0.05$ and $\lambda = 0.1$) for (a) $uniform$ distribution, (b) $normal$ distribution, (c) $exponential$ distribution, (d) $Chi - Square$ distribution.



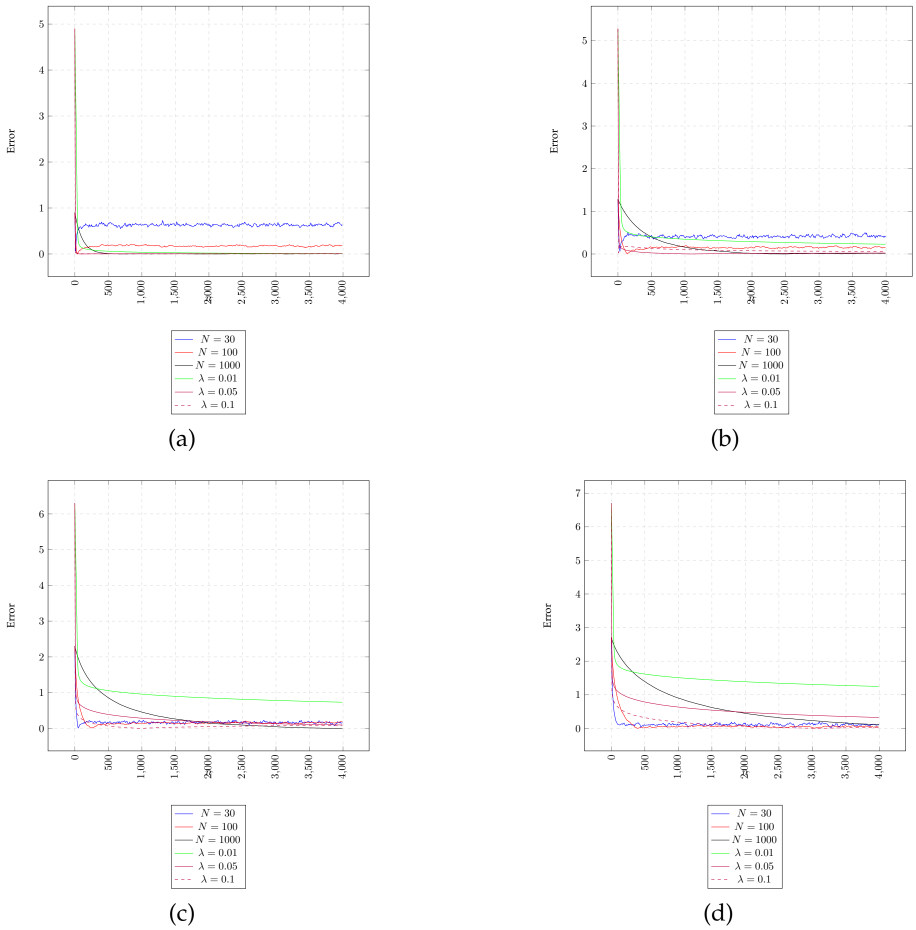
(a)

(b)

(c)

(d)

**Fig. 2.** This figure depicts the variation of the estimation error with time $n$ for the quantile of 70% for the DQE ($N = 30$, $N = 100$ and $N = 1000$) and for the EWSA ($\lambda = 0.01$, $\lambda = 0.05$ and $\lambda = 0.1$) for (a) $uniform$ distribution, (b) $normal$ distribution, (c) $exponential$ distribution, (d) $Chi - Square$ distribution.



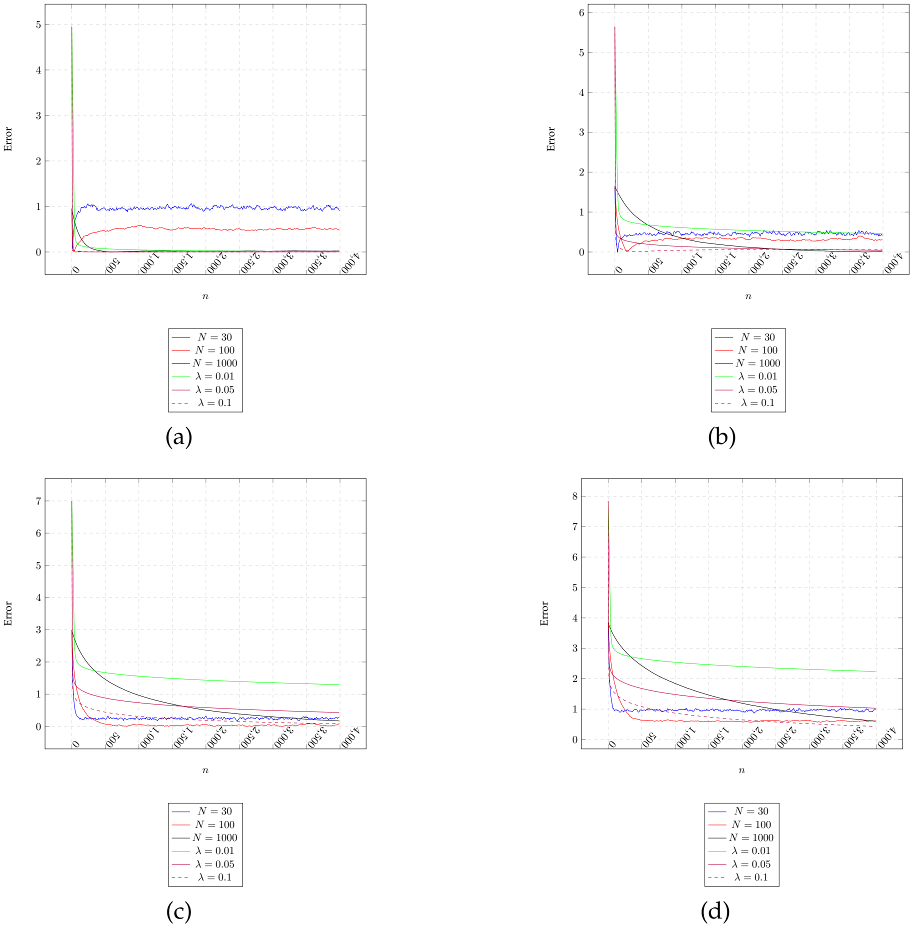
(a)

(b)

(c)

(d)

**Fig. 3.** This figure depicts the variation of the estimation error with time $n$ for the quantile of 90% for the DQE ($N = 30$, $N = 100$ and $N = 1000$) and for the EWSA ($\lambda = 0.01$, $\lambda = 0.05$ and $\lambda = 0.1$) for (a) $uniform$ distribution, (b) $normal$ distribution, (c) $exponential$ distribution, (d) $Chi - Square$ distribution.



(a)

(b)

(c)

(d)

**Fig. 4.** This figure depicts the variation of the estimation error with time $n$ for the quantile of 95% for the DQE ($N = 30$, $N = 100$ and $N = 1000$) and for the EWSA ($\lambda = 0.01$, $\lambda = 0.05$ and $\lambda = 0.1$) for (a) $uniform$ distribution, (b) $normal$ distribution, (c) $exponential$ distribution, (d) $Chi - Square$ distribution.

## References

1. M. Agache and B. J. Oommen. Generalized pursuit learning schemes: New families of continuous and discretized learning automata. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 32(6):738–749, December 2002.
2. A. Arasu and G. S. Manku. Approximate counts and quantiles over sliding windows. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 286–296. ACM, 2004.
3. J. Cao, L. Li, A. Chen, and T. Bu. Tracking quantiles of network data streams with dynamic operations. In *IEEE INFOCOM*, pages 1–5. IEEE, 2010.
4. J. M. Chambers, D. A. James, D. Lambert, and S. V. Wiel. Monitoring networked applications with incremental quantile estimation. *Statistical Science*, pages 463–475, 2006.
5. F. Chen, D. Lambert, and J. C. Pinheiro. Incremental quantile estimation for massive tracking. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 516–522. ACM, 2000.
6. G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.
7. M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *ACM SIGMOD Record*, volume 30, pages 58–66. ACM, 2001.
8. D.-S. Huang and W. Jiang. A general cpl-ads methodology for fixing dynamic parameters in dual environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(5):1489 –1500, October 2012.
9. J. K. Lanctôt and B. J. Oommen. Discretized estimator learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-22(6):1473–1483, November/December 1992.
10. Q. Ma, S. Muthukrishnan, and M. Sandler. Frugal streaming for estimating quantiles. In *Space-Efficient Data Structures, Streams, and Algorithms*, pages 77–96. Springer, 2013.
11. J. I. Munro and M. S. Paterson. Selection and sorting with limited storage. *Theoretical computer science*, 12(3):315–323, 1980.
12. B. Oommen and M. Agache. Continuous and discretized pursuit learning schemes: various algorithms and their comparison. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 31(3):277 –287, June 2001.
13. B. J. Oommen. Absorbing and ergodic discretized two-action learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-16:282–293, March/April 1986.
14. B. J. Oommen. Stochastic searching on the line and its applications to parameter learning in nonlinear optimization. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-27B:733–739, 1997.
15. B. J. Oommen and E. Hansen. The asymptotic optimality of discretized linear reward-inaction learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-14(3), May/June 1986.
16. B. J. Oommen and J. K. Lanctôt. Discretized pursuit learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-20(4):931–938, July/August 1990.
17. B. J. Oommen and G. Raghunath. Automata learning and intelligent tertiary searching for stochastic point location. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-28B:947–954, 1998.
18. B. J. Oommen, G. Raghunath, and B. Kuipers. Parameter learning from stochastic teachers and stochastic compulsive liars. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-36B:820–836, 2006.

19. M. A. L. Thathachar and B. J. Oommen. Discretized reward-inaction learning automata. *Journal of Cybernetics and Information Science*, pages 24–29, Spring 1979.

20. L. Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4(4):706–711, 1983.

21. B. Weide. Space-efficient on-line selection algorithms. In *Computer Science and Statistics: Proceedings of the Eleventh Annual Symposium on the Interface*, pages 308–311, 1978.

22. A. Yazidi, O. Granmo, B. John Oommen, and M. Goodwin. A novel strategy for solving the stochastic point location problem using a hierarchical searching scheme. *IEEE Transactions on Cybernetics*, 44(11):2202–2220, Nov 2014.

23. A. Yazidi, O.-C. Granmo, and B. J. Oommen. A stochastic search on the line-based solution to discretized estimation. In *Advanced Research in Applied Artificial Intelligence*, pages 764–773. Springer, 2012.

24. A. Yazidi and H. Hammer. Dicsretized qunatile estimation using the spl theory. *Unabridged journal version of this paper, 2018. To be submitted for publication.*

25. A. Yazidi and H. Hammer. Multiplicative update methods for incremental quantile estimation. *IEEE Transactions on Cybernetics*, PP(99):1–10, 2017.

26. A. Yazidi, H. L. Hammer, and B. J. Oommen. A higher-fidelity frugal quantile estimator. In *International Conference on Advanced Data Mining and Applications*, pages 76–86. Springer, 2017.