Quantile Search with Time-Varying Search Parameter

John Lipor
Department of Electrical and Computer Engineering
Portland State University
Email: lipor@pdx.edu

Gautam Dasarathy
Department of Electrical and Computer Engineering
Rice University
Email: gautamd@rice.edu

***Abstract*—We consider the problem of active learning in the context of spatial sampling, where the sampling cost is a function of both the number of samples taken *and* the distance traveled during the sampling procedure. We present a novel extension of the Quantile Search (QS) algorithm, in which the tuning parameter m varies throughout the search procedure. We analyze the resulting algorithm in terms of both sample complexity and distance traveled. Empirical results show that our proposed method outperforms standard QS with fixed m in all cases considered.**

I. Introduction & Problem Formulation

Level set estimation is an important problem in scientific domains such as water quality monitoring. For example, consider the motivating problem of determining the spatial extent of a hypoxic (low oxygen concentration) region in Lake Erie, which is a strong indicator of lake health [1]–[4]. In this case, all points with oxygen concentration below and above a fixed threshold can be considered as two classes in a binary classification problem, where the goal is to estimate the Bayes Decision Boundary (BDB). Labels are obtained by sending a vehicle to a physical location to take measurements of the phenomenon of interest.

In practice, the hypoxic region is not stationary and spans an area on the order of hundreds of square kilometers [2]. Therefore, to accurately model this region, any sampling procedure must be completed as quickly as possible. Efficiently determining the BDB can be viewed as an *active learning* problem [5], [6], where the sampling cost depends both on the number of samples collected *and* the distance traveled throughout the procedure; importantly the latter quantity depends on the current location of the vehicle.

In [4], the authors demonstrate how this problem can be solved using a series of one-dimensional searches for the change point θ^* of a binary step function. The proposed *quantile search* (QS) algorithm is shown to provide an explicit tradeoff between the number of samples taken and distance traveled during the estimation procedure. QS is a generalization of binary bisection [7]–[9], where rather than sampling at the mid-point of the feasible interval, samples are taken at a location $1/m$ into this interval, where $m \geq 2$ is fixed. By increasing m, more samples are required, but the distance "overshoot" is reduced, providing the desired tradeoff.

In this work, we extend the QS algorithm by allowing m to vary throughout the sampling procedure. We provide a novel update method for m that is dependent on the size of the feasible interval and analyze its convergence properties. We verify our results empirically, then show that our proposed algorithm outperforms QS with fixed m in all scenarios considered.

II. Proposed Algorithm & Analysis

As stated in the introduction, the problem of interest may be reduced to a series of one-dimensional search problems [4], [6]. Define the step function class

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R} : f(x) = \mathbb{1}_{[0,\theta)}(x), \theta \in [0,1]\}$$

where $\mathbb{1}_S(x)$ denotes the set indicator function. In contrast to the standard active learning scenario, our goal is to estimate an unknown θ^* while minimizing the total time required for sampling, a function of both the number of samples taken *and* the distance traveled. Towards this end, we suppose that the sampling vehicle obtains observations $\{Y_n\}_{n=1}^N \in \{0,1\}^N$ from the sample locations $\{X_n\}_{n=1}^N$ in the unit interval according to $Y_n = f_\theta(X_n)$.

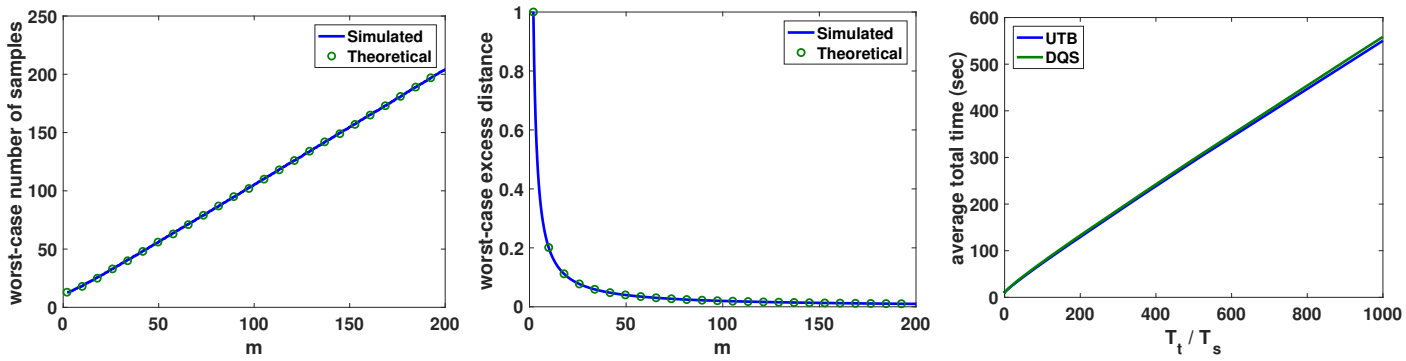
Consider an arbitrary sampling strategy that collects N_s measurements at locations $\{X_n\}_{n=1}^{N_s}$. Let T_s be the time required to obtain a single measurement and T_t be the time required to travel a unit distance. We may then calculate the total sampling time T_{tot} as

$$T_{\text{tot}} = T_s N_s + T_t \left(\sum_{n=1}^{N_s} |x_n - x_{n-1}| \right). \tag{1}$$

This follows since $\sum_{n=1}^{N_s} |x_n - x_{n-1}|$ is the total distance traveled.

A. Algorithm description

Here we describe our approach for varying m throughout the sampling procedure. We refer to the resulting algorithm as *uniform-to-binary* (UTB) search for the following reason. Our algorithm is a simple two-phase procedure; in the first stage, samples are taken uniformly along the unit interval until the remaining hypothesis space is sufficiently small. At this point, the algorithm performs binary search until a specified convergence criterion is met. Pseudocode is given in Algorithm 1, where `BinarySearch`$([a,b],\varepsilon)$ performs the binary bisection algorithm over the interval $[a,b]$ until the estimation error is less than ε.



(a) (b) (c)

Fig. 1. Performance of proposed UTB algorithm for spatial sampling to converge to an estimation error less than $\varepsilon = 1 \times 10^{-4}$. (a) Worst case number of samples taken. (b) Worst-case distance traveled. (c) Average sampling time as a function of the ratio of travel time to sampling time.

Note that if the true change point is at θ^*, any sampling procedure must travel at least a distance θ^*. Therefore, our strategy for designing the search algorithm is to minimize the total number of samples obtained while constraining the amount by which the total distance traveled *overshoots* θ^*. Constraining this overshoot to $1/m$, it is clear that sampling uniformly minimizes the number of samples subject to this constraint. Once the feasible interval is reduced to a size $2/m$, a binary search minimizes the remaining number of samples to estimate θ within an error ε.

The UTB algorithm is a generalization of deterministic QS with a time-varying search parameter. Recall that in the case where $Y_{n-1} = 1$, QS with parameter m' follows the update rule

$$X_n \leftarrow X_{n-1} + \frac{1}{m'}(b-a), \tag{2}$$

where all parameters are the same as those defined in Algorithm 1. In light of this fact, it is easy to see that UTB follows the update rule defined in (2) with $m' = \max\{2, m(b-a)\}$. Note that the condition on Line 3 of Algorithm 1 implements this.

Algorithm 1 Uniform-to-Binary Search (UTB)

1: **Input:** search parameter m, stopping error ε
2: **Initialize:** $X_0 \leftarrow 0$, $Y_0 \leftarrow 1$, $n \leftarrow 1$, $a \leftarrow 0$, $b \leftarrow 1$
Uniform search stage
3: **while** $b - a > \frac{2}{m}$ **do**
4: **if** $Y_{n-1} = 1$ **then**
5: $X_n \leftarrow X_{n-1} + \frac{1}{m}$
6: **else**
7: $X_n \leftarrow X_{n-1} - \frac{1}{m}$
8: **end if**
9: $Y_n \leftarrow f(X_n)$
10: $a = \max\{X_i : Y_i = 1, i \le n\}$
11: $b = \min\{X_i : Y_i = 0, i \le n\}$
12: $\hat{\theta}_n \leftarrow \frac{a+b}{2}$
13: **end while**
Binary search stage
14: $\hat{\theta} \leftarrow \texttt{BinarySearch}([a,b], \varepsilon)$

B. Analysis

In this section we state theoretical results that characterize the performance of UTB in terms of both the number of samples taken and the distance traveled throughout the search procedure.

Theorem 1: Let θ^* denote the change point of a step function on the unit interval. Then the UTB search with parameters m, ε requires

$$N_s \ge \left\lceil \theta^* m + \log_2\left(\frac{1}{m\varepsilon}\right) - 2 \right\rceil \tag{3}$$

samples to achieve an estimation error $\left|\hat{\theta}_{N_s} - \theta^*\right| \le \varepsilon$.

Theorem 1 provides the sample complexity for any fixed value of θ^*. From (3), we immediately obtain the following average-case and worst-case results.

Corollary 1: The UTB search with parameters m, ε and

$$N_s \ge \left\lceil m + \log_2\left(\frac{1}{m\varepsilon}\right) - 2 \right\rceil \tag{4}$$

samples is guaranteed to achieve (a) $\mathbb{E}_{\theta^* \sim \mathrm{Unf}(0,1)} \left|\hat{\theta}_{N_s} - \theta^*\right| \le \varepsilon$, where $\mathrm{Unf}(a,b)$ is the uniform distribution on $[a,b]$, and (b) $\sup_{\theta^* \in [0,1,]} \left|\hat{\theta}_{N_s} - \theta^*\right| \le \varepsilon$.

Finally, the following result provides an upper bound on the excess distance traveled by the vehicle.

Theorem 2: The UTB search with parameters m, ε satisfies

$$\sup_{\theta \in [0,1]} D_e \le \frac{2}{m}. \tag{5}$$

Given these results, one can differentiate (1) with respect to m to choose the optimal value. It is easily verified that UTB reduces to binary search as T_s dominates and continuous sampling as T_t dominates.

III. Simulations

In this section, we verify the performance of the proposed UTB algorithm. To obtain a profile of performance as a function of m, we range θ^* over 1000 uniformly-spaced values in the interval $[0,1]$ for each of 500 values of $m \in [2, 200]$. Fig. 1 (a) and (b) show the resulting worst-case number of

samples needed and excess distance traveled to converge to an estimation error of $\varepsilon = 1 \times 10^{-4}$. The figures demonstrate that UTB obtains a tradeoff between number of samples and distance traveled via the tuning parameter m. Moreover, our theoretical analysis closely matches the empirical results.

We also compare the performance of UTB with QS (where m is fixed over time). We consider the same grid over θ^* and m for 1000 different ratios of T_t/T_s in the range of $1\times10^{-4} - 1\times10^{3}$. Fig. 1 shows that the UTB algorithm outperforms QS for all ratios considered. Although the improvement appears mild, for the practical values of $T_t = 0.5$ m/s and $T_s = 60$ s in the scenario of [4], the resulting improvement amounts to 510 s faster sampling time. This improvement is compounded by the fact that numerous one-dimensional estimations are used to estimate the two-dimensional boundary.

REFERENCES

[1] G. L. E. R. Laboratory, "Lake Erie hypoxia warning system," http://www.glerl.noaa.gov/res/waterQuality/, 2005.

[2] D. Beletsky, D. Schwab, and M. McCormick, "Modeling the 1998-2003 summer circulation and thermal structure in Lake Michigan," *Journal of Geophys. Res.*, vol. 111, 2006.

[3] D. Scavia, J. D. Allan, K. K. Arend, S. Bartell, D. Beletsky, N. S. Bosch, S. B. Brandt, R. D. Briland, I. Daloğlu, J. V. DePinto *et al.*, "Assessing and addressing the re-eutrophication of lake erie: Central basin hypoxia," *Journal of Great Lakes Research*, vol. 40, no. 2, pp. 226–246, 2014.

[4] J. Lipor, B. P. Wong, D. Scavia, B. Kerkez, and L. Balzano, "Distance-penalized active learning using quantile search," *IEEE Transactions on Signal Processing*, vol. 65, no. 20, pp. 5453–5465, 2017.

[5] B. Settles, *Active Learning*. Morgan & Claypool, 2012.

[6] R. Castro and R. Nowak, "Active learning and sampling," in *Foundations and Applications of Sensor Management*, 1st ed. New York, NY: Springer, 2008, ch. 8.

[7] ——, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, pp. 2339–2353, May 2008.

[8] M. Horstein, "Sequential decoding using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 9, 1963.

[9] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problems in Information Transmission*, vol. 10:223-231, 1974, translated from Problemy Peredachi Informatsii, 10(3):51-61, July-September, 1974.