

Sequential Estimation based on Conditional Cost

George V. Moustakides
ECE, University of Patras, Rion, Greece,
& CS, Rutgers University
New Brunswick, NJ, USA.
E-mail: moustaki@upatras.gr

Tony Yaacoub and Yajun Mei
ISyE, Georgia Institute
of Technology
Atlanta, GA, USA.
Email: (tyaacoub, ymei3)@gatech.edu

Abstract—We consider the problem of parameter estimation under a sequential framework. Specifically we assume that an i.i.d. random process is observed sequentially with its common pdf having a random parameter that must be estimated. We are interested in designing a stopping time that will decide when is the best moment to stop sampling the process and an estimator that will use the acquired samples in order to provide the desired estimate. We follow a semi-Bayesian approach where we assign cost to the pair (estimate, true parameter) and our goal is to minimize the average sample size guaranteeing at the same time an average cost below some prescribed level. For our analysis we adopt a conditional average cost which leads to a considerable simplification in the sequential estimation problem, otherwise known to be analytically intractable. We apply our results to a number of examples and compare our method with the optimum fixed sample size but also with existing sequential schemes.

Index Terms—Sequential estimation, Sequential Analysis.

I. INTRODUCTION

Parameter estimation is needed in numerous problems across different scientific fields. In most applications, estimation is primarily based on fixed sample size methodology. However, when we are interested in obtaining a reliable estimate as quickly as possible then it is necessary to resort to sequential techniques. It is well known that in hypothesis testing, sequential methods [1] enjoy *significant* reduction in the number of samples required to reach a reliable decision as compared to fixed sample size alternatives. Therefore, it is only natural to expect that this important advantage will carry over to estimation as well. Before addressing the problem of sequential estimation let us first introduce some necessary background knowledge regarding classical estimation.

We observe a collection of random variables $\{\xi_1, \dots, \xi_t\}$, where $t > 0$ is an integer. For simplicity, we assume $\{\xi_t\}$ is i.i.d. with a common pdf $f(\xi|\theta)$ and parameter θ is considered random with a known prior pdf $\pi(\theta)$. Regarding the process $\{\xi_t\}$, the samples are generated as follows: Nature randomly selects the parameter θ following $\pi(\theta)$; then keeping θ fixed, Nature generates the sequence $\{\xi_t\}$ following $f(\xi|\theta)$. It is therefore clear that the joint pdf of the set of samples $\{\xi_1, \dots, \xi_t\}$ and θ has the following form

$$f_t(\xi_1, \dots, \xi_t, \theta) = \pi(\theta) \cdot f(\xi_1|\theta) \cdots f(\xi_t|\theta). \quad (1)$$

The joint pdf induces a probability measure which we denote by $P(\cdot)$ while we reserve the symbol $E[\cdot]$ for the corresponding expectation. If we also denote with $\mathcal{F}_t = \sigma\{\xi_1, \dots, \xi_t\}$ the

sigma-algebra generated by the first t samples, then we can write the conditional (posterior) pdf of θ given \mathcal{F}_t as

$$f_t(\theta|\mathcal{F}_t) = \frac{\pi(\theta) \prod_{j=1}^t f(\xi_j|\theta)}{\int \pi(\theta) \prod_{j=1}^t f(\xi_j|\theta) d\theta}. \quad (2)$$

Equations (1),(2) describe completely the statistical behavior of our observations. The goal is, using the acquired samples, to estimate the specific realization of θ that generates the data.

When we have a fixed sample size $\{\xi_1, \dots, \xi_t\}$, then the problem of optimum estimation is solved very efficiently by following the Bayesian formulation [2, Pages 142–156]. Specifically let $\hat{\theta}(\xi_1, \dots, \xi_t)$ denote any nonlinear functions of the observations which can serve as a potential estimator of θ . Assume we are given a cost function $C(\hat{\theta}, \theta)$ and consider the average cost $E[C(\hat{\theta}, \theta)]$, where averaging is with respect to observations and θ . We are interested in finding the estimator that minimizes this expression. In other words we would like to perform the minimization $\inf_{\hat{\theta}} E[C(\hat{\theta}, \theta)]$ which leads to the classical Bayes estimator.

To find our estimator, we compute the *conditional* average cost

$$E[C(\hat{\theta}, \theta)|\mathcal{F}_t] = \int C(\hat{\theta}, \theta) f_t(\theta|\mathcal{F}_t) d\theta, \quad (3)$$

where $f_t(\theta|\mathcal{F}_t)$ is defined in (2). Then, it is well known that the optimum Bayes estimator satisfies

$$\hat{\vartheta}_t = \arg \inf_{\hat{\theta}} E[C(\hat{\theta}, \theta)|\mathcal{F}_t] \quad (4)$$

and the corresponding minimum conditional average cost is given by

$$\mathcal{C}_t = \inf_{\hat{\theta}} E[C(\hat{\theta}, \theta)|\mathcal{F}_t] = E[C(\hat{\vartheta}_t, \theta)|\mathcal{F}_t]. \quad (5)$$

Both $\hat{\vartheta}_t$ and \mathcal{C}_t are \mathcal{F}_t -measurable since they are functions of the available observations.

II. SEQUENTIAL ESTIMATION

Under a sequential setup, process $\{\xi_t\}$ is acquired sequentially. At each time t we observe the accumulated information \mathcal{F}_t which grows with time, thus generating the filtration $\{\mathcal{F}_t\}$ and the sequence $\{f_t(\cdot)\}$ of joint pdfs. We use the same symbols $P(\cdot)$ and $E[\cdot]$ to denote the corresponding probability measure and expectation. One would be interested in defining a stopping time T which is adapted to $\{\mathcal{F}_t\}$ and a corresponding

estimator $\hat{\theta}_T$ which is \mathcal{F}_T -measurable (uses the observations that are available up to the time of stopping T) in order to provide an estimate of θ .

Since our goal is to limit the number of samples needed to compute the estimate, we would like to find a pair $(T, \hat{\theta}_T)$ that minimizes the average number of samples $E[T]$ while, at the same time, we control the average estimation cost. To be more precise we would like to consider the following constrained optimization problem for the determination of the optimum pair

$$\inf_{T, \hat{\theta}_T} E[T], \quad \text{subject to: } E[C(\hat{\theta}_T, \theta)] \leq \beta, \quad (6)$$

where β is a level selected by the Scientist. It has been pointed out in the literature [3]–[6] that solving (6) presents computational challenges and this problem is by no means analytically tractable.

A. Alternative Optimization Problem

The analytical difficulties we mentioned before can in fact be circumvented if we are willing to sacrifice part of our performance. We therefore propose to replace the constraint in (6) with the following *conditional* alternative

$$E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] \leq \tilde{\beta}.$$

If for example we select $\tilde{\beta} = \beta$ then the previous conditional version assures validity of the unconditional constraint in (6). The proposed modification in the constraint suggests a corresponding optimization problem

$$\inf_{T, \hat{\theta}_T} E[T], \quad \text{subject to: } E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] \leq \tilde{\beta}, \quad (7)$$

as a replacement of the original one in (6). The formulation of the parameter estimation problem with (7) is along the same lines of the approaches adopted in [5], [6] for Gaussian processes. We should also mention that similar ideas were used for simultaneous detection and estimation for Gaussian [7] and conditionally Gaussian [8] data.

Remark 1. Before continuing with the analysis and solution of our optimization let us discuss the differences between the two approaches depicted in (6) and (7). We observe that in the first we can have realizations of the observation sequence for which, at the time of stopping, the conditional average cost will satisfy $E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] > \beta$. Inequalities in the “wrong” direction tend to require smaller sample sizes, thus contributing towards the reduction of $E[T]$. As we can see, in (7) such inequalities are not permitted since we force the conditional average cost to be always below $\tilde{\beta}$ for every realization of the observations. Therefore if we select $\tilde{\beta} = \beta$ we will end up with a scheme that satisfies the constraint in (6) in the *strict* sense. For this reason we need to increase $\tilde{\beta}$ slightly and select $\tilde{\beta} > \beta$ in order to achieve exact equality.

Remark 2. We should emphasize that even with a value of $\tilde{\beta}$ selected so as to satisfy the constraint in (6) with equality, the scheme we obtain by solving (7) is *not* the optimum for (6). The expectation, however, is that the performance degradation

by solving (7) instead of (6) will not be overly dramatic. In any case, as we mentioned, because of this performance sacrifice, our estimation problem simplifies considerably allowing for the development of an analytic solution.

The optimizations depicted in (6) and (7) require the definition of a pair $(T, \hat{\theta}_T)$. In the sequel, using proper analysis, we are going to design a candidate pair $(\mathcal{T}, \hat{\theta}_{\mathcal{T}})$ and then we will demonstrate that it is in fact the one that solves the optimization problem of interest, namely the problem in (7). We begin the presentation of $(\mathcal{T}, \hat{\theta}_{\mathcal{T}})$ by first introducing our estimator.

B. Candidate Estimator

Let us fix the stopping time T and attempt to find the estimator $\hat{\theta}_T$ that minimizes the conditional average cost $E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T]$. Assuming T stops a.s. we can write

$$\begin{aligned} E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] &= E \left[\sum_{t=0}^{\infty} C(\hat{\theta}_t, \theta) \mathbb{1}_{\{T=t\}} | \mathcal{F}_T \right] \\ &= \sum_{t=0}^{\infty} E \left[C(\hat{\theta}_t, \theta) | \mathcal{F}_t \right] \mathbb{1}_{\{T=t\}} \\ &\geq \sum_{t=0}^{\infty} \inf_{\hat{\theta}} E \left[C(\hat{\theta}, \theta) | \mathcal{F}_t \right] \mathbb{1}_{\{T=t\}} \\ &= \sum_{t=0}^{\infty} \mathcal{C}_t \mathbb{1}_{\{T=t\}} = \mathcal{C}_T. \end{aligned} \quad (8)$$

We note that the indicator function $\mathbb{1}_{\{T=t\}}$ can be moved outside the conditional expectation because it is \mathcal{F}_t -measurable. Furthermore using (4) and (5) for each deterministic value of t , we lower bound the conditional average cost with its minimum value \mathcal{C}_t . It is also clear that the inequality in (8) becomes an equality if we select $\hat{\theta}_t$ to be the optimum Bayes estimator $\hat{\vartheta}_t$.

This result suggests that when we stop at T if we apply the optimum Bayes estimator to the available data \mathcal{F}_T then the conditional expected cost $E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T]$ matches the lower bound \mathcal{C}_T . Consequently, for any stopping time T , we propose as candidate estimator the Bayes estimator $\hat{\vartheta}_T$.

C. Candidate Stopping Time

Let us now turn to the definition of the candidate stopping time. As observations accumulate, at each time instant t we can compute the corresponding Bayes estimate $\hat{\vartheta}_t$ and the resulting minimum conditional average cost \mathcal{C}_t . The sequence $\{\mathcal{C}_t\}$ that is generated by these sequential computations can serve to define our candidate stopping time as follows

$$\mathcal{T} = \inf\{t \geq 0 : \mathcal{C}_t \leq \tilde{\beta}\}. \quad (9)$$

In other words we monitor the sequence of minimum conditional average costs and the *first time* the value of \mathcal{C}_t falls below $\tilde{\beta}$ this is the time we stop.

Combining the two results it is clear that we propose the pair $(\mathcal{T}, \hat{\vartheta}_{\mathcal{T}})$ for stopping and parameter estimation. More precisely we suggest to stop at \mathcal{T} defined in (9) and use the data obtained up to the time of stopping to compute the Bayes estimate $\hat{\vartheta}_{\mathcal{T}}$.

With the next theorem we show that this choice is optimum in the sense that it solves the constrained optimization problem defined in (7).

Theorem: Consider any competing pair $(T, \hat{\theta}_T)$ which satisfies the constraint $E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] \leq \tilde{\beta}$. Assuming that \mathcal{T}, T stop a.s., then for each realization of our observations we have $\mathcal{T} \leq T$.

Proof: Since the pair $(T, \hat{\theta}_T)$ satisfies the constraint in (7) this means that

$$\tilde{\beta} \geq E[C(\hat{\theta}_T, \theta) | \mathcal{F}_T] \geq \mathcal{C}_T.$$

The first inequality is due to our assumption and the second is a consequence of (8) where we fix T and minimize over $\hat{\theta}_T$. We can thus conclude that $\mathcal{C}_T \leq \tilde{\beta}$. But this inequality immediately implies $\mathcal{T} \leq T$. Indeed this is the case because \mathcal{T} is the first time instant for which $\mathcal{C}_t \leq \tilde{\beta}$. We have thus proved that for each realization any competing stopping time T will be no less than the candidate stopping time \mathcal{T} . Clearly this also implies that $E[\mathcal{T}] \leq E[T]$. This argument proves that the proposed pair is the one solving the constrained optimization problem depicted in (7). ■

We must point out that if the constraint is very mild, namely $\tilde{\beta}$ is overly large, then our method can lead to a trivial optimum pair $(T, \hat{\theta}_T)$. Indeed it is possible to stop at $\mathcal{T} = 0$ a.s. and simply use the prior to provide the necessary estimate. This can happen when $\mathcal{C}_0 \leq \tilde{\beta}$, namely

$$\mathcal{C}_0 = \inf_{\hat{\theta}} \int C(\hat{\theta}, \theta) \pi(\theta) d\theta \leq \tilde{\beta},$$

leading to the deterministic estimate

$$\hat{\vartheta}_0 = \arg \inf_{\hat{\theta}} \int C(\hat{\theta}, \theta) \pi(\theta) d\theta.$$

Consequently, in order to avoid such a trivial outcome we must select $\tilde{\beta} < \mathcal{C}_0$.

Remark 3. We should emphasize that the desired problem to solve is (6). It is because of its analytical intractability that we resort to (7) which is possible to solve efficiently. When however we study the performance of the scheme produced by (7) we must test its behavior with respect to the constraint in (6) and *not* the conditional version adopted in (7). In this sense even though the pair $(\mathcal{T}, \hat{\vartheta}_{\mathcal{T}})$ is “optimum”, it should not come as a surprise if its performance, in some cases, turns out to be inferior to the fixed sample size estimator.

III. OPTIMIZING COVERAGE PROBABILITY

Major goal in parameter estimation is, of course, the design of an estimator but also the selection of a sample size that can assure that the estimate is within a prescribed (confidence) interval around the correct value with some minimal guaranteed (coverage) probability. Specifically we would like to find a sample size T , fixed or random (stopping time), and an estimator $\hat{\theta}_T$ of θ assuring that $P(|\hat{\theta}_T - \theta| \leq h) \geq \alpha$. Parameter $h > 0$ denotes the half width of the confidence interval and

$\alpha \in (0, 1)$ the minimal level of the coverage probability. The two quantities h, α are specified by the Scientist.

This problem can be effectively treated using the general framework we introduced in the previous section by selecting $C(\hat{\theta}, \theta) = 1 - \mathbb{1}_{\{|\hat{\theta} - \theta| \leq h\}}$ and $\beta = 1 - \alpha$. The conditional average cost function, using (3), can be written as

$$P(|\hat{\theta}_t - \theta| > h | \mathcal{F}_t) = 1 - \int_{\hat{\theta}_t - h}^{\hat{\theta}_t + h} f_t(\theta | \mathcal{F}_t) d\theta,$$

where in the integration one should take into account the (essential) support of θ as it is dictated by the prior $\pi(\theta)$ (for example if $\theta \geq 0$ a.s. then the lower integration boundary must be replaced by $(\hat{\theta}_t - h)^+$). The Bayes estimator and the corresponding minimum conditional average cost are given by

$$\hat{\vartheta}_t = \arg \sup_{\hat{\theta}} \int_{\hat{\theta} - h}^{\hat{\theta} + h} f_t(\theta | \mathcal{F}_t) d\theta, \quad (10)$$

$$\mathcal{C}_t = 1 - \int_{\hat{\vartheta}_t - h}^{\hat{\vartheta}_t + h} f_t(\theta | \mathcal{F}_t) d\theta. \quad (11)$$

As we will have the chance to verify from the examples that follow, working directly with the coverage probability most often results in estimators and conditional costs that do not have analytic expressions and need to be computed numerically.

Next we present three classical parameter estimation examples where we compute their estimators and stopping times and compare their performances with fixed sample size methods and existing sequential techniques.

A. Mean of a Gaussian

Let us begin by considering the classical problem of estimating the unknown mean of a Gaussian random variable. Suppose our observations $\{\xi_t\}$ are $\xi_t \sim \mathcal{N}(\theta, \sigma^2)$ and for the prior of the mean we have $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$, where $\mu, \sigma^2, \sigma_\theta^2$ are known. The first step in our analysis consists in computing the posterior pdf $f_t(\theta | \mathcal{F}_t)$. By using (2) it is a simple exercise to verify that

$$f_t(\theta | \mathcal{F}_t) = \mathcal{N}(\mu_t, \sigma_t^2)$$

where

$$\mu_t = \frac{\sigma_\theta^2 \sum_{j=1}^t \xi_j + \mu \sigma^2}{\sigma_\theta^2 t + \sigma^2}; \quad \sigma_t^2 = \frac{\sigma_\theta^2 \sigma^2}{\sigma_\theta^2 t + \sigma^2}.$$

From (10),(11) the Bayesian estimator can be found as follows

$$\hat{\vartheta}_t = \arg \sup_{\hat{\theta}} \left\{ \Phi\left(\frac{h + \hat{\theta} - \mu_t}{\sigma_t}\right) - \Phi\left(\frac{-h + \hat{\theta} - \mu_t}{\sigma_t}\right) \right\} = \mu_t,$$

where $\Phi(\cdot)$ denotes the standard Gaussian cdf and

$$\mathcal{C}_t = 2\Phi\left(-\frac{h}{\sigma_t}\right).$$

Since \mathcal{C}_t is purely deterministic it is clear that the resulting stopping time \mathcal{T} in (9) will be *deterministic* as well. Actually for this case we can even solve the original optimization problem (6) and the resulting optimum stopping time is still deterministic [3].

Remark 4. With this simple example we realize that sequential estimation does not necessarily enjoy similar characteristics as sequential hypothesis testing (in fact this is the reason we included this case). We recall that in hypothesis testing when deciding between $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, 1)$, optimum sequential techniques require, on average, *four times less* samples than optimum fixed sample size tests [2, Page 109]. When, however, we estimate the mean of a Gaussian random variable, as we have seen, there is absolutely no gain. Fortunately this conclusion is not universal and in the next two examples we will experience gains that are worth reporting.

B. Bernoulli Trials

Methods that estimate *proportions* accompanied by confidence intervals are being used in many applications as polls; surveys; determination of fractions of people, animals or goods having certain traits/characteristics; etc. In these problems minimizing the number of samples that are necessary to assure estimates of a given quality is, clearly, of paramount importance. The simplest and most common model used to describe the corresponding data is Bernoulli binary sequences, which is also the model we adopt here.

In the literature there are various fixed sample size estimators [9], [10] addressing the question of proportion estimation, but we can also find sequential methods involving stopping times. In particular in [3] the optimization problem defined in (6) for this specific example is treated under an asymptotic regime, while in [11], [12] stopping rules are proposed and numerically compared without being supported by any form of optimality.

Let us analyze this estimation problem using the methodology we introduced in the previous section. Consider an i.i.d. process $\{\xi_t\}$ with binary samples $\xi_t \in \{0, 1\}$ and $P(\xi_t = 1) = \theta \in [0, 1]$. Probability θ is the parameter to be estimated for which we assume to have a symmetric prior $\text{Beta}(\theta, a, a)$ with known parameter $a > 0$. Due to the limits of θ we additionally need to assume that $0 < h < 0.5$.

If we call $S_t = \xi_1 + \dots + \xi_t$ then the conditional pdf $f_t(\theta|\mathcal{F}_t)$ satisfies

$$f_t(\theta|\mathcal{F}_t) = \frac{\theta^{a+S_t-1}(1-\theta)^{a+t-S_t-1}}{B(a+S_t, a+t-S_t)} \quad (12)$$

which is $\text{Beta}(\theta, a+S_t, a+t-S_t)$ distributed. Since $0 \leq \theta \leq 1$ if we apply (10) we can write

$$E[C(\hat{\theta}, \theta)|\mathcal{F}_t] = 1 - \text{Beta_cdf}(\min\{1, \hat{\theta}+h\}, a+S_t, a+t-S_t) + \text{Beta_cdf}((\hat{\theta}-h)^+, a+S_t, a+t-S_t).$$

For $S_t = 1 - a$ the previous expressions is minimized by $\hat{\theta}_t = h$ and for $S_t = a + t - 1$ with $\hat{\theta}_t = 1 - h$. For any other value of S_t finding the Bayes estimator requires the *numerical* solution of the equation

$$\hat{\theta}_t = \arg \left\{ \hat{\theta} : \left(\frac{\hat{\theta}-h}{\hat{\theta}+h} \right)^{a+S_t-1} = \left(\frac{1-h-\hat{\theta}}{1+h-\hat{\theta}} \right)^{a+t-S_t-1} \right\}$$

with $h \leq \hat{\theta} \leq 1 - h$.

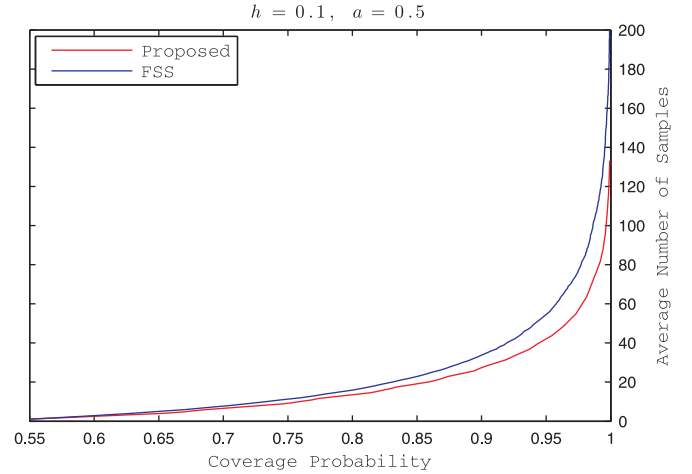


Fig. 1. Average number of samples as a function of coverage probability for the Beta distribution when $h = 0.1$ and $a = 0.5$.

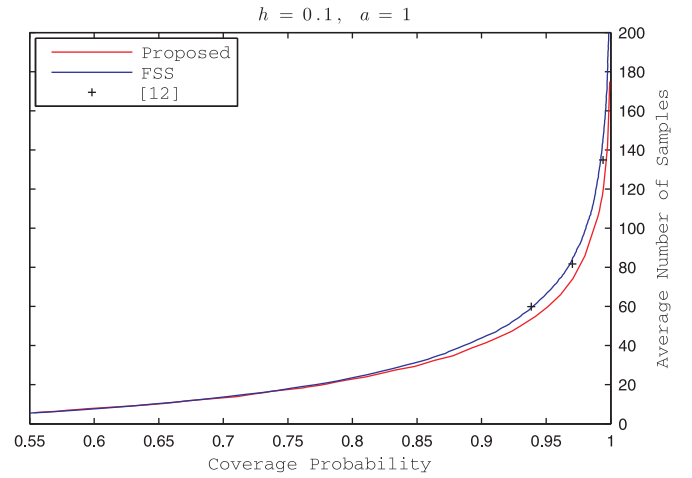


Fig. 2. Average number of samples as a function of coverage probability for the Beta distribution when $h = 0.1$ and $a = 1$.

Fig. 1 depicts the relative performance of the proposed and the fixed sample size (FSS) method and in Fig. 2 we also include the sequential method of [12] whose parameters are tuned for best performance (this is the reason why we have only three points). Graphs are the result of averages of 100000 realizations. As we can see, the proposed method outperforms the fixed sample size and the estimator in [12]. We also observe that, as the coverage probability approaches 1 we enjoy bigger gains in sample size, but the reward is by no means near the levels we experience in hypothesis testing.

C. Exponential Distribution

In the third example we consider samples that are distributed according to the one sided exponential distribution. In particular we assume that their density is

$$f(\xi|\theta) = \theta e^{-\theta\xi}, \quad \theta > 0, \quad \xi \geq 0,$$

while the prior is also exponential of the form

$$\pi(\theta) = a e^{-a\theta}, \quad a > 0,$$

where a is considered known. If we now compute the conditional pdf of θ given \mathcal{F}_t then

$$f_t(\theta|\mathcal{F}_t) = \frac{S_t^{t+1}}{t!} \theta^t e^{-S_t \theta}, \text{ where } S_t = a + \sum_{j=1}^t \xi_j,$$

which is $\text{Gamma}(\theta, t+1, S_t^{-1})$ distributed. From (10)

$$\hat{\vartheta}_t = \frac{h}{\tanh(h \frac{S_t}{t})}$$

and applying (11)

$$\mathcal{C}_t = 1 - \text{Gamma_cdf}(\hat{\vartheta}_t + h, t+1, S_t^{-1}) + \text{Gamma_cdf}(\hat{\vartheta}_t - h, t+1, S_t^{-1}).$$

It is interesting to note that the Bayesian estimator is not consistent since, using the LLN, we have $\frac{S_t}{t} \rightarrow \frac{1}{\theta}$ a.s. and

$$\lim_{t \rightarrow \infty} \hat{\vartheta}_t = \frac{h}{\tanh(\frac{h}{\theta})} \neq \theta.$$

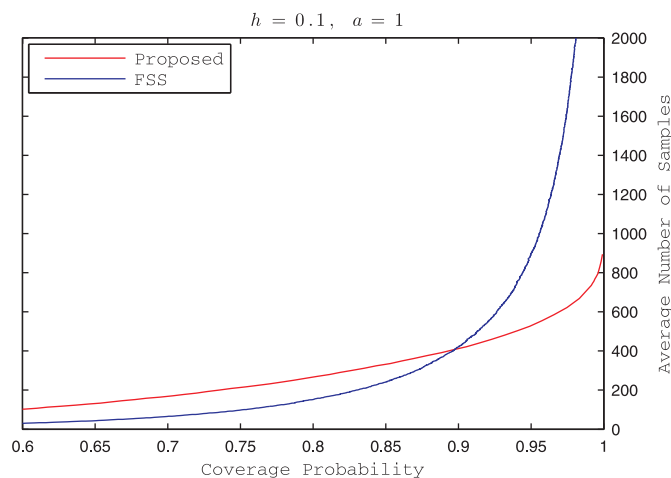


Fig. 3. Average number of samples as a function of coverage probability for the exponential distribution when $h = 0.1$ and $a = 1$.

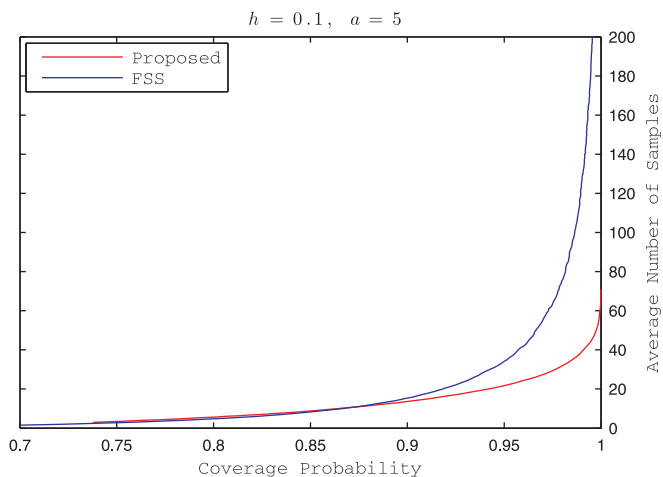


Fig. 4. Average number of samples as a function of coverage probability for the exponential distribution when $h = 0.1$ and $a = 5$.

We can show that this limiting value has, in fact, error which is within the pre-specified confidence interval since

$$\left| \frac{h}{\tanh(\frac{h}{\theta})} - \theta \right| \leq h.$$

From Fig. 3 and 4 we see that, for coverage values larger than 0.9 (which is the practically interesting range), with the proposed method we can enjoy *substantial* gains as compared to the fixed sample size estimator. In particular, if the coverage probability is close to 0.99 the number of samples required by the proposed scheme is at least four times smaller than in the fixed sample size case. On the other hand, for coverage probabilities below 0.9 the fixed sample size prevails.

IV. CONCLUSION

In the examples we presented, the performance of the proposed scheme was not always better than the fixed sample size estimator (although for high coverage probabilities it persistently outperformed it, and some times even considerably). This is because our method is not the solution of the optimization in (6). We are currently working towards the development of the exact solution of (6), but *only* for certain special cases (since the general problem is intractable) that are also of wide interest as, for example, percentage estimation. In fact we anticipate that the results obtained in this work will play a vital role in developing the corresponding long sought, strictly optimum sequential estimation scheme.

ACKNOWLEDGEMENT

This work was supported by the US National Science Foundation under Grant CIF 1513373 through Rutgers University and under Grant CMMI 1362876 through Georgia Institute of Technology.

REFERENCES

- [1] A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, vol. 16, no. 2, pp. 117–186, 1945.
- [2] H. V. Poor. *An Introduction to Signal Detection and Estimation, 2nd Edition*. Springer, New York, 1988.
- [3] M. Ghosh, N. Mukhopadhyay and P. K. Sen. *Sequential Estimation*, Wiley, New York, 1997.
- [4] B. K. Ghosh. On the attainment of the Cramer-Rao bound in the sequential case. *Seq. Anal.*, vol. 6, pp. 267–288, 1987.
- [5] P. Grambsch. Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Ann. Stat.*, vol. 11, pp. 68–77, 1983.
- [6] G. Fellouris. Asymptotically optimal parameter estimation under communication constraints. *Ann. Stat.*, vol. 40, pp. 2239–2265, 2012.
- [7] Y. Yilmaz, G. V. Moustakides and X. Wang. Sequential joint detection and estimation. *SIAM, Theory Probab. Appl.*, vol. 59, no. 3, pp. 452–465, 2015.
- [8] Y. Yilmaz, S. Li and X. Wang. Sequential joint detection and estimation: Optimum tests and applications. *IEEE Trans. Sig. Proc.*, vol. 64, no. 20, pp. 5311–5326, 2016.
- [9] G. Casella. Refining binomial confidence intervals. *Canad. J. Stat.*, vol. 14, pp. 113–129, 1986.
- [10] A. Agresti and B. A. Coull. Approximate is better than “exact” for interval estimation of binomial proportions. *Amer. Stat.*, vol. 52, pp. 119–126, 1998.
- [11] P. Armitage. Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*, vol. 45, pp. 1–15, 1958.
- [12] J. Frey. Fixed-width sequential confidence intervals for a proportion. *Amer. Stat.*, August 2010, vol. 64, no. 3, pp. 242–249, 2010.