How Much Evidence Should One Collect?

Remco Heesen

October 10, 2013

Abstract

This paper focuses on the question how much evidence one should collect before deciding on the truth-value of a proposition. An analysis is given of a model where evidence takes the form of Bernoulli-distributed random variables. From a Bayesian perspective, the optimal strategy depends on the potential loss of drawing the wrong conclusion about the proposition and the cost of collecting evidence. It turns out to be best to collect only small amounts of evidence unless the potential loss is very large relative to the cost of collecting evidence.

1 Introduction

Suppose a scientist wants to learn the truth-value of some proposition. Perhaps because some important decision depends on it, perhaps just because she wants to know. She can gather evidence, but no collection of evidence conclusively settles the truth-value of the proposition. Gathering evidence is costly: it requires time and effort, which could be spent on other pursuits.

How much evidence should the scientist collect in such a scenario? Or in other words, how should the benefits of more evidence be traded off against the costs? This paper analyzes a model where the evidence takes the form of Bernoulli trials and finds the optimal Bayesian strategy for this model. The

results are obtained by applying the sequential probability-ratio test (Wald 1947, Wald and Wolfowitz 1948, DeGroot 2004).

Sections 2 and 3 describe the model and state the results. Sections 4 and 5 discuss the results and draw some conclusions. An appendix contains the proofs.

2 The Model

Let p be a proposition. The scientist is interested in learning the truth-value of p, thus the relevant set of possible worlds is $\Omega = \{p, \neg p\}$, i.e., p is either true or false.

In this model evidence about p takes the form of random variables that are distributed like X, where

$$X \mid p \sim Ber(1-\varepsilon),$$
$$X \mid \neg p \sim Ber(\varepsilon),$$

for some given $\varepsilon \in (0, 1/2)$. So if p is true it is more likely that $X = 1$ than that $X = 0$, while if p is false this is reversed. Realizations of X are assumed to be independent in each of the two possible worlds, so any collection of evidence forms an i.i.d. dataset.

At a cost $c > 0$, the scientist gains one piece of evidence (i.e., one realization of X). Gaining a piece of evidence may reflect an experiment done by the scientist, or it may reflect what the scientist learns through testimony (say, by reading a paper by another scientist).

The scientist is allowed to collect evidence sequentially. That is, the decision whether or not to collect a $k+1$-st piece of evidence may depend on what is learned from the first k pieces of evidence.

Whenever the scientist decides to stop collecting evidence, she has to choose a terminal decision from the set $D = \{d_1, d_2\}$, where d_1 represents the

decision to believe that p is true (and to act on that belief when appropriate), and d_2 represents the decision to believe that p is false.

The scientist is faced with a trade-off. Collecting more evidence reduces the chance of drawing the wrong conclusion about the truth-value of p, but increases the accumulated costs. Collecting less evidence reduces the costs, but increases the chance of drawing the wrong conclusion about p.

In order to mathematically analyze this trade-off, I need some additional assumptions about the way individual scientists make decisions. I assume scientists act as if they were Bayesian statisticians. This means that their decisions can be modeled as follows.

First, at any given time the scientist has a subjective probability $\xi \in [0, 1]$ that reflects how likely she thinks it is that p is true. Second, in response to evidence she updates these beliefs using Bayes' rule. Third, the scientist has a loss function that puts a numerical value on each decision in each possible world. Fourth, the scientist makes decisions that minimize risk, which is the expected value of the loss relative to her subjective beliefs.

In this model, the loss ℓ is zero if the decision is "correct" (d_1 if p and d_2 if $\neg p$), and $\beta > 0$ if the decision is "incorrect" (d_2 if p and d_1 if $\neg p$, see table 1). The total loss is then ℓ plus the number of connections made times c.

$\ell(w, d)$	p	$\neg p$
d_1	0	β
d_2	β	0

Table 1: The loss function ℓ.

As a result, the risk associated with decision d_1 is $(1 - \xi)\beta$ and the risk associated with decision d_2 is $\xi\beta$. By assumption, the scientist chooses the decision with the lowest risk, so the risk associated with the decision is

$$\rho_0(\xi) := \min\{\xi\beta, (1 - \xi)\beta\}.$$

It remains to ask how much evidence the scientist will collect. This is a sequential sampling problem. The scientist wants to learn the true state of the world, which she can do by sampling at a cost c from a probability distribution that depends on the state of the world. Let Δ denote the set of all possible sequential decision procedures the scientist might use. So each $\delta \in \Delta$ is a function that specifies whether the scientist collects an additional piece of evidence as a function of the evidence obtained so far.

Let X_i denote the i-th piece of evidence. Let $\xi(X_1, \ldots, X_n)$ denote the posterior probability that p is true after seeing $X_1, \ldots, X_n$ (assuming the prior was ξ). Let $N(\delta)$ denote the number of connections made under sequential decision procedure $\delta \in \Delta$ (in general, this is a random variable). Then the risk of a sequential decision procedure $\delta \in \Delta$ is

$$\rho(\xi, \delta) := \mathbb{E}\left[\rho_0(\xi(X_1, \ldots, X_{N(\delta)})) + cN(\delta)\right].$$

By assumption the scientist chooses the procedure with the lowest risk, i.e., the procedure $\delta^* \in \Delta$ that satisfies

$$\rho(\xi, \delta^*) = \inf_{\delta \in \Delta} \rho(\xi, \delta).$$

The existence of a procedure δ^* that satisfies this equation is guaranteed by Chow and Robbins (1963, theorem 1). The next section is dedicated to specifying δ^*.

3 The Results

The problem that the scientist needs to solve is that of finding an optimal stopping rule. DeGroot (2004, sections 12.14–12.16) provides an analysis of this situation.

Let ξ denote the scientist's prior before seeing any evidence and let $c > 0$ be the cost of one observation. The observations are i.i.d. with distribution

$$f_1(x) := \Pr(X_i = x \mid p) = \varepsilon^{1-x}(1-\varepsilon)^x, \qquad x = 0, 1$$

if p is true, and

$$f_2(x) := \Pr(X_i = x \mid \neg p) = \varepsilon^{x}(1-\varepsilon)^{1-x}, \qquad x = 0, 1$$

if p is false. Let

$$Z_i := \log \frac{f_2(X_i)}{f_1(X_i)} = (1 - 2X_i) \log \frac{1-\varepsilon}{\varepsilon}.$$

Consider the sequential decision procedure $\delta(a, b)$ that continues to take observations as long as

$$a < \sum_{i=1}^{N} Z_i < b,$$

for some $a < 0$ and $b > 0$. Note that each Z_i can take only two possible values: $\log \frac{1-\varepsilon}{\varepsilon}$ if $X_i = 0$ and $-\log \frac{1-\varepsilon}{\varepsilon}$ if $X_i = 1$. Thus $\sum_{i=1}^{N} Z_i$ can only take values that are integer multiples of $\log \frac{1-\varepsilon}{\varepsilon}$. So without loss of generality a and b can be rounded to integer multiples of $\log \frac{1-\varepsilon}{\varepsilon}$. In that case $\sum_{i=1}^{N} Z_i$ must be exactly equal to either a or b when $\delta(a, b)$ takes no further observations.

Proposition 1 (DeGroot (2004))**.** *Suppose the random variables Z_i can only take the values z and $-z$ for some z and a and b are integer multiples of z. Then the risk of the sequential decision procedure $\delta(a, b)$ is*

$$\begin{aligned}\rho(\xi, \delta(a, b)) = \xi\beta\frac{1-e^a}{e^b-e^a} + (1-\xi)\beta\frac{e^a(e^b-1)}{e^b-e^a} + c\xi\frac{a(e^b-1)+b(1-e^a)}{(e^b-e^a)\mathbb{E}[Z_i \mid p]} \\ + c(1-\xi)\frac{ae^a(e^b-1)+be^b(1-e^a)}{(e^b-e^a)\mathbb{E}[Z_i \mid \neg p]} \qquad (1)\end{aligned}$$

and the optimal sequential decision procedure among those that take at least one observation is $\delta(a^*, b^*)$ where $a^* < 0$ and $b^* > 0$ are the values that minimize (1)*.*

So the optimal sequential decision procedure in the decision problem under consideration (assuming at least one observation is taken) takes the form

$$\delta_{m,n} := \delta\left(-m\log\frac{1-\varepsilon}{\varepsilon}, n\log\frac{1-\varepsilon}{\varepsilon}\right),$$

where m and n are positive integers. The scientist considers the difference between the number of X_i so far observed that took the value zero and the number of X_i so far observed that took the value one. The procedure then tells her to continue to take observations as long as that difference is strictly between $-m$ and n. If the difference hits $-m$ she stops taking observations and chooses decision d_1, and if the difference hits n she stops and chooses decision d_2.

Let g_k be defined by

$$g_k(\varepsilon) = \frac{(1-\varepsilon)^{2k+1} - \varepsilon^{2k+1}}{(1-2\varepsilon)^2\varepsilon^k(1-\varepsilon)^k} + \frac{2k+1}{1-2\varepsilon},$$

for all non-negative integers k and $\varepsilon \in (0, 1/2)$. Since $g_{k+1}(\varepsilon) > g_k(\varepsilon)$ for all k and ε, there is a unique k^* such that

$$g_{k^*-1}(\varepsilon) < \frac{\beta}{c} \leq g_{k^*}(\varepsilon).$$

(Unless $\beta/c \leq g_0(\varepsilon)$; in that case define $k^* = 0$. See also tabel 2)

Proposition 2. *If $\xi = 1/2$, the optimal sequential decision procedure is δ_{k^*,k^*}.*

This proposition determines the optimal procedure for a scientist who starts out thinking p is equally likely to be true or false. What if the scientist has a different prior?

Proposition 3. *Let $d \in \mathbb{Z}$. If*

$$\xi = \frac{\varepsilon^d}{\varepsilon^d + (1-\varepsilon)^d},$$

the optimal sequential decision procedure is δ_{k^*+d,k^*-d} (where the optimal procedure takes no observations if $k^* + d \leq 0$ or $k^* - d \leq 0$).

k^*	β/c
0	$(0, g_0(\varepsilon)]$
1	$(g_0(\varepsilon), g_1(\varepsilon)]$
$\vdots$	$\vdots$
k	$(g_{k-1}(\varepsilon), g_k(\varepsilon)]$
$k+1$	$(g_k(\varepsilon), g_{k+1}(\varepsilon)]$
$\vdots$	$\vdots$

Table 2: k^* is determined by finding an interval of the form $(g_{k-1}(\varepsilon), g_k(\varepsilon)]$ such that β/c is in that interval.

Corollary 4. *For any $\xi \in (0,1)$ not covered by proposition 3 there must be a $d \in \mathbb{Z}$ such that*

$$\frac{\varepsilon^d}{\varepsilon^d + (1-\varepsilon)^d} < \xi < \frac{\varepsilon^{d-1}}{\varepsilon^{d-1} + (1-\varepsilon)^{d-1}}.$$

Then the optimal sequential decision procedure is one of δ_{k^*+d,k^*-d}, δ_{k^*+d-1,k^*-d+1}, δ_{k^*+d-1,k^*-d}, or δ_{k^*+d,k^*-d+1}.

One can derive general inequalities to determine which of these four procedures is optimal for given values of ξ, β, c, and ε, but this is not important for my purposes here.

What proposition 3 and its corollary show is that in general a larger value of k^* indicates that more observations will be needed to come to a decision on the truth-value of p. The value of ξ biases the process towards one conclusion or the other but it does not change this general level k^*. I will focus on the value of k^* in the remainder of this paper.

4 Discussion

In the previous section I identified k^* as a function of the parameters β, c, and ε (see figure 1). If the prior is $\xi = 1/2$, the optimal sequential decision procedure may be characterized straightforwardly in terms of k^* (if $\xi \neq 1/2$, this characterization holds except for a bias towards one conclusion or the other).

Where $k^* = 0$, the optimal procedure is to take no observations. Where $k^* = 1$, the optimal procedure is to take exactly one observation. Where $k^* > 1$, the optimal procedure is to take observations until the absolute difference between the number of observed X_i that take the value one and the number that take the value zero is k^*.

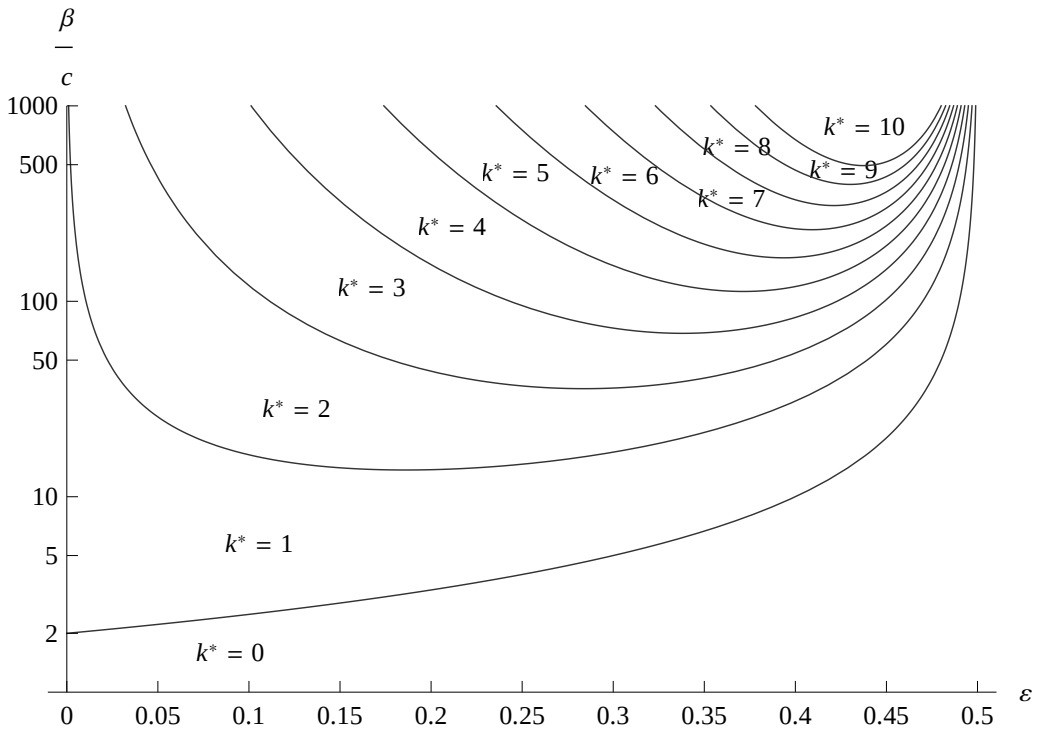

Figure 1: k^* when $\beta/c \leq 1000$ and $0 < \varepsilon < 1/2$. The indifference curves are the functions $g_k(\varepsilon)$. Note that the β/c-axis is logarithmic.

How does k^* respond to changes in the parameter values? Consider a change along the vertical axis of figure 1. All else being equal, if β increases the scientist takes more observations before making a decision. This is reasonable, because an increase of β means that coming to the wrong conclusion gives a higher loss, and taking more observations helps decrease the chance of that. Conversely, if c increases the scientist takes less observations before making a decision (all else being equal). This is also reasonable, because increased costs of observations give the scientist an incentive to come to a decision quickly, even if this increases the chance of making the wrong decision.

Now consider a change in the reliability of the evidence (the horizontal axis of figure 1). In the limit as ε goes to $1/2$, k^* decreases to the point where it is optimal to take no observations at all. This is because at such a high value of ε, observations provide no meaningful information: the two possible outcomes are almost equally likely in either of the two possible worlds. Given that nothing is learned from them anyway, it is unreasonable to pay any cost to see the value of these random variables, no matter the value of c. This is why $g_k(\varepsilon)$ goes to infinity as ε goes to $1/2$ (for any k).

In the limit as ε goes to 0, it is optimal to take at most one observation. At such a low value of ε, one observation is enough for the scientist to learn which world she is in with near-certainty. So whatever the value of c, there is no point in paying it more than once. The only question is whether one or zero observations should be taken. This question is equivalent to asking whether the cost to the scientist of guessing which world she is in (with a $1/2$ chance of being correct, this cost is $\beta/2$) or the cost of the one observation needed to learn which world she is in (i.e., c) is lower. Clearly, taking one observation is better if $\beta/c > 2$, while taking no observations is better if $\beta/c < 2$. In accordance with this result, $g_0(\varepsilon)$ goes to 2 as ε goes to 0, while $g_k(\varepsilon)$ goes to infinity for all $k > 0$.

For moderate values of ε, there is some more interesting behavior. As long as ε is not too close to its limits, the value of β/c actually matters. At values

of β/c greater than 13.7, more complicated decision procedures than "decide immediately" or "take one observation and then decide" start appearing. Noting that the vertical axis in figure 1 is logarithmic, it is worth mentioning that quite large values of β/c are needed before procedures that wait for a larger difference than a few between the number of observations favoring p's truth and p's falsity come into the picture. For instance, if $\beta/c \leq 100$, it is never optimal to wait for a larger difference than 4, whatever the value of ε.

5 Conclusion

I analyzed a model of a scientist trying to learn the truth-value of a proposition by observing evidence in the form of Bernoulli-distributed random variables. I asked and answered the question how much evidence a Bayesian scientist should want to see, given that each observation comes at a cost c, and the loss for an incorrect conclusion is β.

Qualitatively, the results are as expected. A higher loss β or a lower cost c leads to a higher number of observations, and vice versa.

Quantitatively, the results are perhaps a little more surprising. If the loss is no higher than the cost of thirteen observations ($\beta \leq 13c$) then it is optimal to take no more than one observation. Even if the loss is as high as the cost of a hundred observations it is not optimal to wait for a difference larger than four between the number of observations favoring one conclusion and the number of observations favoring the other.

This suggests that only the most important propositions (where the results of having the wrong belief about it are many times worse than the costs of collecting additional evidence) merit extensive investigation. For less important propositions collecting a single piece of evidence (or simply guessing the truth-value based on no evidence at all) is often the best strategy.

The model used here lends itself to extension in various respects. It might be interesting to use different distributions for the evidence and compare the results with the case analyzed here. Such variations would remain within

the basic framework of the sequential probability-ratio test and could thus be analyzed on similar lines. Less straightforward extensions might consider more complicated problems (e.g., learning about multiple propositions) and different forms of the loss function.

A Proofs

From proposition 1 it follows that the optimal procedure that takes at least one observation takes the form $\delta(a, b)$, where a is a negative integer multiple of $\log \frac{1-\varepsilon}{\varepsilon}$ and b is a positive integer multiple of $\log \frac{1-\varepsilon}{\varepsilon}$.

If $\xi = 1/2$, the symmetry of the problem (the loss for a wrong decision β and the cost per observation c are the same whether p is true or false) implies that $a = -b$. So the optimal procedure that takes at least one observation is of the form

$$\delta_{k,k} := \delta\left(-k \log \frac{1-\varepsilon}{\varepsilon}, k \log \frac{1-\varepsilon}{\varepsilon}\right),$$

for some positive integer k. Note also that

$$\mathbb{E}[Z_i \mid \neg p] = (1-2\varepsilon) \log \frac{1-\varepsilon}{\varepsilon} = -\mathbb{E}[Z_i \mid p].$$

Next I apply equation (1) to $\delta_{k,k}$, plugging in the expected values of Z_i, and using some algebra to simplify the resulting expression. This yields

$$\rho\left(\frac{1}{2}, \delta_{k,k}\right) = \beta \frac{\varepsilon^k}{(1-\varepsilon)^k + \varepsilon^k} + c \frac{k}{1-2\varepsilon} \frac{(1-\varepsilon)^k - \varepsilon^k}{(1-\varepsilon)^k + \varepsilon^k}.$$

Now I can compare the risk of different procedures, for example $\delta_{k,k}$ and $\delta_{k+1,k+1}$. This way I find

$$\begin{aligned}\rho\left(\frac{1}{2}, \delta_{k+1,k+1}\right) - \rho\left(\frac{1}{2}, \delta_{k,k}\right) &= \beta \frac{\varepsilon^k (1-\varepsilon)^k (2\varepsilon - 1)}{((1-\varepsilon)^{k+1} + \varepsilon^{k+1})((1-\varepsilon)^k + \varepsilon^k)} \\ &+ c \frac{1}{1-2\varepsilon} \frac{(2k+1)(1-\varepsilon)^k \varepsilon^k (1-2\varepsilon) + (1-\varepsilon)^{2k+1} - \varepsilon^{2k+1}}{((1-\varepsilon)^{k+1} + \varepsilon^{k+1})((1-\varepsilon)^k + \varepsilon^k)}.\end{aligned}$$

So $\rho(1/2, \delta_{k+1,k+1}) < \rho(1/2, \delta_{k,k})$ if and only if

$$\frac{\beta}{c} > g_k(\varepsilon) = \frac{(1-\varepsilon)^{2k+1} - \varepsilon^{2k+1}}{(1-2\varepsilon)^2 \varepsilon^k (1-\varepsilon)^k} + \frac{2k+1}{1-2\varepsilon}.$$

For a given value of ε, $g_{k+1}(\varepsilon) > g_k(\varepsilon)$ for all $k \geq 0$ because:

$$\frac{\partial g_k(\varepsilon)}{\partial k} = \log \frac{1-\varepsilon}{\varepsilon} \left(\frac{(1-\varepsilon)^{2k+1} + \varepsilon^{2k+1}}{(1-2\varepsilon)^2 \varepsilon^k (1-\varepsilon)^k} \right) + \frac{2}{1-2\varepsilon} > 0,$$

for all $k \geq 0$, $0 < \varepsilon < 1/2$.

So there is a unique positive integer k^* such that

$$g_{k^*-1}(\varepsilon) < \frac{\beta}{c} \leq g_{k^*}(\varepsilon).$$

(Unless $\beta/c \leq g_1(\varepsilon)$; in that case set $k^* = 1$.) Moreover, δ_{k^*,k^*} is the optimal sequential decision procedure that takes at least one observation:

$$\rho\left(\frac{1}{2}, \delta_{k^*,k^*}\right) \leq \rho\left(\frac{1}{2}, \delta_{k,k}\right)$$

for all positive integers k (with equality only if either $k = k^*$ or $\beta/c = g_{k^*}(\varepsilon)$ and $k = k^* + 1$).

So far, I have focused on determining an optimal procedure under the assumption that at least one observation is taken. It remains to be determined whether $\rho(1/2, k^*, k^*) < \rho_0(1/2)$, that is whether the optimal procedure that takes at least one observation is better than taking no observations at all. First consider whether taking one observation is better than taking none. Since

$$\rho\left(\frac{1}{2}, 1, 1\right) = \beta\varepsilon + c,$$
$$\rho_0\left(\frac{1}{2}\right) = \frac{\beta}{2},$$

it follows that $\rho(1/2, 1, 1) < \rho_0(1/2)$ if and only if

$$\frac{\beta}{c} > \frac{2}{1-2\varepsilon} = g_0(\varepsilon).$$

It turns out that the criterion is exactly $\beta/c > g_0(\varepsilon)$. So if taking zero observations is better than taking one observation $\beta/c \leq g_0(\varepsilon) < g_k(\varepsilon)$ for all $k > 0$, so in that case taking zero observations is better than taking any number of observations.

So I can simply extend the definition of k^* to be the unique positive integer such that

$$g_{k^*-1}(\varepsilon) < \frac{\beta}{c} \leq g_{k^*}(\varepsilon),$$

unless $\beta/c \leq g_0(\varepsilon)$, in which case k^* is defined to be zero. The optimal sequential decision procedure when the prior is $1/2$ is δ_{k^*,k^*}. This proves proposition 2.

Now consider a prior of the form

$$\xi_d = \frac{\varepsilon^d}{\varepsilon^d + (1-\varepsilon)^d}$$

for some $d \in \mathbb{Z}$. This might be called a conjugate prior for this decision problem since conditioning on evidence yields posterior probabilities of the same form: $\xi_d(1) = \xi_{d-1}$ and $\xi_d(0) = \xi_{d+1}$ (recall that $\xi(x)$ denotes the result of using Bayes' rule with prior ξ and evidence $X_1 = x$).

Note that $\xi_0 = 1/2$ so the optimal sequential decision procedure for ξ_0 is δ_{k^*,k^*}. But in light of the above that means that it is optimal to continue taking observations as long as the posterior is between ξ_{k^*} and ξ_{-k^*}, to stop and choose decision d_2 if the posterior hits ξ_{k^*}, and to stop and choose decision d_1 if the posterior hits ξ_{-k^*}.

These optimal stopping points do not depend on the prior. Thus for any prior ξ_d it is optimal to continue taking observations as long as the posterior remains between ξ_{k^*} and ξ_{-k^*}. But this is exactly the sequential decision procedure δ_{k^*+d,k^*-d} (assuming $\xi_{k^*} < \xi_d < \xi_{-k^*}$, i.e., $k^* + d > 0$ and $k^* - d > 0$, otherwise it is optimal to decide immediately). This proves proposition 3.

If $\xi_d < \xi < \xi_{d-1}$ then observing $X_i = 0$ k^*-d+1 times forces the posterior to be less than ξ_{k^*}, at which point it is optimal to stop taking observations.

Observing $X_i = 0$ less than $k^* - d$ times forces the posterior to be larger than ξ_{k^*-1}, so continuing to take observations is optimal.

Similarly, observing $X_i = 1$ $k^* + d$ times forces the posterior to be greater than ξ_{-k^*}, at which point it is optimal to stop taking observations. Observing $X_i = 1$ less than $k^* + d - 1$ times forces the posterior to be less than ξ_{-k^*+1}, so continuing to take observations is optimal.

Hence one of δ_{k^*+d,k^*-d}, δ_{k^*+d-1,k^*-d+1}, δ_{k^*+d-1,k^*-d}, or δ_{k^*+d,k^*-d+1} is the optimal sequential decision procedure. This proves the corollary.

References

Y.S. Chow and Herbert Robbins. On optimal stopping rules. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(1):33–49, 1963. ISSN 0044-3719. doi: 10.1007/BF00535296. URL `http://dx.doi.org/10.1007/BF00535296`.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New Jersey, 2004.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3): 326–339, 1948. ISSN 00034851. URL `http://www.jstor.org/stable/2235638`.